UNIVERSITY OF BOLOGNA

Ph. D. Program in
" Biodiversity and Evolution"
(XIX cycle)
BIO/08 - Anthropology

MOLECULAR VARIABILITY OF LACTASE PERSISTENCE IN EURASIAN POPULATIONS

Cristina Fabbri

Coordinator:

Prof. Giovanni Cristofolini

Supervisor:

Prof. Davide Pettener

Academic Dissertation, 2007

TABLE OF CONTENTS

INTRODUCTION

1.1 Lactase persistence

The ability to digest milk lactose depends on the activity of lactase-phlorizin hydrolase (LPH) (EC 3.2.1.23/62) enzyme that is exclusively expressed in the small intestine (Mantei et al. 1988). Lactose is the major carbohydrate in mammalian milk and it has to be hydrolyzed in the monosaccharide glucose and galactose to be easily assimilated into the bloodstream and used as a source of energy (Figure 1.1). Human milk has the highest lactose content (7%), cow milk contains 4.8% of lactose.

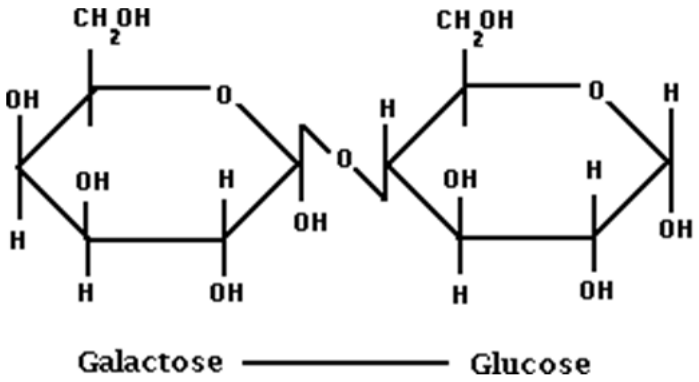


Figure 1.1 Lactose molecule composed of a galactose unit and a glucose unit

In most mammals, lactase activity is maximum after birth and remains high during infancy; it declines rapidly after the weaning phase and this down regulation determines a reduced capacity to digest lactose and has been interpreted as a shifts towards adult diet (Flatz 1987).

The ability to digest lactose is a variable genetic trait in human populations, in fact two distinct phenotypes are present: non-persistence and persistence. The ancestral lactase non-persistence phenotype is characterized by lactase activity decline and it is common to most human populations. The derived lactase persistence phenotype presents high levels of activity also during adulthood (Flatz 1987).

Lactase-persistence subjects are able to consume large amount of fresh milk without complications; on the contrary people with low levels of lactase activity may develop adverse symptoms, such as abdominal pain, flatulence and diarrhoea. This occurs because lactose is not efficiently hydrolysed in the small intestine and reaches the large intestine where it is fermented by enteric bacteria. However, many non-

persistence individuals can drink milk without experiencing symptoms associated with milk intolerance; they can adjust the dietary intake of milk to their individual tolerance threshold or they can eat milk products, like cheese and yoghurt, containing low level of lactose, or even bacteria that secrete lactase themselves. Vice versa there are some co-factors that may influence the severity of symptoms: velocity of gastric emptying, individual differences in prostaglandin synthesis, colonic bacteria and colonic irritability (Flatz 1987).

1.2 Lactase persistence distribution and adaptative hypotheses

Lactase persistence frequency varies considerably among human populations.

The phenotype distribution has been determined directly by small-intestinal biopsies or indirectly by physiological lactose tolerant tests that measure individual lactose digestion capacity; the "Breath Hydrogen Determination" test (BH_2T) measures the level of hydrogen derived from metabolism of lactose by colonic bacteria in non-persistence subjects, and the "Blood Glucose Determination" test measures the increase in blood glucose concentration, both after the oral assumption of a lactose load of 50 g. (Dahlqvist 1964; Sahi 1974; Metz et al. 1975; Arola 1994).

The persistence trait is highly prevalent among Northern Europeans, with values >90%, and gradually decreases toward the South and the East. Values are intermediate (30-70%) in the Middle East, around the Mediterranean and in South and Central Asia and low in native Americans, Pacific islanders and in Southeast Asia. Of particular interest are pastoralist groups having a milk-drink culture of sub-Saharian Africa and Arabia, which typically have higher frequencies of lactase persistence than their neighbouring non-pastoralist groups (Flatz 1987; Sahi 1994; Swallow 2003) (Figure 1.2).

The unusual geographic distribution of this trait , as well as its association with the cultural habit of consuming milk (Holden & Mace 1997), has led to the proposal of several selective hypotheses. Natural selection could has played a major role in shaping the actual persistence distribution since the development of cattle domestication during the Neolithic Transition ~7,500-9,000 years ago.

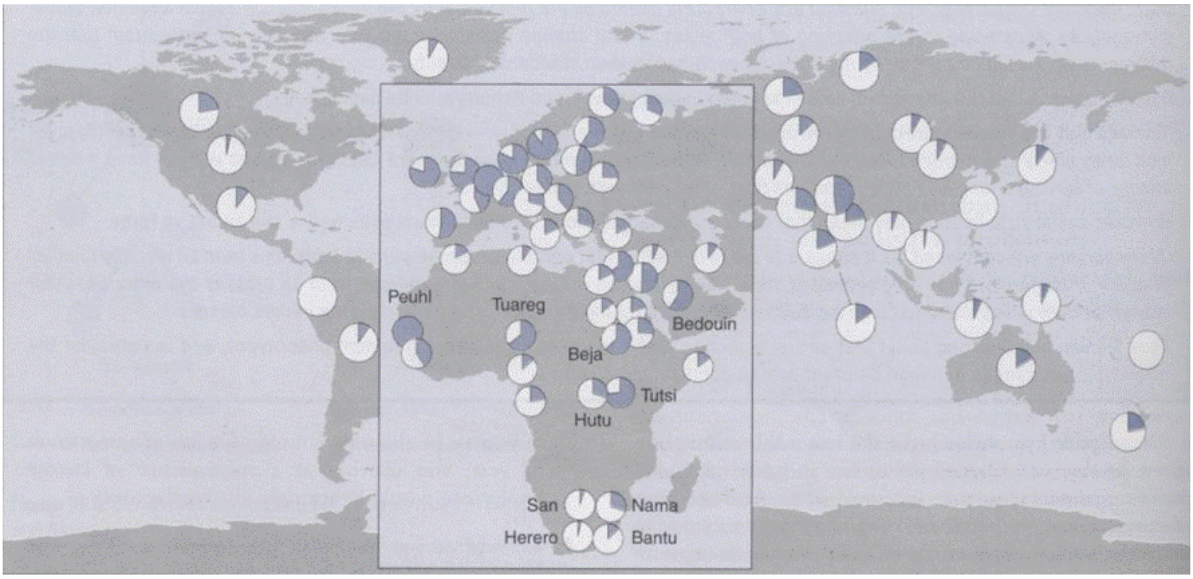


Figure 1.2 Worldwide lactase persistence (in blue) and non-persistence (in white) frequency distribution, based on physiological data (reviewed in Jobling et al. 2003)

Different hypotheses are summarised below:

1) *The gene-culture coevolutionary hypothesis*: lactase persistence distribution is the result of a recent selection pressure associated with the pastoralism advantages of the assumption of fresh milk in adulthood. Individuals with lactase persistence were able to use all the nutrients of milk throughout life, therefore, they were stronger, better equipped to survive and possibly had more children (Simoons 1969; Simoons 1970; Johnson et al. 1974; Flatz & Rotthauwe 1977).

2) *Calcium absorption*: this hypothesis has been proposed to explain the high value of lactase persistence in Northern Europe (Flatz & Rotthauwe 1973). Cause of the low solar irradiation and the low nutritional supply of vitamin D, a major absorption of calcium contained in milk could reduce rickets and pelvic deformities in persistence subjects. Nevertheless, Simoons (2001) showed that lactase non-persistence people can absorb calcium as lactase persistence one.

3) *Source of water in desert zone*: milk provides important additional fluid particularly in arid region; for some desert nomads milk may be the only source of water, for this reason lactose intolerance symptoms can be disadvantageous (Cook, 1978).

In contrast, the "*reverse cause argument*" suggests dairying was adopted precisely by those populations that could digest lactose, in this case the genetic change would have preceded the development of pastoralism (reviewed in Aoki 2001).

The *gene-culture coevolution hypothesis* was strongly supported when a high diversity in cattle milk protein genes and lactase persistence distribution was demonstrated to coincide in Europe (Beja-Pereira et al. 2003). The study found substantial geographic coincidence between high diversity between cattle milk genes and high frequency of present-day lactase persistence in the areas where the first European Neolithic milk dependent groups have been developed (>5,000 years ago) (Zvelebil 2000) (Figure 1.3).

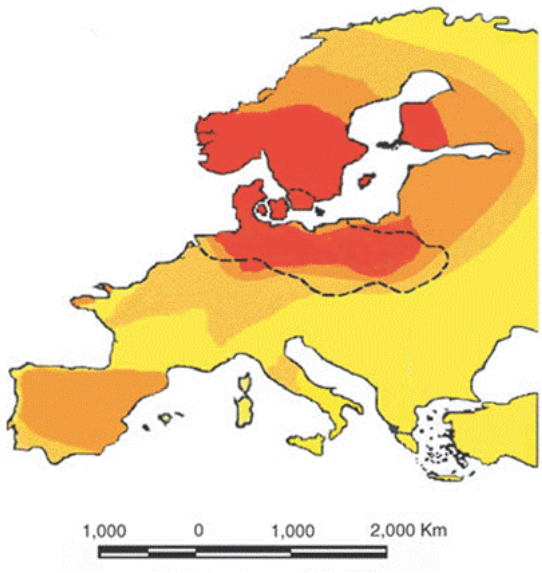


Figure 1.3 Geographic distribution of lactase persistence mutation in contemporary Europeans; the hotter the colour the highest the frequency. The dashed black line indicates the limits of the geographic distribution of early Neolithic cattle pastoralist inferred from archaeological data (Beja-Pereira et al. 2003).

On the other hand a recent study identified two genes TRPV6 and TPRV5, on chromosome 7q, that have been under strong selection pressure in Northern Europe. The fact these genes play an important role in kidney, intestine and placenta calcium absorption, together with the same lactase persistence frequency distribution, suggest that increased calcium absorption may have been the driving force behind selection for lactase persistence in Northern Europe (Akey et al. 2004)

1.3 Molecular basis of lactase persistence

Lactase persistence is inherited as an autosomal dominant trait. Activity levels of lactase in adults show a trimodal distribution pattern that suggests three genotypes: homozygotes for the persistence allele having lactase persistence, recessive homozygotes having lactase non-persistence and heterozygotes having lactase persistence but with intermediate level of enzyme activity.

Lactase enzyme is codified by LCT gene (NM_002299), mapped to the long arm of chromosome 2q21, composed of 17 exons and approximately 50 kb long (Kruse et al. 1988; Harvey et al. 1993) (Figure 1.4). There is one kb promoter region preceding LCT (Boll et al. 1991). The gene encodes an mRNA transcript of 6,274bp and the primary translation product is 1,927 amino acids long (Mantei et al. 1988).

Many single nucleotide polymorphisms (SNPs) have been discovered within LCT gene and in the surrounding region but none of them was shown to be the causative variant of lactase persistence; however four globally common haplotypes A, B, C and U have been identified (Boll et al. 1991; Lloyd et al.1992; Harvey et al. 1995; Hollox et al. 2001; Poulter et al. 2003). Haplotype A is the most common in Northern Europe (86%) and decreases toward South (36%) following lactase persistence distribution. This association is not only geographic, it has been confirmed by physiological tests and estimations of mRNA transcript in the small intestine of lactase persistence individuals that resulted belonging to haplotype A (Harvey et al. 1998; Swallow 2003).

Only few years ago, in a study of linkage disequilibrium (LD) in Finnish families, two polymorphisms were discovered upstream the lactase gene: C/T-13910 (rs4988235) and G/A-22018 (rs182549). Both substitutions are located within introns of the adjacent MCM6 gene, involved in the cell cycle, respectively 14kb and 22kb upstream from the initiation codon of LCT gene. The derived alleles T-13910 and A-22018 were found, respectively, totally and highly associated to lactase persistence phenotype (Enattah et al. 2002) (Figure 1.4). This different association depends on the genealogical relation between the two polymorphisms, that were originated according to C-G, C-A, T-A phylogenetic sequence (Poulter et al. 2003; Swallow 2003) (Figure 1.5).

Further studies demonstrated both T-13910 and A-22018 reside within a very large region of Linkage Disequilibrium (see next section for more details on LD) spanning 1Mb and particularly they occur only on the background of a very extended haplotype A (Poulter et al. 2003; Swallow 2003).

The presence of T-13910 allele on extended region of LD, makes it possible that it could be the causative mutation or it could be linked to an undetected variant located in the same region (Poulter et al. 2003; Bersaglieri et al. 2004).

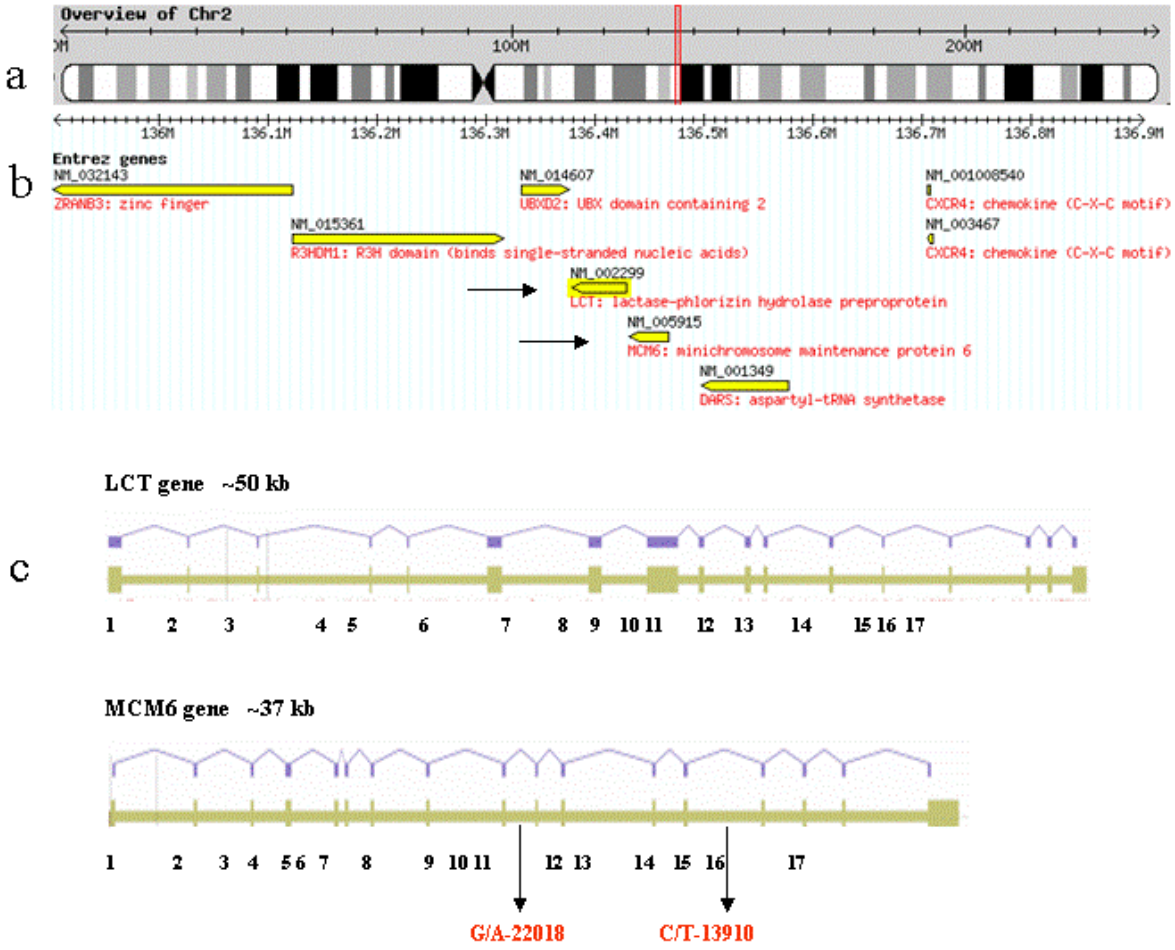


Figure 1.4 (a) Overview of chromosome 2, the position of the genes of interest is indicated in red. **(b)** Enlargement of one Mb region (chr2:135903814..136903813) containing LCT, MCM6 and other neighbour genes; for each gene the complete name, the NM reference number, and the direction of transcription (yellow arrows) are reported. This picture is from the Hapmap site, www.hapmap.org, release 21a/phasell Jan2007. **(c)** LCT and MCM6 gene structures; the number indicate the exons and the above violet line the final protein. These pictures are from http://genewindow.nci.nih.gov . The positions of the two mutations, C/T-13910 and G/A-22018 associated to lactase persistence, are indicated by black arrows; they lie respectively within the 13rd and 9th introns of MCM6 gene.

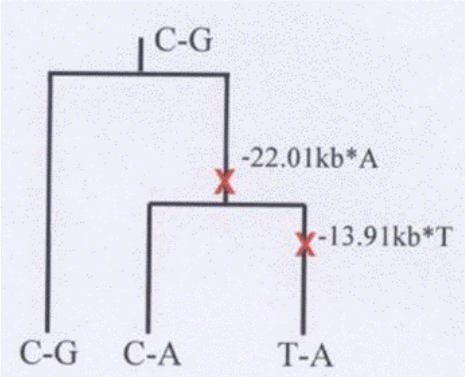


Figure 1.5 Possible genealogical relations among the three haplotypes formed by the combination of C/T-13910 and G/A-22018 polymorphisms (picture from Coelho et al. 2005).

1.4 Genetic evidence of selection

Both population demographic history and natural selection can shape patterns of DNA variation, but the former affects patterns of variation at all loci in the genome in a similar manner, whereas natural selection acts upon specific loci (Cavalli-Sforza 1966; Nielsen 2001).

The important role of natural selection in shaping lactase persistence distribution has been confirmed by genetic analyses.

Natural selection leaves signatures on the selected gene and its surrounding region (Ronald & Akey 2005; Sabeti et al. 2006), such as a large allele frequency distribution among populations, determined by differential pressures of selection that cause alleles to rise dramatically in frequency in some but not all of the populations. Large differences in lactase persistence distribution were previous shown by phenotype data and more recently confirmed by C/T-13910 and G/A-22018 allele distribution and analyses of measures of population differentiation (Flatz 1987; Bersaglieri et al. 2004).

Also a common haplotype that remains intact over unusually long distances, indicates the presence of a selected allele; in response to strong positive selection the selected allele rises rapidly to high frequency and the frequency of the haplotype on which the allele occurs will increase too, because there is insufficient time for recombination to disrupt the haplotype. Many studies confirm LCT gene lies in a region with extensive Linkage Disequilibrium (LD). LD is the non-random association of alleles at two or more loci; combinations of alleles or genetic markers in LD occur more frequently in a population than would be expected from a random formation of haplotypes based on allelic frequencies (review in Wall & Pritchard, 2003).

Harvey et al (1995) identified strong linkage disequilibrium across a 70kb haplotype spanning the lactase gene and Poulter et al (2003) reported the existence of 1Mb stretch of LD, in the same region, containing the T-13910 and A-22018 alleles. Moreover, Bersaglieri et al (2004) identified a region spanning >800kb characterized by long stretch of homozigosity that confirms a recent origin for the haplotype containing the T-13910 and A-22018 alleles, because there where no time for recombination to act.

Age estimation based on the decay of LD in either directions from the LCT core region and intragenic recombination, that occurs as a function of time, established the T-13910 allele began to rise in frequency between 2,188-20,650 years ago in American-Europeans and more recently, between 1,625-3,188, in Scandinavian populations (Bersaglieri et al. 2004). These estimations suggest the mutation has appeared after the separation between Euro-Asians and Africans and are consistent with the absence of the T-13910 allele in Africa (Bersaglieri et al. 2004; Mulcare et al. 2004). The more recent estimation in Northern Europe is consistent with the origin of farming in this area <9000 years ago (Simoons 1970; Kretchmer 1971; Scrimshaw & Murray, 1988). A recent estimation based on microsatellite diversity within the persistence lineage confirms an age of the T-13910 allele between 7,000 and 12,000 years (Coelho et al. 2005).

