

Multiple Rare Variants as a Cause of a Common Phenotype: Several Different Lactase Persistence Associated Alleles in a Single Ethnic Group

Catherine J. E. Ingram · Tamiru Oljira Raga · Ayele Tarekegn · Sarah L. Browning · Mohamed F. Elamin · Endashaw Bekele · Mark G. Thomas · Michael E. Weale · Neil Bradman · Dallas M. Swallow

Received: 23 June 2009 / Accepted: 30 October 2009 / Published online: 24 November 2009
© Springer Science+Business Media, LLC 2009

Abstract Persistence of intestinal lactase into adulthood allows humans to use milk from other mammals as a source of food and water. This genetic trait has arisen by convergent evolution and the derived alleles of at least three different single nucleotide polymorphisms ($-13910C>T$, $-13915T>G$, $-14010G>C$) are associated with lactase persistence in different populations. Each allele occurs on an extended haplotype, consistent with positive directional selection. The SNPs are located in an ‘enhancer’ sequence in an intron of a neighboring gene (*MCM6*) and modulate lactase transcription in vitro. However, a number of lactase persistent individuals carry none of these alleles, but other low-frequency single nucleotide polymorphisms have been

observed in the same region. Here we examine a cohort of 107 milk-drinking Somali camel-herders from Ethiopia. Eight polymorphic sites are identified in the enhancer. -13915^*G and -13907^*G (a previously reported candidate) are each significantly associated with lactase persistence. A new allele, -14009^*G , has borderline association with lactase persistence, but loses significance after correction for multiple testing. Sequence diversity of the enhancer is significantly higher in the lactase persistent members of this and a second cohort compared with non-persistent members of the two groups ($P = 7.7 \times 10^{-9}$ and 1.0×10^{-3}). By comparing other loci, we show that this difference is not due to population sub-structure, demonstrating that increased diversity can accompany selection. This contrasts with the well-documented observation that positive selection decreases diversity by driving up the frequency of a single advantageous allele, and has implications for association studies.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9301-y) contains supplementary material, which is available to authorized users.

C. J. E. Ingram · M. G. Thomas · D. M. Swallow (✉)
Research Department of Genetics, Evolution and Environment (GEE), University College London, Wolfson House,
4 Stephenson Way, London NW1 2HE, UK
e-mail: d.swallow@ucl.ac.uk

C. J. E. Ingram
e-mail: catherine.ingram@ucl.ac.uk

A. Tarekegn · S. L. Browning · N. Bradman
The Centre for Genetic Anthropology, GEE,
University College London, London, UK

T. O. Raga · A. Tarekegn · E. Bekele
Addis Ababa University, Addis Ababa, Ethiopia

M. F. Elamin
Elrazi College of Medical Sciences, Khartoum, Sudan

M. E. Weale
Department of Medical & Molecular Genetics, King’s College
London, Tower Wing, Guy’s Hospital, London, UK

Keywords Lactase persistence · Population genetics · Evolutionary genetics · Gene-culture co-evolution · Selection · Africa · Milk · Humans

Introduction

Persistence of small-intestinal lactase production into adult life in humans is caused by genetic differences *cis*-acting to the lactase gene, *LCT* (Wang et al. 1995), which enable some alleles to escape the developmental down-regulation characteristic of the ancestral state, in which lactase expression declines and restricts the ability to digest lactose in milk after childhood. Lactase persistent individuals (lactose digesters) can readily consume large amounts of milk as adults, and there is considerable evidence to

suggest that the trait is subject to strong positive selection in humans (reviewed in Ingram et al. 2009). To date, three different nucleotide polymorphisms have been identified that are associated with increased lactose digestion capacity. The first allele to be identified, -13910^*T (rs4988235) was discovered in a Finnish sample (Enattah et al. 2002), and is tightly associated with high adult lactase expression throughout Europe. -13910^*T is present on the background of a single very extended haplotype previously defined ‘A’ (Hollox et al. 2001; Poulter et al. 2003; Bersaglieri et al. 2004; Coelho et al. 2005). More recently, -13915^*G (rs41380347) was found to be associated with increased lactose digestion capacity (Ingram et al. 2007; Tishkoff et al. 2007) and greater lactase activity (Imtiaz et al. 2007) in Sudanese and Middle Eastern populations respectively, and a third allele, -14010^*C , was found to be associated with increased lactose digestion capacity in Kenyan and Tanzanian populations (Tishkoff et al. 2007). -13915^*G and -14010^*C are also present on single extended haplotypes that are distinct from the A haplotype and from each other (Tishkoff et al. 2007; Enattah et al. 2008). -13910^*T , -13915^*G and -14010^*C are all located within an intron of *MCM6* (the upstream neighbor of *LCT*), in a sequence that acts as an enhancer for lactase expression in vitro (Troelsen et al. 2003). Each of the alleles increases transcription compared with the ancestral allele in promoter-reporter constructs (Troelsen et al. 2003; Tishkoff et al. 2007), although the effects are small and there is no simple change in transcription factor binding (Ingram et al. 2007; Enattah et al. 2008). The pattern of association of each of these alleles with lactose digestion status is not absolute (Ingram et al. 2007; Tishkoff et al. 2007), and although there is an intrinsic error rate in the lactose tolerance test (LTT) (Mulcare et al. 2004), many persistent individuals can be confidently identified who do not carry -13910^*T , -13915^*G or -14010^*C . Two other putative functional alleles have been identified, -13913^*C (rs41456145) and -13907^*G (rs41525747), although these occur at very low frequencies in the groups tested so far, resulting in insufficient power to examine their association with lactose digestion (Ingram et al. 2007; Tishkoff et al. 2007; Enattah et al. 2008).

