Human lactase persistence:

Evaluation of concordance between the breath hydrogen test and molecular genotyping

Analysis of evolutionary history using a microsatellite approach

Margarida Cardoso Reis Sá Coelho

Porto 2005

Tex-4/93a

Fac.

24 Fev 05

Margarida Cardoso Reis Sá Coelho

Human lactase persistence:

Evaluation of concordance between the breath hydrogen test and molecular genotyping

Analysis of evolutionary history using a microsatellite approach

PORTO

2005

Dissertação de Mestrado em
Biodiversidade e Recursos Genéticos
apresentada à Faculdade de Ciências
da Universidade do Porto

Agradecimentos/ Acknowledgments

Ao meu orientador Professor Jorge Rocha. Pela presença amiga, pela motivação, pela dedicação, pelo exemplo humano, pelos mundos fascinantes que me tem dado a descobrir a cada passo e por tudo o que não cabe em palavras mas que guardo no coração.

Ao Professor Sobrinho Simões pelo caloroso acolhimento no IPATIMUP, pela oportunidade de ver um sonho tornado realidade.

Ao Professor Nuno Ferrand, coordenador do curso de Mestrado em Biodiversidade e Recursos Genéticos, pelo interesse prestado ao trabalho e pela ajuda no decorrer deste.

Da Unidade de Gastrenterologia Pediátrica do Hospital Santa Maria: à Drª Ana Isabel Lopes pela colheita de dados clínicos, análise e interpretação da prova de hidrogénio, pelo entusiasmo e estímulo; à Enfermeira Ima Figueiredo pela dedicação na execução da prova de hidrogénio expirado e na colheita das amostras biológicas; ao Prof. Dr. Evangelista Rocha pelas sugestões na análise dos resultados.

Ao Luís Pedro Resende, que iniciou comigo este trabalho, pela colaboração na parte laboratorial do trabalho e pelo entusiasmo sempre demonstrado.

From Universitá La Sapienza, I would like to thank to Cinzia Battagia for the sympathy and the participation in typing the Italian and Fulbe samples and to Professor Giovanni Destro Bisol for the comments and suggestions during the work and for the Italian and Fulbe DNA samples.

From Universita di Bologna, to Cristina Fabbri for the sympathy and the assistance in typing the Italian samples.

Da Universidade Pedagógica de Moçambique ao Prof. Dr. António Prista e a Sílvio Saranga pela colaboração na colheita de amostras de Moçambique.

À Sílvia Pereira e aos meus colegas de mestrado, pelo companheirismo e espírito de entreajuda.

A todos os meus colegas do IPATIMUP por contribuírem, das mais variadas formas, para a realização deste trabalho. Um agradecimento especial à Susana Seixas, pela paciência e disponibilidade com que sempre me ajudou nos (muitos) momentos que precisei, por ter sempre a solução certa na manga, pela amiga que se tornou. A todos aqueles que estiveram mais próximo (Rita Quental, Susana Cristina, Gil Tomás, Sofia Quental, Raquel Matos, Luísa Azevedo, Pedro Soares, Filipe Pereira, Sandra Martins, Alexandra Lopes, Sandra Beleza), pelo privilégio da partilha das dificuldades e alegrias do dia-a-dia, pela boa disposição, pelo intercâmbio dos nossos mundos científicos.

Aos colegas da informática agradeço as ajudas preciosas nos pequenos dramas informáticos. Um agradecimento especial ao Wagner pela elaboração de um programa que foi de grande importância na fase em que o trabalho se encontrava.

À Sociedade Portuguesa de Gastrenterologia pelo financiamento do projecto: "Má absorção de lactose: estudo de prevalência e análise da correlação entre o diagnóstico molecular e a prova do hidrogénio expirado", sem o qual não teria sido possível concretizar este trabalho.

Ao Nelson, à Lurdes, ao Filipe, ao Ricardo, à Olinda e à Sara... por tudo. Por serem pessoas tão bonitas.

Mito da criação

No princípio existia uma enorme gota de leite.
Então chegou Doondari e criou a pedra.
A pedra criou o ferro;
E o ferro criou o fogo;
E o fogo criou a água ;
E a água criou o ar.
Então Doondari desceu pela segunda vez.
Juntou os cinco elementos
E moldou-os num homem,
Mas o homem era orgulhoso.
Então Doondari criou a cegueira e a cegueira derrotou o homem.
Mas quando a cegueira se tornou demasiado orgulhosa,
Doondari criou o sono, e o sono derrotou a cegueira;
Mas quando o sono se tornou demasiado orgulhoso,
Doondari criou a preocupação, e a preocupação derrotou o sono;
Mas quando a preocupação se tornou demasiado orgulhosa,
Doondari criou a morte, e a morte derrotou a preocupação.
Quando a morte se tornou demasiado orgulhosa,
Doondari desceu pela terceira vez.
E ele veio como Gueno, o Eterno,
E Gueno derrotou a morte.

Mali, Fulani
(Trad.: Vasco David)
(Rosa do Mundo-
2001 poemas para o futuro; 3ª edição, Assírio e Alvim)

Table of contents

Summary 9

Resumo 11

1. Introduction 13

2. Material and methods 19

2.1 Sample composition 20

2.2 Identification of the lactase persistence status 21

2.2.1 Breath hydrogen test (BH2test) 21

2.2.2 Molecular test 21

2.2.3 Comparison between the results of the tests 23

2.3 Haplotype characterization 23

2.3.1 Haplotype inference 24

2.3.2 Allele age determination 24

2.3.3 Neutrality test 26

3. Results and discussion 27

3.1 Concordance study 28

3.1.1 Results from the breath hydrogen (BH2) and the molecular tests 29

3.1.2 Comparison between classifications 31

3.2. Haplotype diversity and evolution of the lactase persistence polymorphism 38

3.2.1 Frequency of the SNP-haplotypes in different populations 39

3.2.2 Microsatellite variation within the SNP haplotypes 41

3.2.3 Estimation of the age of the -13.91kb*T allele 47

3.2.4 Neutrality tests 50

4. Conclusions 53

4.1 Concordance study 54

4.2 Evolutionary history of the lactase persistence polymorphism 55

4.2.1 Frequency of lactase persistence in different populations 55

4.2.2 Microsatellite variation and evolutionary history of the lactase persistence polymorphism 56

5. References 58

Summary

The ability to digest lactose in human adults is a hereditary condition caused by the persistence of lactase activity after weaning, which may represent one of the most impressive examples of genetic adaptation to modifications in dietary habits. Two recently described single nucleotide polymorphisms (SNPs), -13.91kbC/T and -22.01kbG/A, were shown to be highly associated with the lactase restriction/persistence variation, providing an important tool to study lactose digestion capacity in human populations.

This work is a contribution to the understanding of the natural history of lactase persistence: its origin and the factors that might have influenced its present geographic distribution.

The work is divided in two parts. In the first part, the concordance between diagnoses based on the breath hydrogen test (BH2 test) and the molecular results obtained through -13.91kbT/A and -22.01kbG/A genotyping was evaluated in 68 Portuguese individuals. A level of concordance as high as 93% was observed between the results from the BH2 test and the classifications based on the -13.91kbC/T polymorphism. This shows that -13.91kbC/T genotyping can be a valid alternative to physiological tests in the characterization of the lactase activity profile of the Portuguese population. Due to its relative simplicity, non-invasiveness and high level of automation, the molecular test is particularly suitable for large scale studies. Furthermore molecular testing of lactase persistence may also be useful in the differential diagnosis of abdominal complaints, without the shortcomings associated with inter-individual variability in physiological response.

In the second part of the work, the distributions of the -13.91kbC/T and -22.01kbG/A polymorphisms were studied in samples from Portugal, Italy, São Tomé Island, Mozambique and in the Fulbe ethnic group from Cameroon. In addition, the levels of diversity associated with the "core haplotypes" defined by the -13.91kbC/T and -22.01kbG/A SNPs were assessed through the analysis of 4 fast evolving microsatellites loci (D2S3010, D2S3013, D2S3015 and D2S3016). The prevalences of lactase persistence predicted from the frequencies of the -13.91kb*T allele, more closely associated with the trait, were found to vary considerably between populations. The frequencies in Portugal (60%), here determined for the first time, lie within the prevalence range previously reported for Southern France and Northern Spain. The 24% frequency of lactase persistence in Italy is compatible with previous studies based on physiologic tests. In the African samples, the

estimates for lactase persistence were higher in the pastoralist Fulbe population (38%) than in São Tomé (7,8%) and Mozambique (4,0%), which are not associated with dairying traditions. The frequency of the -13.91kb*T allele appears to be a good predictor of the prevalence of lactase persistence in these populations.

The survey of microsatellite diversity revealed a substantial reduction of haplotype variation in the chromosomes bearing the -13.91kb*T allele across all populations, suggesting that lactase persistence had a relatively recent origin.

Age estimates based on the intra-allelic microsatellite variation indicate that the -13.91kb*T allele originated only after the separation between European and African populations and may be as recent as 12500-7500 years.

The use of a neutrality test based on the comparison of the frequency of the -13.91kb*T allele and its observed levels of intra-allelic variability has shown that this variant is too recent to have reached its current frequencies without the influence of positive selection. This evidence supports the hypothesis that the -13.91kb*T allele arose in Eurasia and reached its present distribution in a relatively short time due to the selective advantage of lifelong unrestricted use of milk.

Taken together, the results show that even a limited number of microsatellite loci may provide sufficient resolution to reconstruct the most important aspects of the evolutionary history of human lactase persistence.

Resumo

A capacidade de digerir a lactose na idade adulta é uma característica hereditária causada pela persistência da actividade da lactase após desmame, que pode considerar-se um um exemplo de adaptação genética a modificações nos hábitos nutricionais. Recentemente foram identificados dois polimorfismos genéticos, -13.91kbC/T e -22.01kbG/A, fortemente associados à variação na restrição/persistência da lactase, que passaram a constituir um importante instrumento de estudo da capacidade de digestão da lactose nas populações humanas.

Este trabalho é uma contribuição para a compreensão da história natural da persistência da lactase, da sua origem e dos factores que podem ter influenciado a sua actual distribuição geográfica.

O trabalho está dividido em duas partes. Na primeira parte, avaliou-se a concordância entre os diagnósticos obtidos com o Teste do Hidrogénio expirado (BH2) e os resultados moleculares da genotipagem dos polimorfismos -13.91kbC/T e -22.01kbG/A em 68 indivíduos portugueses. Um nível elevado de concordância (93%) foi observado entre os resultados do teste do BH2 e as classificações baseadas no polimorfismo -13.91kbC/T. Este resultado mostra que a genotipagem do -13.91kbC/T constitui uma alternativa válida aos testes fisiológicos na população portuguesa. Dada a sua relativa simplicidade, não invasividade, e alto nível de automatização, o teste molecular é particularmente apropriado em estudos de larga escala, podendo também ser útil no diagnóstico diferencial de patologias com sintomas abdominais sem as desvantagens associadas à variação inter-individual na resposta fisiológica.

Na segunda parte do trabalho, as distribuições dos polimorfismos -13.91kbC/T e -22.01kbG/A foram estudadas em amostras de Portugal, Itália, ilha de São Tomé, Moçambique e no grupo étnico Fulbe dos Camarões. Procedeu-se também à caracterização dos níveis de diversidade haplotípica associados às linhagens definidas pelos 2 polimorfismos através da análise de 4 microssatélites (D2S3010, D2S3013, D2S3015 e D2S3016). Verificou-se que as prevalências da persistência da lactase calculadas a partir das frequências do alelo -13.91kb*T, que tem maior associação com a característica, variam consideravelmente entre populações. A frequência em Portugal, aqui determinada pela primeira vez, enquadra-se no intervalo de valores anteriormente descritos para o Sul de França e Norte de Espanha. A frequência de 24% da persistência da lactase em Itália é

compatível com estudos anteriores baseados em testes fisiológicos. Nas amostras africanas, a população Fulbe, com forte tradição de consumo de leite, apresenta um valor de prevalência da persistência da lactase (38%) claramente superior ao das populações de São Tomé (7,8%) e Moçambique (4,0%), que descendem de sociedades onde a produção e consumo de lacticínios não tem tido uma grande importância nas respectivas economias de subsistência. A frequência do alelo -13.91kb*T revelou-se um bom indicador da prevalência da persistência da lactase nestas populações.

