Herders of Indian and European Cattle Share Their Predominant Allele for Lactase Persistence

Irene Gallego Romero,*†[1] Chandana Basu Mallick,[2] Anke Liebert,[3] Federica Crivellaro,[1] Gyaneshwer Chaubey,[2] Yuval Itan,‡[3,4] Mait Metspalu,[2] Muthukrishnan Eaaswarkhanth,§[5] Ramasamy Pitchappan,[6] Richard Villems,[2] David Reich,[7,8] Lalji Singh,[5,9] Kumarasamy Thangaraj,*[5] Mark G. Thomas,[3,10] Dallas M. Swallow,[3] Marta Mirazón Lahr,[1] and Toomas Kivisild[1]

[1]Department of Biological Anthropology, Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, United Kingdom

[2]Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu and Estonian Biocentre, Tartu, Estonia

[3]Research Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[4]Centre for Mathematics & Physics in the Life Sciences and Experimental Biology, University College London, London, United Kingdom

[5]Centre for Cellular and Molecular Biology, Hyderabad, India

[6]Chettinad Academy of Research & Education, Chettinad Health City, Chennai, India

[7]Department of Genetics, Harvard Medical School

[8]Broad Institute of Harvard and MIT

[9]Genome Foundation, Hyderabad, India

[10]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

†Present address: Department of Human Genetics, University of Chicago.

‡Present address: St Giles Laboratory of Human Genetics of Infectious Diseases, the Rockefeller University.

§Present address: New York State Department of Health Wadsworth Center, Albany, New York.

***Corresponding author:** E-mail: ireneg@uchicago.edu; thangs@ccmb.res.in.

Associate editor: Sarah Tishkoff

Abstract

Milk consumption and lactose digestion after weaning are exclusively human traits made possible by the continued production of the enzyme lactase in adulthood. Multiple independent mutations in a 100-bp region—part of an enhancer—approximately 14-kb upstream of the *LCT* gene are associated with this trait in Europeans and pastoralists from Saudi Arabia and Africa. However, a single mutation of purported western Eurasian origin accounts for much of observed lactase persistence outside Africa. Given the high levels of present-day milk consumption in India, together with archaeological and genetic evidence for the independent domestication of cattle in the Indus valley roughly 7,000 years ago, we sought to determine whether lactase persistence has evolved independently in the subcontinent. Here, we present the results of the first comprehensive survey of the *LCT* enhancer region in south Asia. Having genotyped 2,284 DNA samples from across the Indian subcontinent, we find that the previously described west Eurasian -13910 C>T mutation accounts for nearly all the genetic variation we observed in the 400- to 700-bp *LCT* regulatory region that we sequenced. Geography is a significant predictor of -13910*T allele frequency, and consistent with other genomic loci, its distribution in India follows a general northwest to southeast declining pattern, although frequencies among certain neighboring populations vary substantially. We confirm that the mutation is identical by descent to the European allele and is associated with the same >1 Mb extended haplotype in both populations.

Key words: positive selection, India, haplotype, lactase persistence, pastoralism.

Introduction

Lactase persistence (MIM #223100) is an exclusively human trait (Johnson et al. 1974), inherited in a dominant Mendelian fashion (Ferguson and Maxwell 1967; Sahi et al. 1973), and defined as the ability to maintain post-weaning production of the enzyme lactase–phlorizin hydrolase (LPH), which hydrolyzes the milk sugar lactose into its component monosaccharides. In all other mammals tested, LPH production diminishes significantly after weaning and is never resumed (Plimmer 1906; Sebastio et al. 1989; Buller et al. 1990; Lacey et al. 1994). Initially considered to have a worldwide distribution, data collected from the mid 1960s and onward revealed that lactase persistence is geographically restricted and not the ancestral condition in humans. This trait is most common in northwestern Europe (Cuatrecasas et al. 1965; Bayless and

Rosensweig 1966; Simoons 1978; Davey Smith et al. 2009; Ingram, Mulcare, et al. 2009) and decreases in frequency east and south of there, reaching near-zero frequencies east of the Indian subcontinent (Ingram, Mulcare, et al. 2009; Itan et al. 2010). It is also present at moderately high frequencies in certain ethnic groups from sub-Saharan Africa, the Near East and the Mongolian plateau that have a history of milk consumption and pastoralism (Simoons 1970a; McCracken 1971; Durham 1991).

West Eurasian lactase persistence has been attributed to a single nucleotide change located approximately 14-kb upstream of the *LCT* gene (rs4988235, henceforth -13910 C>T; Enattah et al. 2002; Troelsen et al. 2003). In populations examined so far, the -13910*T allele is associated with a single extended haplotype that spans about 1 Mb (Poulter et al. 2003; Bersaglieri et al. 2004). The only exceptions to this are some rare chromosomes present in a group of populations living in the Eastern European Plain, where an argument has been made for the convergent origin of the -13910*T substitution (Enattah et al. 2007), although in this case, the data are also compatible with recombination or errors in phase inference. In spite of this strong genotype–phenotype association in Europe, -13910 C>T does not explain the observed patterns of lactose tolerance in Africa (Mulcare et al. 2004) and some other regions of the Old World (Itan et al. 2010). Recent studies have instead identified other variants of probable functional significance in the same small (approximately 100 bp) genomic region with known enhancer function (Olds and Sibley 2003; Lewinsky et al. 2005) among pastoralist, milk-drinking groups in sub-Saharan Africa (Ingram et al. 2007; Tishkoff et al. 2007; Ingram, Raga, et al. 2009) and Saudi Arabia (Enattah et al. 2008). Compared with the wide geographic distribution of -13910 C>T, the African and Arabian variants exhibit somewhat narrower and overlapping geographic distributions (Itan et al. 2010).

These multiple instances of phenotypic convergence imply the existence of a strong selective advantage to being lactase persistent in cultures where milk is commonly available. Indeed, genomic scans for signatures of natural selection have identified the 1-Mb region on chromosome 2 surrounding -13910*T as harboring signals of recent strong positive selection centered around -13910*T itself (Bersaglieri et al. 2004; Voight et al. 2006; Sabeti et al. 2007); the same is true of the >2 Mb surrounding -14010*C, one of the African variants (Tishkoff et al. 2007). In addition, variation in linked microsatellites (Coelho et al. 2005), an absence of the -13910*T allele in a sample of early Neolithic central European skeletons (Burger et al. 2007) and demic (Gerbault et al. 2009), as well as spatially explicit computer simulations (Itan et al. 2009) have all added support to the case for a strong selective advantage. Other studies have documented the occurrence of several different mutations in lactose-digesting members of a single ethnic group, suggestive of a recent soft selective sweep (Ingram, Raga, et al. 2009).