The strong selection pressure was estimated to have acted during the past 10,000 years, 400 generations, coinciding with the dairy culture development (Bersaglieri et al. 2004; Coelho et al. 2005; Myles et al. 2005; Enattah 2005, unpublished data). The selection power was calculated to be between 14-15 % in European derived populations and between 9-19% in Northern Europe (Bersaglieri et al. 2004) in agreement with previous estimations (1-7%) (Cavalli-Sforza 1973; Flatz & Rotthauwe 1977; Aoki 1986; Flatz 1987).

Coelho et al. (2005) estimated the role of natural selection in the rapid increase in frequency of lactase persistence using intra-allelic variability at microsatellite loci under a wide range of demographic models. The probability values to find the observed intra-allelic diversity under neutrality resulted very low.

1.5 Lactase regulation

It has been suggested that the regulation of LCT gene expression involves both transcriptional and posttranscriptional controls. However in the majority of cases, the LCT mRNA level have been shown to correlate with lactase activity, so the transcriptional regulation is probably the most important factor affecting the level of the enzyme (Escher et al. 1992; Wang et al. 1995; Rossi et al. 1997; Wang et al. 1998). It has been already demonstrated that lactase persistence is caused by cis-acting element located closely to LCT gene in European populations (Wang et al. 1995).

Conservative analyses of the lactase gene promoter revealed a well conserved 150bp region in rat, mouse, rabbit, pig and human, containing bind transcription factors (CE1a, CE2c and GATA-site) very important for lactase expression (reviewed in Troelsen, 2005). Moreover, the activity of this region is modulated by transcriptional enhancers and silencers located upstream the lactase gene (Olds & Sibley 2003; Troelsen et al. 2003). Even if distal regulatory regions are not conserved in pig, rat and human the same type of pattern does exist.

The discovery of polymorphisms associated to lactase persistence in Euro-Asian populations led to great developments in understanding the regulation of LCT gene. Both SNPs, C/T-13910 and G/A-22018, were found in introns of MCM6, a mammalian homologue of mis5 of yeast, that functions as a cell cycle factor (Takahashi et al. 1994). MCM6 is expressed in a variety of human tissues including intestine; intestinal expressions of MCM6 and LCT were not observed to correlate with lactase persistence or non-persistence subjects, suggesting that these genes are independently regulated. The influence of the MCM6 gene on adult-type hypolactasia seems to be only structural (Harvey et al. 1996).

Studies *in vitro* in Caco-2 cell lines of human and rat, demonstrated both the C-13910 and T-13910 variants enhance LCT promoter activity; but the region containing the T-13910 allele was demonstrated to be more effective in increasing gene transcription than the one containing the C-13910 allele. On the contrary, the –22018 region has a weak silencer activity not influenced by the G/A–22018 polymorphism (Troelsen et al. 2003; Olds & Sibley 2003).

Subsequent *in vitro* analyses of the region, identified Oct-1 as the transcription factor binding strongly to the T-13910 allele. The sequence in which the T-13910 allele is located binds the transcriptional enhancer protein Oct-1, while the sequence containing the ancestral C-13910 allele binds only poorly. Oct-1, together with GATA-

6, HNF4α and Fox sites, are needed for full T-13910 enhancer activity (Lewinsky et al. 2005).

The post-weaning decline of lactase in mammals is the result of a decrease in the lactase promoter activity. The decline could depend on a decrease in the recruitment of transcriptional activators or an increased recruitment of repressors of LCT promoter. Both T-13910 and C-13910 variants act as enhancers, but the presence of Oct-1 at the T-13910 enhancer during the post-weaning decline could result in the inability to down-regulate the lactase expression due to increased recruitment of transcriptional activators or by preventing the action of repressors (Lewinsky et al. 2005).

Oct-1 has been reported to recruit co-factors for chromatin structure enhancing or silencing genes depending on the tissue type and promoter architecture (Zabel et al. 2002; Rememyi et al. 2004). For this reason Lewinsky et al (2005) proposed Oct-1 binding to the T-13910 allele *in vivo* could induce chromatin changes close to LCT gene that are involved in the lactase-persistence phenotype.

1.6 Multiple origins of lactase persistence

The T-13910 variant has been found in European Caucasians or their descendants in agreement with epidemiological data (Enattah et al. 2002; Bersaglieri et al. 2004; Mulcare et al. 2004; Coelho et al. 2005; Enattah 2005, unpublished data). Outside of Europe the T-13910 allele was strongly represented in Pakistan and Algeria, lower frequencies were observed in Middle Eastern populations (Bersaglieri et al. 2004; Enattah 2005, unpublished data). For Eurasian populations the C/T-13910 polymorphism can be regarded as a suitable first-stage screening test for adult-type hypolactasia in adults (Rasinperä et al. 2004; Swallow 2006).

African origin

However the T-13910 allele is highly associated with lactase persistence in Euro-Asiatic populations, this allele is rare in many African milk drinking pastoralist groups where lactase persistence phenotype has been reported at high frequency (Mulcare et al. 2004). The allele is so rare that it can not explain the frequency of the lactase-persistence phenotype through Africa. This suggested that there may be more than one cause of lactase persistence, or that the T-13910 allele is not functional and the true causal mutation is located elsewhere on the extended haplotype A.

Recently, two studies on African populations provided evidence that the T–13910 allele is not the worldwide cause of lactase persistence (Ingram et al. 2006; Tishkoff et al. 2006).

From a genotype-phenotype analyses in Kenyan and Tanzanian populations three new substitutions have been discovered; in particular a G/C-14010 mutation that regulates LCT gene expression. The derived alleles exist on haplotype backgrounds different from each others and from the common European haplotype A. The data provide that two independent causal variants, respectively in Eurasia and East Africa, are the result of convergent evolution related to a strong selective force, represented by adult milk consumption. The age of C-14010 allele (~2,700-19,200 years) supports a recent spread of pastoralism in East Africa (Tishkoff et al. 2006).

Euro-Asiatic origin

The frequencies and diversity patterns of haplotypes associated to lactase persistence and non-persistence could help to better understand the origin of lactase persistence trait in Euro-Asian populations, characterized by the T-13910 allele presence. To this aim a subset of 9 SNPs located in a 30kb region flanking the lactase variants (including the C/T-13910 and G/A-22018 polymorphisms) have been analysed in 37 populations from Eurasia and Africa (Enattah 2005, unpublished data).

Four major lactase non-persistence haplotypes were identified (Hap 1-4): haplotypes 1-3 are highly divergent from the main lactase persistence Hap5, while Hap4 differ from Hap5 just for the two lactase associated mutations at -13910 and –22018 loci. Hap4 could be considered as the allelic background on which lactase persistence mutation(s) occurred and it showed the highest frequency among Trans-Ural populations. Among Ob-Ugric speakers, in the Eastern part of Urals, Hap4 reaches values of 33% while lactase persistence has less than 5% of frequency; in the Western part of Ural, the frequency of Hap4 is 35% among Komi, Udmurtains, Lapp Saami. High frequency of Hap4 were found also in Han Chinese (36%), where the lactase persistence allele is absent.

There are other common lactase persistence haplotypes, such as Hap9 and Hap10, where lactase persistence variants occurred on an insertion polymorphism background, and they showed the highest frequencies (5-15%) always in populations of Western of Ural mountains like Komi, Udmurtians, Mokshas and Erzayas.

These results suggest that lactase persistence in Eurasia could have occurred through two different lineages in human history: one very early on the background of Hap1 and Hap2 resulting in the less frequent lactase persistence Hap9 and Hap10; the other more recent lineage originated from the background of lactase non persistence Hap4, resulting in the most common lactase persistence Hap5.

Cause of the high frequencies of pre-lactase persistence Hap4 and lactase persistence haplotype diversity, it has been hypothesised the lactase persistence mutation has occurred among pastoral nomad groups in Trans-Ural region. Thanks to nomad migrations the mutation reached the Western slope of Ural and spread to region between the Volga and the Urals, to North of the Caucasus and to the Black Sea.

These data are consistent with the westward expansion, from the Kurgan area above the Caucasus, of pastoral nomads speaking Indo-European languages between 2,500 and 1,500 B.C. (Cavalli-Sforza 1994).

1.7 Lactase persistence in Italian populations

Italy lies at the South boundary of the North-South decreasing gradient of lactase persistence in Europe.

Previous studies reviewed in Flatz 1987, based on physiological estimations (mainly on breath hydrogen analyses) on 592 healthy Italian subjects, reported higher values of lactase persistence in the North (50%) and lower in the South (28%) and in Sicily (29%) (Marenco et al. 1970; Arrigo et al. 1980; Zuccato et al. 1983; Burgio et al. 1984; Rinaldi et al. 1984). The level of non-persistence in Italians is unusually high for an European population, but very similar to other southern Mediterranean ones, probably due to the migration barrier of Alps (Flatz, 1987).

On the other hand Cavalli-Sforza et al. (1987) didn't confirm the general tendency of lactase phenotype distribution. They found persistence frequencies between 48% in the North and 59% in the South with the highest value in the centre of Italy (81%). They suggested an heterogeneity in frequency of lactase persistence depending on composition of sampling subjects. Italians are an admixed population with various degree of admixture of ethnic groups from northern Europe, where lactase persistence prevails, and from Mediterranean area, where high frequency of non-persistence exists. They proposed the situation in central Italy samples is consistent with the origin of people in this region that are descendants of population with a pastoral or mixed agro-pastoral culture (Grosso 1934).

AIM OF THE STUDY

This thesis aim to analyze the molecular variability associated with lactase persistence in Eurasian populations.

First we want to improve data on molecular estimations of lactase persistence typing the two biallelic polymorphisms totally and highly associated to the trait, the C/T-13910 and the G/A-22018 discovered upstream the LCT gene by Enattah et al. (2002).

We realized an extend sample collection in Italy, considering populations from different geographic areas and with different isolation degrees. A total of 749 individuals were typed in order to analyzed lactase persistence distribution across the country because different patterns have been observed by previous physiological estimations. Great part of physiological data revealed the presence of a decreasing gradient toward South of persistence (Marenco et al. 1970; Arrigo et al. 1980; Zuccato et al. 1983; Burgio et al. 1984; Rinaldi et al. 1984; Flatz 1987), consistent with the European distribution. On the other hand the results of Cavalli-Sforza et al. (1987) suggested the absence of the gradient and a more microgeographic variability related to the history of each population.

Lactase persistence frequencies were estimated in two Asian groups too: in Kurds and Persians from Middle East, a particularly interesting region for the development of the Neolithic Transition, and four Central Asian populations with a nomadic pastoral origin (two Kirghiz groups from Sary-Tash and Talas, Uighurs and Kazakhs).

In the second part we developed a phylogeographic analysis of lactase persistence lineages in the three population groups from Eurasia. We tried to understand the genetic background on which the mutation associated to lactase persistence (C/T-13910) occurred and to contribute to the reconstruction of lactase persistence evolution, localizing a possible place of origin for the mutation and its spread of diffusion across Eurasia.

To this aim a subset of samples were typed for three microsatellite markers, fast evolving markers able to capture genetic variability in a genomic region where positive natural selection seems to have reduced the genetic variability. Several analyses have been realized: first we analyzed the phylogenetic relationships among the persistent haplotypes and their distribution across the three populations groups and other populations from literature; second, we measured intra- and inter-population genetic variability; and third,

we estimated the age of T-13910 allele using to intra-allelic accumulation of microsatellite diversity.

MATERIALS AND METHODS

3.1 Population samples

In this study we analysed samples from Italy (N=749), Middle East (N=71) and Central Asia (N=297), for a total of 1117 subjects. Samples were collected during different periods and by different methods, but apparently healthy and unrelated subjects were always chosen.

Italian sample: samples were collected to obtain a complete screening of lactase persistence of the country. To the aim we considered 4 geographic areas (Figure 3.1 and Table 3.1): *North, Centre-North, Centre-South* and *South* (this last one comprising of islands); within each area we sampled populations with different isolation degrees to better analyse microgeographic variability. We sampled 18 populations, for a total of 749 individuals.

Figure 3.1 Geographic distribution of the 18 Italian populations sampled and subdivision in the 4 geographic areas: *North, Centre-North, Centre-South* and *South.*

The isolation degree was established depending on geographic, economic and social features of samples; socially and culturally isolated ethnic groups, such as Ladini of Trentino, or villages in close valleys, Val di Scalve (BG) and Alfero (FC), or in mountainous areas, Albidona (CS), have been considered high isolated samples. Cities,

with a higher gene flow, have been considered low isolated samples. The geographic position and the isolation degree for each sample analysed are reported in Figure 3.1 and Table 3.1.

Area		Population	N	Isolation degree
North	1	Val di Scalve (BG)	102	high
	2	Ladini	30	high
	3	Udine	53	low
	4	Castelmassa (RO)	34	medium
		tot	219	
Centre-North	5	Bologna	95	low
	6	Palazzuolo sul Senio (FI)	33	medium
	7	Alfero (FC)	41	high
	8	Alta Val Savio (FC)	37	medium
		tot	206	
Centre-South	9	Tocco da Casauria (PE)	38	medium
	10	Roma	30	low
	11	Norma (LT)	30	medium
	12	Benevento	30	low
		tot	128	
South and islands	13	Messapi	23	medium
	14	Albidona (CS)	52	high
	15	Roseto Capo Spulico (CS)	27	medium
	16	Belvedere Marittimo (CS)	28	medium
	17	Siracusa	29	low
	18	Gallura	37	medium
		tot	196	
		total Italy	749	

Table 3.1 List of the Italian populations sampled, grouped for geographical area; for each population the sample size (N) and the isolation degree are indicated.

DNA was obtained from buccal swabs using a brush or from 3 cc. of blood, thanks to local AVIS (Volunteer Italian Blood donors Association) groups' collaboration for Val di Scalve (BG), Udine, Castelmassa (RO), Bologna, Palazzuolo sul Senio (FI), Alfero (FC), Alta Val Savio (FC), Albidona (CS), Roseto Capo Spulico (CS), Siracusa and Gallura (Sardinia) samples. The origin of each sample was verified by bio-demographic information on the last three generations.

Ladini of Trentino, Norma (LT), Benevento, Messapi of Puglia, Belvedere Marittimo (CS), Tocco da Casauria (PE) and Roma samples have been collected by prof G. Destro-Bisol's group ("La Sapienza" University, Rome); Tocco da Casauria and Roma data have been already published in Coelho et al. (2005).

The **Middle East** sample consist of 49 Kurdish subjects coming from Iraq, Iran, Syria and Turkey that actually live in Italy, and of 22 subjects of Persian ethnicity; Persian samples were collected in the South-West region of Iran called Fars (Figure 3.2 and Table 3.2).

Sample collection was realized in 2003-2004 and bio-demographic information was obtained for each subject. DNA was obtained from buccal swabs using a brush.

Figure 3.2 Geographic position of Middle Eastern samples. The extension of the Kurdish region across Iraq, Iran, Syria and Turkey is represented in green. The red point indicates the South-West region of Iran where Persian samples were collected.

The **Central Asia** sample is composed of two Kirghiz groups from Sary-Tash (65) and Talas (79), 48 Uighurs and 105 Kazakhs, for a total of 297 samples (Figure 3.3 and Table 3.2). Sary-Tash is an isolated high-altitude village (3,200 m above sea level) in the Pamir mountains in the South of the country, while the Talas valley (900 m above sea level) is in the Northern part of Kirghizstan, close to Kazakhstan and Uzbekistan. The Uighurs were sampled in the village of Penjim (600 m above sea level) in the East of Kazakhstan, close to Chinese border. Kazakh samples were collected in villages of the Kegen valley, a high plain 2,100 m above sea level. Their subsistence economy is based mainly on goat, sheep, horse and yak breeding and agriculture.

Blood-samples were collected within the CAHAP (Central Asia High Altitude People) project during two expeditions (1993, 1994) in Kirghizstan and Kazakhstan.

Figure 3.3 Geographic position of Central Asia samples: the Kirghiz groups from Sary-Tash (in light blue) and Talas (in green), the Uighurs (in yellow) and the Kazakhs (in red).

Area	Population	N
Middle East	Kurds (Kur)	49
	Persians (Per)	22
	tot	71
Central Asia:		
Kirghizstan	Kirghizes of SaryTash (KSa)	65
Kirghizstan	Kirghizes of Talas (KTa)	79
Kazakhstan	Uighurs (Uig)	48
Kazakhstan	Kazakhs (Kaz)	105
	tot	297

Table 3.2 List of the Middle Eastern and Central Asia samples with relative sample sizes (N); the short names used in the test are gives in parentheses.

3.2 Lab methods

3.2.1 DNA extraction

DNA was extracted from blood following "Salting Out" (A) protocol, while for extraction from buccal swabs we follow "kit Qiamp" (Qiagen) protocol (B) or a modified "Salting Out" protocol (C).

A. "Salting Out" protocol for DNA extraction from blood:

- Move unfrozen blood to sterile 15 ml tubes, add 12 ml **RCLB** (Red Cell Lising Buffer), vortex, centrifuge 10 min at 3,000 rpm, discard supernatant playing attention to conserve the pellet; repeat 3-4 times till the pellet not release colour.
- Add 3 ml **WCLB** (White Cell Lising Buffer), vortex, add 25 µl **Proteinase K** (20 mg/ml) and 25 µl **SDS** 20%; vortex and incubate 1 hour at 55°C.
- Add 1.7 ml **Sodium Acetate** 3M pH 5.2, agitate manually and centrifuge 10 min at 3000 rpm; move supernatant to new 15 ml tubes.
- Add the same volume of **Isopropyl alcohol,** agitate softly; you could see the DNA "jelly-fish", the finally amount of sterile water depends on jelly-fish dimension. Centrifuge 10 min at 3,000 rpm.
- Discard supernatant and add 3 ml **ethanol 80%**, vortex and centrifuge 10 min at 3,000 rpm.
- Discard supernatant, turn upside down the tubes and leave dry on absorbent paper for 2 hours.
- Resuspend DNA in sterile water.

B. protocol for DNA extraction from buccal swabs by "Kit QIAamp DNA miniKit 50", QIAGEN

This method is based on specific columns (QIAamp Spin Column) with a silica-gel membrane able to absorb DNA.

- Place the brush in a 1.5 ml tube, add 1 ml sterile water and agitate the brush before discarding it. Centrifuge 3-4 min at 13,000 rpm. Remove supernatant without the pellet.
- Add 200 µl **sterile water**, 20 µl **Proteinase K** and 200 µl **buffer AL**. Vortex and incubate 10 min at 56° C.
- Add 200 µl **ethanol 96-100%,** briefly centrifuge and move the solution to the QIAamp Spin Column (in a 2 ml collection tube).

- Add 500 µl **buffer AW1,** centrifuge 1 min at 8,000 rpm, discard the tube containing the filtrate and add a new 2 ml collection tube.
- Add 500 µl **buffer AW2** centrifuge 3 min at 13,000 rpm and discard the tube containing the filtrate.
- Place the QIAamp Spin Column in a clean 1.5 ml tube, add 200 µl **buffer AE**, incubate at room temperature for 1 min and centrifuge 1 min at 8,000 rpm.

C. modified "Salting Out" protocol for DNA extraction from buccal swabs

Buccal swabs were collected using a brush conserved in 1.5 ml tube with ethylic alcohol (90-100%). This protocol was suggested by prof. Rocha J., University of Porto.

- Centrifuge the tubes with cheeks scraps 5 min at 14,000 rpm. Discard supernatant.
- Resuspend pellet in 200 µl **High TE**; vortex the tubes.
- Add 270 µl **Madisen Lysis Solution** and 10 µl **Proteinase K**; vortex the tubes.
- Incubate overnight at 37°C.
- Add 120 µl **NaCl solution 6M**; vortex the tubes and centrifuge 20 min at 14,000 rpm.
- Remove supernatant to a new tube and discard the pellet.
- Add 600 µl chilled **ethanol** (100%).
- Let the tubes 30 to 60 min at -20°C.
- Centrifuge 20 min at 14,000 rpm (if possible at 4°C) and discard supernatant.
- Add 200 µl **ethanol** (70%) and centrifuge 20 min at 14,000 rpm (if possible at 4°C) and discard supernatant. Repeat twice.
- Dry the empty tubes 1 hour at 60°C and add 40-60 µl water.