A análise da diversidade dos microssatélites revelou uma redução substancial da diversidade haplotípica associada à mutação 13.91kb*T nas diferentes populações, o que indica que a persistência da lactase é relativamente recente. As estimativas da idade da mutação baseadas na variação dos microssatélites sugerem que o alelo -13.91kb*T se terá originado após a separação entre as populações Africanas e Europeias e situam a idade mínima do alelo no intervalo entre 12500-7500 anos. A utilização de um teste de neutralidade baseado na comparação entre a frequência do alelo -13.91kb*T e os seus níveis observados de diversidade alélica mostrou que este variante é demasiado recente para que as suas frequências actuais possam ter sido atingidas sem favorecimento selectivo. No seu conjunto, a evidência recolhida apoia a hipótese de que o alelo -13.91kb*T surgiu na Eurasia e atingiu a sua distribuição num período de tempo relativamente curto devido à vantagem selectiva conferida pelo uso do leite ao longo de toda a vida.

Estes resultados mostram que mesmo um número limitado de microssatélites oferece resolução suficiente para a reconstrução dos aspectos mais importantes da história evolutiva da persistência da lactase.

1. Introduction

The ability to digest milk lactose in adulthood is a polymorphic human trait due to variation in the persistence of lactase activity in the small intestine. Except for rare cases of congenital lactase deficiency, intestinal lactase activity reaches a peak shortly after birth and remains invariably high during infancy and early childhood. In non-human Mammals, lactase activity always declines after the weaning phase as part of the developmental regulation of the lactase gene and this synchronized down-regulation has been interpreted by some as an evolutionary adaptation promoting the shift towards adult diet and optimizing the spacing of offspring (Flatz, 1987).

In contrast to the invariant nature of lactase decline in other mammals, two phenotypes can be distinguished in humans on the basis of lactase activity profiles: the ancestral lactase restriction phenotype, in which lactase activity declines after weaning, and the derived lactase persistence phenotype characterized by the maintenance of high levels of enzyme activity throughout adulthood (Flatz, 1987). Family studies have shown that this phenotypic variation in lactase activity profiles is genetically controlled and that lactase persistence is due to a dominant mutation in an autosomal locus (Sahi, 1994).

Curiously, the perception of this variation seems to be deeply rooted in medical tradition since Galen himself is generally acknowledged to have written that: "With regard to milk, it should not be given to all, but only to those who digest it well and perceive no symptoms in the right hypochondrium."

Lactose is the major disaccharide present in milk and consists of the monosaccharides glucose and galactose. It is a carbohydrate of considerable nutritional importance during the suckling phase that acts as a major supplier of energy. In the absence of high lactase activity, lactose cannot be broken into its components and will remain unabsorbed during its passage through the small intestine. The unbroken lactose induces osmotic changes in the gut and leads to the build up of lactic acid, short chain carbonic acids, carbon dioxide, hydrogen gas and methane, as a result of fermentation by enteric bacteria in the large intestine. As a consequence, individuals may feel abdominal pain, flatulence and diarrhoea among other symptoms of milk intolerance. However, absence of lactase persistence is not always associated with gastrointestinal discomfort. Many subjects with lactase restriction can drink milk without experiencing any intolerance symptoms (Flatz, 1987).Therefore the notion that lactase restriction is invariably linked to milk/lactose intolerance may be misleading. People may adjust the dietary intake of milk and dairy products to their individual tolerance threshold taking no notice of any trouble. There may also be cultural adaptation to

milk/lactose intolerance through the consumption of milk products that are soured or otherwise treated, like cheese and yogurts. These products cause few problems because they contain relatively low levels of lactose, or even contain bacteria that secrete lactases themselves (eg. *Lactobacillus acidophilus*). Furthermore, lactase activity doesn't seem to be the only variable determining the pattern and severity of symptoms. Velocity of gastric emptying, individual differences in prostaglandin synthesis (which stimulate peristalsis in small intestine), colonic bacteria and colonic irritability may also play an additional role (Flatz, 1987). Nevertheless, "lactose intolerance" is still a useful expression to describe the symptoms that ensue if the individual threshold of digestible lactose is exceeded. On the other hand, the fact still remains that lactase restriction does limit the use of large quantities of fresh milk in adulthood.

The diagnosis of lactase activity phenotypes may be done with direct or indirect methods (reviewd in Arola, 1994). Direct methods are based on the determination of lactase activity in intestinal biopsies and provide a reliable diagnosis, but are very invasive and therefore not suitable for large scale population studies. Indirect methods, also known as lactose tolerance tests (LTT), are based on the quantification of metabolites that are formed when lactose is not digested. One of the most popular LTT is the breath hydrogen test (BH_2T), which measures the hydrogen content in the proband's breath that results from lactose fermentation by colonic bacteria and is excreted through the lungs after entering the bloodstream. The concentration of breath hydrogen determined by gas chromatography after a lactose load is then used to classify individuals according to their lactase phenotype (Flatz, 1987). Since the load of lactose is high enough to ensure that most individual tolerance threshold are exceeded, the terms "lactase restriction" and "lactose intolerance" are used interchangeably in the context of LTTs classifications. Although LTTs, and especially BH_2T, are considered to be convenient, non-invasive methods yielding unequivocal resolution, there are a number of sources of error that may led to incorrect phenotypic classification (Arola, 1988). For example, changes in colonic flora may cause inability to ferment undigested lactose and prevent the identification of some cases of lactase restriction (false negatives). Conversely, secondary loss of lactase activity due to damages in the intestinal epithelium may result in the wrong inclusion of tolerant individuals in the lactase restriction category (false positives).

The prevalence of the lactase persistence phenotype is highly variable and appears to be positively correlated with the milk drinking habits of different human populations.

Lactase persistence is most frequent in Northern Europe, where milk dependent cattle pastoralism might have been developed as early as 5000 years ago (Beja-Pereira *et al*, 2003), and gradually decreases towards the south and east of the continent (Fig.1). Outside Europe, lactase restriction is the predominant phenotype in most world populations, except in African and Arabic nomadic pastoralist which typically show higher frequencies of lactase persistence than their neighbouring non-pastoralist communities (Sahi, 1994). According to the so called "culture-historical" hypothesis (Simoons, 1970 and McCracken, 1971), this correlation is caused by a recent selective pressure associated with the advantages of drinking milk that led to a rapid increase of the lactase persistence mutation in populations with milk drinking habits. Such a rapid increase of lactase persistence would have been possible only under special conditions of milk dependency in human societies where fresh milk was the major source of essential nutrients that could not be obtained in other foods available (Flatz, 1987). An alternative to the "culture-historical" hypothesis called the "reverse cause argument" suggests that the increase of lactase persistence in some human populations is unrelated to milk use and that dairying was adopted precisely by those populations that could tolerate lactose (reviewed in Aoki, 2001). Thus, the two major proposals (culture-historical hypothesis and reverse cause argument) differ in the temporal priority given to cultural or genetic change.

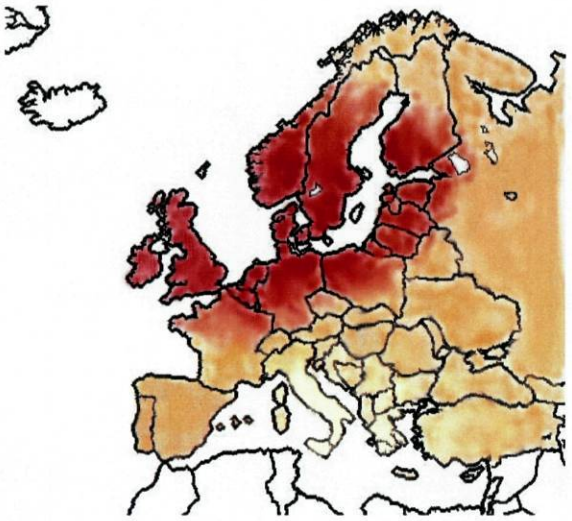

Fig.1 Distribution of the lactase persistence mutation in present-day Europeans. The hotter the color the highest the frequency (Adapted from Beja-Pereira *et al*, 2003).

The molecular basis for variation in lactase persistence has remained unclear for a long time. The lactase gene is well characterized only since 1991 (Boll *et al*, 1991). It is located on chromosome 2q21, comprises 17 exons and covers approximately 49kb, with a corresponding mRNA of slightly more than 6kb that encodes a 1927 aminoacid polypeptide. Initial efforts to identify the cause of the persistence/restriction variation led to the identification of several DNA polymorphisms within the LCT gene and its neighbouring promoter regions, but none showed the appropriate phenotype-genotype correlation (reviewed in Swallow, 2003).

More recently, a T allele of a C/T polymorphism located in a potential regulatory site 13910kb upstream the lactase gene (-13.91kb*T allele) was found to be completely associated with lactase persistence in Northern Europeans (Enattah *et al*, 2002). A similar, but less strong association, was also found with a G/A single nucleotide polymorphism located at -22018bp (-22.01kb), about 8kb further upstream. Both the -13.91kb and -22.01kb Single Nucleotide Polymorphisms (SNPs) are located within introns of the neighbouring gene MCM6, which is involved in the cell cycle, and illustrate the complexity of the regulation of the LCT gene by showing how apparently silent DNA variants in non-coding regions of the genome may play unexpected functional roles many kilobases away.

Kuokkanen *et al* (2003) have demonstrated that LCT mRNA levels are consistently more elevated in chromosomes bearing the -13.91kb*T allele indicating that it is associated with the *cis* transcriptional regulation of LCT. More recent transfection studies have suggested that the -13.91kb*T allele might be indeed the causative factor directly involved in the enhancement of LCT transcription (Olds and Sibley, 2003; Troelsen *et al*, 2003).

The finding of the -13.91kb*T mutation has also provided a basis for addressing major issues concerning the evolutionary history of lactase persistence. Using several linked SNPs, Bersaglieri *et al* (2004) found that the -13.91kb*T allele is associated with remarkable high levels of haplotype homogeneity in Northern Europeans, indicating that the high frequencies of lactase persistence might have been reached in a short time frame due to the action of natural selection. However, since there is no data on linked haplotype variation in other populations, the effect of population history in the generation of this linkage disequilibrium is still not fully appreciated. On the other hand the possibility that haplotype homogeneity might be caused by dominant suppression of recombination over Mb distances could not be completely ruled out.

The analysis of the critical -13.91kb C/T polymorphism in Africa has shown that the -13.91kb*T allele is very rare in most tested populations, including pastoralist communities where higher frequencies of lactase persistence were previously reported on the basis of physiological tests (Mulcare *et al*, 2004). Only in the Fulbe and Hausa pastoralists from Cameroon was the frequency of -13.91kb*T consistent with the levels of lactase persistence predicted by the physiological evaluations. This result implies that either the -13.91kb*T is not the direct cause of lactase persistence or that there is further genetic heterogeneity underlying this condition.

In order to have a wider perspective of the evolutionary history of the lactase persistence polymorphism it is important to: a) understand whether lactase persistence arose multiple times in human populations or it was caused by a single mutation; b) to assess the concordance between persistence candidate mutations and physiological tests in different populations; c) to determine the age of lactase persistence candidate mutation; d) to perform genetic tests of natural selection in a broad range of populations; e) to reconstruct the major population movements that could have spread the persistence- candidate mutations in different areas.

This work aims to contribute to these general goals and presents an analysis of the genetic polymorphisms of lactase persistence that is divided into two parts: in the first part, the concordance between lactase phenotypes determined by the breath hydrogen test and by the -13.91kb C/T polymorphism genotypes was evaluated in Portuguese family samples with probands that had lactose tolerance tested for differential diagnosis. In the second part, the evolutionary history of lactase persistence was studied through the characterization of the haplotype variation associated with the -13.91kb C/T and -22.01kb G/A polymorphisms by using fast evolving microsatellite loci in geographically diverse populations with different subsistence patterns from Portugal, Italy, São Tomé island (W Africa), Mozambique and the Fulbe ethnic group from Cameroon.