The likely selective advantage has been argued to be conditional on postweaning access to abundant sources of milk (Simoons 1970a; McCracken 1971; Durham 1991), something that would not have been widely available to humans prior to the domestication of cattle, sheep, or goats and the development of dairying practices. It has therefore been argued that the generalized adoption in Europe of a dairying culture during the Middle Neolithic, after the domestication of taurine cattle in the Near East, led to positive selection for lactase persistence and thus to the trait's present-day high frequency in populations from the region (Clutton-Brock 1999; Troy et al. 2001; Evershed et al. 2008; Itan et al. 2009). Its presence in sub-Saharan Africa may at least in part be explained by the adoption of pastoralism and dairying by north African groups that subsequently expanded south as indicated by the archaeological record and the pattern of expansion of their cattle (Hanotte et al. 2002), but the occurrence of several novel alleles in East Africa and the Arabian peninsula indicates convergent evolution. One with wider distribution, -13915*G, is most frequent in Saudi Arabian camel herders, who also exhibit a high frequency of lactase persistence (Enattah et al. 2008), suggesting that it may have originated in the Arabian peninsula, although the other known mutations (13907*G and 14010*C) are likely to have originated in Africa (Ingram et al. 2007; Tishkoff et al. 2007). The Indian subcontinent, at the eastern extreme of the phenotypic distribution (Simoons 1970b), remains a glaring lacuna in current knowledge. India is the world's largest milk producer today (Food and Agriculture Organization of the United Nations 2009), and zebu cattle, *Bos indicus*, a species distinct from west Eurasian *Bos taurus*, was probably domesticated in the Indus Valley around 7,000 years before present (YBP) (Meadow 1993; Loftus et al. 1994; Baig et al. 2005; Chen et al. 2010). A second milking animal, the water buffalo (*Bubalus bubalis*), accounts for over 50% of the milk produced in India (Kumar et al. 2007) and has similar levels of milk lactose to cattle (Jensen 1995), although specific details of its domestication time and archaeological context in India are lacking. Yet, despite being host to over 1 billion people and having a well-documented history of consumption of dairy products (Simoons 1970b), little is known about the distribution of lactase persistence beyond a flurry of studies conducted in the 1970s (Desai et al. 1970; Swaminathan et al. 1970; Gupta et al. 1971; Reddy and Pershad 1972; Tandon et al. 1981). These data suggest that phenotypic lactase persistence could be as high as 0.73 in some north Indian populations and rare or absent from some groups in the south of the country, predicting an overall north–south phenotypic frequency gradient, as well as a moderately high countrywide frequency of the phenotype. With the exception of a single recent study in which variation at the -13910 position was typed in two mixed urban populations (Babu et al. 2010), there exists no genotypic data related to the lactase persistence phenotype in India.

We have undertaken the first countrywide screening of DNA samples from all major language groups and geographic regions of India ($n = 2284$) and sequenced the region around -13910 to identify polymorphisms that are known to be or might be associated with lactase persistence.

In order to distinguish between shared and independent origins of mutations and to determine the extent of potentially positive selection, we genotyped a subset of 199 samples for over 1,000 single nucleotide polymorphisms (SNPs) surrounding the *LCT* locus on chromosome 2. These data are used to address four aspects of the genotypic basis for milk drinking among south Asians—first, what is the distribution of the west Eurasian -13910*T allele as well as that of any other mutations in the 700-bp region surrounding -13910 C>T that may be segregating in the country; second, how does this distribution fit previously reported frequencies of the lactase persistence phenotype; third, is the trait associated with linguistic, geographical, or ancestral subsistence factors; and lastly, is there genetic evidence for positive selection acting on lactase persistence alleles in the Indian subcontinent.

Materials and Methods

Sample Composition

Our sample consisted of 2,264 individuals from 105 distinct tribal and caste populations representing all five major language families and 22 of 28 Indian states, and a single Union territory, as well as a sample of 20 Tharu individuals from Nepal, for a total sample size of 2,284 unrelated individuals. States were grouped by previously described geographic regions (Sahoo et al. 2006); the Nepali samples were grouped with north India. A more detailed description of the samples is provided in supplementary table 1(Supplementary Material online). All samples were collected with the informed written consent of the donors, and the study was approved by the Institutional Ethical Committee of the Centre for Cellular and Molecular Biology in Hyderabad.

Data Sequencing and Genotyping

Polymerase chain reaction (PCR) amplification was carried out using previously described conditions and primer pairs MCM6i13-LAC14CL2 and MCM6i13-MCM778, which respectively yield 427- and 706-bp amplicons (Ingram et al. 2007). We sequenced the amplified region surrounding the -13910 C>T locus in all 2,284 individuals using an ABI Prism 3730XL DNA Analyzer and BigDye terminator chemistry (Applied Biosystems, Foster City, CA). Sequence data were analyzed using Sequencher 4.8 (Genecodes, Ann Harbor, MI); all sequences were inspected visually and scored at all identified variable sites. All newly determined polymorphic positions and singletons were confirmed by reamplification and sequencing from the opposite strand. Similarly, the allelic status of individuals with the -13910*T allele was confirmed by an independent PCR. Population-, geographic-, linguistic-, and continental-level exact tests of Hardy–Weinberg equilibrium were carried out using a previously described method (Wigginton et al. 2005). Wilcoxon rank sum and Kruskal–Wallis tests were used to search for significant effects in population allele frequencies of subsistence, linguistic, and geographic factors. Chi-square tests were used to confirm associations between the presence/absence of lactase persistence alleles and geographic and linguistic factors and Mantel tests to test for interactions between geographic origin and linguistic affiliation known to covary across India. The linguistic distance matrix ranked populations that spoke languages from the same language family as having a distance of 0, whereas populations who spoke languages from different families had a distance of 1. We deemed the 0–1 approach most suited to ensuring the independence of the two matrices as the inclusion of more detailed linguistic information (i.e., subbranches) would force the introduction of additional arbitrary parameters when calculating distances between subbranches of different language families.

In all cases, we estimated a predicted phenotypic frequency of lactase persistence due to the given mutation assuming dominance using the Hardy–Weinberg formula $2pq + p^2$, where p is the frequency of the lactase persistence–associated allele and $q = 1 - p$, observing no deviations from expectations in any of the individual population groups. To reduce noise due to small sample size effects, however, we excluded from all population-level analyses 25 groups that had sample sizes <10. Data from a total of 81 populations were therefore included in downstream analyses. Then, we used the GenoPheno Monte Carlo method (Mulcare et al. 2004; Itan et al. 2010) to test whether our genotypic data could explain observed (from published literature) phenotype frequencies throughout the subcontinent, after factoring in sampling errors at both the genotypic and the phenotypic level, as well as type 1 and type 2 error rates in phenotype testing methods (usually breath hydrogen or blood glucose) as described by Mulcare et al. (2004) and Itan et al. (2010). All statistical analyses were performed using R 2.10 (R Core Development Team 2009). Published phenotype frequencies based on lactose tolerance testing of Indian populations without genotype information were retrieved from the online GLAD database (http://www.ucl.ac.uk/mace-lab/resources/glad, 2011 August 15). Surface interpolation and maps were generated by the PyNGL Python module (http://www.pyngl.ucar.edu/, 2011 August 15).

Haplotype Data Analysis

One hundred and ninety-nine Indian individuals in our sample (supplementary table 1, Supplementary Material online) were also genotyped using either the Illumina Infinium 650K ($n = 152$) or the 610K SNP ($n = 47$) platforms (Illumina, San Diego, CA). We retrieved unphased SNP data for the 5 Mb surrounding -13910 C>T (hg18 chr2:134,000,000–139,000,000) for these individuals as well as for the 506 unrelated individuals of European ($n = 156$), Near Eastern (n = 160), or Pakistani ($n = 190$) ancestry included in the Human Genome Diversity Panel—Centre d'Étude du Polymorphisme Humain (HGDP-CEPH), who have been previously genotyped on the 650K platform (Li et al. 2008). Data were phased using Beagle 3.1 (Browning and Browning 2007). Although -13910 C>T is not included in either the 650K or the 610K platforms, the HGDP-CEPH populations have been previously genotyped at this position (Bersaglieri et al. 2004), so we were able to include

it in our data set. The 5-Mb region spanned 1,152 SNP loci present in both SNP chips genotyped in 705 individuals.

By the use of a solid spine method (Barrett et al. 2005), we identified in the phased data a 60-kb linkage block spanning the lactase enhancer, containing 15 SNPs, which we used to define short-range haplotypes in the region surrounding the enhancer (supplementary table 2, Supplementary Material online). We then used the program NETWORK 4.5.1.6 to create an unrooted median joining network (Bandelt et al. 1999) from these SNPs (fig. 2). There is no overlap between markers included in the Illumina 650K and those originally employed to define the core *LCT* gene haplotypes (Hollox et al. 2001). To associate the 60-kb linkage block haplotypes with the established core haplotype nomenclature, we used the trio samples from HapMap CEU individuals, who have been genotyped for both sets of markers as had other members of their two- or three-generation families, which had allowed us to determine accurately the phased haplotypes from family structure.