Extraction products were visualized and quantified by electrophoresis on 1% agarose gel.

3.2.2 SNPs genotyping

All the polymorphisms considered in this study are located in a wide region of ~165kb encompassing LCT gene on chromosome 2q21; marker positions and the relative distance between them are illustrated in Figure 3.4. We analysed 2 Single Nucleotide Polymorphisms (SNPs), C/T-13910 and G/A-22018, and 3 Short Tandem Repeat polymorphisms (STRs) D2S3010, D2S3013 and D2S3015. These markers have

different mutation rates, the multiallelic STRs are fast evolving markers there will be used to study the level of diversity within the haplotypes defined by the two more stable SNPs.

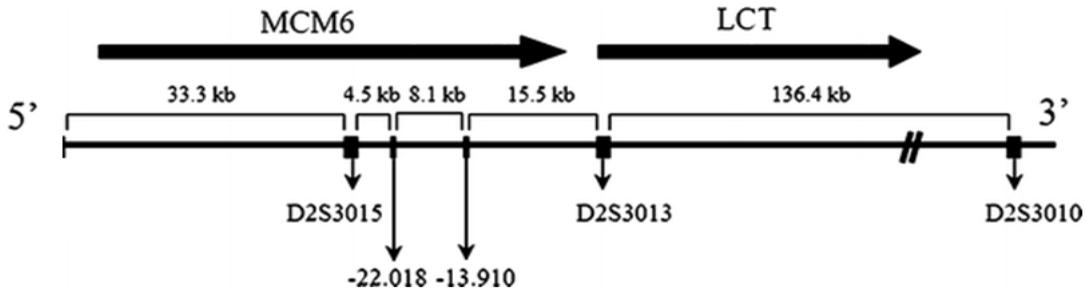


Figure 3.4 Schematic representation of the genomic region including LCT and MCM6 genes with the relative positions of the 2 SNPs (C/T13910 and G/A22018) and the 3 microsatellite (D2S3010, D2S3013 and D2S3015) typed in this study.

The first part of the study is based on the genotyping of the two SNPs C/T-13910 and G/A-22018, associated to lactase persistence in Eurasian populations. Both biallelic markers were typed in Italian, Middle Eastern and Central Asian samples using Polymerase Chain Reaction (PCR), followed by digestion with specific restriction enzymes (Coelho et al. 2005).

Here we reported the amplified sequences including the polymorphic loci (in red), forward and reverse primers are in bold characters and short sequences recognised by enzymes are underlined.

C/T-13910

5'.**GCAGGGCTCAAAGAACAATC**TAAAAATCAAACATTATACAAATGCAACCTAAGGAGGA
GAGTTCCTTTGAGGCCAYGGgCTACATTATCTTATCTGTATTGCC**AGCGCAGAGGCCTACTA**
GTACA. 3'

G/A-22018

5'.**CTCAGTGATCCTCCCACCTC**AGCCTCTTGAGTAGCTGGGACCACAAGCACCCGCCACCA
TGCCCGGCTAATTTTTGTATTTTTAGTAGAGAAAATGGGTTTTCGCCATGTTGGCCAGGCTGG
TCTCGAACTCCTGACCTCAGGTGATCCACCCACCTCGGCTTCCCAAAGTACTGGGACAAAGG
TGTGAGCCACCGcGCCCAGCTGAGAATGCTGTTTTTAAGGACATCTTTTTAATGGTAACTTAT
AG**GCCTTTACTGATAGGGTAGGGG**. 3'

The PCR reaction mix components and their final concentrations are described in Table 3.3; the mix has a final volume of 25 µl.

	concentration	volume (µl)
H_2O		
dNTPs	0,08mM	0,8
primer F	0,1 µM	0,5
primer R	0,1 µM	0,5
$MgCl_2$	1mM	1,0
BUFFER	1X	2,5
Taq	1U/µl	0,2

Table 3.3 Final concentration and volume for each component of the PCR mix.

The *thermal cycler* conditions are the following: initial denaturation of DNA for 5 min at 94°C, than 35 cycles of denaturation at 94°C for 1 min, primers annealing at 60°C for 1 min and extension by Taq polymerase activity at 72°C for 1 min, and a final extension at 72°C for 10 min. PCR conditions are the same for both markers. The amplification products for the C/T-13910 and G/A-22018 markers are respectively 125bp and 271bp long.

The C/T-13910 sequence was digested by FaqI/FinI enzyme (2U/µl) through incubation at 37°C for 2 hours. When the ancestral C allele is present the fragment is 125bp, when the T allele is present the enzyme cuts the sequence and two fragments of 80bp and 45bp are visible (Figure 3.5 A). CT and TT genotypes were classified as lactase persistent (Enattah et al. 2002).

The enzymatic digestion for the G/A-22018 sequence was realized by BsmI enzyme (1U/µl) at 37°C for 2 hours. The enzymatic restriction produces 196bp and 75bp fragments if individuals have the G allele, while there is no restriction when A allele is present (271bp) (Figure 3.5 B). GA and AA genotypes were classified as lactase persistent (Enattah et al. 2002).

Restriction patterns have been visualized by electrophoresis on agarose gel and were visible on ultraviolet rays thanks to previous coloration through Ethidium Bromide. The gel was more concentrated (3.5%) for the C/T-13910 sequence, cause of the smallest DNA fragment dimensions, while it was only 3% concentrated for the other marker.

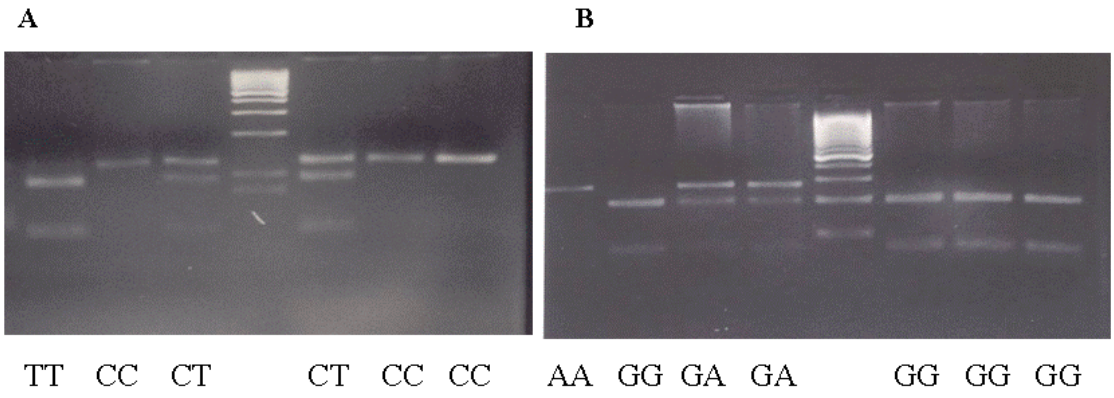


Figure 3.5 Restriction patterns for the C/T-13910 (A) and the G/A-22018 (B) polymorphisms. We used a GeneRuler 100bp ladder as reference.

3.2.3 STRs genotyping

To better investigate the genetic diversity associated to the core haplotype defined with the two biallelic markers (C/T-13910 and G/A-22018), we typed 3 fast evolving markers D2S3010, D2S3013 and D2S3015 (Figure 3.4). The 3 STRs were typed in six Italian populations: in particular in a subset of Val di Scalve group (N=42) composed by all the subjects with at least one T-13910 allele plus some CC-13910 homozygous, in the whole Palazzuolo (N=33), Tocco (N=37), Roma (N=30) and Alfero (N=41) samples and in great part of Albidona sample (N=42). Tocco and Roma data were already published (Coelho et al. 2005). All Kurdish (N=49) and Persian (N=22) samples of Middle East were typed and a subgroup of the four Central Asian populations including all subjects with at least a T-13910 chromosome plus some CC-13910 homozygous (KSa N= 28; KTa N=30; Uig N=29; Kaz N=30).

D2S3013 and D2S3015 microsatellite were typed in a duplex PCR (Table 3.4 A) and D2S3010 in a simplex PCR (Table 3.4 B) with fluorescently labelled primers, followed by separation of amplification products by capillary electrophoreses on automatic sequencer (ABI PRISM® 310 Genetic Analyser) (Figure 3.6).

The primers used are listened above:

D2S3010-F	5'-TTA GGC CCT CTC TTC GAA TGA T-3'
D2S3010-R	5'-GAT TTA GGT GGA GAC ACA C-3'
D2S3013-F	5'-GAG AAT ATA GTC ATA AAC TAT GTT-3'
D2S3013-R	5'-ATT TTG GAT TAT ATA TGC TTT CTT G-3'
D2S3015-F	5'-CCT GTA GTC CCA GCT AAT TTC-3'
D2S3015-R	5'-CAG AGA AGT TTT GTT TGT GGA-3'

A	concentration	volume (µl)
H_2O		4.276
dNTPs	0.2 mM	1
D2S3013 F	0.5 µM	1.25
D2S3013 R	0.5 µM	1.25
D2S3015 F	0.1µM	0.25
D2S3015 R	0.1 µM	0.25
$MgCl_2$	1.5 mM	0.375
BUFFER	1X	1.25
Taq	2U	0.1
tot		10

B	concentration	volume (µl)
H_2O		4.776
dNTPs	0.2 mM	1
D2S3010 F	0.5 µM	1.25
D2S3010 R	0.5 µM	1.25
$MgCl_2$	1.5 mM	0.375
BUFFER	1X	1.25
Taq	2U	0.1
tot		10

Table 3.4 Final concentrations and volumes of the PCR mix components, for the D2S3013/D2S3015 duplex PCR (A) and for D2S3010 simplex PCR (B).

In both PCRs, samples were denaturated for 5 min at 94°C, followed by 35 cycles of 94°C for 1 min, 56°C for 1 min and 72°C for 1 min, followed by 10 min a final extension at 72°C.

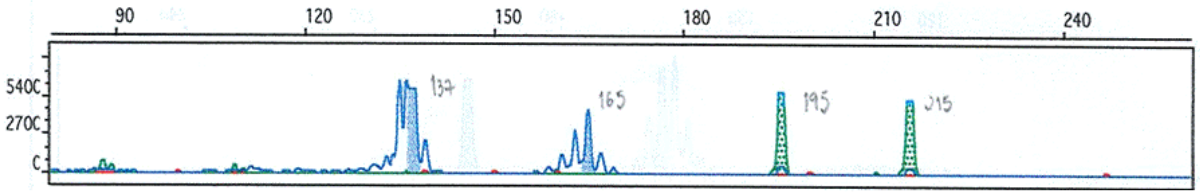


Figure 3.6 An example of pattern obtained after the separation of an amplification product of the D2S3013/D2S3015 duplex PCR by ABI PRISM® 310 Genetic Analyser. In blue the alleles at locus D2S3013; in green the alleles at locus D2S3015.

3.3 Statistic analyses

3.3.1 Allele and genotype frequencies

Both allele and genotype frequencies were determined by direct counting.

Assuming Hardy-Weinberg equilibrium, in absence of the action of evolutive forces, such as mutation, gene flow, genetic drift and natural selection, genotypic frequencies are constant through generations. The null hypothesis of equilibrium is tested; the p value express the probability that differences between the expected and observed frequencies were random. When $p<0.05$ the null hypothesis is rejected, because differences are too strong to be random. Genotype frequencies for each biallelic marker were tested for deviation from Hardy-Weinberg equilibrium using a test analogous to Fisher's exact test, implemented in the software *Arlequin 3.01* (Excoffier et al. 2005)

Bonferroni correction was applied to non-equilibrium loci; it is a statistical correction for multiple testing: α/K, where α is the level of "type I error" considered to reject the null hypothesis (0.05 in our case) and K is the number of tests done.

3.3.2 RxC exact test

RxC test was used to compare allele and genotype frequency distribution among populations. It is a Windows program for the analysis of contingency tables. Rather than relying on asymptotic methods such as Chi-square or log-likelihood methods, RxC employs the metropolis algorithm to obtain an unbiased estimate of the exact p-value associated to the observed differences (Raymond & Rousset 1995).

3.3.3 Haplotype inferences

The gametic phase can be inferred starting from genotype data, using statistical methods. To this aim we used the software *Phase v.2.1* that implements a Bayesian method for reconstructing haplotypes comparing unresolved haplotypes, with two or more heterozygous loci, to similar resolved haplotypes, that have no ambiguous loci (Stephens and Donnelly 2003).

The output file consists of a list of haplotypes and frequency and for each sample the "best" haplotype reconstruction is reported. Moreover, there is a list of the confidence probabilities associated with each phase call.

When the number of samples is very small there is little information in the data to allow the haplotypes to be inferred and this is reflected in a relatively low phase calling

probability. Larger data sets will typically be much more informative about a large proportion of phase calls, for these reason we ran all population data together. We inferred five-locus haplotypes, considering both the SNPs and STRs markers together.

Haplotype frequencies were obtained by direct counting.

3.3.4 Intra-population genetic variability

We calculated several standard indexes of genetic variability.

Genetic variability within each microsatellite locus was estimated through Heterozygosity (H) value, the probability that two alleles sampled at random from the population are different in length. We calculated an average H across all loci too.

Haplotype diversity is defined as the probability that two randomly chosen haplotypes are different in the sample:

$$H = [n/(n-1)] \left(1 - \sum_{i=1}^{k} pi^2\right)$$

where n is the number of gene copies in the sample, k is the number of haplotypes, pi is the sample frequency of the i-th haplotype. Haplotype diversity is a direct estimation of population variability, it will be higher in a population with a great number of alleles and lower in a population with few alleles or with only one common allele. This index was calculated using the software *Arlequin ver 3.01* (Excoffier et al. 2005).

3.3.5 Phylogenetic analysis

Phylogenetic relationships among the different five-locus haplotypes were explored by calculating a Median Joining using *Network 4.2* (Bandelt et al. 1999).

The method is based on Kruskal's *minimum spanning tree algorithm* and Farris' *maximum parsimony algorithm*. The first one realizes all possible trees with the minimal value of sum of branch length (the shortest distance among the haplotypes observed); the second creates consensus haplotypes (median vectors) to joint the trees.

The network summarizes all parsimony trees and all possible evolutionary pathways are represented through cycles.

To better compare the haplotype distribution among populations with different sample sizes, we calibrated the networks considering all populations of the same size and recalculating haplotype frequencies

3.3.6 Inter-population genetic variability

Total inter-populations diversity (Fst) was calculated from haplotype data using the infinite-allele model, *Arlequin ver 3.01* (Excoffier et al. 2005). Fst will be 0, when all meta-populations have the same allele frequencies (Wright 1951).

We performed an Analyses of Molecular Variance (AMOVA) to investigate the genetic structure of populations using information on the allelic content of haplotypes, as well as their frequencies (Excoffier et al. 1992; Schneider et al. 1999). It is possible to test an eventual genetic structure measuring the variability at several hierarchical level: among populations (Fst), among groups (Fct), within groups of populations (Fsc). The number of permutations used to test the significance of covariance components and fixation indices was 10,000, *Arlequin ver 3.01* (Excoffier et al. 2005).

3.3.7 Genetic distances

Fst index was used to calculated pair-wise genetic distances between the populations. It is an index analogue to Fst of Wright (1951) based on pair-wise differences among halotypes. Pairwise Fst was calculated applying the Nei's equation (1987), *Arlequin ver 3.01* (Excoffier et al. 2005). Fst statistical significance was assessed using 10,000 bootstrap replications. The matrix obtained was standardized by elimination of negative values summing the higher negative value to each other value.

The distance matrix was synthetically represented by a non-metric Multidimensional Scaling (MDS). MDS is a non-metric multivariate analysis that allows to reproduce many data, on many populations and many loci, in two or three dimensions, reducing the loss of information. We generate MDS using the software StatSoft ver 6 (StatSoft Italia srl, 2001) and choosing a final bi-dimensional representation; the program gives stress values indicating the goodness of approximation.

3.3.8 Age estimation of T-13910 allele

The method used to estimate the Time of the Most Recent Common Ancestor (TMRCA) of the T-13910 allele is based on the simulation of the overtime decay in the frequency of the allele originally associated with the T-13910 allele in each microsatellite locus (Seixas et al. 2001). The modal allele length at each microsatellite locus in the pooled sample was considered to be the ancestral. The method is based on the intra-allelic accumulation of microsatellite diversity, assuming a stepwise mutation

model and using a 25-year generation time. Recombination can be taken into account, according to the formula:

$$p_{(g,i)} = p_{(g-1,i)}\,(1-\mu-r) + rq_i + (\mu/2)\,[p_{(g-1,\,i-1)} + p_{(g-1,\,i+1)}]$$

where $p_{(g,i)}$ is the frequency of a marker microsatellite allele with i repeats in g generation within the T-13910 allele, q_i is the frequency of the allele in the whole population, r is the recombination fraction between the T-13910 site and each microsatellite locus and μ is the microsatellite mutation rate. 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 3 STRs loci (Table 3.5).

Confidence intervals were calculated assuming a rapid population growth according to Goldstein et al. (1999).

Two sets of mutation rates (μ) were used. The first 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 homozygosis under the single-step stepwise mutation model. We assumed Ne=10,000 (Takahata 1993) and homozygosis was estimated from the microsatellite allele frequency distribution in a population where frequencies are less probably determined by selection (Coelho et al. 2005) (Table 3.5).

The second set was derived from the average 0.001 value obtained from observed mutations in pedigrees (Weber & Wong 1993). Locus specific mutation rate were calculated by apportioning this average according to the ratios of the locus-specific estimates calculated by the indirect approach (Table 3.5).

		D2S3010	D2S3013	D2S3015
m1	**μ**	0.0009	0.0005	0.000095
m2	**μ**	0.0023	0.0013	0.0002
m3	**μ**	0.0009	0.0005	0.000095
	r	0.0015	0.00016	0.00013
m4	**μ**	0.0023	0.0013	0.0002
	r	0.0015	0.00016	0.00013

Table 3.5 Mutation (μ) and recombination (r) rates used for the age estimations of T-13910 allele. m1, assuming suppression of recombination and microsatellite indirect estimation of mutation rates; m2, assuming suppression of recombination and microsatellite mutation rates calculated from a 0.001 direct average estimate; m3, mutation rates as in m1 and assuming recombination rates; m4 mutation rates as in m2 and assuming recombination rates.

3.3.9 Neutrality tests

Neutrality tests for T-13910 allele were performed using the method of Slatkin and Bertorelle (2001), which evaluates if the observed frequency of an allele (T-13910) is consistent with its level of variability (estimates using the three STR loci) under a given demographic pattern, assuming neutrality. Intra-allelic variability was measured as the minimum number of mutations (S_0) observed at linked microsatellite marker loci (Slatkin & Bertorelle 2001; Slatkin 2002) and was inferred using median-joining networks, *Network 4.2* (Bandelt et al. 1999).

For each population the number of chromosomes bearing the T-13910 allele and the total number of chromosomes in the sample were used.

We used both sets of mutation rates used for age calculations, considering simultaneously the 3 STR loci.

We test two different demographic models. D1 assumes a constant exponential growth rate r=0.008 starting 900 generations ago (t) from an initial population of (N_1) 10^3 to a final one of (N_0) 1,340,000 (Pritchard et al. 1999). D2 assumes that the effective population size increased exponentially (r=0.0146) from (N_1) 10^4 to (N_0) $5x10^9$ starting (t) 900 generations ago (Kruglyak 1999).

Neutrality is rejected when the probabilities of finding a number of mutations $\leq S_0$ in the microsatellite loci linked to T-13910 allele is lower than 0.001.

RESULTS

4.1 Allele and genotype frequencies at C/T–13910 and G/A-22118 loci in Italian populations

Genotype frequencies at -13910 and -22018 loci observed in 749 samples belonging to the 18 Italian populations are showed in Table 4.1. Allelic frequencies are reported in the same table and the map (Figure 4.1) clearly represents C and T allelic (at locus –13910) distribution within each population and within each geographic group.