2. Material and methods

2.1 Sample composition

The concordance study between molecular and physiological tests was undertaken in 68 paediatric cases referred for abdominal symptoms suggestive of lactase restriction and their first degree family members recruited in Paediatrics Gastroenterology Unit of University Hospital Santa Maria (age range: 5 to 68 years; mean age 18.4 years).

The haplotype diversity study was carried in random samples of unrelated individuals from geographically and ethnically diverse populations: Northern Portugal (N=90), Central Italy (N=68, 37 from Tocco and 30 from Rome), São Tomé island (N=142, from different locations in the island), the Fulbe ethnic group from Cameroon (N=51) 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 18th 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 tradition of pastoralism nor dairy practices. 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 15th 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 and has an estimated European admixture of 11% (Tomás *et al*, 2002).

DNA was obtained from buccal swabs using standard extraction methods.

2.2 Identification of the lactase persistence status

2.2.1 Breath hydrogen test (BH_2T)

The hydrogen breath test was performed according to standard methodology: after a minimum of 6 hours fasting, a 20% lactose aqueous solution was ingested, at a dosage of 2g/Kg body weight up to a maximum of 50g. Expired air samples were collected at minute 0 (previously to the lactose administration) and at 30-minute intervals for 3 hours after ingestion of the lactose source. Diagnosis was made on the basis of hydrogen concentration in expired air as measured by gas chromatography. The test was considered positive if the maximal increase in BH2 was higher than 20ppm (Flatz, 1987).

Gastrointestinal symptoms during the test were determined by asking the subjects to recount them. All subjects also completed a questionnaire about quantity and frequency of milk consumption.

2.2.2 Molecular test

The recently described -13.91kb C/T polymorphism reported to be highly associated with lactase restriction/persistence trait and the 22.01kb G/A SNP, with a similar but less strong association, were the basis of the molecular diagnosis (Fig.3). Individuals were genotyped for the -13.91kb C/T variant using the polymerase chain reaction (PCR) followed by digestion with BsmFI restriction enzyme (2U/μL). In the PCR the following primers were used: -13.91kbF (5'- GCAGGGCTCAAAGAACAATC- 3') and -13.91kbR (5'-TGTACTAGTAGGCCTCTGCGCT-3'). The reaction mixture contained 0.5 μM of each primer, 0.2mM of each deoxynucleotide triphosphate (dNTP), 10mM Tris-HCl (pH 8.8), 50mM KCl, 0.08% Nonidet, 1.5 mM $MgCl_2$ 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. Amplification produces a 125-bp product, and enzymatic restriction produces 80- and 45-bp fragments if the individual has the T allele and a 125bp fragment in individuals with the C allele. DNA fragments were visualized by silver staining after non-denaturing electrophoresis separation in 9%

polyacrylamide gels. For the -22.01kb G/A variant, a similar approach was used. The primers used to the amplification were: -22.01kbF (5'-CTCAGTGATCCTCCCACCTC-3') and -22.01kbR (5'- CCCCTACCCTATCAGTAAAGGC-3'). The reaction mixture differed only in the $MgCl_2$ concentration (1,0 mM). The annealing time was reduced to 30 seg. Amplification produces a 271-bp product, and enzymatic restriction with Hin 6I (1U/μL) produces 196- and 75-bp fragments if the individual has the G allele and a 271bp fragment in individuals with the A allele.

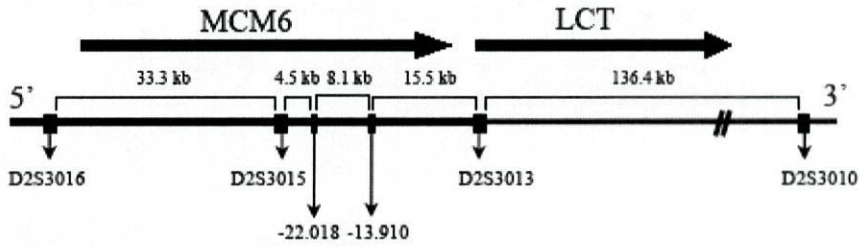

Figure 3: Schematic representation of the genetic interval including the Lactase locus (LCT) and the neighbour gene for the human homolog of a yeast gene involved in the cell cycle (MCM6), with the relative locations of the two SNPs (-13.91kb and -22.01kb) and the four microsatellite (D2S3010, D2S3013, D2S3015, D2S3016) used to characterize the haplotype diversity associated with the lactase restriction/persistence polymorphism. Distances are as in BAC clone RP11-34L23 (GenBankAcession no. AC011893.7).

2.2.3 Comparison between the results of the tests

The results obtained from the BH_2 test, the recorded symptoms and the molecular test were contrasted against each other. Specificity, sensitivity and positive and negative predictive values were evaluated for each test by using the alternative tests as the golden standard, as follows:

		Gold standard		
		positive	Negative	
Test	positive	A	b	a + b
	negative	C	d	c + d
		a + c	b + d	a + b + c + d

Sensitivity= a/(a+c),

Specificity= d/(b+d),

Positive predictive value= a/(a+b),

Negative predictive value= d/(c+d),

2.3 Haplotype characterization

To assess the genetic diversity associated with the lactase persistence polymorphism we used combined haplotypes defined by the -13.91kb and -22.01kb SNPs and four fast evolving linked microsatellites loci: D2S3010, D2S3013, D2S3015 and D2S3016. The four microsatellites were typed by PCR amplification in two duplex PCRs with fluorescently labelled primers, 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'-GAGAATATAGTCATAAACTATGTT-3' and 5'-ATTTTGGATTATATATGCTTTCTTG-3' (labelled with FAM fluorescence)) and D2S3015 (5'- CCTGTAGTCCCAGCTAATTTC-3' and 3'-CAGAGAAGTTTTGTTTGTGGA-5' (labelled with TET fluorescence)) at 0.5 μM and 0.075 μM concentrations respectively. The second duplex reaction included the primers for D2S3010

(5'- TTAGGCCTCTCTTCGAATGAT-3' and 5'- GATTTAGGTGGAGACACAC-3' (labelled with FAM fluorescence)) and D2S3016 (5'- GAGAAAAATTAGGTGTGAAACCA-3' and 5'- CCCTTTAGCTGCCTGAACTG-3' (labelled with TET fluorescence)) at 0.5 μM and 0.075 μM concentrations, respectively. Each duplex reaction mixture contained the primers, 0.2mM of each deoxynucleotide triphosphate (dNTP), 10mM Tris-HCl (pH 8.8), 50mM KCl, 0.08% Nonidet, 1.5 mM $MgCl_2$ and 1 U *Taq* polymerase. 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.

2.3.1 Haplotype inference

Six-locus haplotypes combining the two -13.91kb and -22.01kb SNPs and the four microsatellites were determined from the combined genotype data for all the populations by statistical inference with the software package PHASE, version 2.0.2. (Stephens, Smith and Donnelly, 2001; Stephens and Donnelly, 2003). The haplotype frequencies were calculated by direct counting after resolution of each individual haplotype phase.

2.3.2 Allele age determination

To estimate the time of the most recent common ancestor (TMRCA) of the -13.91kb*T allele associated with lactase persistence, we used three different methods based on the intra-allelic 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 that is independent of population demography was calculated by $T=\bar{\Delta} / \bar{\mu}$, where $\bar{\Delta}$ is the average squared difference in repeat number (ASD) between each sampled -13.91kb*T haplotype and the root haplotype, averaged over loci, and $\bar{\mu}$ is the microsatellite mutation rate averaged over loci (Stumpf and Goldstein, 2001). The root of the -13.91kb*T clade (10-21-4-2) was obtained by combining together the modal allele lengths at each microsatellite locus in the pooled sample. The TMRCA central estimates and confidence intervals were calculated using the program Ytime (Behar *et al*, 2003).

The second method, which also assumes no recombination, is based on the accumulated variance in microsatellite repeat number ($\bar{V}$), averaged over loci, whose relationship with time may be approximated by $T=-2Ne\ \ln(1-\bar{V}/Ne\bar{\mu})$ or $T=\bar{V}/\bar{\mu}$, depending on the assumption of a constant population size (Ne) or a rapid population growth, respectively (Slatkin, 1995, Goldstein *et al*, 1996, Su *et al*, 1999). Estimates performed with this method were done either assuming different constant Ne values or rapid population growth. Confidence intervals for the estimates with constant population size were calculated according to Goldstein *et al* (1996).

The third method is based on the simulation of the overtime decay in the frequency of the allele originally associated with -13.91kb*T in each microsatellite locus (Seixas *et al*, 2001). Unlike the other two methods, this approach allows for recombination to be taken into account, according to the relation $p_{(g,i)}= p_{(g-1,i)}(1-\mu-r) + rq_i + (\mu/2)\ [p_{(g-1,\ i-1)} + p_{(g-1,\ i+1)}]$, where $p_{(g,i)}$ is the frequency of a marker microsatellite allele with i repeats in generation g within the -13.91kb*T allele, q_i is the frequency of that allele in the whole population, r is the recombination fraction between the -13.91kb site and each microsatellite locus and μ stands for the microsatellite mutation rate. The modal allele length at each microsatellite locus in the pooled sample was considered to be the ancestral. 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 1cM=1Mb, according to the approximate estimates provided by Kong *et al* (2002) for the region encompassing the four microsatellite loci. Confidence intervals were calculated by using ±2x the standard deviation of $p_{(g,i)}$ (Goldstein *et al*, 1999).

For each age estimation method we used two sets of microsatellite mutation rates (μ). The first set was derived indirectly from the parameter $\theta=4Ne\mu$, assuming mutation-drift equilibrium and using the unbiased θ estimator proposed by Xu and Fu (2004), based on the sample homozygosity under the stepwise mutation model. We assume Ne =10000 (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.

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.

2.3.3 Neutrality test

The possibility of occurrence of selection acting on the -13.91kb*T allele was studied by using the method developed by Slatkin and Bertorelle (2001) which evaluates whether the observed frequency of a lineage is consistent with its levels of variability under a given demographic pattern, assuming neutrality. We used the test modality that measures the intra-lineage variability by the minimum number of mutations (S_0) observed at linked microsatellite marker loci, assuming the infinite sites model and no recombination. The assumption of an infinite alleles model for microsatellite loci discards information from the dataset but has the advantage of avoiding the introduction of additional uncertainty about the mutation process (Slatkin and Bertorelle, 2001; Slatkin, 2002).

The tests were performed by considering the simultaneous combination of all four microsatellites with the -13.91kb*T allele. S_0 was calculated by using Median-Joining networks (Bandelt *et al*, 1999) to infer the minimum number of mutations necessary to generate the observed haplotypes. The networks were calculated using the program NETWORK 4.1.0.0. (http://www.fluxus-engineering.com). To allow for the possibility that these estimates may be to small, given the occurrence of recurrent mutations, we doubled all the values of S_0, as suggested by Slatkin and Bertorelle (2001). Different combinations of two global demographic models and the two sets of microsatellite mutation rates were considered in the calculation of the TMRCA of the -13.91kb*T allele (Table 3). 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 10^3. 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 5×10^9, also starting 900 generations ago. The data consists of the full haplotypes combining all four microsatellite markers linked to the -13.91kb*T allele.

3. Results and discussion

3.1 Concordance study

3.1.1 Results from the breath hydrogen (BH2) and the molecular tests

Figures 4 A and B present examples of negative and positive BH2 tests, respectively. When an increase higher than 20 ppm occurred, individuals were classified as lactase persistent (Flatz, 1987).

Figures 5 and 6 present typical results of the -13.91kb C/T and -22.01kb G/A genotyping, respectively. The -13.91kb polymorphism was genotyped by analysing the presence of a BsmFI restriction site. The genotyping of the -22.01kb G/A polymorphism was made by analysing the presence of a Hin6I restriction site. Homozygous individuals for the -13.91kb*C allele genotypes were classified as lactase restrictors; CT and TT genotypes were classified as lactase persistents. In the -22.01kb polymorphism, homozygous for the G allele genotypes were classified as lactase restrictors; GA and AA genotypes were classified as persistent (Enattah *et al*, 2002).