We employed the program Sweep (Sabeti et al. 2007) to calculate EHH and relative EHH (REHH) scores for the 15-SNP–long core haplotypes in our four continental population groups after exclusion of 106 SNPs not annotated in the UCSC human genome build 16, which is the genomic build employed by Sweep. We further chose to calculate separate scores for Indo-European speakers and non-Indo-European speakers to test for differences between the two groups in light of a possible association between geography, linguistic affiliation, and the frequency of the lactase persistence allele uncovered by our analyses.

Results

-13910 C>T in India

Sequencing of the 427- and 706-bp PCR amplicons, approximately 14-kb upstream of the *LCT* gene in a total of 2,284 individuals, revealed that the derived -13910*T allele has the highest frequency among the observed mutations as well as the widest distribution throughout the Indian subcontinent (table 1, supplementary table 1, Supplementary Material online). Its frequency ranges from 0.8% among the Tibeto-Burman speakers to 18.4% among Indo-European speakers, with west India showing the highest frequency of the derived allele; it has an average countrywide frequency of 10.3%. Seven other segregating polymorphic sites were also observed, one of which (-13779 G>C) has an overall frequency of 2.4% in India (supplementary table 3, Supplementary Material online), whereas the remaining six mutations combined have an overall frequency of roughly 1%.

A broad geographic distribution map of -13910*T allele frequency in south Asia is presented in figure 1. Despite the general north to south and west to east frequency gradient, the occurrence of the -13910*T allele showed large differences between some neighboring groups (supplementary table 1, Supplementary Material online). Population-level mean allele frequency is 0.09 (standard deviation [SD] ± 0.118); it is absent in almost a third of the population samples that met our criteria for minimum sample size of 10 and peaks in the Ror of Haryana (0.489, $n = 46$), a traditional milk-drinking people who herd water buffalo and a predicted frequency of the lactase persistence phenotype of 0.739. Of further curious note is the single Great Andamanese individual with a C/T genotype, suggestive of recent mainland Indian admixture (Thangaraj et al. 2003; Reich et al. 2009).

No single population under study showed significant departures from Hardy–Weinberg expectations (table 1). However, when pooled either by linguistic or by geographic criteria, disequilibrium is apparent in south and west India and in Dravidian and Indo-European speakers, indicating a Wahlund effect (Wahlund 1928) due to nonnegligible population structure. When considering all India as a single population, deviations from equilibrium due to the excess of homozygotes are pronounced (exact $P = 2.83 \times 10^{-13}$). This deviation is more pronounced than in 156 unrelated European HGDP-CEPH samples (exact $P = 0.0007$). This finding is consistent with genome-wide patterns of generally higher population differentiation in India than in Europe (Reich et al. 2009) as it stems primarily from an excess of homozygotes of both genotypes and corresponding shortage of heterozygous individuals.

We grouped populations by their linguistic or geographic affiliation (table 1, supplementary table 1, Supplementary Material online) and carried out Kruskal–Wallis tests to assess the effect of these factors on -13910*T allele frequencies. Both geography ($P = 1.853 \times 10^{-6}$, $n = 81$) and language ($P = 1.74 \times 10^{-6}$, $n = 81$) exhibit significantly nonrandom association with allelic frequencies, whereby groups from west India, or those that speak Indo-European languages are more likely to carry the -13910*T allele than any other group. It is difficult to quantify the degree of autocorrelation between our two main explanatory variables, but they are well known to covary (supplementary table 1, Supplementary Material online); a further Kruskal–Wallis test applied only on data from Indo-European speakers showed that geography remains significant ($P < 0.01$) even when controlling for language. Because there is no geographic region covered by our data with a sufficiently large number of populations of diverse linguistic affiliations, we cannot test the reverse hypothesis. Contingency table χ^2 tests further support the lack of independence between the presence of lactose tolerance in populations (i.e., freq T > 0) and linguistic affiliation ($P = 6.15 \times 10^{-5}$, degrees of freedom [df] = 5) or geographic origin ($P = 3.24 \times 10^{-4}$, $df = 6$); they also strongly support the lack of independence between our two main predictors ($P = 3.85 \times 10^{-29}$, df = 30). Finally, Mantel tests suggest that geographic origin is having a significant effect on population frequencies of -13910*T ($P = 0.005$), although the effect of linguistic affiliation is negligible ($P = 0.31$); a partial Mantel test for the effects of linguistic affiliation holding geography constant further supports this conclusion ($P = 0.707$).

Other Enhancer Region Alleles Identified

Apart from -13910 C>T, we found seven other segregating sites in our samples with a combined frequency of the derived alleles 0.035. Four have been previously described; the

Table 1. Sample-Wide Allele Frequencies of Previously Published Lactase Persistence Associated Variants by Linguistic and Geographic Affiliation.

		Observed Allele Frequency					
	n	−14010*C[a]	−13915*G rs41380347	−13910*T rs4988235	−13907*G rs41525747	Other Observed Variants[b]	HWE *P* Value −13910*T
Linguistic affiliation[c]							
Andamanese	24	0	0	0.021	0	0	1
Austroasiatic	331	0	0	0.011	0	0	1
Dravidian	853	0	0	0.076	0	0.046	1.47×10^{-4}
Indo-Europeans	896	0	0	0.184	0	0.016	3.24×10^{-4}
Isolate	60	0	0	0.025	0	0.042	1
Tibeto-Burman	120	0	0	0.008	0	0	1
Total	2284	0	0	0.103	0	0.025	2.83×10^{-13}
Geographic region[d]							
West	468	0	0	0.213	0	0.019	2.71×10^{-4}
North	290	0	0	0.159	0	0.022	0.268
East	310	0	0	0.018	0	0.003	1
North East	139	0	0	0.007	0	0	1
Central	179	0	0	0.056	0	0.003	0.430
South	864	0	0	0.085	0	0.047	5.99×10^{-4}
Andaman and Nicobar islands	34	0	0	0.015	0	0	1
Total	2284	0	0	0.103	0	0.025	2.83×10^{-13}

NOTE.—[a]The rs id associated with this position, rs4988233, has alleles C and T, the derived T allele was not detected in our sample.
[b] More details on the sites grouped under "other observed variants" are given in the text and in supplementary tables 3 and 4 (Supplementary Material online).
[c] Linguistic affiliations of the 106 sampled populations are given in supplementary table 1 (Supplementary Material online).
[d] Geographic regions are defined as in Sahoo et al. (2006).

remaining three are novel (table 1; supplementary tables 3 and 4, Supplementary Material online). Of these seven sites, the mutation -13779 G>C, previously reported in a single lactose intolerant Somali individual (Ingram, Raga, et al. 2009), is the most common, with a countrywide sample frequency of 0.024. Its distribution is primarily restricted to eastern Kerala, where it reaches its highest frequency in the hunter-gatherer group Kattunaikkan (0.146, $n = 41$) and in the nearby Nilgiri Hills of Tamil Nadu. Its frequency in the Toda, who are pastoralists and herders of water buffaloes, is 0.091 ($n = 55$), much lower than the observed value for -13910*T for the same group (0.445) but still worthy of note. Also of interest is the -13915 T>C allele, which we have previously found in Muslim populations from southern India (Eaaswarkhanth et al. 2010) and which we report here at a low frequency in other populations from the same region, again including the Toda. This is distinct from the previously described -13915 T>G mutation, which was not found in any of the samples.