Since -13907^*G gave suggestive evidence of function in vitro (Tishkoff et al. 2007) and had been found previously in Ethiopians (Ingram et al. 2007), we initially collected samples from an Ethiopian pastoralist population with a view to testing the association of this allele with lactase persistence. Resequencing of the enhancer region in this cohort revealed remarkable heterogeneity, including previously reported, as well as novel variant alleles, which show a very marked difference in distribution with respect to lactose digestion status. The significance of this diversity is evaluated by testing for population substructure at other

loci and re-examining our previously tested phenotyped cohort which shows the same phenomenon. We report haplotype associations of each allele using a large pooled data set ($n = 746$) for increased accuracy of statistical inference.

Subjects and Methods

Test Cohort and Lactose Tolerance Testing

The ethnic group selected and tested were Somali, of whom about 3 million are resident in Ethiopia. The Somali were selected because of their documented history as a traditional pastoralist population (Blench 1999). Volunteers were recruited in Shinile (9.6833 N, 41.8500 E; approximately 10 km from Dire Dawa), in the Somali region of eastern Ethiopia. Individuals over the age of 18 and of self-declared Somali ethnicity were invited to participate in the study on the day prior to testing, when the purpose and possible side effects of the lactose tolerance test were clearly explained by a local interpreter and nurse. Each person who consented to participate agreed to observe an overnight (8 h) fast in preparation for the test. A local interpreter with personal knowledge of the participants was employed to recruit volunteers who were unrelated at least at grandparental level. Each sample donor was asked to complete a questionnaire which recorded whether they had taken antibiotics recently or experienced any gastro-intestinal illness. We also recorded their milk-drinking habits and that of their parents, and their grandparents. The age range of sample donors was from 19 to 70 years old, with 80% being between 20 and 50. Buccal cell samples were collected as described by Freeman et al. (2003). DNA samples were linked to questionnaire and lactose digestion data by code, but names were not taken. Local approval for this study was obtained from Addis Ababa University and the genetic part of the work was conducted in London under UCLH 99/0196 and 01/0236 ethics approvals.

Lactose tolerance testing was conducted as follows: breath hydrogen baseline readings were obtained using a MicroH₂ meter (Micro Medical Ltd) and all eligible individuals (i.e. breath hydrogen $\geq 0 \leq 20$ ppm) were given 50 g lactose dissolved in 250 ml water at room temperature and were requested to stay for the entire test duration (3 h). Breath hydrogen readings were taken at 30-min intervals during the test. In total, 107 samples were collected. Participants were classified into four mutually exclusive categories, two unambiguous (lactose digesters, D, and lactose non-digesters, ND) and two ambiguous (intermediate, I, and hydrogen non-producers, H₂NP). Assignment to these categories used the following criteria: D, a rise in

breath hydrogen of not more than 20 ppm for the entire test duration; ND, a rise in breath hydrogen of greater than 20 ppm or more for at least two consecutive readings; I, fluctuations in breath hydrogen occur throughout the test, not giving a clear sustained rise above 20 ppm nor a clear sustained flat line; H₂NP, a baseline reading of zero which remained at zero for the test duration. H₂NP individuals do not release hydrogen in their breath following the lactose load either because they have an absence or low number of hydrogen producing bacteria in their gut flora (Gilat et al. 1978) or because they are genuine lactose digesters. It is impossible to distinguish these two phenotypes without conducting further tests for the presence/absence of hydrogen producing bacteria, and so here they have been included in an ambiguous category which reflects this uncertainty.

Details of our earlier collection of DNA samples with associated lactose digestion phenotype, collected from the Jaali population residing in Shendi in central Sudan ($n = 86$), are described in Ingram et al. (2007).