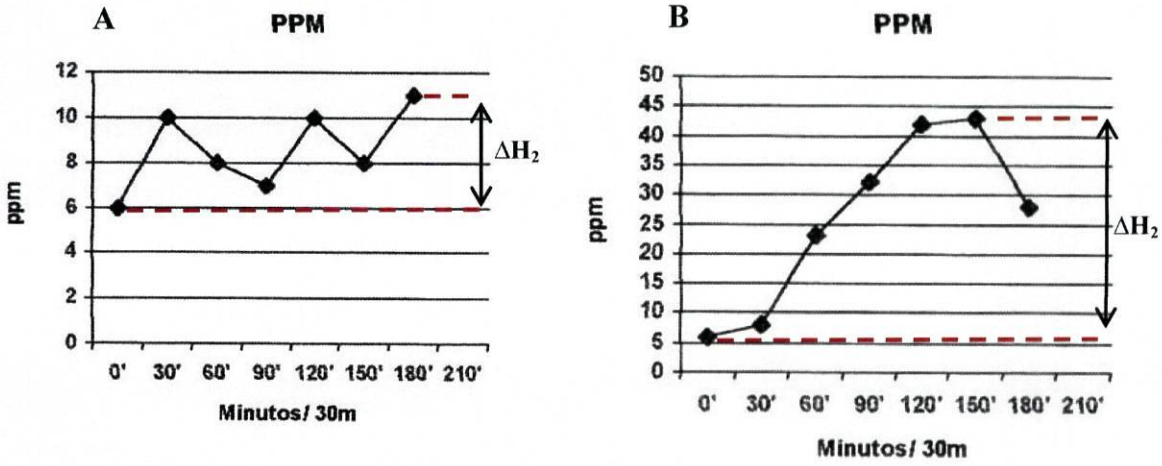

Figure 4: Variation of the breath hydrogen concentration measured during the 3hour period that ensued the lactose load. **A**: Individual classified as lactase persistent (Δ [H_2] < 20ppm); B: Individual classified as lactase restrictor(Δ [H_2]> 20ppm).

C/T -13910

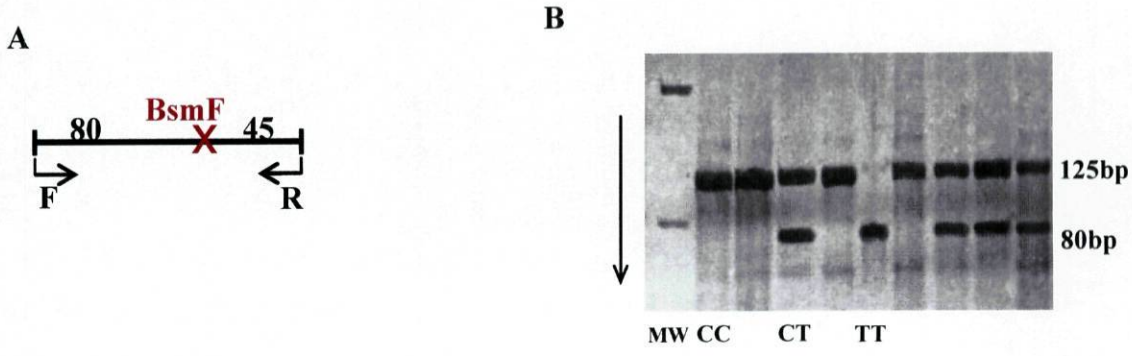

Figure 5: A- Position of the polymorphic BsmFI restriction site within the 125bp fragment containing the -13.91kb C/T variation; **B**- Electrophoretic separation of the fragments resulted from BsmFI digestion in 9 individuals. The 45bp band is not shown. MW- molecular weight marker.

A/G -22018

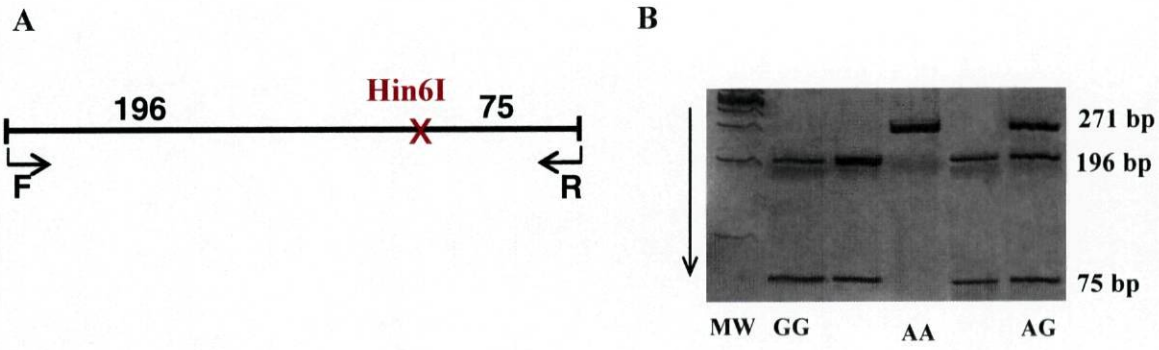

Figure 6: A- Position of the polymorphic Hin6I restriction site within 271 bp fragment containing the -22.01kb G/A variation; **B** – Electrophoretic separation of the fragments resulted from Hin6I digestion in 5 individuals. MW- molecular weight marker.

3.1.2 Comparison between classifications

The comparison between the results from different methods was made by estimating the following indicators: sensitivity, specificity, positive predictive value and negative predictive value. Table 1 synthesizes all the estimated values.

Table 1: Comparison of the results from the different tests. In each column a specific test is taken as the "gold standard".

N=56*		"Gold Standard"			
		-13910	**-22018**	**BH2 test**	**Symptoms**
Sensitivity	-13.91		1,00	0,969	0,765
	-22.01	0,971		0,938	0,735
	BH2 test	0,912	0,909		0,794
	Symptoms	0,765	0,758	0,844	
Specificity	-13.91		0,957	0,875	0,636
	-22.01	1,000		0,875	0,636
	BH2 test	0,955	0,913		0,773
	Symptoms	0,636	0,609	0,708	
Positive predictive value	-13.91		0,971	0,912	0,765
	-22.01	1,000		0,909	0,758
	BH2 test	0,969	0,938		0,844
	Symptoms	0,765	0,735	0,794	
Negative predictive Value	-13.91		1,000	0,955	0,636
	-22.01	0,957		0,913	0,609
	BH2 test	0,875	0,875		0,708
	Symptoms	0,636	0,636	0,773	
Concordance	-13.91		0,982	0,929	0,714
	-22.01	0,982		0,911	0,696
	BH2 test	0,929	0,911		0,786
	Symptoms	0,714	0,696	0,786	

* From the original 68 cases, only 56 were typed for all criteria.

Lactase-persistent individuals are conventionally described as "negative" and the restrictors as "positive".

The lack of sensitivity of a test will be reflected in the presence of false negative (FN) individuals (wrongly considered to be lactase persistent), and the lack of specificity will be reflected in the presence of false positive (FP) individuals (wrongly considered to be non-persistent).

Symptoms present the weakest correlation in relation to all other tests (Table 2). The unreliability of the symptoms to predict the lactase persistence status is widely recognized. Besides the capacity to digest the lactose, there are many other functional reasons that explain the presence or absence of symptoms during the BH2 test, including the individual threshold response to intestinal gas, intestinal flora or gastrointestinal transit time (Flatz, 1987).

A high correlation is observed between the two SNPs: in all but one individual there was concordance between -13.91kbC/T and -22.01kbG/A diagnosis (Table 2). The discordant case was considered to be lactase persistent with the -22.01kb and lactase restrictor with the -13.91kb polymorphism. The analysis of the results obtained in the same individual by using the breath hydrogen test and the record of the symptoms gives support to the positive diagnosis of the -13.91kb SNP and suggest that the -13.91kbC/T is more strongly associated to the lactase persistent trait than the -22.01kbG/A variant. This observation is in agreement with the recent results of Enattah *et al* (2002) who reported a higher association between the -13.91kb SNP and the lactase persistent trait (as assessed by direct lactase assay from jejunal biopsy) than with the -22.01kb.

It may be useful to interpret the patterns of discrepancy between the two SNPs and the BH2 test in a phylogenetic context. Figure 7 presents a schematic representation of the hypothetical evolutionary relationship between the haplotypes defined by -13.91kb and -22.01kb polymorphisms. The genotype of the discordant case (-13.91kbCC/-22.01kbGA) would correspond in Figure 7A to the combination of haplotypes (a) and (b). The (a) and (c) haplotypes are more commonly found. The (a) haplotype is highly associated with lactase restriction and the (c) haplotype is highly associated with the lactase persistence. In an evolutionary perspective, the (b) haplotype seems to be intermediate. The association between the -22.01kb*A allele and the lactase persistence is disrupted in this haplotype, giving rise to a false negative and decreasing the sensitivity of this molecular marker.

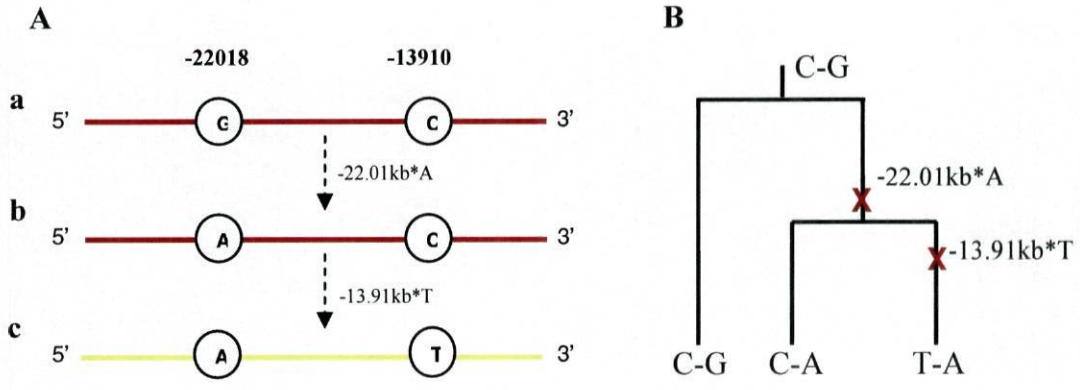

Figure 7: A- Schematic representation of the possible evolutionary steps involved in the origin of the three haplotypes formed by the combination of the -13.91kbC/T and 22.01kbG/A SNPs. The red line corresponds to the putative lactase restrictor haplotypes and the yellow line to the lactase persistent one. **B-** Tree showing the possible genealogical relations among the three haplotypes formed by the combination of the -13.91kbC/T and 22.01kbG/A SNPs.

In this respect it is interesting to note how the temporal order of persistence-associated mutations determines the intensity of this association and how it is reflected in some indicators of concordance. If the older mutation is taken as the golden-rule, the sensitivity and negative-predictive value of the younger mutation should be 100%. Conversely, if the younger mutation is the golden-rule, it is the specificity and the positive predictive value that should be 100%.

Figure 8 depicts the results of the BH2 test in individuals classified as lactase persistent/restrictor with the molecular classification based on the -13.91kb polymorphism, which is the SNP that presents the highest correlation with the physiological method.

The interpretation of the causes of discrepancies between the two tests could give further insights into the genetic and physiologic understanding of the lactase persistence polymorphism. Mulcare *et al*, 2004 combined the results of five studies and estimated to the breath hydrogen test the following rate errors: FN=9/132 (0,06818) and FP=5/120 (0,04166). These values are not very different from the values found in this study: FN=3/56 (0,0535) and FP=1/56 (0,01785).

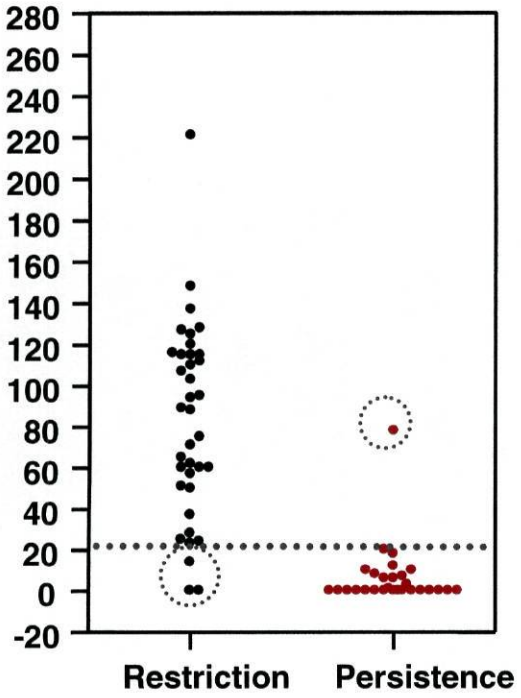

Figure 8: Expired maximum increase in H_2 concentrations during the breath hydrogen test in individuals classified as lactase persistent/ restrictor with -13.91kb polymorphism. The horizontal dotted line indicates the borderline value (20ppm). Incompatible results are shown with a dotted circle.