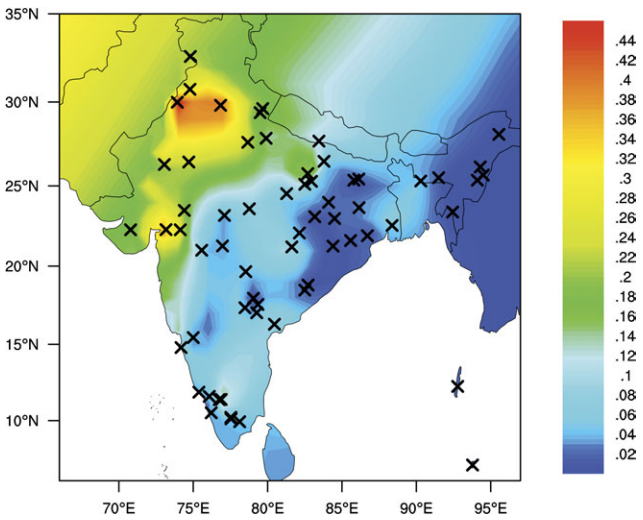

FIG. 1. Distribution of -13910 C>T in India. Crosses represent sample locations. Colors and color key show the frequency of the -13910*T allele estimated by surface interpolation.

Reconciling Genotype and Phenotype

On the basis of previously collected lactase persistence phenotype data (Desai et al. 1970; Swaminathan et al. 1970; Gupta et al. 1971; Reddy and Pershad 1972; Tandon et al. 1981), the average observed frequency of lactase persistence in India appears to be roughly 0.40 (supplementary fig. 1, Supplementary Material online). However, based on our -13910*T allele frequency data, the predicted countrywide phenotype frequency in our sample is only 0.196. The same is true when we calculate the mean predicted phenotypic frequency at the population group level, 0.158 (SD ± 0.191)—the median population value is likewise low, 0.089. In our data, point estimates of phenotype frequency explained by -13910*T allele frequency only reach or exceed 0.40 in 11 of the 81 sampled groups and are less than 0.1 in 45 of the 81 groups. Given that we observe an absence of -13910*T alleles in many of these samples, phenotype frequencies could in many cases be close to zero; 95% confidence intervals for all populations are given in supplementary table 1 (Supplementary Material online). Even when we assume that all observed mutations in the genetic region under study are causative and dominant, our predicted countrywide phenotype frequency only increases to 0.24, again substantially lower than 0.40 (supplementary table 5, Supplementary Material online). These results suggest either that there are other major genetic variants

Table 2. The Effect of Pastoralism on HWE.

	Pastoralists Included[a]	Pastoralists Excluded
Dravidians	1.47×10^{-04}	0.134
Indo-Europeans	3.24×10^{-04}	0.084
South India	5.99×10^{-04}	0.154
West India	2.71×10^{-04}	0.153
All India	2.83×10^{-13}	3.68×10^{-5}

NOTE.—The table shows *P* values for HWE for all populations grouped by geographic region (a) with pastoralists included and (b) with pastoralists excluded.

[a] Only those groups that originally included pastoralists are shown in this comparison. All values are from exact HWE tests.

outside our sequencing range that are causative of the lactase persistence phenotype or that previous studies have overestimated the actual phenotype frequency in India. However, our GenoPheno analyses (Mulcare et al. 2004; Itan et al. 2010) using interpolated allele and phenotype frequency distributions (supplementary fig. 2, Supplementary Material online) identify only small regions of India where the difference between genotypic and phenotypic frequency remains significant after taking into account sampling and phenotype measurement errors.

Lactase Persistence Genotypes and Pastoralism in India

Several other mutations in the region surrounding -13910 C>T have been found in populations with a tradition of dairying, suggesting repeated convergent evolution of the lactase persistence phenotype. We therefore included five groups in our sample—Ror, Toda, Jat Muslim, Gawli, and Yadava—to represent traditional pastoralist populations of India as we expected these to be the most likely populations to harbor previously unknown mutations. However, even in these populations, -13910 C>T is the main segregating polymorphism, with other variants making relatively small contributions to the diversity in this genomic region. We cannot, however, exclude that these groups have additional alleles in regulatory regions outside the region sequenced. Although at a countrywide level pastoralism is significantly associated with elevated frequencies of -13910*T (Wilcoxon rank sum $P = 0.0065$), the number of pastoralist groups included in our calculation is too small for the test to have strong explanatory power. We note that three of these populations (Ror, Jat Muslim, and Toda) have the first, third, and fourth highest frequencies of the -13910*T allele in the sample; these values are greater than the overall sample mean by at least 2 SD (supplementary table 1, Supplementary Material online).

We therefore examined whether these three clear outliers with much higher -13910*T frequencies could be behind the observed deviations from Hardy–Weinberg expectations in the combined sample (table 1). Indeed, following the removal of all pastoralist individuals from the pooled data sets, HWE *P* values become nonsignificant in west and south India and Indo-European and Dravidian speakers (table 2), although once again when all India is treated as a single group, disequilibrium remains highly significant (exact $P = 3.60 \times 10^{-5}$). More interestingly, excluding all pastoralist groups from our Kruskal–Wallis tests for nonrandom associations of geographic and linguistic factors on allele frequency levels reveals that geography ($P = 3.65 \times 10^{-6}$) is slightly more strongly associated with population frequencies of -13910 C>T than language ($P = 4.36 \times 10^{-6}$).

Origins of Indian Lactase Persistence and Evidence for Positive Selection

In light of other instances of convergent evolution of the lactase persistence phenotype in human populations, a key question is whether the Indian -13910*T allele is identical by descent to the European one or the product of convergent evolution. To this end, we examined a 60 kb 15-SNP–long core region (supplementary table 2, Supplementary Material online) and identified 21 distinct haplotypes in 705 individuals of western Eurasian ancestry for whom we have genome-wide SNP data (table 3, fig. 2). This sample includes 506 unrelated individuals from Europe, Pakistan, and the Middle East taken from the HGDP-CEPH and 199 Indian individuals from our own sample of 41 populations (supplementary table 1, Supplementary Material online), none of which have a history of pastoralism. All but four of the identified haplotypes have overall sample frequency <0.05; the remaining ones can be easily mapped to haplotypes A*T, A, B, and C as previously defined (Hollox et al. 2001; Poulter et al. 2003). All -13910*T alleles in our southern Asian sample (fig. 2) are found on the previously defined European A haplotype background, strongly arguing for a single common origin. Our network clearly recapitulates the previously established relationships between, and distribution of, the high frequency haplotypes, and most of the uncommon haplotypes can be easily identified as recombinants of high frequency ones.

The -13910*T allele has been identified as having undergone strong and recent positive selection in European populations (Bersaglieri et al. 2004). We searched for evidence of selection in our Indian sample by calculating the extended haplotype homozygosity score (EHH; Sabeti et al. 2002) in a 5-Mb region surrounding the -13910 C>T site. Figure 3*A–D* plots the decay of EHH against genetic distance from the 15-SNP core in all four continental population groups under study; results are shown only for the four observed high-frequency haplotypes. The -13910*T associated haplotype (shown in purple in fig. 3) consistently exhibits markedly greater EHH on both sides of the 15-SNP core than any other haplotype. Similarly, REHH scores associated with the haplotype are markedly greater than those associated with any other haplotype (supplementary fig. 3, Supplementary Material online). High homozygosity encompasses both the *LCT* and the *MCM6* genes and extends nearly 1 Mb on the *LCT* side (i.e., toward the centromere), although it drops off quickly in the other direction in all groups under study, a finding consistent with previous reports (Poulter et al. 2003; Bersaglieri et al. 2004; Voight et al. 2006; Sabeti et al. 2007). The structure of this extended haplotype block is fundamentally the same in all populations as a joint analysis of all samples yields a single

Table 3. Haplotype Frequencies by Region.