			N	CC	CT	TT	C	T	GG	GA	AA	G	A
North	1	Val di Scalve	102	0.578	0.373	0.049	0.765	0.235	0.549	0.392	0.059	0.745	0.255
	2	Ladini	30	0.433	0.467	0.100	0.667	0.333	0.400	0.333	0.267	0.567	0.433
	3	Udine	53	0.528	0.453	0.019	0.755	0.245	0.528	0.453	0.019	0.755	0.245
	4	Castelmassa	34	0.735	0.235	0.029	0.853	0.147	0.618	0.353	0.029	0.794	0.206
		tot	219	0.571	0.384	0.046	0.763	0.237	0.534	0.393	0.073	0.731	0.269
Centre-North	5	Bologna	95	0.789	0.211	0.000	0.895	0.105	0.737	0.253	0.011	0.863	0.137
	6	Palazzuolo	33	0.727	0.242	0.030	0.848	0.152	0.667	0.303	0.030	0.818	0.182
	7	Alfero	41	0.805	0.171	0.024	0.890	0.110	0.732	0.244	0.024	0.854	0.146
	8	alta Val Savio	37	0.595	0.378	0.027	0.784	0.216	0.595	0.378	0.027	0.784	0.216
		tot	206	0.748	0.238	0.015	0.867	0.133	0.699	0.282	0.019	0.840	0.160
Centre-South	9	Tocco	38	0.842	0.105	0.053	0.895	0.105	0.842	0.105	0.053	0.895	0.105
	10	Roma	30	0.733	0.233	0.033	0.850	0.150	0.733	0.233	0.033	0.850	0.150
	11	Norma	30	0.833	0.167	0.000	0.917	0.083	0.353	0.074	0.015	0.883	0.117
	12	Benevento	30	0.833	0.167	0.000	0.917	0.083	0.800	0.200	0.000	0.900	0.100
		tot	128	0.813	0.164	0.023	0.895	0.105	0.797	0.172	0.031	0.883	0.117
South and	13	Messapi	23	0.870	0.130	0.000	0.935	0.065	0.826	0.174	0.000	0.913	0.087
islands	14	Albidona	52	0.788	0.192	0.019	0.885	0.115	0.750	0.231	0.019	0.865	0.135
	15	Roseto	27	0.926	0.074	0.000	0.963	0.037	0.889	0.111	0.000	0.944	0.056
	16	Belvedere	28	0.893	0.107	0.000	0.946	0.054	0.857	0.143	0.000	0.929	0.071
	17	Siracusa	29	0.828	0.172	0.000	0.914	0.086	0.793	0.207	0.000	0.897	0.103
	18	Gallura	37	0.730	0.270	0.000	0.865	0.135	0.676	0.324	0.000	0.838	0.162
		tot	196	0.827	0.168	0.005	0.911	0.089	0.786	0.209	0.005	0.890	0.110
		total Italy	749	0.728	0.250	0.023	0.852	0.148	0.690	0.276	0.033	0.828	0.172

Table 4.1 Relative genotype (in grey colour) and allelic (in white colour) frequencies at C/T-13910 and G/A-22118 loci observed in the 18 Italian populations. Populations are grouped for geographic areas and the sample sizes are indicated by N.

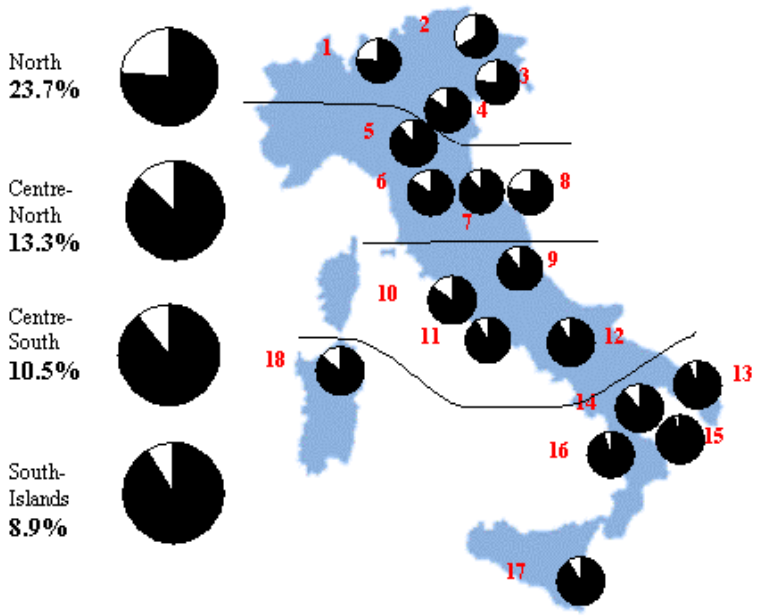


Figure 4.1 Map of allelic frequencies for C and T alleles at –13910 locus in Italian populations. Numbers indicate the populations following the order in Table 4.1. The four geographic areas are separated by black lines and the total frequencies for each area are on left side of the map.

The values indicate a North-South decreasing gradient in Italy for both T-13910 and A-22018 alleles, the derived alleles associated to lactase persistence in Euro-Asiatic populations (Enattah et al. 2002; Bersaglieri et al. 2004). This is consistent with great part of previous estimations of lactase persistence based on physiological test, that are reported in table 4.2 A (Marenco et al. 1970; Arrigo et al. 1980; Zuccato et al. 1983; Burgio et al. 1984; Rinaldi et al. 1984).

		N	Lactase Persistence
A	**North**	383	50%
	South	109	28%
	Sicily	100	29%
	tot	592	
B	**North**	89	48%
	Centre	65	81%
	South	51	59%
	tot	205	

Table 4.2 Lactase persistence estimations based on physiological tests: in the first part (A) data confirming the North-South gradient (Marenco et al. 1970; Arrigo et al. 1980; Zuccato et al. 1983; Burgio et al. 1984; Rinaldi et al. 1984), in the second part (B) data showing the absence of the gradient and a more microgeographic variability (Cavalli-Sforza et al. 1987).

As we can see, both molecular and physiological data confirm the decline in frequency toward South typical of the European continent (Flatz 1987; Swallow 2003), but lactase persistence frequencies are higher for phenotype estimations.

The T-13910 allele exhibits frequencies ranging from 33% in Ladini, in the North of Italy, to 3.7% in Roseto (CS), in the South, for an average estimation of lactase persistence in Italy of 15%. Generally A-22018 allele frequencies are higher because it represents a more ancient mutation, even if the two mutations are temporarily closed (Poulter et al, 2003; Swallow 2003). Always in Ladini samples we found the highest number of TT-13910 and AA-22018 homozygous.

There were no significant departures from Hardy-Weinberg equilibrium for both loci (data not shown) in any populations, excepted in Tocco sample that resulted in equilibrium after the Bonferroni correction for multiple testing.

4.2 Relationship between T-13910 and A-22018 alleles

The relationships between T-13910 and A-22018 allele distribution within each population were represented by a simple linear regression (Figure 4.2). We found a substantial correspondence between the frequencies of the derived alleles, the population positions in the graphic reflect their geographic positions and the T-13910 and A-22018 allele distribution that declines from North to South.

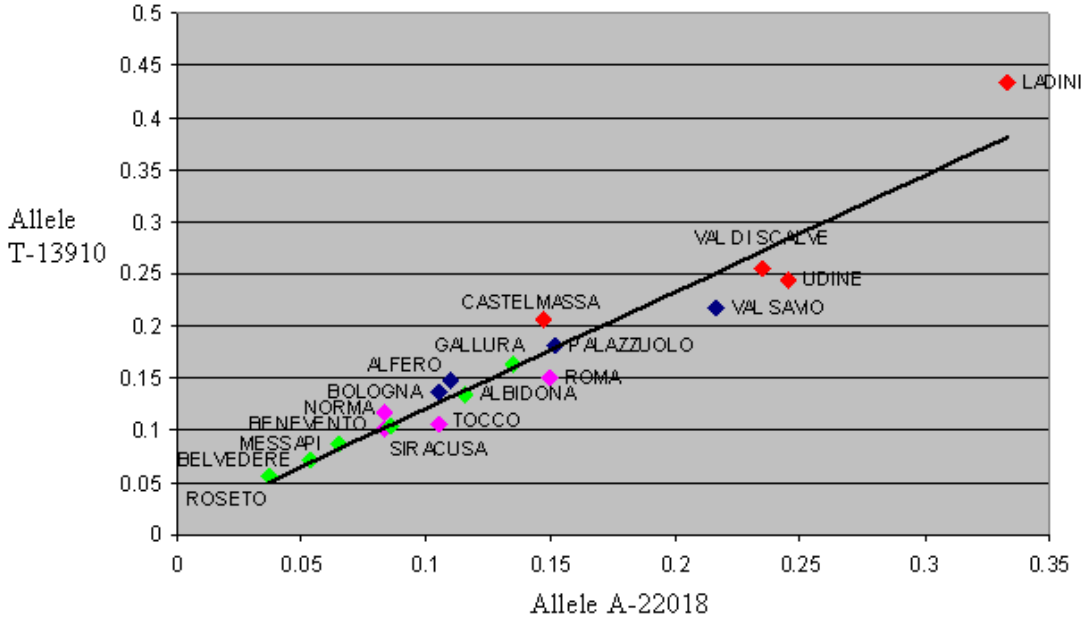


Figure 4.2 Simple linear regression between the frequencies of T-13910 (on the Y axis) and A-22018 (on the X axis) derived alleles. Northern populations are in red, Centre-Northern ones are in blue, Centre-Southern ones are in pink and Southern ones are in green.

Interesting when we consider the entire Italian sample, the distribution of frequencies is statistically different for both markers (p=0, RxC test), but excluding Northern samples (Ladini, Val di Scalve, Udine and Castelmassa) the difference is no more significant (p=0.24 for –13910 and p=0.46 for -22018, RxC test). South and Central Italy populations result more homogeneous among them.

Considering the four geographic groups, populations within Centre-North, Centre-South and South groups confirm their similarity (p>0,05, RxC test), while in the Northern ones the allelic distribution at -22018 locus results significantly different (p=0.02, RxC test). The data evidence Northern populations are more different among them, but respect to other Italian populations too, as we can see also from their position in the Figure 4.2.

4.3 Allele and genotype frequencies at C/T–13910 and G/A-22118 loci in Asian populations

Table 4.3 shows the distribution of allele and genotype frequencies in the two Asian groups considered in this study. In Middle East lactase persistence ranges from 6.8% to 14.3%; Central Asian values are lower ranging from 3.1% to 6.2% with the highest values in the two highland populations, the Kirghiz from Sary-Tash and the Kazakhs.

Homozygous subjects are rare: two TT-13910/AA-22018 homozygous subjects were found in the Kurdish sample, while 3 AA-22018 homozygous are present in the Kazakh sample.

Frequencies distribution within the two groups result homogeneous (P>0.05, RxC). In Central Asia, no significant differences were found either among different ethnic groups (Kirghizes- Uighurs- Kazakhs) or between low- and high-land populations (Kirghiz from Talas and Uighurs - Kirghiz from Sary-Tash and Kazakhs).

		N	CC	CT	TT	C	T	N	GG	GA	AA	G	A
Middle East	Kurds	49	0.755	0.204	0.041	0.857	0.143	48	0.750	0.208	0.042	0.854	0.146
	Persians	22	0.864	0.136	0.000	0.932	0.068	22	0.773	0.227	0.000	0.886	0.114
	tot	71	0.789	0.183	0.028	0.880	0.120	70	0.757	0.214	0.029	0.864	0.136
Central Asia	Kir-SaryTash	65	0.877	0.123	0.000	0.938	0.062	28*	0.714	0.286	0.000	0.857	0.143
	Kir-Talas	79	0.924	0.076	0.000	0.962	0.038	30*	0.800	0.200	0.000	0.900	0.100
	Uighurs	48	0.938	0.063	0.000	0.969	0.031	28*	0.857	0.143	0.000	0.929	0.071
	Kazakhs	105	0.886	0.114	0.000	0.943	0.057	30*	0.467	0.433	0.100	0.683	0.317
	tot	297	0.902	0.098	0.000	0.951	0.049	116*	0.707	0.267	0.026	0.841	0.159

Table 4.3 Relative genotype (in grey colour) and allelic (in white colour) frequencies at C/T–13910 and G/A-22118 loci observed in 2 Middle Eastern and 4 Central Asia populations; N indicates the sample sizes;* indicates a subsample.

4.4 Frequency of the SNP and STR haplotypes

A subset of samples was typed for additional 3 microsatellite polymorphisms D2S3010, D2S3013 and D2S3015, in order to analyse the genetic variation within the core haplotypes defined by C/T–13910 and G/A-22118 markers. A total of 413 subjects from Italy (N=225), Middle East (N=71) and Central Asia (N=117) have been typed, for a more precise sample description see *materials and method* section 3.2.3.

To infer 5-loci haplotypes we used the Bayesian method implemented in Phase v.2.1 software (Stephens & Donnelly 2003), that reports the "best" haplotype reconstruction for each sample and a list of the confidence probabilities associated with each phase call. 177 haplotypes were inferred and were jointed to Roma and Tocco ones (Coelho et al. 2005), for a total of 194 haplotypes; the haplotype list with the absolute frequencies in each population is reported in *supplementary material*, Table S1. *Phase callings* have confidence values >0.88, the probabilities associated to D2S3010 and D2S3013 loci are lower because they present a wider spectrum of alleles.

The frequencies of the core haplotypes defined by C/T-13910 and G/A-22018 polymorphisms are shown in Table 4.4. In Italy, the frequency of lactase persistence TA haplotype varies between 10-15%, excluded Val di Scalve sample (49%) that is not representative of the whole population, because we preferentially chose T-13910 allele chromosomes in order to deeply investigate the persistence lineages. Similar range have been found in Middle Eastern populations (9-14%). In Central Asia we found the lower frequency of TA lineages among the Uighurs, 5.2% and the highest among the Kazakhs, 20%; also for these populations we constructed a subset reducing the number of subjects and including all subjects having at least a T-13910 allele.

Population	N of chromosomes	Haplotypes		
		CG	CA	TA
SC*	84	0.500	0.012	0.488
PAL	66	0.818	0.030	0.152
ALF	82	0.854	0.037	0.110
TO	74	0.892	0.000	0.108
RM	60	0.850	0.000	0.150
ACS*	84	0.833	0.024	0.143
Italy	450	0.784	0.018	0.198
Ku	98	0.847	0.010	0.143
Per	44	0.864	0.045	0.091
Mid. East	142	0.852	0.021	0.127
Ksa*	56	0.857	0.036	0.107
Kta*	60	0.900	0.000	0.100
Uig*	58	0.931	0.017	0.052
Kaz*	60	0.683	0.117	0.200
Cent. Asia	234	0.842	0.043	0.115
total	826	0.812	0.025	0.162

Table 4.4 Frequencies of the core haplotypes defined by C/T-13910 and G/A-22018 polymorphisms in Italian, Middle Eastern and Central Asian populations; populations with * are subsamples.

The low frequency of CA intermediate haplotypes, that as the ancestral CG haplotypes are associated to lactase non-persistence, is related to the short time intercourse between G/A-22018 and C/T-13910 mutations (Poulter et al. 2003; Swallow 2003). The polymorphisms were originated according to C-G, C-A, T-A phylogenetic sequence; the absence of TG haplotypes depends on the absence of recombination between the two loci.

4.5 Microsatellite variation within the SNPs haplotypes

Microsatellite allele frequency distribution for the three population groups are shown in Figure 4.3; we separately analyzed the distribution within the CG non-persistence and TA persistence lineages. The distribution within each population is represented in supplementary material, Figure S1. The estimated sizes of allele 1 in each microsatellite are: 184bp for D2S3010; 125bp for D2S3013 and 175bp for D2S3015. Table 4.5 A and B show heterozygosity estimations (H) for each locus and the average for all loci within each populations and within each groups.

In general, the three microsatellite loci show different patterns of variation determined by different mutation rates, but distribution patterns and heterozygosity (H) for each locus are similar among the different population groups. CG lineages show higher variability than TA lineages, as expected for a relative recent origin for the C-T mutation at 13910 locus (Poulter et al. 2003; Swallow et al. 2003).

D2S3010 locus shows a bell shaped distribution and similar heterozygosity among the populations. Within the persistence TA haplotypes the range of variability is reduce and allele 10 is the most frequent in both Italian and Middle Eastern populations, while allele 11 is dominant in Central Asia. Allele 11 is the modal allele in Roma sample too.

D2S3013 has a bimodal distribution with a common 7 allele within CG lineages. The same bimodality is not observed within TA haplotypes, where the allele 21 is modal. Interesting we found a low heterozygosity in Val di Scalve likely determined by founder effect, allele 21 reaches a frequency of 85%.

D2S3015 locus is characterized by a lower mutation rate and presents a small number of alleles; allele 4 is the most common within CG lineages and it is the only one within TA lineages with the exception of allele 7, 2 chromosomes in Albidona sample, maybe determined by recombination events.

In contrast to a previous study (Coelho et al. 2005), we found the presence of recombinants within TA haplotype in both D2S3013 and D2S3015 loci, even if at very low frequencies. In spite of that, all populations have the same reduced pattern within TA lineages suggesting a single origin for all TA chromosomes from different geographic regions. The unimodal distribution at the D2S3013 locus, determined by a stepwise accumulation of mutations around allele 21, and the monomorphism at the D2S3015 locus (allele 4), suggest the mutation associated to persistence occurred on a 21-4 background (Coelho et al. 2005). It is more difficult to identified the ancestral allele at D2S3010 locus because it is more variable also within TA lineages.

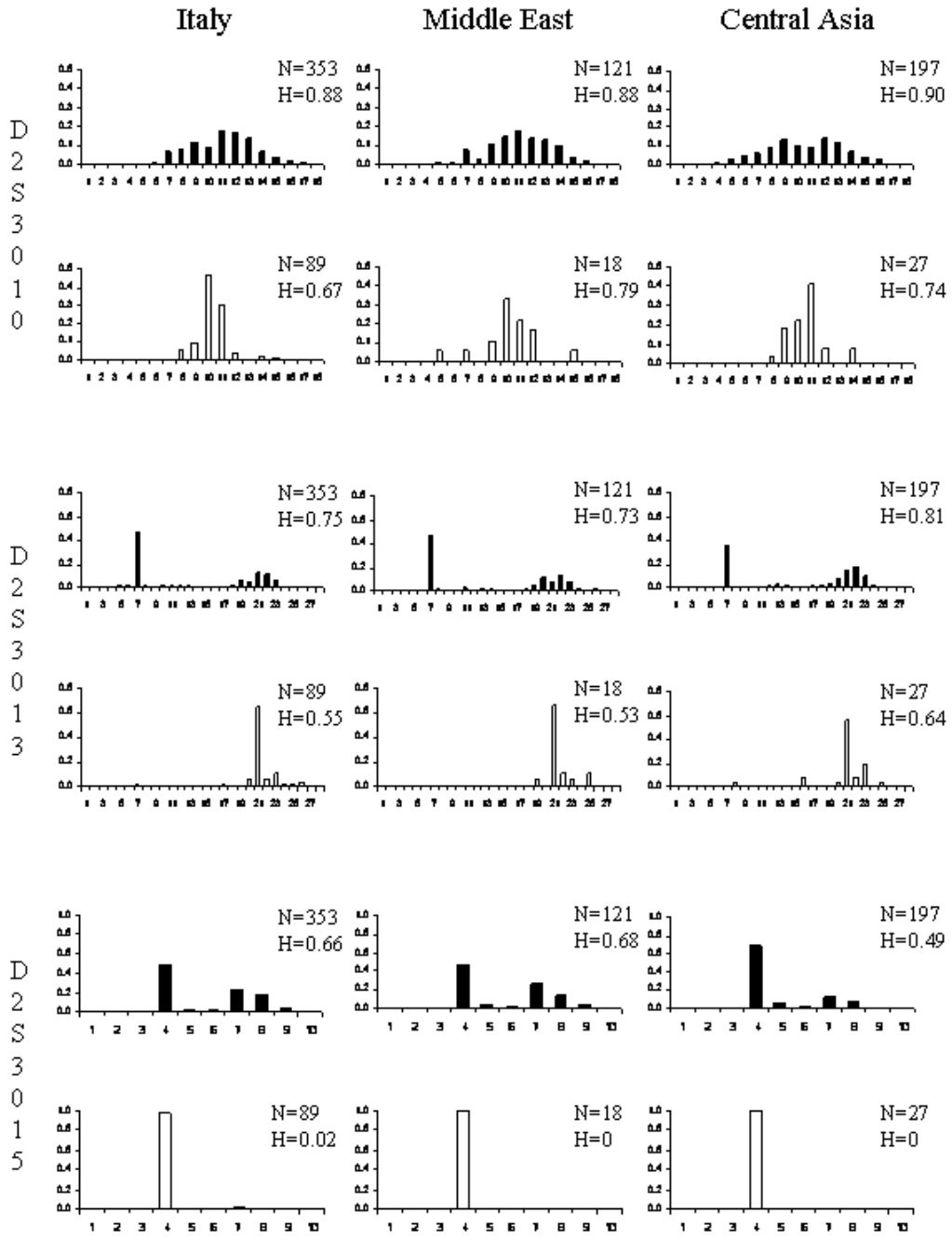


Figure 4.3 Micorsatellite allele frequency distribution within lactase non-persistence CG (in black) and lactase persistence TA (in white) (-13910/-22018 loci) haplotypes in Italian, Middle Eastern and Central Asian populations. N= number of chromosomes; H= heterozygosity.