If the discrepancies are attributed to the recognized limitations of the BH2 test, then the occurrence of false negatives (insensitivity) can be explained by several factors. The type of colonic bacteria could influence the quantity of hydrogen detected. For example, hydrogen-utilizing flora producing methane and colonic flora unable to produce hydrogen (Arola, 1988) can lead to an underestimate of the hydrogen production and would require a parallel methane measurement and a breath test with lactulose, respectively. There are also certain antibiotics (e.g. metronidazole) that suppress the colonic hydrogen formation (Flatz, 1987). On the other hand, lack of specificity (presence of false positives) of the breath hydrogen test has also some possible justifications: congenital lactase deficiency, secondary lactose malabsorption due to epithelial damage or the increase of hydrogen excretion caused by certain antibiotics (e.g. neomycin) (Flatz, 1987).

However, if the discordant results persist after the exclusion of the conditions associated with the insensitivity/inespecificity of the breath hydrogen test, alternative genetic explanations could be claimed, and the BH2 test should be otherwise considered the golden rule.

Figure 9 presents a schematic representation of the genetic implications of the discordant results between the BH2 test and the -13.91kb molecular test. The presence of

false positives in the molecular test could be explained if the lactase persistence is caused by other mutations besides the -13.91kb*T. Alternatively, if we assume that the -13.91kb polymorphism is not causal, false positives would be present if the -13.91kb*T allele arises after the causal mutation (Fig.9B). On the other hand, false negatives could be an indication that the -13.91kb*T occurred before the causal mutation (Fig.9C).

The three scenarios illustrated in figure 9 are mutually exclusive. If we assume that the -13.91kb*T mutation is causal (Fig.9A) there is no genetic explanation to the presence of the false negatives. If the -13.91kb*T is just a marker for lactase persistence (figures 9B and 9C), the simultaneous observation of false positives and false negatives is not expected unless recombination is considered. Figure 10 illustrates the implications of recombination in the case of non-causality of the -13.91kb*T mutation. In the two scenarios illustrated in figure 10, there are two major haplotypes: one containing the -13.91kb*C allele and without the causal mutation and other containing the -13.91kb*T allele and the causal mutation. Recombination between these two haplotypes leads to the occurrence of both false negatives (*FN*) and false positives (*FP*).

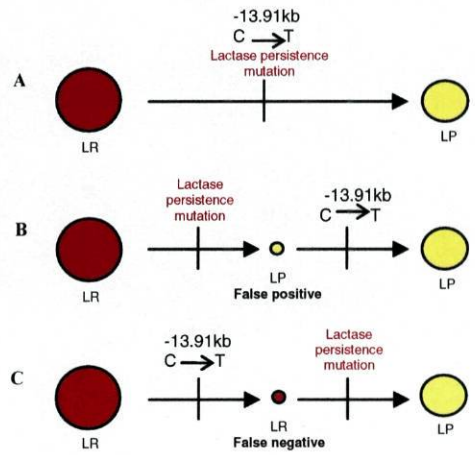

Figure 9: Schematic representation of the genetic implications of the discordant results between the BH2 test and the -13.91kb molecular classification. **A** - The -13.91kb*T is the causal mutation of the lactase persistence; **B** - The -13.91kb*T mutation postdates the causal event; **C** - The -13.91kb*T mutation antedates the causal event. LR- lactase restriction; LP- lactase persistence.

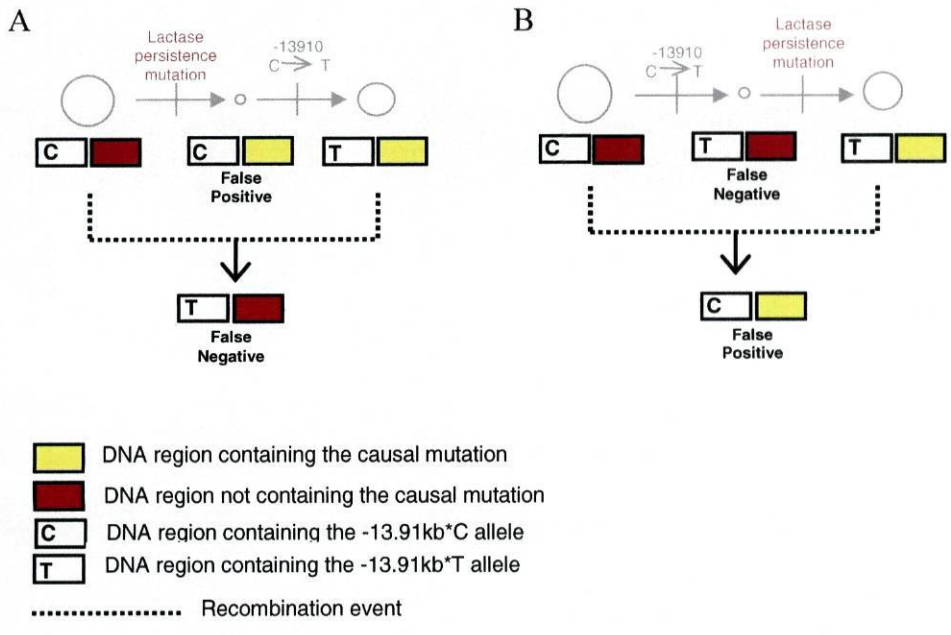

Figure 10: Schematic representation of the implications of the recombination in the scenario of a non-causality of the -13.91kb*T mutation. **A** - The -13.91kb*T mutation postdates the causal event; **B** - The -13.91kb*T mutation antedates the causal event.

Recently, Mulcare *et al* (2004) reported a very low frequency of the T allele in African populations that have high prevalences of lactase persistence according to physiological tests. The genetic consequences of this observation are illustrated in figure 11. If non-causality of the -13.91kb*T is assumed (Fig.11A) the causal mutation is likely to have occurred before the -13.91kb*C/T mutation. In this context, the high association between the -13.91kb*T and the causal mutation would occur only in some regions and would be less useful as a predictor of the lactase persistence trait in Africa. The possibility of occurrence of at least two different causative mutations (Fig.11B), one in Africa and other in Europe, is also plausible since recent functional studies have suggested that the -13.91kb*T allele might be indeed the causative factor directly involved in the enhancement of LCT transcription (Olds and Sibley, 2003; Troelsen *et al*, 2003).

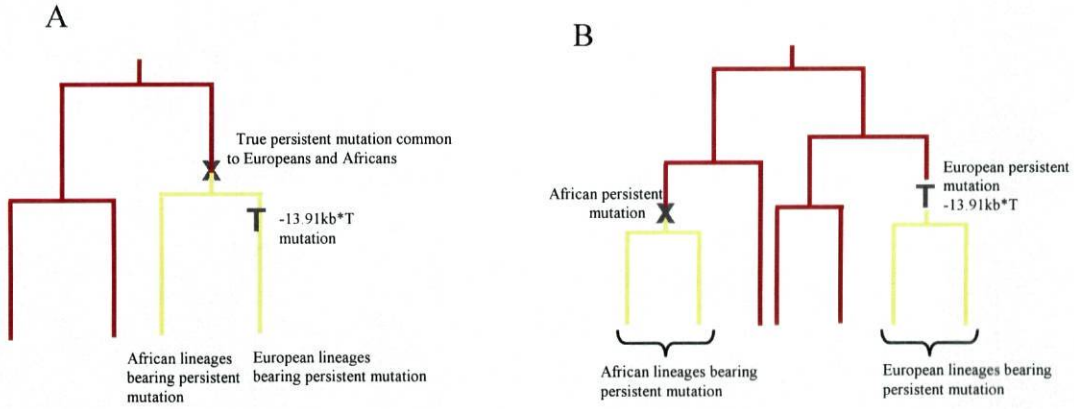

Figure 11: Phylogeograhic implications of different hypothesis about the causative role of the -13.91kb*T mutation in lactase persistence that may explain why this mutation is absent from African populations with high frequencies of lactase persistence. **A-** If -13.91kb*T is not the causal mutation then the absence of the mutation in African populations with high frequency of lactase persistence can be explained by assuming that the real lactase persistent mutation occurred before -13.91kb*T. **B-** If -13.91kb*T is the causal mutation, then there is genetic heterogeneity (lactase persistence in Europe and Africa are caused by different mutations).

Taken together, these results show that important inferences about the evolutionary history of the lactase persistence polymorphism can be made by comparing the classical physiological tests with molecular tests based in the presence/absence of candidate mutations. However, to completely clarify the discrepancies found, further studies comparing the different tests against the more reliable intestinal disaccharidases assay in different populations are required. This is an important goal also for applied purposes. The implementation of a genetic test for lactose intolerance will be useful for health practitioners as a screening tool in differential diagnosis that may substitute or complement more tedious and time consuming physiological tests. An accurate diagnosis of lactose intolerance is very important to exclude some serious gastrointestinal disorders (e.g. functional bowel disorders, dysmotility-type dyspepsia) with symptoms that resemble those experienced by lactase restrictor individuals. On the other hand, the combined use of the -13.91kb molecular test and the breath hydrogen test may represent an additional tool in differentiating the lactase non-persistence from secondary lactase deficiency. Finally, the information on the prevalence of lactose intolerance from population surveys may assist in programming targeted nutritional interventions and public health educational initiatives.

3.2 Haplotype diversity and evolution of the lactase persistence polymorphism

The analysis of haplotype diversity associated with the lactase persistence candidate mutation(s) is essential to have a full appreciation of the major factors that might have influenced the origin, evolution and spread of this trait. Due to their high mutation rates, microsatellites may be markers of choice to study the recent evolutionary episodes that influenced the distribution of lactase persistence, providing complementary information to studies based on SNP analysis, especially in situations where recombination might have been suppressed.

In this section we present an analysis of the genetic variation based on the assessment of microsatellite variation within core haplotypes defined by the -13.91kb C/T and -22.01kb G/A. The analysis addresses three major points. First, the distribution of the SNP -13.91kb and -22.01kb haplotypes was assessed in different populations. Second, different methods, based on the microsatellite variability linked to a specific allele, were used to estimate the age of the -13.91kb mutation. Finally, the role of natural selection in shaping the distribution of the -13.91kb*T mutation allele was assessed by analysing the compatibility between the levels of variation associated with the mutation and its frequency in each population.

3.2.1 Frequency of the SNP haplotypes in different populations

The frequencies of the core haplotypes defined by the -13.91kb C/T and -22.01kb G/A SNPs and the expected prevalences of lactase persistence in different populations are shown in Table 2.

The frequencies of lactase persistence estimated on the basis of the -13.91kb*T allele frequencies vary considerably between populations: 60% in Portugal, 24% in Italy, 7,8% in São Tomé, 4,0% in Mozambique and 38% in Fulbe. The estimate from North Portugal is within the frequency range previously reported for samples from southern France and Northern Spain on the basis of physiological tests (Flatz, 1987; Leis *et al* 1997; Swallow, 2003). The value found for the Italian sample is similar to previous estimates from the southern part of the country and lower than in samples from northern Italy, confirming previous observations on a high level of geographic microdifferentiation among Italian populations (Flatz, 1987; Swallow, 2003). The estimates from São Tomé and Mozambique are within the range observed for the majority of African non-pastoralists populations (Flatz,

1987; Swallow, 2003). It is likely that lactase persistence in these two populations is due to recent admixture with Europeans. In S. Tomé, for example, the frequency of the -13.91kb*T allele is very close to that expected from a reported 11% level of admixture with the Portuguese colonists (Tomás *et al*, 2002). Finally, the 0.21 frequency of the -13.91kb*T allele in the Fulbe is higher than a previous 0.11 estimate in another sample from Cameroon (Mulcare *et al,* 2004), but this difference is not significant (p=0.09) according to the exact test of population differentiation of Raymond and Rousset (1995) implemented in the Arlequin 2.1 software (Schneider *et al*, 2000). Both values are consistent with previous estimates of lactase persistence in the Fulbe based on physiological tests, which range from 29% in Nigeria (Kretchmer *et al*, 1971) to 100% in Senegal (Arnold *et al*, 1980).