	India						
Haplotype[a]	Indo-European Speakers (138)	Non-Indo-European Speakers (61)	All India[b] (199)	Europe (156)	Near East (160)	Pakistan (190)	Overall (705)
1 (A*T)	0.149	0.049	0.118	0.343	0.069	0.232	0.187
2 (A)	0.239	0.295	0.256	0.147	0.200	0.184	0.200
3	*0.004*	*0.008*	0.005				0.001
4					*0.003*	*0.003*	0.001
5	0.025		0.018	0.006		0.024	0.013
6	0.014	0.033	0.020	0.038	0.025	0.016	0.024
7				*0.003*	*0.003*		0.001
8	0.007	*0.008*	0.008				0.002
9					*0.003*		*0.001*
10					*0.003*		*0.001*
11				*0.003*			*0.001*
12 (B)	0.217	0.148	0.196	0.285	0.384	0.234	0.269
13	0.022	0.016	0.020		*0.003*	0.005	0.008
14					*0.003*	*0.003*	0.001
15	0.011		0.008			0.005	0.004
16 (C)	0.304	0.418	0.339	0.170	0.288	0.258	0.268
17					*0.003*		*0.001*
18	0.007	0.025	0.013			0.011	0.006
19						0.005	0.001
20					0.009		0.002
21				*0.003*	*0.003*	0.021	0.007

NOTE.—Sample sizes are shown in parentheses; italicized cells are haplotypes found in a single individual and thus likely recombinants. Haplotype numbering corresponds to figure 2.

[a] Haplotype names from Hollox et al. (2001) are given in parentheses.

[b] All India is the sum of Indo-European and non Indo-European speakers.

long-range haplotype associated with the T allele, with EHH scores that fall between those observed in the different continental groups (supplementary fig. 4, Supplementary Material online).

However, although the >1-Mb haplotype is found in all -13910*T-carrying populations, the decay patterns are not identical across populations, with the Indian samples exhibiting a markedly ladder-like pattern of decay, whereas Europe, the Near East, and Pakistan are all characterized by a single steep drop of EHH roughly 800 kb from the core region at around position 135.5 Mb on chromosome 2 (NCBI human genome build 36). Also striking is a clearly visible 500-kb–long region of relatively high EHH associated with haplotype B in all populations. This region is centered around 136 Mb, where two genes, *ZRANB3* and *R3HDM1*, are located. Although the first gene is a zinc-finger protein of unknown function, this second gene, which encodes an R3H domain–containing protein, has been implicated in food conversion efficiency in cattle (Barendse et al. 2007). *R3HDM1* also shows some evidence of having come under positive selection during the domestication process (Gibbs et al. 2009), suggesting that there may be other nearby functional variants under selection in humans. It is interesting to note that Tishkoff et al. (2007) similarly detected evidence of a secondary sweep centered around *R3HDM1*, and it is possible that the ancestral haplotype for -14010*C on which Tishkoff et al. base their observations may be closely related to the one we are detecting.

With the exception of haplotype A, which shows markedly greater conservation in the proximal direction in the Near Eastern sample than in any other region, the remaining haplotypes display similar degrees of decay across populations. Finally, because Indo-European–speaking groups were repeatedly identified as being significantly associated with lactase persistence in our data, we examined EHH in Indo-European speakers alone (supplementary fig. 5, Supplementary Material online). There is a sizable overrepresentation of Indo-Europeans ($n = 138$) in our genomic Indian data set ($n = 199$), which explains the similar decay patterns in figure 3*A* and supplementary figure 3*a* (see Supplementary Material online). Of the 122 non-Indo-European chromosomes in our genome-wide data set, only 6 carry the -13910*T allele, and thus, we did not examine them independently.

Discussion

The populations of Europe, Arabia, and parts of Africa that adopted subsistence strategies that included dairying have high frequencies of individuals who can digest lactose after weaning. This phenotype has been shown to be associated with multiple functional mutations in an *LCT* enhancer region, roughly 14-kb upstream from the transcription start site (reviewed in Ingram, Mulcare, et al. 2009). At the genetic level, lactase persistence has evolved independently at least four times (Enattah et al. 2002; Tishkoff et al. 2007; Enattah et al. 2008); in two of these cases (-13910*T and -14010*C), the alleles have demonstrably come under the action of positive selection (Bersaglieri et al. 2004; Coelho et al. 2005; Tishkoff et al. 2007; Enattah et al. 2008; Itan et al. 2009). Therefore, the appearance and subsequent expansion of an agropastoral subsistence strategy

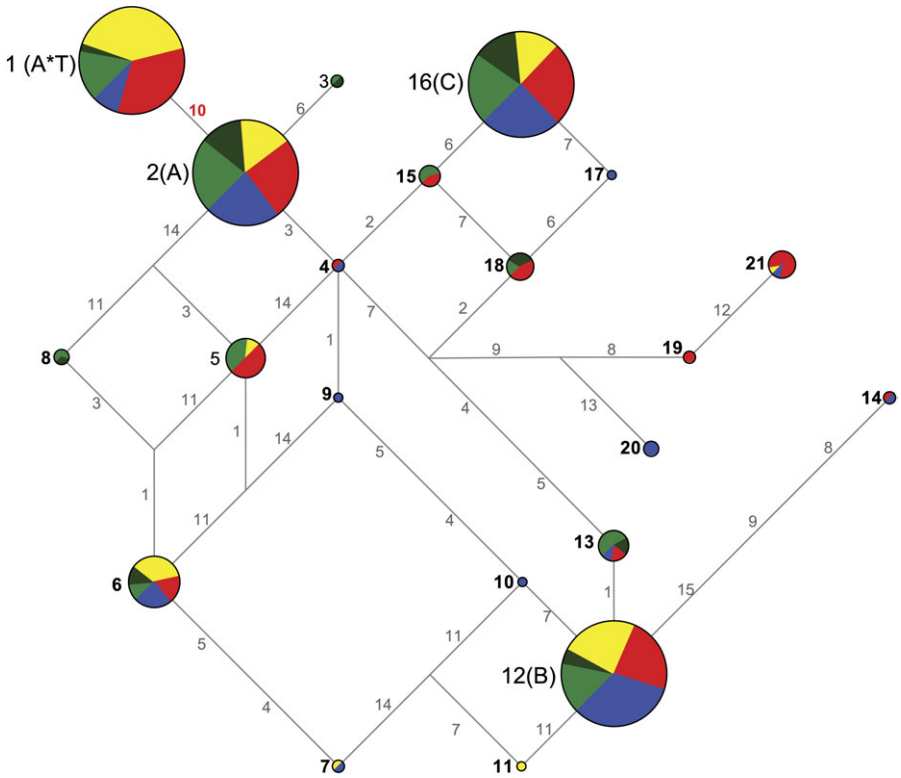

FIG. 2. Maximum parsimony neighbor joining network of 21 15-SNP–long haplotypes identified in 705 samples of western Eurasian origin. Node nomenclature corresponds to that in table 3, mutations correspond to those in supplementary table 2 (Supplementary Material online). Nodes are proportional to haplotype frequency; geographic distribution of each haplotype is coded by color: Yellow, Europe; red: Near East; blue: Pakistan; light green: Indo-European speakers from India; and dark green: non-Indo-European speakers from India.

over much of India during the last 7,000 years (Meadow 1993; Loftus et al. 1994; Possehl 1997; Fuller 2003) lead to expectations of a high frequency of lactase persistence in the subcontinent. Indeed, previous studies have suggested that lactase persistence is a common phenotype in Pakistan and northern India (Desai et al. 1970; Swaminathan et al. 1970; Gupta et al. 1971; Reddy and Pershad 1972; Bersaglieri et al. 2004; Babu et al. 2010). However, the genetic basis of adult lactase persistence in India has remained largely unexplored, leaving major questions about the origins, diversity, and history of this major cultural and nutritional adaptation unanswered. We have conducted a systematic search by sequencing the region in which the key mutations known to cause lactase persistence are found, as well as looking for new mutations, in over 2,284 individuals from 106 populations living in the Indian subcontinent.