A	SC	PAL	ALF	TO	RM	ACS	Italy
N of TA chr.	41	10	9	8	9	12	89
H D2S3010	0.689	0.740	0.494	0.594	0.444	0.625	0.671
H D2S3013	0.263	0.760	0.519	0.750	0.593	0.708	0.552
H D2S3015	0.000	0.000	0.000	0.000	0.000	0.153	0.022
avarage H	0.317	0.500	0.337	0.448	0.346	0.495	0.415
N of CG chr.	42	54	70	66	51	70	353
H D2S3010	0.762	0.872	0.881	0.888	0.880	0.863	0.880
H D2S3013	0.718	0.820	0.751	0.685	0.776	0.683	0.748
H D2S3015	0.670	0.585	0.628	0.716	0.613	0.678	0.659
avarage H	0.717	0.759	0.753	0.763	0.756	0.741	0.762

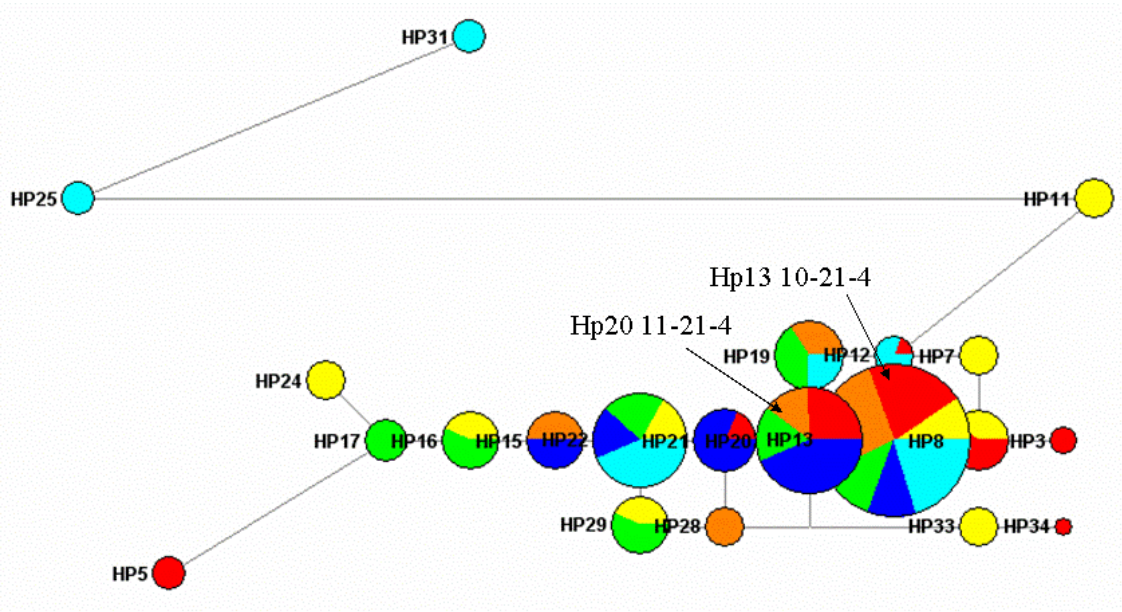
B	Ku	Per	Mid. East	KSa	KTa	Uig	Kaz	Asia
N of TA chr.	14	4	18	6	6	3	12	27
H D2S3010	0.786	0.625	0.790	0.667	0.722	0.444	0.667	0.738
H D2S3013	0.459	0.625	0.525	0.278	0.722	0.444	0.667	0.642
H D2S3015	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
avarage H	0.415	0.417	0.438	0.315	0.481	0.296	0.444	0.460
N of CG chr.	83	38	121	48	54	54	41	197
H D2S3010	0.865	0.863	0.876	0.893	0.881	0.899	0.886	0.902
H D2S3013	0.724	0.733	0.733	0.820	0.753	0.793	0.816	0.808
H D2S3015	0.661	0.723	0.684	0.294	0.492	0.583	0.535	0.489
avarage H	0.750	0.773	0.764	0.669	0.709	0.758	0.746	0.733

Table 4.5 Number of lactase persistence TA and lactase non-persistence CG (-13910/-22018 loci) haplotypes, heterozygosity for each locus and average heterozygosity. (A) In Italian population: Val di Scalve SC, Palazzuolo PAL, Alfero ALF, Tocco TO, Roma RM and Albidona ACS. (B) in Asian populations: Kurds Ku, Persians Per, Kirghiz of Sary Tash KSa, Kirghiz of Talas KTa, Uighurs Uig and Kazakhs Kaz.

4.6 Phylogenetic analysis of TA lineages

We calculated median joining (MJ) networks that show the relationships between the observed TA haplotypes and their distribution across populations. We realized separate networks for the three continental groups (Figure 4.4, 4.5 and 4.6) and than we compare these haplotypes to Portugal, Sao Tomè and Fulbe from Cameroon, the only data in literature that are based on the same markers (Coelho et al. 2005) (Figure 4.7). All TA haplotypes used in this analyses are listed in *supplementary material,* Table S2. To observe the distribution patterns without being influenced by the sample sizes, we calibrated the networks: we recalculated haplotype frequencies considering the same population size for all populations in the network.

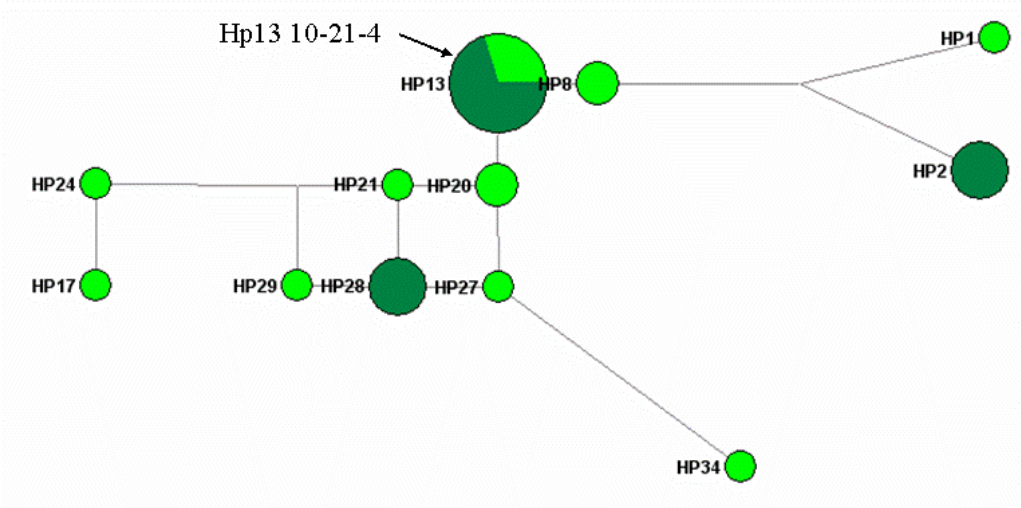
In Italian populations haplotype **Hp13**, determined by the modal allele for each microsatellite **10-21-4**, is the most common (38%) and it is the only one presents in all six samples (Figure 4.4). **Hp 20** differs for a single mutational step at locus D2S3010 (**11-21-4**) and it is mainly observed in Roma sample. Haplotype diversity vary among populations with a medium value of 0.823 ±0.034.



	N of TA chr.	N of TA Hp	polimorphic sites	haplotype diversity	S.D.
SC	41	8	2	0.755	0.052
PAL	10	9	2	0.978	0.054
ALF	9	5	2	0.722	0.016
TO	8	7	2	0.964	0.077
RM	9	5	2	0.861	0.087
ACS	12	6	3	0.803	0.096
Italy	89	21	3	0.823	0.034

Figure 4.4 Calibrated Median Joining Network of the TA haplotypes in six Italian populations. Each circle represents a different haplotype, its size is proportional to its relative frequency and the presence in each population is indicated with a different color. The two most common haplotypes are shown. Standard diversity indexes are reported in the table.

Hp 13 is the most frequent in Middle East too (27%) and it is the only one shared between Kurds and Persians (Figure 4.5).

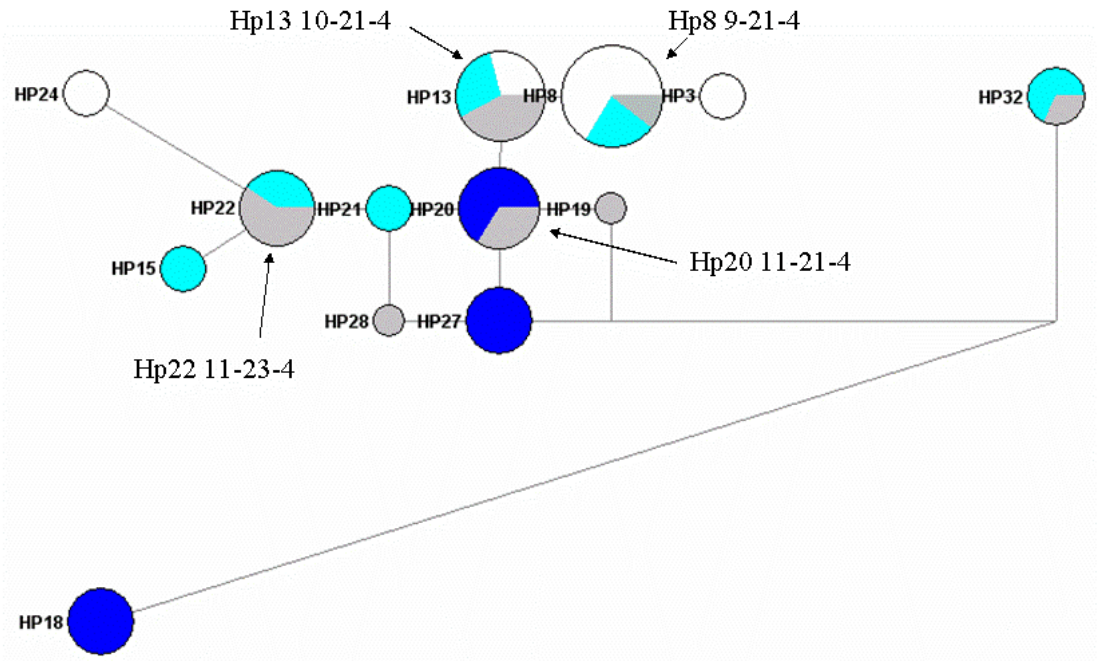


	N of TA chr.	N of TA Hp	polimorphic sites	haplotype diversity	S.D.
Ku	14	10	2	0.945	0.045
Ir	4	3	2	0.833	0.222
Mid. East	18	12	2	0.922	0.051

Figure 4.5 Calibrated Median Joining Network of the TA haplotypes in Middle Eastern populations, each population is indicated with a different color. The most common haplotype is shown. Standard diversity indexes are reported in the table.

.

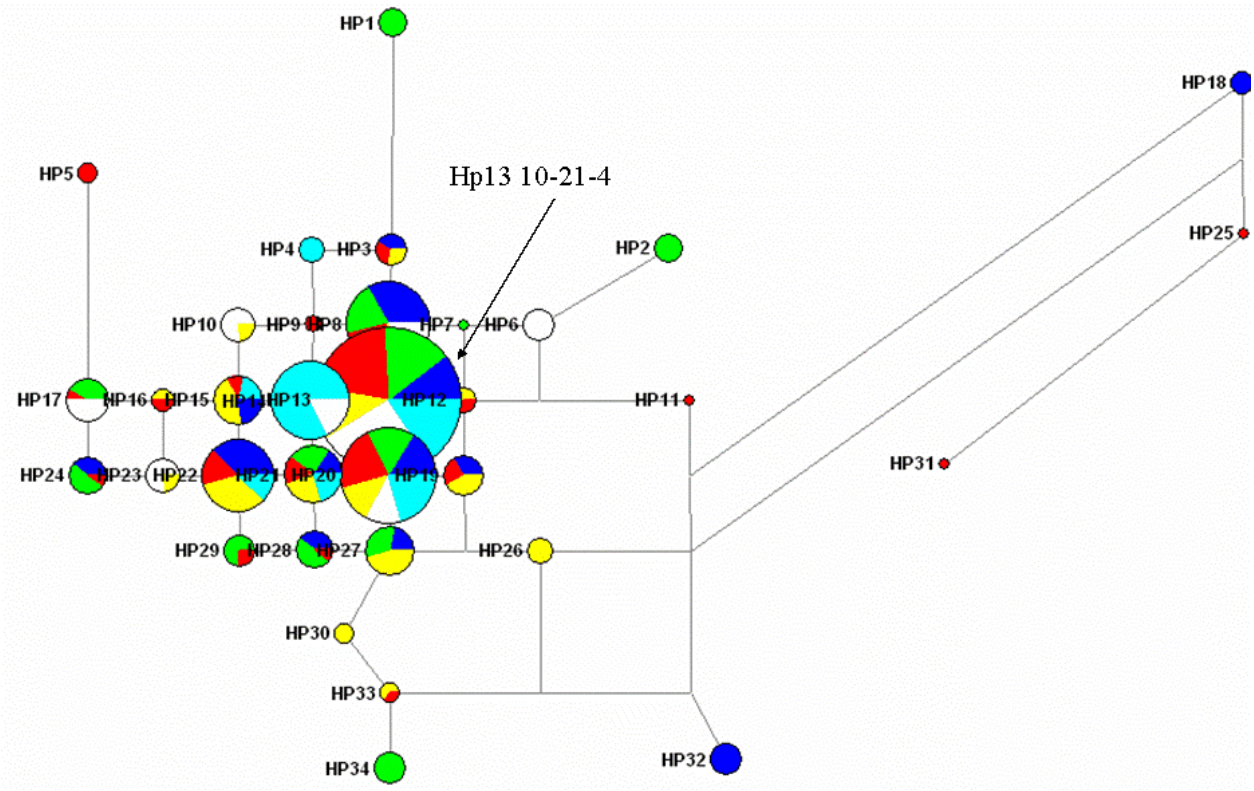
The examination of Figure 4.6 reveals that **Hp8** (**9-21-4**) and **Hp13** are the most frequent in Central Asia populations; both haplotypes are absent in the Uighurs. Asian samples present an higher haplotype diversity (HD=0.915 ±0.029) determined by a relative higher number of different haplotypes with similar frequencies across populations: 9-21-4 (19%), 10-21-4 (19%), 11-23-4 (15%) and 11-21-4 (11%).



	N of TA chr.	N of TA Hp	polimorphic sites	haplotype diversity	S.D.
KSa	6	4	2	0.800	0.172
KTa	6	6	2	1.000	0.096
Uig	3	3	2	1.000	0.272
Kaz	12	7	2	0.894	0.063
Cent. Asia	27	13	2	0.915	0.029

Figure 4.6 Calibrated Median Joining Network of the TA haplotypes in Central Asian populations, each population is indicated with a different color. The most common haplotypes are shown. Standard diversity indexes are reported in the table.

We compare our haplotype data to Portugal (60% of lactase persistence estimated on the basis of C/T-13910 and G/A-22018 markers) Sao Tomè (7.8% of lactase persistence likely due to recent admixture with Europeans) and Fulbe haplotypes (38% of lactase persistence; this pastoralist population from Cameroon is among the few African populations where the T-13910 allele is present, maybe introduced through retromigrations from North-Africa) (Mulcare et al. 2004; Coelho et al. 2005). **Hp13** is confirmed to be the worldwide most frequent haplotype with an ubiquitous distribution, followed by Hp20 and Hp8. In the Fulbe sample Hp13 reaches high frequencies (29%) but the most common haplotypes in this African group is Hp14 10-22-4 (38%) (Figure 4.7).



	N of TA chr.	N of TA Hp	polimorphic sites	haplotype diversity	S.D.
Fulbe	21	7	2	0.781	0.062
Sao Tomè	13	8	2	0.808	0.113
Portugal	66	17	2	0.909	0.017
Italy	89	21	3	0.823	0.034
Europe	155	27	3	0.867	0.019
Mid. East	18	12	2	0.922	0.051
Cent. Asia	27	13	2	0.915	0.029

Figure 4.7 Calibrated Median Joining Network of the TA haplotypes in the three population groups considered in this study (Italy, Middle East and Central Asia) and in Portugal, Sao Tomè and Fulbe from literature (Coelho et al. 2005). Each population is indicated with a different color. The most common haplotype is shown. Standard diversity indexes are reported in the table.

Cause of its frequency and distribution Hp13 (10-21-4) could be considered the ancestral haplotype for lactase persistence, consistent with the results of Coelho et al (2005). On the other hand in Central Asia there are other common haplotypes, persistence mutation could have appeared on a haplotype different from the haplotype that further spread to other continents and reached great frequencies.

Sorted by decreasing haplotype diversity, the group of Asian populations (both Middle East and Central Asia ones) was followed by European (including Portugal and Italian chromosomes), the admixed Sao Tome and Fulbe from Africa. Small differences could be related to the fast increasing of persistence trait frequency cause of its great selective advantage.

4.7 Analyses of inter-population variability within TA lineages

Genetic pair-wise differences among the 12 populations analyzed in this study have been visualized by a no-metric Multidimensional Scaling (MDS) (Figure 4.8). Populations cluster for geographic area, except for ACS sample from Italy that clusters closer to Asian populations likely for the presence of recombinant chromosomes. We found no differences among population within Middle East and Central Asia; Italian populations resulted less homogeneous among them (Fst=0.045, p=0.031) and this diversity is mainly determined by SC sample: excluding this sample from North Italy the genetic difference was no more significant (Table 4.6).

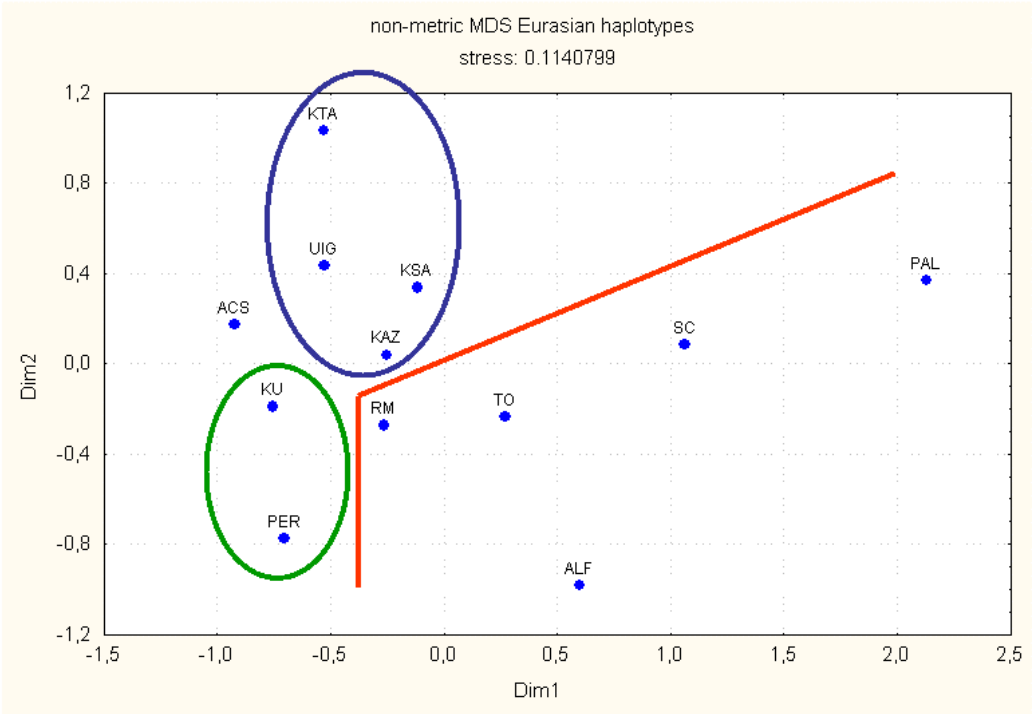


Figure 4.8 Non-metric Multidimensional Scaling (MDS) representation of pair-wise F_{ST} values among the 12 populations. The genetic distance matrix was realized from haplotype data.