Since, as previously shown (Poulter *et al*, 2003; Swallow, 2003), the -13.91kb and -22.01kb polymorphisms were originated according to a C-G→ C-A→T-A phylogenetic sequence, the low frequency of C-A intermediate haplotypes indicates that the -22.01kb G→A mutation might have occurred only shortly before the -13.91kb C→T mutation. As previously discussed (see 3.1.3), it is the occasional occurrence of this C-A haplotype that may lead to the wrong identification of lactase persistence on the basis of the -22.01kb genotyping.

Table 2: Frequencies of the haplotypes defined by -13.91kb C/T and -22.01kb G/A polymorphisms and predicted prevalences of lactase persistence in the different populations.

Haplotype		Populations				
-13.91kb	-22.01kb	Portugal (N=90)	Italy (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[a]		0.62	0.24	0.38	0.08	0.02

[a]Frequency of -13.91kb CT + TT genotypes assuming Hardy-Weinberg equilibrium

3.2.2 Microsatellite variation within the SNP haplotypes

Figures 11 and 12 present the patterns of the D2S3013/D2S3015 and D2S3010/D2S3016 duplex reaction products, respectively.

In figure 13, the microsatellite allele frequency distributions within the common C-G and T-A -13.91kb/-22.01kb SNP core haplotypes are shown for a pooled sample that combines the data from all populations. Figure 14 shows equivalent distributions in each population. Allele frequency distributions in Finland were retrieved from the original data from Enattah *et al* (neither CG haplotypes nor D2S3010 data were available). Differences related to the pattern of variation observed across microsatellite loci are observed.

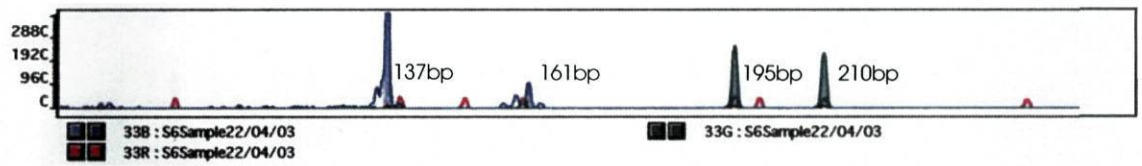

Figure 11: Pattern obtained after the separation of an amplification product of the D2S3013/D2S3015 duplex PCR in the automated genetic analyzer ABI PRISM® 310. In blue (D2S3013): heterozygote 137/161 bp; in green (D2S3015): heterozygote 195/210 bp.

Figure 12: Pattern obtained after the separation of an amplification product of the D2S3010/D2S3016 duplex PCR in the automated genetic analyzer ABI PRISM® 310. In green (D2S3016): homozygote 150/150 bp; in blue (D2S3010): heterozygote 220/236 bp.

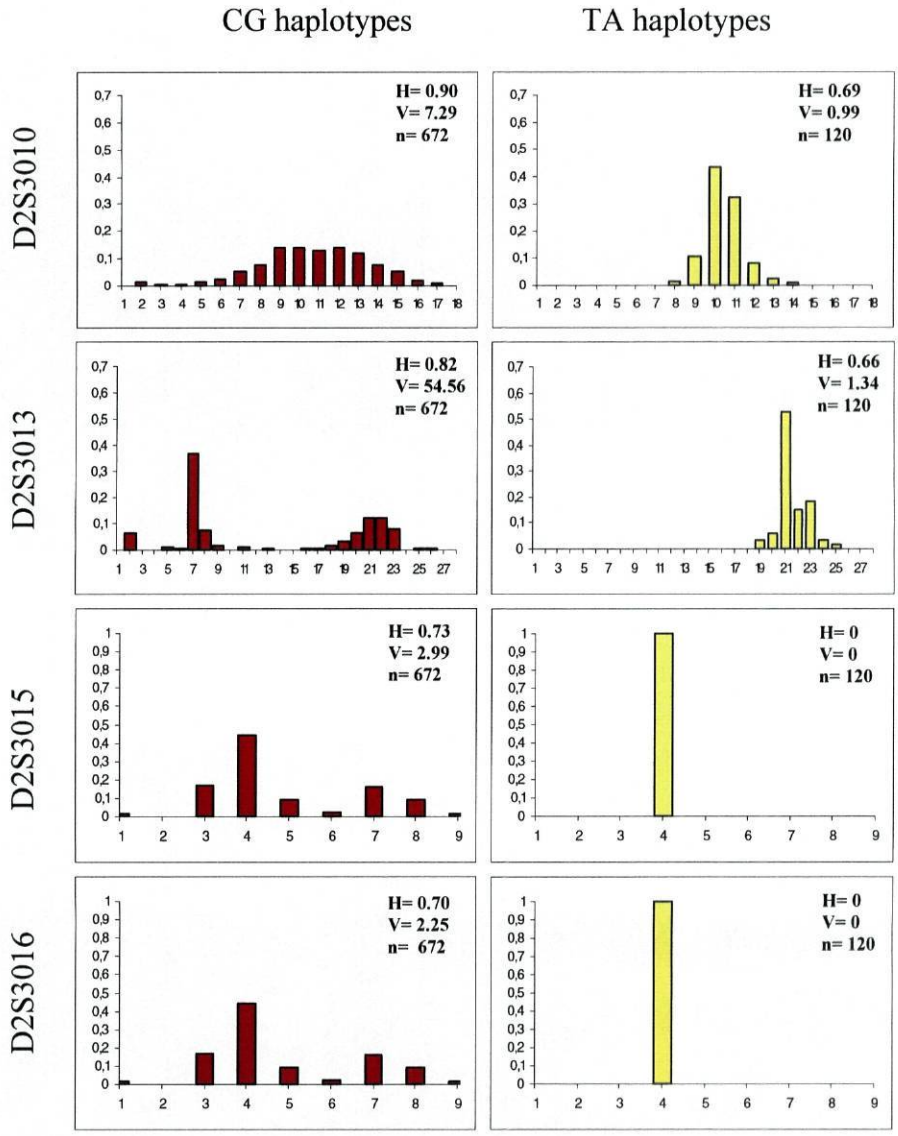

Figure 13: Microsatellite allele frequency distributions within C-G and T-A -13.91kb/-22.01kb SNP haplotypes in a pooled sample with populations from São Tomé, Portugal, Italy, Mozambique and Cameroonian Fulbe. The estimated sizes of allele 1 in each microsatellite are: 188 bp for D2S3010; 133 bp for D2S3013; 190 bp for D2S3015 and 148 for D2S3016 (H, heterozygosity; V, variance in repeat number; n, number of chromosomes).

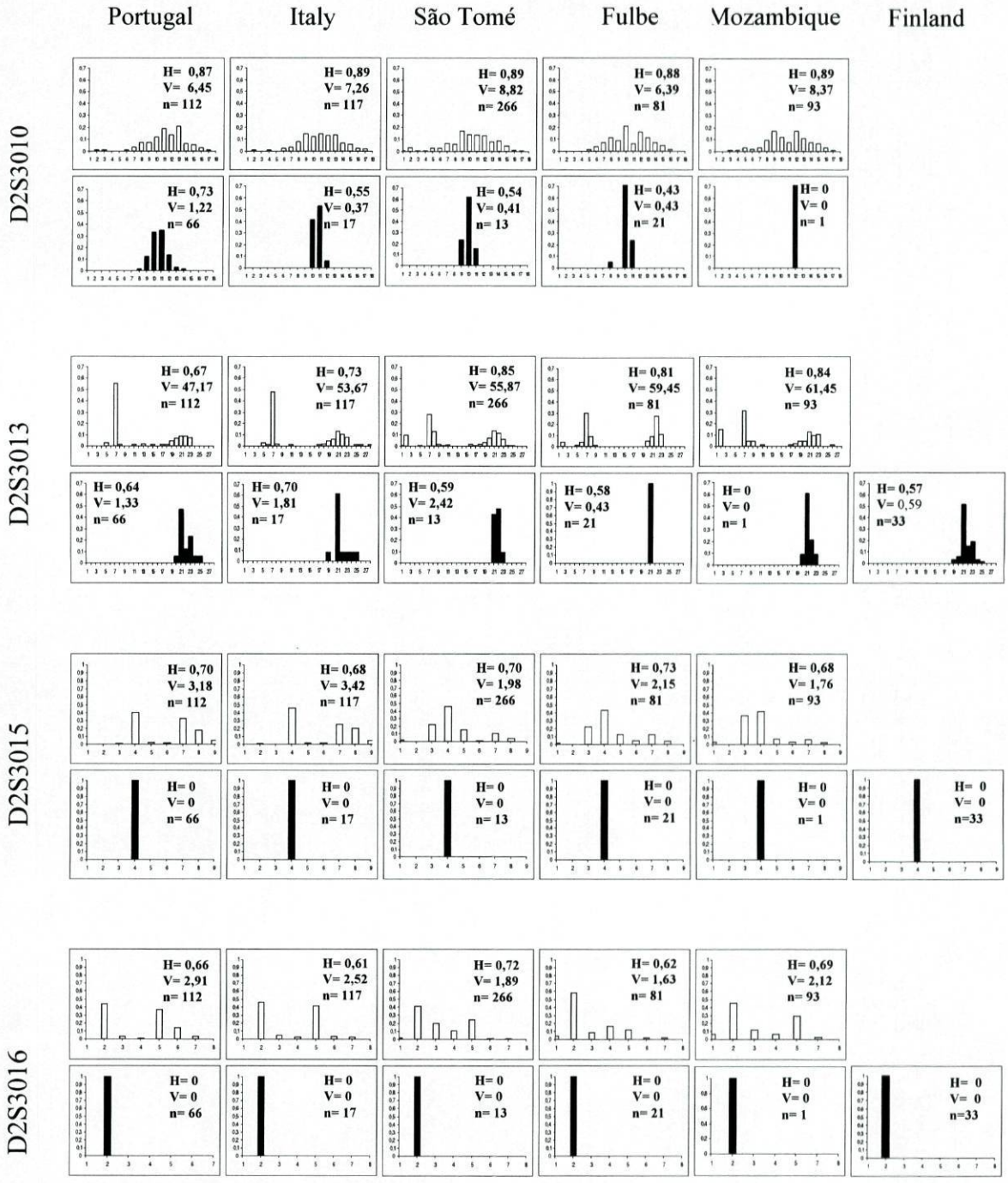

Figure 14: Allele frequency distributions of the D2S3010, D2S3013, D2S3015, D2S3016 loci within the C-G (in white) and T-A (in black) SNP haplotypes from Portugal, Italy, São Tomé and Fulbe (Cameroon), Mozambique. The estimated sizes of allele 1 in each microsatellite are: 188 bp for D2S3010; 133 bp for D2S3013; 190 bp for D2S3015 and 148 for D2S3016. (H: heterozygosity; V: variance in repeat number; n: number of chromosomes).

A clear reduction in microsatellite variation is observed within the T- A haplotypes suggesting a relatively recent origin for the -13.91kb C → T mutation. This result is consistent with the low extended haplotype diversity associated with the -13.91kb*T and -22.01kb*A alleles observed in previous studies based on SNP variation (Poulter *et al*, 2003; Bersaglieri *et al*, 2004). The higher diversity accumulated in D2S3010 and D2S3013 implies that these loci have higher mutation rates than D2S3015 and D2S3016 (Fig.13).

The analysis of each population separately gives additional information about the role of the demographic history in the pattern of variation observed in the global distributions. It is clear that the -13.91kb C → T mutation had a single origin, since T-A chromosomes cluster together despite their geographic location when the microsatellite allele frequencies among SNP core haplotypes from different populations are compared (Figs.14 and 15A).