-13910*T allele frequency predicts lactase persistence with high accuracy in most of Europe, especially in the northwestern half of the continent (Itan et al. 2010). The first study of its incidence in India showed that, at least in two mixed urban populations, this mutation occurs at frequencies >30% (Babu et al. 2010). The results presented herein not only confirm the widespread presence of the -13910*T allele in India but also the absence of other alleles in the genomic region under study that could make significant contributions to lactase persistence frequencies. Population-level -13910*T allele frequencies within our samples range from 0 to 0.489 and predicted phenotype frequencies from 0 to 0.739. Nonetheless, the predicted countrywide phenotype frequency in our sample, obtained from pooling all data points, is substantially lower (0.196) than estimated in previous interpolation-based studies with considerably less genetic data (~0.40) (Ingram, Mulcare, et al. 2009; Itan et al. 2010). GenoPheno analyses (Mulcare et al. 2004; Itan et al. 2010) on interpolated phenotype and allele frequencies indicate that with the exception of some small regions, this difference is not statistically significant. It should, however, be noted that there is a paucity of phenotypic data.

Short- and long-range haplotype data point strongly toward a single shared origin of the -13910 C>T mutation in Europe and India (fig. 3). Given the distribution of the -13910*T allele, the observed degree of haplotype conservation indicates a recent origin followed by a rapid rise in frequency, most likely driven by strong positive selection. There are three a priori scenarios that could explain this sharing: 1) The LRH originated somewhere in south Asia and spread westward; 2) the LRH originated in Europe and spread to India either through the Near East or the Central Asian steppe; and 3) the initial sweep that led to the establishment of the LRH occurred neither in Europe nor India but somewhere in between, either in the Central Asian steppe or in the Fertile Crescent, and spread both east and west from there. It is commonly held that -13910*T and its associated phenotype only became truly advantageous following the domestication of cattle and the adoption of a dairying culture (Simoons 1970a; McCracken

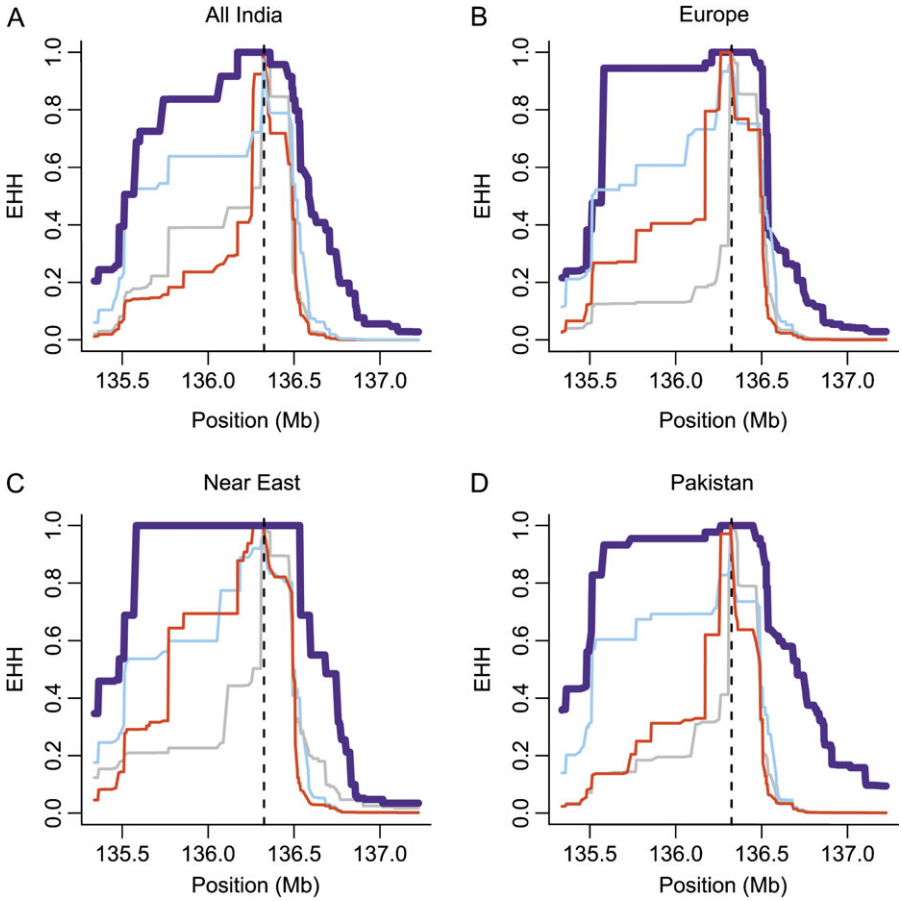

Fig. 3. EHH decay at the continental level. Haplotypes are coded by color; names correspond to those in table 3 and figure 2: purple, haplotype 1 (A*T), associated with lactase persistence; orange: haplotype 2 (*A*); pale blue: haplotype 12 (*B*); and gray: haplotype 16 (*C*). The dashed line marks the location of the 15-SNP core relative to which EHH scores were calculated; it spans hg18 positions 136,278,205–136,339,424. LCT itself is located at positions 136,261,885–136,311,220, whereas MCM6 is at chr2: 136,313,666–136,350,481. The centromere is located at roughly 95 Mb; the telomere at 242 Mb. The genomic coordinates are given by Human Genome Build 36/hg18.

1971; Durham 1991). Comparison of the European and Indian history of cattle keeping suggests that the context for the selective sweep was in place in Europe roughly 2,000 years earlier than in south Asia (Copley et al. 2003; Copley, Berstan, Dudd, et al. 2005; Copley, Berstan, Mukherjee, et al. 2005; Evershed et al. 2008; Balaresque et al. 2010). The earliest evidence of cattle herding in south Asia comes from the Indus River Valley site of Mehrgarh and is dated to ~7,000 YBP (Meadow 1993), thus postdating the predicted start of the European Neolithic sweep ~7,500 YBP (Burger et al. 2007; Itan et al. 2009). It is therefore plausible that from Europe, the allele subsequently spread into Central Asia, the Near East, Pakistan, and India, although not necessarily by the same route or the same demographic event.

If the -13910*T allele was introduced into south Asia from the west, the question arises as to whether this introduction was part of a major migratory event or a minor gene flow event potentially followed by a selective sweep within the subcontinent. A recent study has argued for a sizeable contribution from western Eurasia to the ancestral north Indian gene pool (Reich et al. 2009), although it has not provided a date or a precise origin location for this event, and the observation is compatible with low continuous gene flow over a long time period. However, there is a paucity of mitochondrial DNA (mtDNA) and Y chromosome lineages in the Indian gene pool that have been assigned a Central Asian origin (Metspalu et al. 2004; Sahoo et al. 2006), suggesting that the west Eurasian genetic contribution identified by Reich et al. (2009) principally reflects gene flow from Iran and the Middle East. It is therefore likely that the -13910*T allele was introduced to the subcontinent via a small number of migrants in the wake of the well-documented sweep for lactase persistence in the European middle Neolithic (Bersaglieri et al. 2004; Burger et al. 2007; Ingram, Mulcare, et al. 2009; Itan et al. 2009). Interestingly, a similar small-scale introgression event may also explain the presence of the 13910*T allele in central African populations (Mulcare et al. 2004; Coelho et al. 2005). The overall low frequency of the -13910*T allele in the subcontinent, and its absence from roughly a third of the population samples, suggests that the selective advantage it confers has been largely restricted to pastoralist groups. Three of the five sampled pastoralist populations have -13910*T allele frequencies vastly greater than the sample mean; of those, the Ror of Haryana are responsible for the frequency peaks visible in figure 1, a pattern suggestive of

adaptation to a lifestyle tightly associated with the consumption of milk products by these populations. A similar high-frequency peak associated with the Toda in south India, however, is obscured by the presence in the sample of three nearby populations with very low -13910*T frequencies. Further studies with more pastoralist populations would be required to explore the link between the lactase persistence allele frequencies and lifestyle in India.