	Fst	p value
Mid. East	-0.019	0.561
Cent. Asia	-0.002	0.488
Italy	0.045	0.031
no SC	-0.027	0.772
no PAL	0.055	0.024
no ALF	0.060	0.013
no TO	0.054	0.014
no RM	0.039	0.053
no ACS	0.053	0.029
Global	0.028	0.003

Table 4.6 Analyses of inter-population variability using Fst. In grey, Fst estimation in the complete Italian sample and excluding one population a time.

The analyses of molecular variance (AMOVA) in the three population groups suggest that the portion of genetic variance allocated among populations (F_{ST}) is 0.029 (P=0.051), while no difference has been found among groups. Furthermore, the analysis of genetic structure shows that: first the largest differences are among Italian and Central Asian populations (Fst=0.045, p=0.028); second Middle East populations are similar to Central Asian ones (Fst=0) and little divergent from Italian one (Fst=0.019, p=0.057) (Table 4.7).

AMOVA			p value
Italy-Mid. East-Cent. Asia	Fct	-0.003	0.678
	Fst	0.029	0.051
	Fsc	0.032	0.073
Italy-Mid. East	Fct	-0.021	0.844
	Fst	0.019	0.057
	Fsc	0.039	0.040
Italy-Cent. Asia	Fct	0.009	0.462
	Fst	0.045	0.028
	Fsc	0.036	0.064
Mid. East-Cent.Asia	Fct	-0.001	0.464
	Fst	-0.005	0.541
	Fsc	-0.005	0.524

Table 4.7 Analyses of molecular variance (AMOVA) in the three population groups.

Global (Central Asia, Middle East, Italy, Portugal, Sao Tomè and Fulbe samples) estimation of Fst (0.028, p=0.003) are very lower than the average F_{ST} observed at genomic level for human populations (12-13%) (Jorde et al. 2000; International

HapMap Consortium, 2005); obviously we are analysing a genomic region under strong positive selection that reduces genetic variability (Table 4.6).

4.8 Estimation of the age of the T-13910 allele

Intra-allelic accumulation of microsatellite diversity was used to estimate the age of T-13910 allele associated to lactase persistence in Eurasian populations. The method is based on the decrease in frequency of modal allele on which the mutation appeared; the modal allele at each microsatellite locus in the pooled sample was considered to be the ancestral, 10-21-4 respectively for D2S3010, D2S3013 and D2S3015 loci. Age estimations were performed considering different mutation rates and the presence or absence of recombination (m1, m2 m3 and m4, Table 4.8) (Coelho et al, 2005).

Considering intra-allelic diversity has been shaped only by mutation, estimates are between 11,469 and 25,136 years in the pooled sample (m1 and m2). Assuming both mutation and recombination, estimates are between 7,374 and 12,162 years: in this case the observed haplotype homogeneity can be only explained by a very recent origin of the T-13910 allele (m3 and m4). The latter estimations are more realistic.

	m1	**m2**	**m3**	**m4**
Italy (N=89)	21,793 (15,115-29,086)	9,689 (6,301-13,483)	10,261 (7,037-14,331)	6,184 (4,148-8,990)
Mid. East (N=18)	27,642 (12,853-55,470)	12,353 (5,464-25,176)	14,186 (6,304-48,301)	8,798 (3,658-23,090)
Cent. Asia (N=27)	38,508 (23,015-70,092)	17,509 (10,439-32,166)	19,501 (11,263-63,309)	12,481 (6,794-34,565)
Total (N=134)	25,136 (19,187-32,333)	11,469 (8,441-14,457)	12,162 (9,109-15,901)	7,374 (5,422-10,084)

Table 4.8 Age estimations of the most recent common ancestor of the T-13910 allele; 95% confidence intervals are gives in parentheses. Estimations were calculated assuming different mutation rates and the absence (m1 and m2) or presence (m3 and m4) of recombination.

Age estimations indicate the time since intra-allelic diversity began to accumulate due to an increase in frequency or population expansion. The mutation results older in Central Asia, followed by Middle East and Italy. SC sample has shown to be different from the other Italian samples, we estimate the age without SC chromosomes and Italian estimations results anyway younger than Middle Eastern ones (data not shown).

4.9 Neutrality test

To figure out if the observed patterns of variability within TA-lineages were consistent with the standard neutral model of evolution, we performed neutrality tests assuming different demographic models and different microsatellite mutational rates (see *materials and method* section 3.3.9).

Neutrality is rejected when the probabilities of finding a number of mutations $\leq S_0$ in the microsatellite loci linked to T-13910 allele is lower than 0.001. Neutrality is always rejected in the Italian sample, while in Middle Eastern and Central Asian populations neutrality is rejected only when an higher mutational rate is assumed (m2) for both demographic models (Table 4.9).

				D1		D2	
Population	ni	n	S0	m1	m2	m1	m2
Italy	89	450	40	1.21E-05	1.64E-35	1.46E-17	2.87E-80
Mid. East	18	142	19	0.603	1.97E-05	0.088	1.13E-11
Cent. Asia	27	234	25	0.508	2.23E-07	0.015	9.59E-18

Table 4.9 Neutrality test. Probability of finding a number of mutations $\leq S_0$ in the microsatellite loci linked to T-13910 allele under two demographic models (D1-D2) and two mutational rates (m1-m2). n_i is the number of chromosomes with the T-13910 allele; n the total number of chromosomes in the sample; S_0 the minimum number of mutations.

DISCUSSION AND CONCLUSIONS

Lactase persistence is a subject of great interest because the genomic region including LCT gene (chr 2q21), which encodes the enzyme lactase-phlorizin hydrolase, is largely considered one of the strongest examples of positive selection pressure (Bersaglieri et al. 2004). It also represents a good example of genetic-environment coevolution, since the genetic trait seems to have increased in frequency starting from the Neolithic Transitions, together with the diffusion of dairy farming and the habit of drink milk in adulthood, that gives nutritional advantages (Holden & Mace 1997).

From a molecular point of view, the discovery of two polymorphisms, the C/T-13910 and the G/A-22018 - whose derived alleles are respectively totally and highly associated to lactase persistence in Eurasian populations - represented an important new tool of investigation (Enattah et al. 2002).

The subject of this thesis is a molecular screening of lactase persistence in three population groups: Italians, Middle Easterns and Central Asians; followed by a phylogeographic analysis of results in order to contribute to the reconstruction of the evolution of lactase persistence trait across Eurasian populations.

To this aim, genetic markers with different mutational rates, located in the region surrounding LCT gene (chr. 2), were used. Biallelic markers (SNPs) associated to lactase persistence were important to identify lactase persistence lineages, on the other hand microsatellite markers (STRs) were fundamental to capture the variability within the lineages (Coelho et al. 2005). In fact, STRs are fast evolving polymorphisms that need less time to accumulate intra-allelic variability, an important feature in a region where positive selection has drastically reduced the genetic variability. When an advantageous mutation rises in frequency also the surrounding region rises in frequency; under selection pressure this process is so fast that neither mutation or recombination has enough time to disrupt the haplotype. The consequence is a large region of linkage disequilibrium between alleles; considering only SNPs a larger number of markers, covering a wide genomic region would be necessary to evidence some sources of variability.

The sampling strategy

In the first part of the study we assessed the distribution of lactase persistence, typing the C/T-13910 and the G/A-22018 markers, in a total of 1117 subjects coming from geographically and ethnically different populations within Eurasia. In this way we contributed to improve data available in literature on molecular persistence distribution across human populations.

We realized an extend sample collection in Italy, dividing the country in four geographic areas and sampling 4-6 populations within each area for a total of 749 individuals. We also considered the level of isolation of samples and verified the provenience of each subject collecting bio-demographic information on the last three generations.

The Asian sample consists of two populations from Middle East (Kurds and Persian from Iran), a particularly interesting region where the Neolithic Transition process that mainly involved Europe has developed. The Central Asian sample is composed of two Kirghiz groups from Sary-Tash and Talas, Uighurs and Kazakhs; all populations have a nomadic pastoral origin and their economy is based mainly on pastoralism.

Lactase persistence variability in Italy

The wide Italian sample collection was necessary to improve avaible molecular data on our country and to better analyze lactase persistence distribution, because in previous physiological estimations different patterns were observed.

Our findings, based on T-13910 allele frequencies, were in agreement with the physiological data that support a decreasing cline of lactase persistence from North to South (Marenco et al. 1970; Arrigo et al. 1980; Zuccato et al. 1983; Burgio et al. 1984; Rinaldi et al. 1984). Both kinds of data revealed the same trend, but physiological estimations of lactase persistence are higher. This could be related to the lack of sensitivity of physiological tests that identify false negative individuals, wrongly considered to be lactase persistent. For example in the “Breath Hydrogen Determination” test, the level of hydrogen derived from metabolism of lactose by colonic bacteria in non-persistent subjects could depend on the type of colonic bacteria, that influences the quantity of hydrogen produced. Moreover, there are some antibiotics suppressing the colonic hydrogen formation (Arola 1988; Flatz 1987).

The decline of lactase persistence toward South is typical of the European continent (Flatz 1987; Swallow 2003), but Italian populations are among the European populations with the lowest prevalence of lactase persistence (Flatz 1987). With an average molecular estimation of 15% (frequency of T-13910 allele), Italy is similar to other Mediterranean populations (from Morocco to Greece with a physiological range between 9-47%; Flatz 1987). Neighbouring European populations have physiological level of lactase persistence higher than 63% (Spain 85%, France 77%, Swiss 84%, South Germany 77%, Austria 80% and ex Yugoslavia 63%; Flatz 1987). Lactase persistence frequencies are very high in North and Central Europe where milk consumption has been a greater biological advantage, both as a source of energy and as a supply of calcium (Flatz & Rotthauwe 1973). The strong decrease in frequency in Italy could be the result of the migration barrier of Alps (Flatz 1987) and of a more intensive gene flow with other Mediterranean populations.

Although we didn't reveal the so marked differences in lactase persistence distribution found by Cavalli-Sforza (1987), we evidenced some microgeographic variability within Italian populations. As expected, Northern populations showed higher level of T-13910 and A-22018 alleles, but they resulted significantly different among them and when compared to other Italian populations. One explanation could be the presence of "genetic isolates", such as the ethnic group of Ladini, that are linguistically isolated, or the Val di Scalve sample, located in a geographically isolated valley close to Bergamo. Genetic drift action, reducing genetic variability within a population and increasing differences among populations in absence of gene flow, could have contributed to increase the frequencies of the derived alleles. But following our criteria of sample collection, isolated populations have been included within each geographic area, for example Alfero (FC) that shows lower level of persistence than the neighbouring Alta Val Savio sample (FC) and Albidona (CS) with higher values of persistence respect to other Southern populations; anyway the allelic distribution within these groups are homogeneous. An additional explanation could be a greater gene flow with other European countries, where lactase persistence is more common as described above. It's largely known that from a genetic point of view, Northern Italian populations are similar to Central Europe populations, while Central and Southern populations are closer to populations from Greece and other Mediterranean countries (Cavalli-Sforza et al. 1994). Furthermore,

the highest frequencies of the derived alleles within the Ladini samples are probably related to their South Germany origin.

Lactase persistence in Asia

The frequency of lactase persistence established typing the C/T-13910 polymorphism varied widely among the two Asian groups considered in this study, but it is in line with the decreasing frequency of the trait, that is very common in Europe and almost non existent among East Asians (Flatz 1987; Sahi et al. 1994; Bersaglieri et al. 2004; Enattah 2005, unpublished data).

In Middle East the frequency ranges from 14.3% in Kurds to 6.8% in Persians, consistent with the low frequency of lactase persistent allele observed in the region (Bersaglieri et al. 2004; Enattah 2005, unpublished data; Ingram et al. 2006). The high value in Kurds is within the range of the region; a previous study demonstrated that, also for other molecular markers, Kurds and other Middle Eastern populations show similar patterns of genetic variability (Useli 2006, unpublished data).

Finally, in the four Central Asian groups we found persistence values between 3.1-6.2%. No significant differences were found either among different ethnic groups (Kirghizes-Uighurs-Kazakhs) or between low and high altitude populations (Kirghiz from Talas and Uighurs - Kirghiz from Sary-Tash and Kazakhs), even if lactase persistence reaches higher frequencies in highland populations, probably as a consequence of a strong founder effect (Perez-Lezaun et al. 1999).

Phylogeographic analyses

The T-13910 allele is associated to lactase persistence in all Eurasian population till now tested, in agreement with epidemiological data (Enattah et al. 2002; Bersaglieri et al. 2004; Mulcare et al. 2004; Coelho et al. 2005).

In order to understand the origins and evolution of lactase persistence, we analysed the genetic variability within the persistence lineages, defined by the derived alleles at –13910 and –22018 loci, in three population groups coming from different geographic areas of Eurasia, using the approach developed by Coelho et al. (2005). This approach is based on the use of fast evolving markers (STRs) able to capture the variability within TA lineages because they need less time to accumulate alleles, an important feature considering they lie in a genomic region where natural

selection has reduced the genetic variability (Harvey et al. 1995; Poulter et al. 2003; Bersaglieri et al. 2004; Coelho et al. 2005).

Within the persistence TA lineages the range of intra-allelic variability was reduced respect to non-persistence CG lineages, as expected for a relative recent origin for the T-13910 allele (Poulter et al. 2003; Swallow et al. 2003) and a consequent reduced amount of time to accumulate new alleles. The allelic pattern was the same within TA chromosomes from different geographic regions suggesting a single origin for all of them. A common origin for all persistence chromosomes was suggested by Bersaglieri (2004), that observed the T-13910 allele was found both in Europe than in Middle East and Pakistan samples on the same genetic background, the haplotype A.

Modal allele for each microsatellite has been clearly identified for D2S3013 (21) and D2S3015 (4) loci, while D2S3010 locus shows a higher variability and different modal alleles in Italy and Middle East (10) and in Central Asia (11). Phylogenetic analyses revealed the presence of a common haplotype 10-21-4 in all populations studied in this thesis and in all European and African populations from literature analyzed for the same set of markers (Portugal, Sao Tomè and Fulbe) (Coelho et al. 2005). Hp 10-21-4 could be the ancestral haplotype, the allelic background on which lactase persistence mutation occurred as already proposed by Coelho et al. (2005). On the other hand, in Central Asian populations there are many haplotypes with similar frequencies 9-21-4 (19%), 10-21-4 (19%), 11-23-4 (15%) and 11-21-4 (11%); we can not exclude that the ancestral haplotype could be different from the one (10-21-4) that further has reached the highest frequency in other populations.

According to intra- and inter-population variability and the age estimations of the T-13910 allele, suggesting an older presence of the C/T-13910 mutation in Central Asia, we proposed a possible Central Asian origin for the mutation associated to lactase persistence. The diffusion in Central Asian samples of many haplotypes with similar frequencies and a great haplotype diversity could be determined by an older presence of the persistent lineages. Usually ancestral populations show a higher level of genetic diversity because the action of evolutionary forces, such as mutation and recombination, is proportional to time

elapsed. Haplotype diversity values were intermediate in Europe and lower in African samples (Coelho et al. 2005); in the latter the T-13910 allele presence is related to a strong European admix in the Sao Tomè sample and likely introduced through retro-migration from North Africa in the Fulbe ethnic group of Cameroon (Mulcare et al. 2004; Coelho et al. 2005).

Interestingly, Enattah (2005, unpublished data) analysed the genetic variability associated to lactase persistence in 37 populations from Eurasia and Africa origin using 9 biallelic markers, and found a high frequency of the pre-persistence haplotype (hap 4) and maximal diversity of persistence haplotypes in Central Asia (East slope of the Urals). He proposed, using a different set of markers, the same region (Central Asia) as the possible place of origin of the mutation associated to lactase persistence.

Our age estimations suggest a pre-Neolithic origin of the T-13910 allele in Central Asia, between 12,400 and 19,500 years B.P., considering more realistic estimations that assume recombination. These estimations confirm the T-13910 allele appeared after the separation from African populations between 100,000-50,000 years ago - explaining the absence of the T-13910 allele in Africa (Mulcare et al. 2004; Tishkoff et al. 2006) - and before the Neolithic Transition occurred approximately 10,000 years ago (Bersaglieri et al. 2004; Coelho et al. 2005).

The migration of individuals carrying the T allele from Central Asia to Europe could be responsible for the diffusion across different populations of the persistent trait. The intermediate genetic position of Middle East populations revealed by AMOVA could depends on its intermediate geographic position and consequently gene flow from the Mediterranean area and Central Asia. On the other hand the genetic features of Middle East sample could be related to the genetic gradient determined by the spread of diffusion of T-13910 allele from Asia to Europe. We could hypothesize that lactase persistence has reached Europe following the Neolithic Transition diffusion. The agriculture and dairying diffusion from Middle East toward West started around 10,000 years ago; it is subject of debate if the Neolithic expansion was a cultural diffusion of new technologies or a demic diffusion caused by a strong demographic development (Ammerman & Cavalli-Sforza 1984). The analysis of the first principle component by Cavalli-Sforza et al

(1996) shows a genetic gradient across Europe with a centre of origin in Middle East, supporting the demic diffusion hypothesis. The gradient is consistent with archaeological data, such as the age estimations of cereal arrival in Europe (Diamond 1998).

The age of the T-allele in the Middle East sample (8,800-14,100 B.P.) brackets the development of agriculture and dairying, and in Italy (6,200-10,200 B.P.) is consistent with the introduction of pastoralism, as demonstrated by archaeological studies (Cavalli-Sforza et al. 1994).

This possible relationship between Neolithic diffusion and lactase persistence diffusion, was underlined by Myles et al. (2005) too. They studied allele T-13910 distribution in North-African populations and found the lactase persistence was probably introduced together with the domestication of ovicaprids, from Near East 7,000 years B.P., as confirmed by archaeological, linguistic and Y-chromosomal data evidences (Holl 1998; Blench 2001; Arredi et al. 2004).

Our age estimation of the T-13910 allele are also consistent with previous ones obtained using the same method, the intra-allelic accumulation of microsatellite diversity, (7,440-12,300 years B.P.) (Coelho et al. 2005) and by the decay of LD in either direction from the LCT core region and intragenic recombination (2,188-20,650 years B.P. in American European and 1,625-3,188 years B.P. in Scandinavian populations) (Bersaglieri et al. 2004).

A recent study on ancient DNA (Burger et al. 2007) tested the C/T-13910 polymorphism in 8 early Neolithic (mostly 7,800-7,000 years ago) and 1 Mesolithic (4,200-2,100 years ago) skeletons from Central Northeast and Southeast Europe; they found only the ancestral allele C and concluded (even if the sample size is small) the allele T was not present at high frequencies in Europe in early Neolithic, before the diffusion of dairying and the development of advantageous condition. Bersaglieri (2004) supported the strong positive selection occurred in European populations after the separation from Asian and African American samples, because even if the persistence mutation lies on the same haplotypes (the haplotype A), the haplotype homozigosity is higher in European populations. From the neutrality tests performed in this study, neutrality is always rejected in the Italian sample, while in Middle Eastern and Central Asian populations neutrality is rejected only when an higher mutational rate is assumed for both demographic models considered.