Non-Phenotyped Cohorts Tested

Non-Phenotyped samples used for distribution data and haplotype inference ($n = 553$) consist of 89 European samples including 23 newly collected samples, 20 CEPHS and 46 samples previously described (Harvey et al. 1998); 89 Cameroonian samples including Fulani, Mambila and Shuwa Arabs; 96 Ethiopian samples including Amhara, Afar and Somali (from a separate collection in Jijiga); 96 Sudanese samples including Beni Amer, Donglawi and Shaigi; and 183 samples from the Middle East (including Bedouin and non-Bedouin Arabs).

Genotyping

Sequencing

PCR of the *MCM6* enhancer sequence was carried out in 15 μ l total volume of 1 \times reaction buffer IV (Abgene), 0.25 units of Taq DNA polymerase (Abgene), 0.2 mM dNTPs, approximately 10–20 ng genomic DNA and 0.5 μ M of each primer (MCM6ex13 5'-ATTTCCAAAGAGTCAG AGGACTTC-3' and MCM6778 5'-CCTGTGGGATA AAAGTAGTGATTG-3'). Cycling conditions were 95°C for 5 min, followed by 34 cycles of 95°C for 30 s, 58°C for 30 s and 72°C for 1 min. PCR products were cleaned by PEG precipitation and sequenced using Big-Dye Terminator chemistry (Applied Biosystems, Foster City, CA, USA) with the MCM6ex13 primer. Novel variants were confirmed by sequencing the reverse complement with the MCM6778 primer. Sequencing of the hypervariable segment 1 (HVS-1) of the mitochondrial DNA (mtDNA) was

carried out as previously described (Thomas et al. 2002). All sequence fragments were electrophoresed on an ABI 3730xl genetic analyser, and chromatograms examined using ChromasPro software (Technelysium).

LCT Haplotypic Markers

–942/3TC>AA, –678A>G, 666G>A (rs3754689) and 5579T>C (rs2278544) were genotyped as described elsewhere (Ingram et al. 2007). rs309180, rs4954490, and rs3769005 were typed by PCR-RFLP and rs4954493 was typed by tetra-primer arms PCR. Primers and PCR conditions can be found in Supplementary Table 1. Genotypes were inferred from agarose gel phenotype assuming no silent alleles.

Y Chromosome

Typing of six unique event polymorphism (UEP) markers (92R7, SRY+465, SRY4064, sY81, Tat and YAP) and six microsatellite markers (DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393) was carried out using the technique previously described in Thomas et al. (1999).

Autosomal Short Tandem Repeat (STR) Markers

Fifteen unlinked autosomal STRs, commonly used for forensic differentiation (Penta E, D18S51, D21S11, TH01, D3S1358, FGA, TPOX, D8S1179, vWA, Penta D, CSF 1P0, D16S539, D7S820, D13S317 and D5S818), were genotyped using the PowerPlex 16 kit (Promega) modifying the manufacturer's instructions only to reduce reaction volume to 10 μ l and enzyme units to 0.5 U/ μ l. Amplified fragments were detected using a 3730xl genetic analyzer (ABI) and analyzed using GeneMapper v4.0 software (Applied Biosystems).

Statistical Analyses

Hardy–Weinberg Equilibrium

Tests for deviation from Hardy–Weinberg Equilibrium were conducted for each genotyped locus within each population using the exact test for HWE based on a Markov chain method implemented either in the program GENEPOP (Raymond and Rousset 1995b) or in Arlequin (Excoffier et al. 2005).

Haplotype Inference

Haplotypes (i.e. intron 13 variations plus *LCT* core haplotype tagging SNPs) were inferred using the Bayesian algorithm implemented in Phase (Stephens et al. 2001).

Haplotypes were inferred for each individual twice; once within their own population group and once for the pooled data set of 746 individuals (providing greater power for inferring the haplotypic associations of rare alleles (Andres et al. 2007). Outputs were inspected for agreement, and in the few cases where discrepancies were present a decision was made based on visual inspection of the haplotype pairs and consideration of the Phase probabilities generated by the software. Such decisions were only necessary in the case of two individuals in the Somali data set. Haplotype inference within the Somali population assigned the single occurrence of -14010^*C to a **C** haplotype, but when inferred for the pooled data set this allele was assigned to the **B** haplotype in all three individuals in whom it occurred. The second discrepancy involved assignment of -13915^*G on to a recombinant haplotype both within the Somali population and in the pooled data set. Upon visual inspection, genotype data for this sample (DD-084) was resolved into common **C** and **U** haplotypes.