A Median Joining network relating the compound SNP-microsatellites haplotypes in the pooled sample is shown in figure 15B. The network has two main branches that reflect the bimodality of the D2S3013 microsatellite allele frequency distributions within C-G core haplotypes (Figs 13 and 14). In contrast with the high variability associated with C-G chromosomes, T-A haplotypes are tightly clustered within one of the two main branches as expected from a unique, relatively recent, origin.

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. 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.13). If recombination had played a major role in the generation of the 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 (Fig. 14). Due to a less clear difference between the shape of the D2S3010 microsatellite allele frequency distributions among C-G and T-A haplotype distributions, it is more difficult to evaluate the role of recombination in the regeneration of diversity in this locus. As previously noted by Poulter *et al* (2003), the relative weights of mutation and recombination in the generation of microsatellite diversity may have important implications for the identification of potential candidate regions where a causal mutation for lactase persistence may lie. For example, Enattah *et al* (2002) considered that variation at the

D2S3013 locus in chromosomes associated with lactase persistence was caused by recombination and used this marker to delimit a 47kb region where -13.91kb C → T was subsequently identified as the best candidate mutation. However, since D2S3013 diversity seems to have been mainly originated through mutation, it is likely that the candidate region is longer than 47kb, implying that the -13.91kb*T allele may just be a marker closely associated with the true, as yet unknown, causal mutation for lactase persistence (Poulter *et al*, 2003).

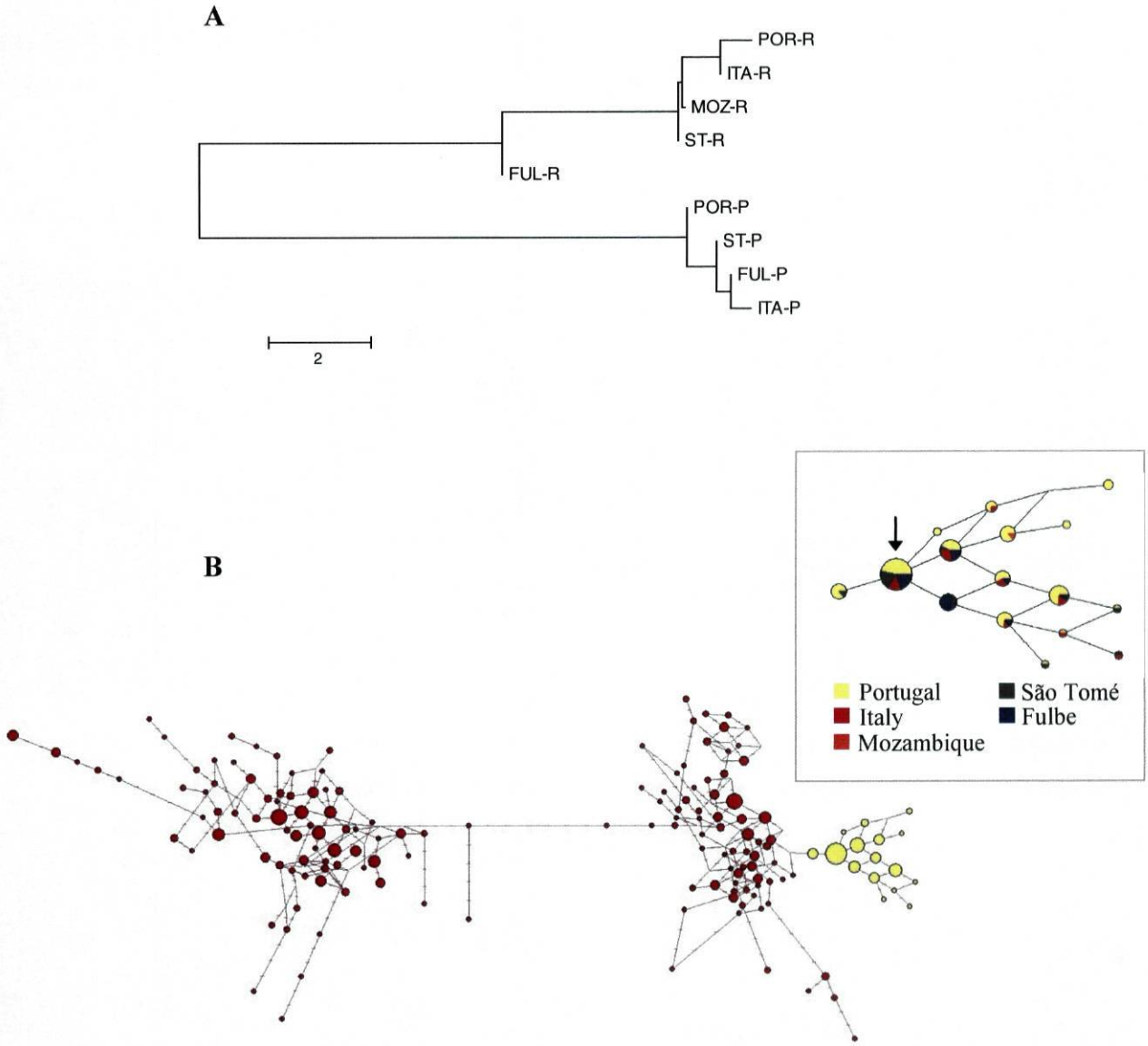

Figure 15: A. Neighbor-joining tree based on allele frequencies of four microsatellites within T-A and C-G -13.91kb/-22.01kb SNP haplotypes from different populations. Allele frequencies were compared by the $(\delta\mu)^2$ genetic distance (Goldstein *et al*, 1995) calculated with the program MICROSAT (http://hpgl.standford.edu/projects/microsat/). POR= Portugal; ITA= Italy; FUL= Fulbe; ST= São Tomé; MOZ= Mozambique. R= lactase-restriction-associated C-G haplotypes; P= lactase-persistence associated T-A haplotypes. **B.** Median-joining network (Bandelt, Forster and Rohl, 1999) representing the compound SNP- microsatellite haplotype variation in a pooled sample with populations 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 the T-A chromosomes haplotypes through the different populations is shown in inset. Haplotypes are represented by circles, with areas proportional to the number of individuals harbouring 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 4.1.0.8, 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.

3.2.3 Estimation of the age of the -13.91kb*T allele

The diversity accumulated in microsatellite loci was used to estimate the age of the -13.91kb C/T candidate mutation for lactase persistence.

Table 3 presents the estimates of the TMRCAs of the -13.91kb*T allele calculated in different samples under various sets of assumptions regarding mutation and recombination rates (m1, m2, m3 and m4). Calculations for the Finnish sample were performed with data taken from Enattah *et al* (2002) and do not include locus D2S3010. Although the confidence intervals (CIs) are wide, there is a good agreement between the different methods for each set of assumptions. In general the average square distance method leads to higher age estimates than calculations based on the accumulated variance (V) or the decrease in frequency of the modal microsatellite allele (p). This discrepancy is particularly noticeable in the Italian sample and may be due to the inadequacy of the assumptions made by the V and p methods in this sample. Confidence intervals for the average square distance method (Δ), which contrary to central estimates do depend on demography (Stumpf and Goldstein, 2001), were calculated under no population growth and may be too wide. If demographic models assuming star-like genealogies are used, the 95% CIs are substantially reduced (results not shown). Calculations based on the accumulated variance under constant population sizes lead to higher TMRCAs than under rapid growth. However, comparison of estimates based on different constant population sizes (with a minimum Ne of 5000) and rapid growth shows that the assumption of a growing population leads at most to a 10% decrease in the age estimation (data not shown). This suggests that the -13.91kb*T allele is recent enough for a strong deviation from a linear relation between age and variance to be observed.

In the pooled sample, estimates that do not take recombination into account (m1 and m2) are within 45000- 30000 or 17500-11750 year ranges, depending on the use of indirect or direct estimates of microsatellite mutation rates, respectively. If we assumed that both recombination and mutation contribute to the intra-allelic diversity, the observed haplotype homogeneity can be only explained by a very recent origin of the -13.91kb*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: 12500 and 7500 years in the pooled sample, under assumption sets m3 and m4, respectively (table 4). The ages calculated with assumption m3 are noticeably similar to those obtained with the same method assuming the set of highest mutation rates and no recombination (assumption m2).

Age estimates in the different populations give an idea of the time passed by since the mutation arrived to each place. The Fulbe and Finnish samples present consistently lower age estimates than Portugal and Italy suggesting that the mutation reached the former populations lately. The relative position of the age estimates obtained for S.Tomé switch as different age estimates are used: the lowest, intermediate and the highest age estimates are observed by using the p, Δ and the V methods respectively. This lack of coherence may reflect the fact that, in S. Tomé, microsatellite diversity observed linked to the -13.91kb*T mutation could be the result of the haplotype diversity of the colonists rather than the step-wise accumulation of mutations.

Taken together, age estimates using different methods and assumptions have shown that the TMRCA of the -13.91kb*T allele is unlikely to be older than 45000 years, implying that the -13.91kb C⟶T mutation occurred only after the separation between European and African populations 100000-50000 years ago (Klein, 2000; Relethford, 2001). More realistic assumptions lead to 12500-7500 years estimates that place the coalescent time of the -13.91kb*T variant more near the Neolithic. This is close to the TMRCA ranges provided by Bersaglieri *et al* (2004), who obtained ages between 2188 and 20650 years for the -13.91kb*T allele in European derived samples using recombinational decay of SNP haplotype homogeneity.

Table 3: Estimates of the number of years to the most recent common ancestral of the -13.91kb*T allele (95% confidence intervals are given in parentheses)

Population	Age estimation methods							
	Average square distance Δ		Accumulated variance V[a]		Decrease in frequency of modal allele P			
	m1[b]	m2[c]	m1	m2	m1	m2	m3[d]	m4[e]
Portugal (n=66)[f]	48370 (9910-127870)[g]	18930 (3870-53620)	43450 (25870-70400)	16525 (12025-22250)	38950 (23690-75500)	15250 (11690-39440)	15560 (10000-28440)	9370 (5940-17190)
Italy (n=17)	52220 (10990-142000)	20440 (4310-57000)	36675 (22400-58025)	14050 (10375-17560)	34625 (13750-140000)	13560 (5250-54440)	135560 (5750-42190)	8315 (3440-27690)
Finland[h] (n=33)	23640 (0-88125)	9250 (0-34000)	21950 (11700-38625)	8400 (5700-12000)	20750 (9625-39875)	8125 (3750-15625)	nd[i]	nd[i]
Fulbe (n=21)	20025 (1475-60075)	7840 (575-26125)	13950 (9175-20425)	5500 (4250-7025)	16920 (7833-54130)	9310 (3625-28250)	10125 (4000-26250)	6060 (2440-16810)
São Tomé (n=13)	46750 (9625-131810)	18290 (3750-54440)	48900 (28700-80750)	18500 (13350-25125)	17560 (3940-51690)	6875 (1560-20250)	7375 (1750-19000)	4400 (1000-11940)
Pooled[j] (n=117)	44610 (10040-110000)	17460 (4125-48300)	39375 (23825-62975)	15060 (11050-20100)	30000 (21125-43690)	11750 (8250-17125)	12300 (89400-17125)	7450 (5375-10440)

a Assuming Ne=5000

b Assuming suppression of recombination and microsatellite mutation rates estimated by the indirect approach: μ1(D2S3010)=0.0009; μ2(D2S3013)=0.0005; μ3(D2S3015)=0.000095; μ4(D2S3016)=0.00011

c Assuming suppression of recombination and microsatellite mutation rates calculated from a 0.001 direct average estimate: μ1(D2S3010)=0.0023; μ2(D2S3013)=0.0013; μ3(D2S3015)=0.0002; μ4(D2S3016)=0.0003

d Mutation rates as in m1 and assuming the following recombination rates between the -13.91kb site and each microsatellite locus: r1(D2S3010)=0.0015; r2(D2S3013)=0.00016; r3(D2S3015)=0.00013; r4(D2S3016)=0.00046

e Mutation rates as in m2 and recombination rates as in m3

f n= number of chromosomes bearing the -13.91kb*T allele

g 95% confidence intervals are given in parentheses; confidence intervals for the Δ method were calculated assuming no population growth

h Based on the data from Enattah et al (2002), not including locus D2S3010

i Not done, due to unavailable distribution of microsatellite allele frequencies in the general population

j Excluding Finland

3.2.4 Neutrality test

Figures 13 and 14 clearly show a low diversity associated with T-A haplotypes. This low variability can be a casual event, within the range of expected values under neutrality (Bamshad and Wooding, 2003). However the wide distribution of the -13.91kb*T allele and its observed high frequencies in some regions raise the question of whether natural selection had also played a role in the present diversity pattern observed.