Following our hypothesis the T-13910 allele reached Middle East from Central Asia, with a further diffusion across Europe in correspondence of the Neolithic diffusion of agriculture and dairying. In contrast to this hypothesis, Enattah (2005, unpublished data) suggested a later diffusion of the lactase persistence mutation: he hypothesized an origin of the mutation among nomadic groups in the East slope of the Urals. Due to nomadic migrations the mutation reached the Western slope of Ural and spread to region between the Volga and the Urals, north of the Caucasus and the Black Sea. He hypothesized this diffusion in Europe was likely related to the expansion of pastoral nomads speaking Indo-European languages from the Kurgan area (above the Caucasus) between 4,500-3,500 B.P (Cavalli-Sforza 1994).

Concluding remarks

The present study confirmed the presence of a decreasing gradient from North to South of lactase persistence in Italian populations consistent with physiological data and the persistence distribution pattern in Europe. Italian populations are among the European populations with the lowest prevalence of lactase persistence (15%), likely determined by the migration barrier of Alps and a more intensive gene flow with other Mediterranean populations

The frequency of lactase persistence decreases in Middle Eastern and Central Asian populations, in agreement with previous physiological and molecular estimations.

Due to intra- and inter-population variability and the age estimations of the T-allele, obtained using fast evolving markers (STRs), we proposed a possible Central Asian origin for lactase persistence, in agreement with Enattah's findings. The jointed results could imply lactase persistence arrived in Europe from Central Asia through at least two ways: the first during the Neolithic and lactase persistence spread across Europe together with the agriculture and dairying diffusion; the second one during the Bronze Age and is related to the expansion in Europe of pastoral nomads speaking Indo-European languages.

REFERENCES

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol., 2(10):e286.

Ammerman AJ, Cavalli-Sforza LL (1984). The Neolithic Transition and the genetics of populations in Europe. Princeton University Press, Princeton, N.J.

Aoki K (1986). A stochastic model of gene-culture coevolution suggested by the "culture historical hypothesis" for the evolution of adult lactose absorption in humans. Proc Natl Acad Sci U S A, 83:2929-33.

Aoki K (2001). Theoretical and empirical aspects of gene-culture coevolution. Theor Popul Biol 59:253–261.

Arola H, Koivula T, Jokela H, Jauhiainen M, Keyrilainene O, Ahola T, Uusitalo A, Isokoski M (1988). Comparisons 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.

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

Arrigo L, Ciangarotti S, Lantiere PB, Capponetto A, Rovida S, Bozzatti D, Detoni T, Duillo MT (1980). Indagine epidemiologica sul grado di incidenza dell'ipolattasia e dell'intolleranza al lattosio, in campioni di popolazione adulta e pediatrica, in Liguria. Rev. Soc. Ital. Sci. Aliment., 9:391-400.

Bandelt HJ, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, 16:37-48. http://www.fluxus-engineering.com

Beja–Pereia 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(4):311-3.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet., 74:1111-1120.

Blench RM (2001). Types of language spread and their archaeological correlates: the example of Berber. Origini, 23:169–189.

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.

Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG (2007). Absence of the lactase-persistence-associated allele in early Neolithic Europeans. PNAS, 104(10):3726-3741.

Burgio GR, Flatz G, Barbera C, Patane R, Boner A, Cajozzo C, Flatz S (1984). Prevalence of primary adult lactose malabsorption and awareness of milk intolerance in Italy. Am. J. Clin. Nutr., 39:100-104.

Cavalli-Sforza LL. (1966). Population structure and human evolution. Proc R Soc Lond B Biol Sci, 164(995):362-79.

Cavalli-Sforza LL (1973). Analytic review: some current problems of human population genetics. Am J Hum Genet, 25:82-104.

Cavalli-Sforza LL (1996). Geni, popoli e lingue. Adelphi.

Cavalli-Sforza LL, Strata A, Barone A, Cucurachi L (1987). Primary adult lactose malabsorption in Italy: regional difference in prevalence and relationship to lactose intolerance and milk consumption. Am. J. Clin. Nutr., 45:748-54.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). The History and Geography of Human Genes. Princeton University Press, Princeton, New Jersey.

Coelho M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, Destro-Bisol G, Rocha J (2005). Microsatellite variation and evolution of human lactase persistence. Hum Genet, 117: 329–339.

Cook GC (1978). Did persistence of intestinal lactase into adult life originate on the Arabian Peninsula? Man (N.S.), 13:418-27.

Dahlqvist A (1964). Method for Assay of Intestinal Disaccharides. Anal Biochem, 57:18-25.

Diamond J (1998). Armi, acciaio e malattie. Breve storia negli ultimi tredicimila anni. Einaudi, pp139.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I (2002). Identification of a variant associated with adult – type hyptolactasia. Nat. Genet., 30:233-37.

Enattah NS (2005). Molecular genetics of lactase persistence. PhD dissertation, University of Helsinki. Unpublished data.

Escher JC, De Koning ND, Van Engen CG, Arora S, Büller HA, Montgomery RK, Grand RJ (1992). Molecular basis of lactase levels in adult humans. J. Clin. Invest., 89:480-483.

Excoffier L, Smouse P, Quattro J (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics, 131:479-491.

Excoffier, Laval LG, Schneider S (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50. http://cmpg.unibe.ch/software/arlequin3/

Flatz G, Rotthauwe HW (1973). Lactose nutrition and natural selection. Lancet, 2:76-7.

Flatz G, Rotthauwe HW (1977). The human lactase polymorphism: physiology and genetics of lactose absorption and malabsorption. Prog Med Genet, 2:205-49.

Flatz G (1987). Genetics of lactose digestion in humans. Adv Hum Genet, 16:1-77.

Goldstein DB, Reich DE, Bradman N, Usher S, Seligsohn U, Peretz H (1999). Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am J Hum Genet, 64:1071–1075.

Grosso G (1934). Marche: dialetti. In: Gentile G, ed. Eniclopedia Italiana. Roma, Italy: Rizzoli& C, 22:232-3.

Harvey CB, Fox MF, Jeggo PA, Mantei N, Povey S, Swallow DM (1993) Regional localization of the lactase phlorizin hydrolase gene, LCT, to chromosome 2q21. Ann Hum Genet, 57 (Pt 3):179-85.

Harvey CB, Pratt WS, Islam I, Whitehouse DB, Swallow DM (1995). DNA polymorphisms in the lactase gene. Linkage disequilibrium across the 70-kb region. Eur J Hum Genet, 3:27-41.

Harvey CB, Wang Y, Darmoul D, Phillips A, Mantei N, Swallow DM (1996). Characterisation of a human homologue of a yeast cell division cycle gene, MCM6, located adjacent to the 5' end of the lactase gene on chromosome 2q21. FEBS Lett, 398:135-40.

Harvey CB, Hollox EJ, Poulter M, Wang Y, Rossi M, Auricchio S, Iqbal TH, Cooper BT, Barton R, Sarner M, Korpela R, Swallow DM (1998). Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. Ann. Hum. Genet., 62:215-223.

Holden C, Mace R (1997). Phylogenetic analysis of the evolution of lactase digestion in adults. Human Biology, 69:605-628.

Holl AFC (1998). The dawn of African pastoralisms: an introductory note. J Anthropol Archaeol, 17:81–96.

Hollox E. J., Poulter M., Zvarik M., Ferak V., Krause A., Jenkis T., Saha N., Kozlov A.I., Swallow D.M. (2001). Lactase haplotype diversity in the Old World. An. J. Hum. Genet., 68:160-172.

Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N, Swallow DM (2006). A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? Hum Genet.,120(6):779-88.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature, 437:1299-320.

Jobling MA, Hurles ME, Tyler-Smith C (2003). Adaptation to diet: Lactase persistence. Features of lactase persistence. Human evolutionary genetics origins, people & disease: pp 414-421.

Johnson JD, Kretchmer N, Simoons FJ (1974). Lactose malabsorption: its biology and history. Adv Pediatr, 21:197-237.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000). The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data. Am. J. Hum. Genet., 66:979–988.

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 (1971). Lactose and lactase-a historical perspective. Gastroenterology, 61:805-13.

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

Kruse TA, Bolund L, Grzeschik KH, Ropers HH, Sjostrom H, Noren O, Mantei N, Semenza G (1988). The human lactase-phlorizin hydrolase gene is located on chromosome 2. FEBS Lett, 240:123-6.

Lewinsky HR, Jensen TGK, Møller J, Stensballe A, Olsen J, Troelsen JT (2005). T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Human Molecular Genetics,14(24):3945-3953.

Lloyd M, Mevissen G, Fischer M, Olsen W, Goodspeed D, Genini M, Boll W, Semenza G, Mantei N (1992). Regulation of intestinal lactase in adult hypolactasia. J Clin Invest, 89:524-9.

Mantei N, Villa M, Enzler T, Wacker H, Boll W, James P, Hunziker W, Semenza G (1988). Complete primary structure of human and rabbit lactase-phlorizin hydrolase: implications for biosynthesis, membrane anchoring and evolution of the enzyme. Embo J, 7:2705-13.

Marenco G, Ghibaudi D, Meraviglia A (1970). Intolleranza al lattosio nell'adulto. Minerva Pediatr, 22:505-513.

Metz G, Jenkins DJ, Peters TJ, Newman A, Blendis LM (1975). Breath hydrogen as a diagnostic method for hypolactasia. Lancet, 1:1155-7.

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-13.9 kb T) does not predict or cause the lactase-persistence phenotype in Africans. Am. J. Hum. Genet., 74 (6):1102-10.

Myles S, Bouzekri N, Haverfield E, Cherkaoui M, Dugoujon JM, Ward R (2005). Genetic evidence in support to a shared Eurasian-North African dairying origin. Hum Genet, 117:34-42.

Nei M (1987). Molecular Evolutionary Genetics. New York, NY: Columbia University Press, p 177.

Nielsen R. (20019. Statistical tests of selective neutrality in the age of genomics. Heredity, 86:641–647

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, 12:2333-40.

Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martínez-Arias R, Clarimón J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999). Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet, 65(1):208–219.

Poulter M, Hollox EJ, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM (2003). The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Annals of Human Genetics, 67:298-311.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol, 16:1791–1798.

Remenyi A, Scholer HR. Wilmanns M (2004). Combinatorial control of gene expression. Nat. Struct. Mol. Biol., 11:812–815.

Rasinpera HSE, Enattah NS, Kuokkanen M, Totterman N, Lindahl H, Jarvela I, Kolho KL (2004) Genetic test, which can be used to diagnose adult-type hypolactasia in children. Gut, 53:1571-1576.

Raymond ML, Rousset F (1995). An exact test for population differentiation. Evolution, 49:1280-3. http://www.marksgeneticsoftware.net/rxc.htm

Rinaldi E, Albini L, Costagliola C, Derosa G, Auricchio G, Devizia B, Auricchio S (1984). High frequency of lactose absorbers among adults with idiopathic senile and presenile cataract in a population with a high prevalence of primary adult lactose malabsorption. Lancet, 1:355-357.

Ronald J, Akey JM (2005). Genome-wide scans for loci under selection in humans. Hum Genomics, 2(2):113-25.

Rossi M, Maiuri L, Fusco MI, Salvati VM, Fuccio A, Auricchio S, Mantei N, Zecca L, Gloor SM, Semenza G (1997). Lactase persistence versus decline in human adults: multifactorial events are involved in down-regulation after weaning. Gastroenterology, 112:1506-14.

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

Sahi T (1974). The inheritance of selective adult-type lactose malabsorption. Scand J Gastroenterol, Suppl 30:1-73.

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

Schneider S, Excoffier L (1999). Estimation of demographic parameters from the distribution of pair-wise differences when the mutation rates vary among sites: Application to human mitochondrial DNA. Genetics, 152:1079-1089.

Scrimshaw N, Murray E (1998). The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. Am J Clin Nutr, 48:1079-1159.

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 (1969). Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. I. Review of the medical research. Am J Dig Dis, 14:819-36.

Simoons FJ (1970). Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. II. A culture historical hypothesis. Am J Dig Dis, 15:695–710.

Simoons FJ (2001). Persistence of lactase activity among Northern Europeans: a weighing of evidence for the calcium absorption hypothesis. Ecology of Food and Nutrition, 40:397-469.

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

Slatkin M (2002). The age of alleles. In: Slatkin M, Veuille M (eds) Modern developments in theoretical population genetics: the legacy of Gustave Male´ cot. Oxford University Press, New York, pp 233–260.

StatSoft Italia srl (2001). STATISTICA, ver 6. www.statsoft.it.

Stephens M, Donnelly P (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet, 73:1162-9.

Swallow DM (2003). Genetics of lactase persistence and lactose intolerance. Annu. Rev. Genet., 37:197-219.

Swallow DM (2006). DNA test for hypolactasia premature. Gut., 55(1):131.

Takahashi K, Yamada H, Yanagida M (1994) Fission yeast minichromosome loss mutants mis cause lethal aneuploidy and replication abnormality. Mol Biol Cell, 5(10):1145-58.

Takahata N (1993). Allelic genealogy and human evolution. Mol Biol Evol, 10:2–22.

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

Troelsen JT, Olsen J, Moller J, Sjostrom H (2003) An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology, 125:1686-94.

Troelsen JT (2005). Adult-type hypolactasia and regulation of lactase expression. Biochim. Biophys. Acta, 1723:19–32.

Useli A. (2006). Analisi di polimorfismi autosomici, del cromosoma Y e del DMA mitocondriale in un campione di etnia Curda. PhD dissertation, University of Bologna. Unpublished data.

Wall JD, Pritchard JK (2003). Haplotype blocks and linkage disequilibrium in the human genome Nat Rev Genet., 4(8):587-97. Review.

Wang Y, Harvey CB, Pratt WS, Sams VR, Sarner M, Rossi M, Auricchio S, Swallow DM (1995). The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. Hum Mol Genet, 4:657-62.

Wang Y, Harvey CB, Hollox EJ, Phillips AD, Poulter M, Clay P, Walker-Smith JA, Swallow DM (1998).The genetically programmed down-regulation of lactase in children. Gastroenterology, 114:1230-6.

Weber JL, Wong C (1993). Mutation of human short tandem repeats. Hum Mol Genet, 2:1123–8.
Wright S(1951). The genetical structure of populations. Ann eugen, 1:323-334.

Xu H, Fu Y (2004). Estimating effective population size or mutation rate with microsatellites. Genetics, 166:555–563.

Zabel MD, Wheeler W, Weis JJ, Weis JH (2002). Yin Yang 1, Oct1, and NFAT-4 form repeating, cyclosporin-sensitive regulatory modules within the murine CD21 intronic control region. J. Immunol., 168:3341–3350.

Zuccato E, Andreoletti M, Bozzani A, Marcucci F, Velio P, Bianchi P, Mussini E (1983). Respiratory excretion of hydrogen and methane in Italian subjects after ingestion of lactose and milk. Eur. J. Clin. Invest., 13:261-266.

Zvelebil M (2000). Archaeogenetics: DNA and the population history of Europe. 57-70. Ed Boyle K., MacDonald Institute Cambridge, Cambridge.

www.hapmap.org: homepage of Hapmap project

http://genewindow.nci.nih.gov

SUPPLEMENTARY MATERIALS

Table S1: Distribution of phased haplotypes consisting of the C/T-13910 and G/A -22018 SNPs and microsatellites D2S3010, D2S3013 and D2S3015.

Hp	D2S3010	D2S3013	-13910	-22018	D2S3015	Ksa	Kta	Uig	Kaz	Ku	Per	SC	PAL	ALF	To	RM	ACS	tot
1	2	19	C	G	4											1		1
2	4	7	C	G	4		1											1
3	4	7	C	G	7											1		1
4	4	23	C	G	4		1											1
5	5	7	C	G	4	2	1		3									6
6	5	19	C	G	4					1								1
7	5	21	T	A	4					1								1
8	5	22	C	G	4			1										1
9	6	7	C	G	4		3	2	1									6
10	6	19	C	G	4												1	1
11	6	20	C	G	7					1								1
12	6	21	C	G	4	1		1										2
13	6	22	C	G	4				1				1		2	1		5
14	6	23	C	G	4		1											1
15	7	7	C	A	4				1									1
16	7	7	C	G	4	3	1	2	2	1				1				10
17	7	7	C	G	5							1						1
18	7	7	C	G	7										1			1
19	7	7	C	G	9						1							1
20	7	10	C	G	4				1			4						5
21	7	18	C	G	4					1	1	1						3
22	7	19	T	A	4						1							1
23	7	19	C	G	4						1		1	2			3	7
24	7	20	C	G	4						2			1	1		1	5
25	7	21	C	G	4	1		1					3	2		1	1	9
26	7	22	C	G	4					1	1				1		1	4
27	7	26	C	G	4	1					1							2
28	8	7	C	G	4		1		1			1						3
29	8	7	C	G	5				1									1
30	8	7	C	G	7		1			1				1		2		5
31	8	7	C	G	8	1												1
32	8	19	C	G	4		1			2			2	4	1	1	1	12
33	8	20	C	G	4	1				1				1				3
34	8	21	T	A	4	1						2						3
35	8	21	C	G	4			1					1	1	1	1		5
36	8	21	C	G	5			1										1
37	8	21	C	G	6			1										1
38	8	22	C	G	4	1		2	2				2	2	1			10
39	8	23	C	G	4	1	1											2
40	8	23	C	G	5				1									1
41	8	24	C	G	4												1	1
42	8	25	C	G	4				1							1	1	3
43	8	26	T	A	4							3						3
44	8	26	C	G	4										1			1
45	9	7	C	G	4		3						1					4

Hp	D2S3010	D2S3013	-13910	-22018	D2S3015	Ksa	Kta	Uig	Kaz	Ku	Per	SC	PAL	ALF	To	RM	ACS	tot
46	9	7	C	G	6		1				1				1			3
47	9	7	C	G	7		1						1					2
48	9	7	C	G	8					1	1				2	2	1	7
49	9	9	C	G	7		1											1
50	9	13	C	G	4								1					1
51	9	18	C	G	4										1	1		2
52	9	19	C	G	4								1	1		1	1	4
53	9	20	T	A	4								1					1
54	9	20	C	G	4			3	1	2		1	2				1	10
55	9	20	C	G	5				1		1			1				3
56	9	21	T	A	4	3	1		1	2		6	1					14
57	9	21	C	A	4				1		1			2			1	5
58	9	21	C	G	2								1					1
59	9	21	C	G	4	1			1					2	1	1	1	7
60	9	21	C	G	5			1			1			1				3
61	9	21	C	G	9	1						1						2
62	9	22	C	A	4				1				1					2
63	9	22	C	G	4	5	1			3	2			2	3		1	17
64	9	22	C	G	5		1											1
65	9	23	C	A	4									1			1	2
66	9	23	C	G	4	2		2	1			1	1		3	1	1	12
67	9	24	C	G	4					1							1	2
68	10	5	C	G	4										1			1
69	10	6	C	G	7											1		1
70	10	7	C	A	7							1						1
71	10	7	C	G	7	1				3			1		1			6
72	10	7	C	G	8			2		3			1	2	1	1	3	13
73	10	8	C	G	8							1						1
74	10	8	C	G	9			1										1
75	10	8	C	G	10					1								1
76	10	11	C	G	4					1	1	1	2	1				6
77	10	12	C	G	4	1	1											2
78	10	17	T	A	4								1					1
79	10	17	C	A	4	2												2
80	10	17	C	G	4			1								1		2
81	10	18	C	G	4	1												1
82	10	19	C	A	4					1								1
83	10	19	C	G	4								2			1		3
84	10	20	T	A	4							1					1	2
85	10	20	C	G	4			1	1		1					2	1	6
86	10	20	C	G	5			1			1							2
87	10	20	C	G	8				1									1
88	10	21	T	A	4	1	1		3	3	2	18	2	5	2	2	5	44
89	10	21	C	A	4			1	1									2
90	10	21	C	G	4	2		1		3	1			1	1	1	1	11
91	10	21	C	G	5			1							1			2
92	10	21	C	G	8				1									1
93	10	22	C	G	2					1								1
94	10	22	C	G	4				1							1		2
95	10	23	T	A	4		1							1		1		3
96	10	23	C	G	4		2			2						1		5
97	10	24	T	A	4								1		1			2
98	10	25	T	A	4					1					1			2
99	11	7	C	G	6				1		1				1			3
100	11	7	C	G	7		2	2		6	1	5	3	1	7	1	4	32