In the case of position -13915, a different substitution, T>G, has been described in East African and Saudi Arabian populations (Ingram et al. 2007; Tishkoff et al. 2007; Enattah et al. 2008) in association with lactase persistence, suggesting that -13915*C, as observed in India, may also be causative of the lactase persistence phenotype, whereas -13779 G>C has not been associated with any known binding sites in the *LCT* promoter region to date (Troelsen et al. 2003; Lewinsky et al. 2005). In published data, the -13779*C allele has been reported only in a single lactose nondigester Somali individual (Ingram, Raga, et al. 2009), although this is not sufficient evidence to rule out an association with lactase persistence. On a broader scale, five of the seven additional variants we have identified surrounding -13910 C>T can be found in southern India, often segregating in multiple groups. Of these groups, the pastoralist Toda are somewhat exceptional as three alleles can be found in the population at appreciable frequency: -13910*T, -13915*C, and -13779*C. Together these findings are suggestive of the beginnings of a soft selective sweep, in a manner similar to that which has been proposed in a Somali cohort where multiple potentially causative mutations occur together in lactose digesters (Ingram, Raga, et al. 2009), although further study of populations in the region and their extended haplotypes would be necessary to confirm this.

These results further indicate that, as in European populations, an association with dairying is the major selective force behind the distribution of lactase persistence in the subcontinent. We nevertheless predict phenotypic frequencies of up to 0.40 in some Indian populations with no traditional links to pastoralism, posing the question of what factors have shaped the countrywide distribution of -13910*T alleles in India among the majority of populations who do not seem to traditionally base their subsistence on dairying (Simoons 1970b). Populations from north or west India that speak Indo-European languages have, in general, substantially higher frequencies of the -13910*T allele than Dravidian speakers from the south. However, the polymorphism is clearly not confined to Indo-European speakers, and within the Indo-European speaking groups, there is a clear west to east frequency decline, whereas there are no abrupt frequency differences between geographically adjacent Indo-European and Dravidian speakers (fig. 1, supplementary table 1, Supplementary Material online). Furthermore, partial Mantel tests showed that this association disappears when geography is controlled for. This geographically contiguous patterning of variation in the locus is consistent with other genetic data from both autosomal and uniparentally inherited marker loci in India (Kivisild et al. 2003; Metspalu et al. 2004; Sahoo et al. 2006; Reich et al. 2009). A particularly striking example of the complexity of the interaction between subsistence strategy, cultural norms, and geographical proximity is found in the Nilgiri Hills, home to the pastoralist Toda, and three additional nonpastoralist groups, who have allelic frequencies ranging from 0.012 to 0.233. When considering socially defined population structure (i.e., castes and tribes), allele frequency patterns within the same endogamous caste can fluctuate vastly across geographic regions. Likewise, some Indo-European and Dravidian tribal communities, like the Bhil, Badaga, or the Toda, have higher allele frequencies than some Indo-European or Dravidian castes.

The pattern observed among speakers of Tibeto-Burman and Austroasiatic languages requires a separate explanation. In both of these groups, the observed -13910*T allele frequencies are close to 0, consistent with the fact that none of the populations studied from among these language families have known histories of pastoralism or significant milk consumption (Simoons 1970b). Here, we do find a clear-cut division between linguistic affiliation and genotype. In the case of Tibeto-Burman groups, this is again in line with previous mtDNA, Y-chromosome, and autosomal studies that identified these populations as recent migrants from eastern Asia (Metspalu et al. 2004; Sahoo et al. 2006; Reich et al. 2009), whereas the rare instances of lactase persistence in Austroasiatic-speaking populations, like the Mahali of Jharkhand, are probably due to recent gene flow from neighboring Indo-European or Dravidian speakers.

Our study has, for the first time, documented the distribution of known alleles in the *LCT* regulatory region across India, as well as the existence of previously unknown alleles. Taken together, our results indicate that the -13910*T allele is responsible for the substantial proportion of lactase persistence in the country. Furthermore, haplotype analyses indicate that the -13910*T allele in India is identical by descent to that found in Europe and western Asia, whereas examination of the pattern of haplotype block structure in the context of the archaeological history of herding across this intercontinental region suggests that the -13910*T allele was introduced to India from the west. However, within India, the lactase persistence phenotype has had a more structured adaptive history, with higher frequencies clustered in those groups that traditionally practice a dairying economy. Lactase persistence remains one of, if not, the best examples of coevolution between cultural and biological innovations, and the historical and socioeconomic complexity of India provides a unique opportunity for exploring the processes that generate human diversity.

Supplementary Material

Supplementary figures 1–5 and tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are very grateful to all the individuals who donated samples for this work and to Dr V. R. Rao and Dr Madhu Bala Sharma at the Anthropological Survey of India and G. Arun Kumar for their collaboration with this work. Dr Jakka Parthasarathy and Anurag Kadian aided in the collection of pastoralist samples. We thank Mari Nelis, Georgi Hudjashov, and Viljo Soo for conducting the autosomal genotyping. This work was supported by the UK-India Education Research Initiative (grant number RG47772), the European Commission Seventh Framework Programme Internal Training Networks (LeCHE, grant number 215362-2), the European Union Seventh Framework Programme Ecogene (grant number 205419), the European Union Regional Development Fund through a Centre of Excellence in Genomics award, and a Bhatnagar Fellowship from Council of Scientific and Industrial Research, Government of India. Web resources: GLAD database: http://www.ucl.ac.uk/mace-lab/resources/glad; NETWORK: http://www.fluxus-engineering.com/sharenet.htm; and OMIM: http://www.ncbi.nlm.nih.gov/omim

References

Babu J, Kumar S, Babu P, Prasad JH, Ghoshal UC. 2010. Frequency of lactose malabsorption among healthy southern and northern Indian populations by genetic analysis and lactose hydrogen breath and tolerance tests. *Am J Clin Nutr*. 91:140–146.

Baig M, Beja-Pereira A, Mohammad R, Kulkarni K, Farah S, Luikart G. 2005. Phylogeography and origin of Indian domestic cattle. *Curr Sci* 89:38–40.

Balaresque P, Bowden GR, Adams SM, et al. (16 co-authors). 2010. A predominantly neolithic origin for European paternal lineages. *PLoS Biol*. 8:e1000285.

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

Barendse W, Reverter A, Bunch RJ, Harrison BE, Barris W, Thomas MB. 2007. A validated whole-genome association study of efficient food conversion in cattle. *Genetics* 176:1893–1905.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.

Bayless TM, Rosensweig NS. 1966. A racial difference in incidence of lactase deficiency. A survey of milk intolerance and lactase deficiency in healthy adult males. *JAMA*. 197:968–972.

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.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 81:1084–1097.

Buller HA, Kothe MJ, Goldman DA, Grubman SA, Sasak WV, Matsudaira PT, Montgomery RK, Grand RJ. 1990. Coordinate expression of lactase-phlorizin hydrolase mRNA and enzyme levels in rat intestine during development. *J Biol Chem*. 265:6978–6983.

Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc Natl Acad Sci U S A*. 104:3736–3741.

Chen S, Lin BZ, Baig M, et al. (23 co-authors). 2010. Zebu cattle are an exclusive legacy of the South Asia Neolithic. *Mol Biol Evol*. 27:1–6.

Clutton-Brock J. 1999. A natural history of domesticated mammals. Cambridge (MA): Cambridge University Press.

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.

Copley MS, Berstan R, Dudd SN, Aillaur S, Mukherjee AJ, Straker V, Payne S, Evershed RP. 2005a. Processing of milk products in pottery vessels through British prehistory. *Antiquity* 79: 895–908.