Table 4 presents the results of the neutrality tests for the -13.91kb*T allele in different samples using the method of Slatkin and Bertorelle (2001). To compare the results of the Finland population with the results of the remaining populations, the method was replicated to all populations but excluding the microsatellite D2S3010 (Table4-3STRs).

Table 4: Probabilities of finding a number of mutations $\leq S_0$ in the microsatellite loci (STRs) linked to the -13.91kb*T allele.

					Demographic models			
					D1[a]		D2[b]	
		i	**n**	**S_0**	**m1[c]**	**m2[d]**	**m1**	**m2**
4 STRs	Portugal	66	180	34	3.94×10^{-7}	1.43×10^{-37}	4.65×10^{-21}	1.85×10^{-82}
	Italy	17	134	16	0.241	3.82×10^{-7}	$8{,}62\times10^{-4}$	7.54×10^{-16}
	Fulbe	21	102	14	0.011	2.21×10^{-12}	3.86×10^{-7}	6.41×10^{-25}
	São Tomé	13	284	22	0.988	0.056	0.467783	3.24×10^{-6}
					m3[e]	**m4[f]**	**m3**	**m4**
3 STRs	Portugal	66	180	10	5.29×10^{-6}	1.94×10^{-22}	3.21×10^{-14}	1.26×10^{-43}
	Italy	17	134	10	0.725	0.005	0.080	6.53×10^{-7}
	Fulbe	21	102	4	0.013	3.73×10^{-8}	1.69×10^{-5}	1.69×10^{-5}
	São Tomé	13	284	4	0.345	0.002	0.035	1.27×10^{-6}
	Finland[g]	33	78	6	5.56×10^{-4}	1.74×10^{-13}	1.43×10^{-9}	4.711×10^{-27}

i, number of chromosomes bearing the -13.91kb*T allele; n, number of chromosomes in the sample; S_0, Double of the minimum number of mutations in linked microsatelllite loci. The four right-hand columns show the tail probabilities, $P = \Pr(Se \geq S_0)$, under the different conditions presented.

[a] Demographic model 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 10^3.

[b] Demographic model that is a variation of the scenarios simulated by Kruglyak (1999) and assumes that the effective population size increased exponentially from 10^4 to 5×10^9 starting at 900 generations ago.

[c, d] Sets of mutation rates as defined in Table4.

[e] Mutation rates as in m1 but not including locus D2S3010.

[f] Mutation rates as in m2 but not including locus D2S3010.

[g] Based on the data from Enattah *et al* (2002), not including locus D2S3010.

Neutrality is rejected at the 0.001 level for the Portuguese and Finnish samples under all assumptions (Table 4).

The results obtained vary substantially with the parameter values assumed. The demographic model "Pritchard" (which implies a lower effective size of the founding population) and the set of mutation rates obtained by the indirect approach (with slow mutation rates) are the conditions that lead to the most conservative tests.

When the 4 microsatellite loci are considered, the Portuguese sample is the only that reject neutrality for all the parameter values analysed. Apart from this population, signatures of selection on the remaining cases depend on the parameter values assumed. In the Italian and Fulbe samples, neutrality cannot be rejected under the most conservative assumption, which combines demographic model D1 with the set of lower mutation rates (m1). In São Tomé, neutrality is rejected only with the least conservative assumption, combining model D2 with higher mutation rates (m2).

The use of only 3 microsatellites decreases the power of the test. Some of the parameters that led to neutrality rejection when 4 microsatellite are used, are compatible with neutrality with this reduced set of microsatellites (Table4). Even so, Finnish and Portugal samples reject neutrality for all the parameters analysed (Table4).

In general, the results reflect the variation in the frequency of the -13.91kb*T allele in different populations (Table 2). Given the observed levels of intra-allelic variability, only in samples with frequencies ≥ 0.35 is the signal of selection robust enough for the conclusions to be insensitive to different values of key parameters.

These results, based on the microsatellite variation, agree with of the ones found by Bersaglieri *et al* (2004) who used the SNP haplotypic homogeneity linked to the -13.91kb*T and -22.01kb*A alleles, providing formal genetic evidences of selection acting on haplotypes associated with lactase persistence.

Taken together the available evidence from studies of microsatellites and SNP variation in human populations and the high correlation between lactase persistence and archaeological sites documenting early European dairying use (Beja-Pereira *et al*, 2003) strengthen the hypothesis that lifelong unrestricted use of milk provided the selective advantage that led to a pronounced increase in the frequency of lactase persistence, at least in some European populations, after the origin of dairying practices in the Neolithic.

However, since the signal of selection may be passively propagated by migration, it remains to be shown whether the current inter-population variation in lactase persistence was

caused by different levels of admixture with early pastoralists after selection in a single area, or if it represents the outcome of multiple independent selection regimes.

4. Conclusions

4.1 Concordance study

A highly satisfactory 93% level of concordance was observed between the results from the breath hydrogen (BH2) test and from the molecular test based on the -13.91kbC/T polymorphism. This concordance shows that molecular diagnosis can be a valid alternative to physiological tests in the identification of the lactase activity profile of the Portuguese population. The method is easily applicable outside the hospital, has a high level of automation and reproducibility and provides age-independent diagnostic results that are particularly suitable for large scale population studies. Furthermore, genetic testing of lactase persistence also provides a less invasive and less tedious alternative for the differential diagnosis of abdominal complaints in patients older than 5 years, which has the added benefit of being insensitive to individual variability in physiological response. Finally, the use of the molecular test in combination with the BH2 test may be important to discriminate between primary and secondary causes of lactose intolerance and to narrow the spectrum of possible diagnoses associated with abdominal discomfort.

4.2 Population distribution and evolutionary history of the lactase persistence polymorphism

4.2.1 Frequency of lactase persistence in different populations

The prevalence of lactase persistence predicted by the -13.91kb*T allele varied widely within and among the European and African populations that were studied. In the European samples, high frequency differences were observed between the Italian and Portuguese populations. The Portuguese frequencies (60%), here determined for the first time, have been shown to lie within the prevalence range previously reported for Southern France and Northern Spain. The Italian frequencies (24%) were consistent with previous studies based on physiologic tests and confirmed that Italy is among the European populations with the lowest prevalence of lactase persistence. In the African samples, lactase persistence was much more frequent in the pastoralist Fulbe population (38%) than in São Tomé (7,8%) and Mozambique (4,0%), which are not associated with dairying traditions. We conclude that the frequency of the -13.91kb*T allele is a good predictor of lactase persistence in the populations analyzed in the present study.

4.2.2 Microsatellite variation and evolutionary history of the lactase persistence polymorphism

The analysis of microsatellite variation linked with the "core haplotype" defined by the -13.91kb and -22.01kb SNPs allowed the inference of key parameters of the evolutionary history of lactase persistence.

The microsatellite allele distributions associated to the chromosomes bearing the -13.91kb*T allele were similar in the different populations suggesting that a single mutation event may be enough to explain the presence of the -13.91kb*T allele in populations geographically as distant as Finland and Cameroons.

The observation of a reduced haplotypic diversity associated to the chromosomes bearing the -13.91kb*T allele suggests that a recent event may have been in the origin of the -13.91kbC⟶T mutation. The application of different methods under various sets of assumptions regarding mutation and recombination rates showed that the TMRCA of the -13.91kb*T allele is unlikely to be much older than ~50000 years, implying that the -13.91kbC⟶T mutation occurred only after the separation between European and African populations 100000-50000 years ago. More recent estimates placed the coalescent time of the -13.91kb*T variant closer to the Neolithic, ~10000 years ago.

The observation that the -13.91kb*T intra-allelic variability was generally lower than expected, given its allele frequencies, suggests that natural selection may have played a role in the present distribution of this allele. The ability of the microsatellite approach to capture the effects of selection was confirmed in the Finnish dataset from Enattah *et al* (2002). Neutrality was further rejected in the Portuguese sample irrespective of the range of tested mutation rates and demographic models. Rejection of neutrality was less independent of demographic and mutation parameters in the Fulbe and, especially, in the Italian samples. However, the effect of migrations and other demographic processes in the geographic patterning of the -13.91kb*T should also be accounted when interpreting the results from the neutrality tests.

Taken together, these results suggest that human lactase persistence is a relatively recent trait that originated in Eurasia and support the hypothesis that the current distribution of this trait is, at least in part, an outcome of the nutritional advantage conferred by the life-long capacity of digesting large quantities of lactose in populations where milk was a critical

part of the diet. The presence of the -13.91kb*T allele in the African populations studied could be explained by introgression from Euro-Asiatic populations.

Our demonstration that a battery of only four microsatellite loci may be sufficiently informative to estimate key micro-evolutionary parameters of the history of lactase persistence highlights the significance of using these faster evolving markers to increase the efficiency of the phylogeographic studies on the ability to digest lactose.

5. References

Aoki K (2001) Theoretical and empirical aspects of Gene-Culture Coevolution. Theor Popul Biol 59: 253-261

Arnold J, Diop M, Kodjovi M, Rozier J (1980) L'intolerance au lactose chez l'adulte au Senegal. C R Seances Soc Biol Fil 174: 983-992

Arola H, Koivula T, Jokela H, Jauhiainen M, Keyrilainen O, Ahola T, Uusitalo A, Isokoski M (1988) Comparison of indirect diagnostic methods for hypolactasia. Scand. J. Gastroenterol. 23: 351-357

Arola H (1994) Diagnosis of hypolactasia and lactose malabsorption. Scand J Gastroenterol 29 (Suppl.202): 26-35

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

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

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

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, Erhardt G (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet 35: 311-313

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

Boll W, Wagner P, Mantei N (1991) Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. Am J Hum Genet 48: 889-902

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, 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 U S A 15: 6723-6727

Goldstein DB, Zhivotovsky LA, Nayar K, Ruiz-Linares A, Cavalli-Sforza LL, Feldman M (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 9: 1213-1218

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

Klein RG (2000) Archeology and the evolution of human behavior. Evol Anthropol 9: 17-36

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir 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, Ransome-Kuti O, Hurwitz R, Dungy C, Alakija W (1971) Intestinal absorption of lactose in Nigerian ethnic groups. Lancet 2: 392-395

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139-144

Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, Järvelä I (2003) Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut 52: 647-652

Leis R, Tojo R, Pavón P, Douwes A (1997) Prevalence of lactose malabsorption in Galicia. J Pediatr Gastroent Nutr 25: 296-300

McCracken RD (1971) Lactase deficiency: an example of dietary evolution. Curr Anthropol 12: 479-517

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

Olds LC, Sibley E (2003) Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet 12: 2333-2340

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM (2003) The causal element for the lactase persistence/nonpersistence polymorphism is located in a 1Mb 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 diferentiation. Evolution 49: 1280-1281

Relethford JH (2001) Genetics and the Search for Modern Human Origins. Wiley-Liss. New York

Sahi T (1994) Genetics and epidemiology of adult-type hypolactasia. Scand J Gastroenterol 29 (Suppl 202): 7-20

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) Arlequim, ver.2.000: a software for population genetics data analysis. University of Geneva, Geneva, Switzerland

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 alpha1-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: 695-71

Slatkin M (1995) Hitchhiking and associative overdominance at a microsatellite locus. Mol Biol Evol 12:473-480

Slatkin M (2002) The age of alleles. Pp 233-260 *in* M. Slatkin and M. Veuille, eds. Modern developments in theoretical population genetics: the legacy of Gustave
Malécot. Oxford University Press, New York

Slatkin M, Bertorelle G (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics 158: 865-874

Spedini G, Destro-Bisol G, Mondovi S, Kaptué L, Taglioli L, Paoli G (1999) The peopling of Sub-Saharian 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

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L. (1999) Y-chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65:1718- 1724

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

Troelsen JT, Olsen J, Møller J, Sjöström H (2003) An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology 125: 1686-1694

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

Cover photo by John Reader