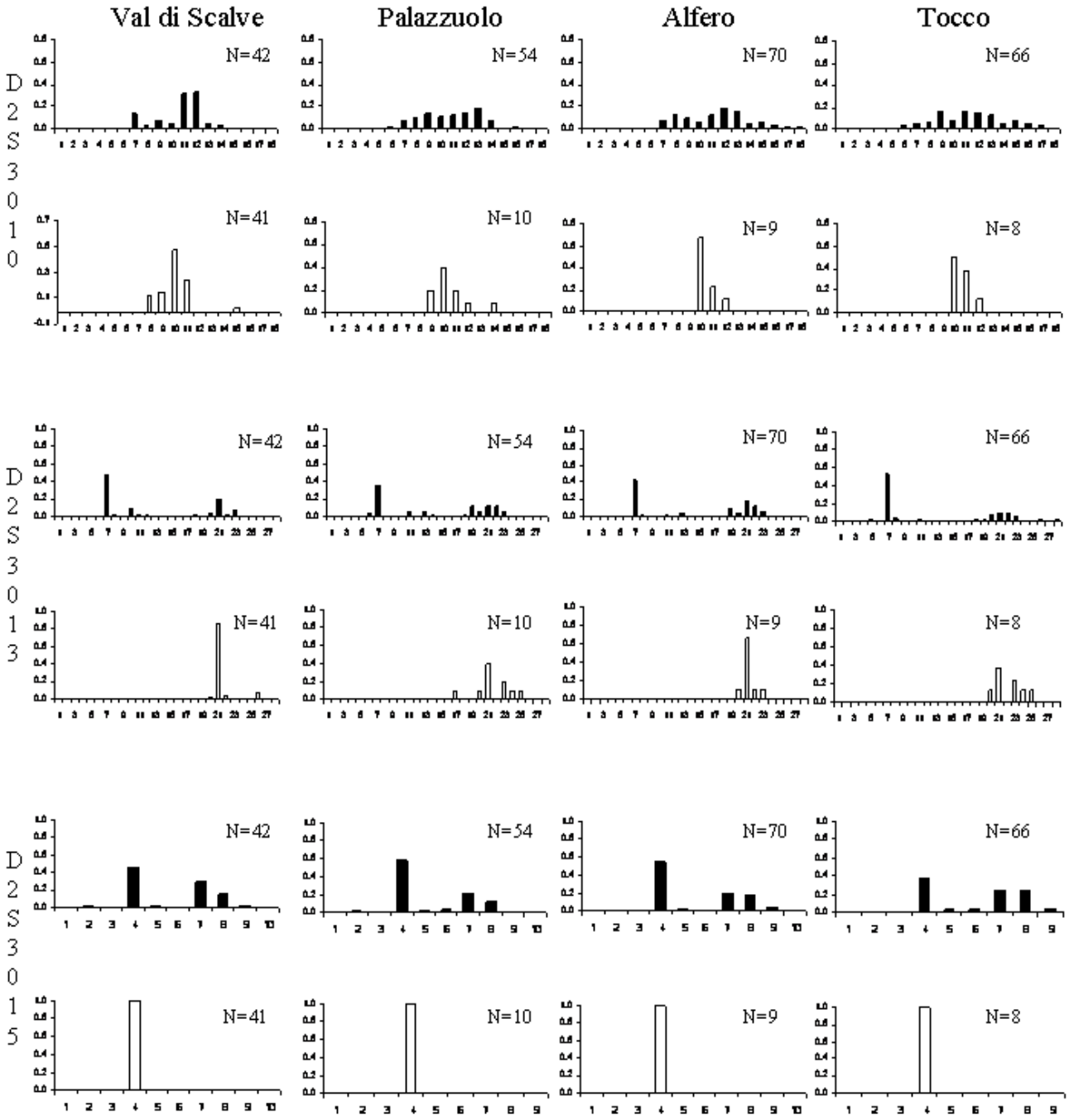
Hp	D2S3010	D2S3013	-13910	-22018	D2S3015	Ksa	Kta	Uig	Kaz	Ku	Per	SC	PAL	ALF	To	RM	ACS	tot
101	11	7	C	G	8										1	2		3
102	11	8	T	A	4			1										1
103	11	11	C	G	4					1								1
104	11	12	C	G	4			1				1					1	3
105	11	13	C	G	4				1				1					2
106	11	15	C	G	3				1									1
107	11	17	C	G	4			1										1
108	11	18	C	A	4				2									2
109	11	19	C	G	4			1		2							1	4
110	11	20	T	A	4				1					1	1		1	4
111	11	20	C	A	4						1							1
112	11	20	C	G	4	1					2						1	4
113	11	21	T	A	4			1	2	2		8		1	1	3		18
114	11	21	C	A	4								1					1
115	11	21	C	G	4				2	2		4		2	1	2	2	15
116	11	22	T	A	4		1			1		2				2		6
117	11	22	C	G	2							1						1
118	11	22	C	G	4	1	1	2		3	1		1	3		1	6	19
119	11	23	T	A	4		1		3				1		1	1	3	10
120	11	23	C	A	4				1									1
121	11	23	C	G	4	1	1			1	2	2	1	3			1	12
122	11	23	C	G	5								1					1
123	11	23	C	G	9										1			1
124	11	25	T	A	4	1				1			1					3
125	12	5	C	G	4											2	1	3
126	12	6	C	G	6								2					2
127	12	7	T	A	4												1	1
128	12	7	C	G	7	1	2		2	9	2	7	2	8	4	1	7	45
129	12	7	C	G	8	1	1	1		2		4	2	3	2	1	2	19
130	12	7	C	G	9						1				1			2
131	12	8	C	G	7										1			1
132	12	11	C	G	4										1			1
133	12	12	C	G	4												1	1
134	12	13	C	G	4	1												1
135	12	14	C	G	4								1					1
136	12	18	C	G	4				1				1					2
137	12	19	C	G	4		1											1
138	12	20	C	G	4	1	2			1		1						5
139	12	21	T	A	4			1		1								2
140	12	21	C	G	4	1	1	2		1		2		1				8
141	12	22	T	A	4				1		1			1				3
142	12	22	C	G	4	1	2	1	3					1		1		9
143	12	23	T	A	4					1			1		1			3
144	12	23	C	G	4		1									1		2
145	12	23	C	G	5				1	1								2
146	12	24	C	G	4				1									1
147	13	7	C	G	6					1							1	2
148	13	7	C	G	7		1	1		2	1	1	1	2	2	3	1	15
149	13	7	C	G	8		3	2		2	1	1	3	4	6	1	4	27
150	13	7	C	G	9					2	1			1		1	3	8
151	13	8	C	G	7									1				1
152	13	11	C	G	4								1					1
153	13	13	C	G	4	1		1					1	1				4
154	13	14	C	G	4		1			1								2
155	13	18	C	G	4		1											1

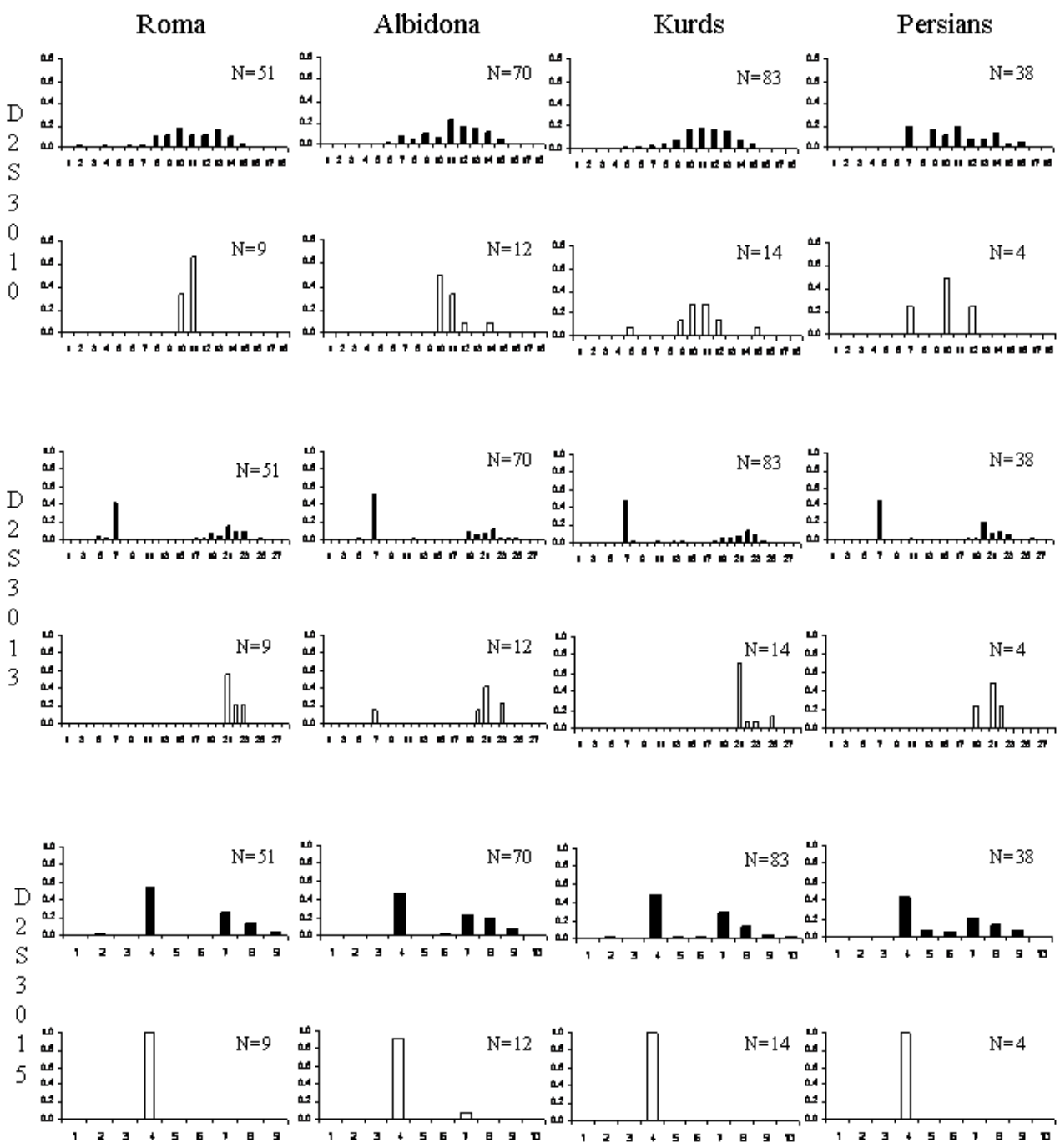
Hp	D2S3010	D2S3013	-13910	-22018	D2S3015	Ksa	Kta	Uig	Kaz	Ku	Per	SC	PAL	ALF	To	RM	ACS	tot
156	13	19	C	G	4		1		1									2
157	13	21	C	G	4	1			1				1	1		1		5
158	13	22	C	G	4	2	1	3		4			3				1	14
159	13	23	C	G	2											1		1
160	13	23	C	G	4	1	1			1				1		1		5
161	13	24	C	G	4		1											1
162	14	7	T	A	7												1	1
163	14	7	C	G	7	1	1	2	2	2	3		4	1	1	4	4	25
164	14	7	C	G	8		1			1	1				1		3	7
165	14	7	C	G	9					1				2			1	4
166	14	14	C	G	4		1	1										2
167	14	16	T	A	4		1		1									2
168	14	21	T	A	4								1					1
169	14	21	C	G	4	1												1
170	14	21	C	G	5			1										1
171	14	21	C	G	8						1	1						2
172	14	22	C	G	4	1		1								1		3
173	14	23	C	G	4	1				3								4
174	14	28	C	G	4										1			1
175	15	7	C	G	8	1			1	3	1			4	1		1	12
176	15	7	C	G	9										1	1	2	4
177	15	8	C	G	8										1			1
178	15	13	C	G	4		1			1								2
179	15	20	C	G	4	1												1
180	15	20	C	G	5										1			1
181	15	21	T	A	4					1		1						2
182	15	21	C	G	4										1	1		2
183	15	21	C	G	5		1											1
184	15	22	C	G	4		1	1										2
185	15	23	C	G	5				1									1
186	16	7	C	G	7			3			1							4
187	16	7	C	G	8	1												1
188	16	13	C	G	4									2				2
189	16	20	C	G	4	1					1		1		2			5
190	16	21	C	G	4		1								1			2
191	17	7	C	G	8			1							1			2
192	17	20	C	G	4										1			1
193	17	21	C	G	4									1				1
194	18	21	C	G	4									1				1
						56	60	58	60	98	44	84	66	82	74	60	84	826

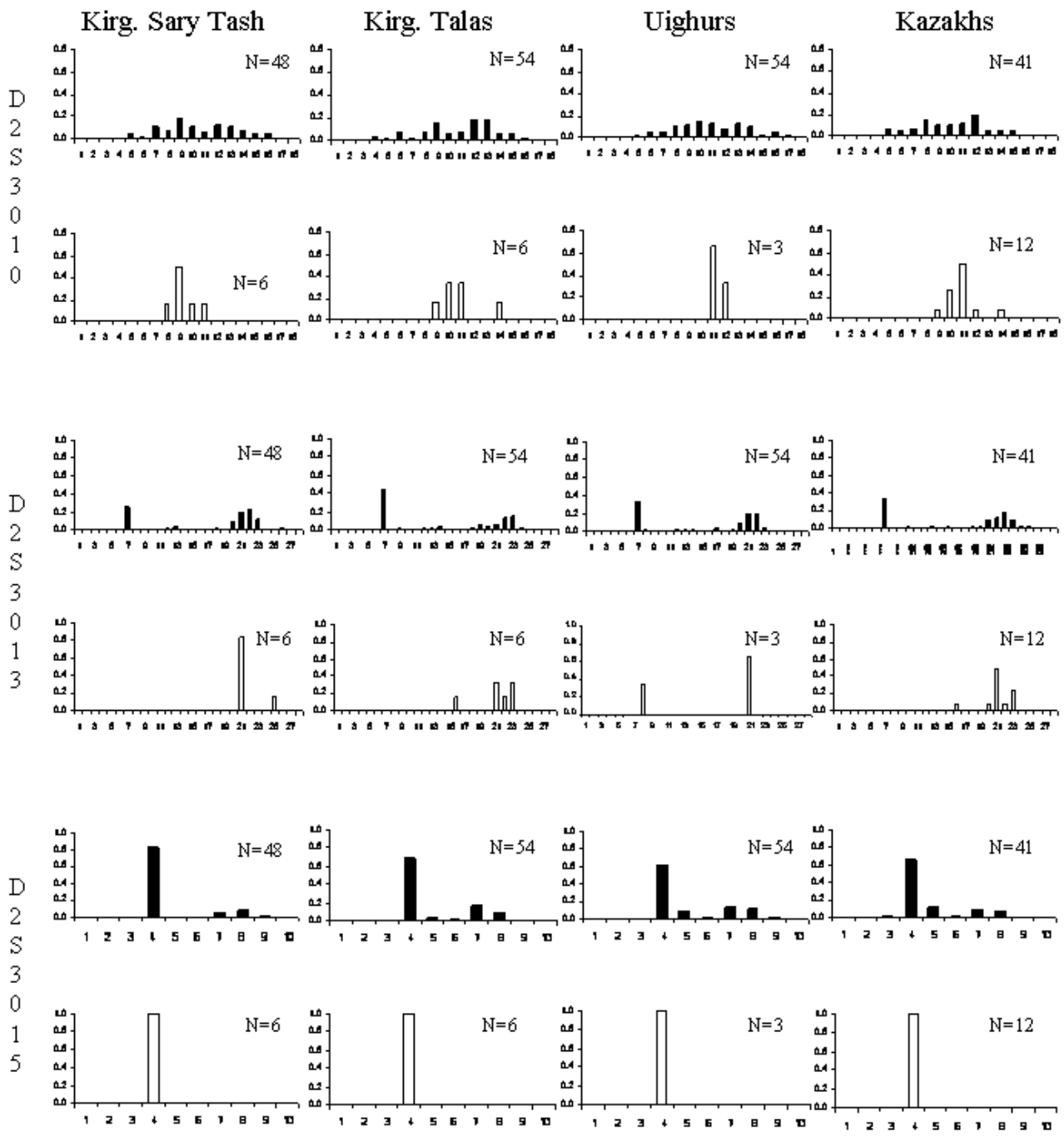
Table S2: Distribution of TA persistence haplotypes within the populations analyzed in this study and within Portugal, Sao Tomè and Fulbe samples from literature (Coelho et al. 2005).

Hp	D2S3010	D2S3013	-13910	-22018	D2S3015	Ksa	Kta	Uig	Kaz	Asia	Ku	Per	MidEast	SC	PAL	ALF	To	RM	ACS	Italy	Portugal	São Tome	Fulbe
Hp1	5	21	T	A	4						1		1										
Hp2	7	19	T	A	4							1	1										
Hp3	8	21	T	A	4	1				1				2						2	1		
Hp4	8	22	T	A	4																		1
Hp5	8	26	T	A	4									3						3			
Hp6	9	19	T	A	4																	1	
Hp7	9	20	T	A	4										1					1			
Hp8	9	21	T	A	4	3	1		1	5	2		2	6	1					7	6	1	
Hp9	9	22	T	A	4																1		
Hp10	9	23	T	A	4																1	1	
Hp11	10	17	T	A	4										1					1			
Hp12	10	20	T	A	4									1					1	2	2		
Hp13	10	21	T	A	4	1	1		3	5	3	2	5	18	2	5	2	2	5	34	14	6	6
Hp14	10	22	T	A	4																	1	8
Hp15	10	23	T	A	4		1			1						1		1		2	5		1
Hp16	10	24	T	A	4										1		1			2	1		
Hp17	10	25	T	A	4						1		1				1			1		1	
Hp18	11	8	T	A	4			1		1													
Hp19	11	20	T	A	4				1	1						1	1		1	3	3		
Hp20	11	21	T	A	4			1	2	3	2		2	8		1	1	3		13	6	1	3
Hp21	11	22	T	A	4		1			1	1		1	2				2		4	4		1
Hp22	11	23	T	A	4		1		3	4					1		1	1	3	6	9		1
Hp23	11	24	T	A	4																1	1	
Hp24	11	25	T	A	4	1				1	1		1		1					1			
Hp25	12	7	T	A	4														1	1			
Hp26	12	19	T	A	4																3		
Hp27	12	21	T	A	4			1		1	1		1								6		
Hp28	12	22	T	A	4				1	1		1	1			1				1			
Hp29	12	23	T	A	4						1		1		1		1			2			
Hp30	13	21	T	A	4																2		
Hp31	14	7	T	A	7														1	1			
Hp32	14	16	T	A	4		1		1	2													
Hp33	14	21	T	A	4										1					1	1		
Hp34	15	21	T	A	4						1		1	1						1			
						6	6	3	12	27	14	4	18	41	10	9	8	9	12	89	66	13	21

Figure S1: Microsatellite allele frequency distribution within the CG non-persistence (in black) and TA persistence lineages (in white), within each population. The estimated sizes of allele 1 in each microsatellite are: 184bp for D2S3010; 125bp for D2S3013 and 175bp for D2S3015. N= number of chromosomes.




Roma
Albidona
Kurds
Persians
D2S3010
D2S3013
D2S3015
N=51
N=70
N=83
N=38
N=9
N=12
N=14
N=4


Kirg. Sary Tash
Kirg. Talas
Uighurs
Kazakhs
D2S3010
D2S3013
D2S3015
N=48
N=54
N=54
N=41
N=6
N=6
N=3
N=12

ACKNOWLEDGMENTS

Thanks to prof. Donata Luiselli and prof. Davide Pettener for the opportunity to work in this field. Thanks to all the other people of the research group: Loredana Castri, Antonella Useli, Alessio Boattini, Marco Sazzini and Graziella Ciani. A particular thanks to the students that helped me in the lab work: Isabella Paganini-Paganelli, Mariangela Laino and Angelo Puglisi.

Thanks to prof. Jorge Rocha and Margarida Coelho from IPATIMUP (Portugal) for the constant help and the important suggestions during the development of the thesis.

Thanks to Prof Giovanni Destro-Bisol, Cinzia Battaggia and all the other people from Università la Sapienza, Rome, for the DNA samples and to have collaborated in typing the samples.

Thanks to all people that kindly offered their DNA for our researches.