Copley MS, Berstan R, Dudd SN, Docherty G, Mukherjee AJ, Straker V, Payne S, Evershed RP. 2003. Direct chemical evidence for widespread dairying in prehistoric Britain. *Proc Natl Acad Sci U S A*. 100:1524–1529.

Copley MS, Berstan R, Mukherjee AJ, Dudd SN, Straker V, Payne S, Evershed RP. 2005b. Dairying in antiquity. III. Evidence from absorbed lipid residues dating to the British Neolithic. *J Archaeol Sci*. 32:523–546.

Cuatrecasas P, Lockwood DH, Caldwell JR. 1965. Lactase deficiency in the adult. A common occurrence. *Lancet*. 1:14–18.

Davey Smith G, Lawlor DA, Timpson NJ, Baban J, Kiessling M, Day IN, Ebrahim S. 2009. Lactase persistence-related genetic variant: population substructure and health outcomes. *Eur J Hum Genet*. 17:357–367.

Desai HG, Gupte UV, Pradhan AG, Thakkar KD, Antia FP. 1970. Incidence of lactase deficiency in control subjects from India. Role of hereditary factors. *Indian J Med Sci*. 24:729–736.

Durham WH. 1991. Cultural mediation: the evolution of adult lactose absorption. In: Coevolution: genes, culture and human diversity. Stanford (CA): Stanford University Press. p 629.

Eaaswarkhanth M, Haque I, Ravesh Z, et al. (12 co-authors). 2010. Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *Eur J Hum Genet*. 18:354–363.

Enattah NS, Jensen TG, Nielsen M, et al. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am J Hum Genet*. 82:57–72.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet*. 30:233–237.

Enattah NS, Trudeau A, Pimenoff V, et al. (22 co-authors). 2007. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am J Hum Genet*. 81:615–625.

Evershed RP, Payne S, Sherratt AG, et al. 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature* 455:528–531.

Ferguson A, Maxwell JD. 1967. Genetic aetiology of lactose intolerance. *Lancet* 2:188–190.

Food and Agriculture Organization of the United Nations. 2009. FAO statistical yearbook. Rome (Italy): Food and Agriculture Organization of the United Nations.

Fuller D. 2003. An agricultural perspective on dravidian historical linguistics: archaeological crop packages, livestock and dravidian crop vocabulary. In: Bellwood P, Renfrew C, editors. Assessing the languaging/farming dispersal hypothesis. Cambridge (MA): McDonald Institute for Archaeological Research Monographs.

Gerbault P, Moret C, Currat M, Sanchez-Mazas A. 2009. Impact of selection and demography on the diffusion of lactase persistence. *PLoS One*. 4:e6369.

Gibbs RA, Taylor JF, Van Tassell CP, et al. (91 co-authors). 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324:528–532.

Gupta PS, Misra RC, Ramachandran KA, Sarin GS, Chutani HK. 1971. Intestinal disaccharidases activity in normal adult population in tropics. *J Trop Med Hyg*. 74:225–229.

Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Rege JE. 2002. African pastoralism: genetic imprints of origins and migrations. *Science* 296:336–339.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM. 2001. Lactase haplotype diversity in the Old World. *Am J Hum Genet*. 68:160–172.

Ingram CJ, Elamin MF, Mulcare CA, et al. (11 co-authors). 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum Genet*. 120:779–788.

Ingram CJ, Mulcare CA, Itan Y, Thomas MG, Swallow DM. 2009a. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet*. 124:579–591.

Ingram CJ, Raga TO, Tarekegn A, Browning SL, Elamin MF, Bekele E, Thomas MG, Weale ME, Bradman N, Swallow DM. 2009b. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *J Mol Evol*. 69:579–588.

Itan Y, Jones BL, Ingram CJ, Swallow DM, Thomas MG. 2010. A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol Biol*. 10:36.

Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG. 2009. The origins of lactase persistence in Europe. *PLoS Comput Biol*. 5:e1000491.

Jensen RG. 1995. Handbook of milk composition. San Diego (CA): Academic Press.

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

Kivisild T, Rootsi S, Metspalu M, et al. (18 co-authors). 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet*. 72:313–332.

Kumar S, Nagarajan M, Sandhu JS, Kumar N, Behl V. 2007. Phylogeography and domestication of Indian river buffalo. *BMC Evol Biol*. 7:186.

Lacey SW, Naim HY, Magness RR, Gething MJ, Sambrook JF. 1994. Expression of lactase-phlorizin hydrolase in sheep is regulated at the RNA level. *Biochem J*. 302(Pt 3):929–935.

Lewinsky RH, Jensen TG, Moller 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. *Hum Mol Genet*. 14:3945–3953.

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

Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. 1994. Evidence for two independent domestications of cattle. *Proc Natl Acad Sci U S A*. 91:2757–2761.

McCracken RD. 1971. Lactase. deficiency: an example of dietary evolution. *Curr Anthropol*. 12:479–517.

Meadow RH. 1993. Animal domestication in the Middle East: a revised view from the eastern margin. In: Possehl G, editor. Harappan civilization. New Delhi (India): Oxford University Press and India Book House. p 295–320.

Metspalu M, Kivisild T, Metspalu E, et al. (16 co-authors). 2004. Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet*. 5:26.

Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, Thomas MG. 2004. The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet*. 74:1102–1110.

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

Plimmer RH. 1906. On the presence of lactase in the intestines of animals and on the adaptation of the intestine to lactose. *J Physiol*. 35:20–31.

Possehl GL. 1997. The transformation of the Indus Civilization. *J World Prehistory*. 11:425–472.

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM. 2003. The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann Hum Genet*. 67:298–311.

R Core Development Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Reddy V, Pershad J. 1972. Lactase deficiency in Indians. *Am J Clin Nutr*. 25:114–119.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L 2009. Reconstructing Indian population history. *Nature* 461:489–494.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sabeti PC, Varilly P, Fry B, et al. (247 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Sahi T, Isokoski M, Jussila J, Launiala K, Pyorala K. 1973. Recessive inheritance of adult-type lactose malabsorption. *Lancet* 2:823–826.

Sahoo S, Singh A, Himabindu G, et al. (12 co-authors). 2006. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc Natl Acad Sci U S A*. 103:843–848.

Sebastio G, Villa M, Sartorio R, Guzzetta V, Poggi V, Auricchio S, Boll W, Mantei N, Semenza G. 1989. Control of lactase in human adult-type hypolactasia and in weaning rabbits and rats. *Am J Hum Genet*. 45:489–497.

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

Simoons FJ. 1970b. The traditional limits of milking and milk use in southern Asia. *Anthropos* 65:547–593.

Simoons FJ. 1978. The geographic hypothesis and lactose malabsorption. A weighing of the evidence. *Am J Dig Dis*. 23: 963–980.

Swaminathan N, Mathan VI, Baker SJ, Radhakrishnan AN. 1970. Disaccharidase levels in jejunal biopsy specimens from American and south Indian control subjects and patients with tropical sprue. *Clin Chim Acta*. 30:707–712.

Tandon RK, Joshi YK, Singh DS, Narendranathan M, Balakrishnan V, Lal K. 1981. Lactose intolerance in North and South Indians. *Am J Clin Nutr*. 34:943–946.

Thangaraj K, Singh L, Reddy AG, Rao VR, Sehgal SC, Underhill PA, Pierson M, Frame IG, Hagelberg E. 2003. Genetic affinities of the Andaman Islanders, a vanishing human population. *Curr Biol*. 13:86–93.

Tishkoff SA, Reed FA, Ranciaro A, et al. (19 co-authors). 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*. 39:31–40.

Troelsen JT, Olsen J, Moller J, Sjostrom H. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology* 125:1686–1694.

Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, Bradley DG. 2001. Genetic evidence for Near-Eastern origins of European cattle. *Nature* 410:1088–1091.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. 4:e72.

Wa19281165106

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 76:887–893.