Population Differentiation

GENEPOP (Raymond and Rousset 1995b) was used to investigate differentiation in STR allele distribution between pairs of sample groups. Contingency tables of alleles were constructed for pairs of populations and tested for independence using a Fishers exact test (Raymond and Rousset 1995a). Genepop uses Fisher's combined probability test to generate an overall P -value for differences in allele distribution between pairs of populations across all loci (Sokal and Rohlf 1994). Population differentiation at the intron 13, Y-chromosome and mtDNA loci was tested using permutation-based AMOVA of Wright's F_{ST} statistic (Excoffier et al. 1992) and an exact test of population differentiation (Raymond and Rousset 1995a) using Arlequin software (Excoffier et al. 2005). Genetic diversity was calculated using the test_h_diff program written by M. Weale (Thomas et al. 2002). The program calculates Nei's h (Nei 1987) for each population and tests for a significant difference in allele distribution between them, based on samples of haplotypes. P values are obtained using both a bootstrapping method and a Z test and the solution for P is equal to the larger of the two.

Results

Lactose Tolerance Test Results

DNA samples and lactose digestion data were collected from 49 females and 58 males of self-declared Somali ethnicity. Eleven were of intermediate lactose digestion status and eight showed no evidence of hydrogen production (Table 1). Unambiguous lactose digestion phenotypes were obtained for 88 of the 107. There were 21 lactose digesters, giving a frequency of 0.24. In this population, lactose digestion capacity did not correlate with milk-drinking behavior ($P = 1.00$ for a Fishers exact test of drinking 500 ml or more of milk per day), and 71% of the participants reported drinking at least half a litre of fresh milk regularly in a day.

Resequencing

The intron 13 enhancer region (-14133 to -13684) was sequenced in all 107 people of the phenotyped Somali cohort. The sequence from exon 13 up to position -14010 was completely invariant, but downstream of -14010 eight polymorphic sites were identified (see Fig. 1 and Table 2). Three of these sites ($-14009T>G$, $-13806A>G$ and $-13779G>C$) had not been previously reported.

Table 2 shows the distribution of each variant allele with respect to lactose digester status. There is a difference in allele distribution between the different groups. Seventeen of the twenty-one lactose digesters showed one or more derived alleles (ascertained by comparison with primate sequence) compared with only 16/67 non-digesters. Only three individuals within the data set are heterozygous for two different derived alleles, and no individual carried more than two derived alleles (Table 3). Each of the novel alleles is associated with different haplotypes, as seen below, so that there is reasonable evidence that the derived alleles are independent. Under this assumption of independence, the difference in prevalence of derived alleles between unambiguous digesters and non-digesters is highly significant ($P = 4.3 \times 10^{-6}$, Fishers exact test for a 2×2 table of variant/ancestral chromosomes and digester status; Table 3).

Table 1 Summary of lactose digestion results obtained from lactose tolerance testing in the Somali cohort

	Lactose digester status			
	Digester	Non-digester	Intermediate	H ₂ non-producer
n individuals	21	67	11	8
Mean H ₂ rise ppm (range)	3 (0–14)	76 (27–172)	20 (16–25)	0

'H₂ rise' refers to breath hydrogen (BH) measured in parts per million (ppm) over the 3-h test duration

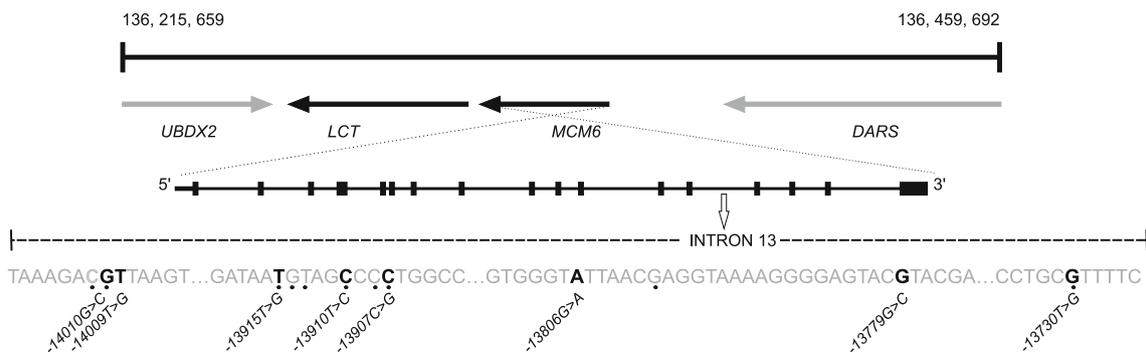


Fig. 1 Map of chr2q21 showing *LCT* and *MCM6* and the location of the SNPs identified within the *LCT* enhancer sequence within intron 13 of *MCM6*. Variant alleles observed in the Somali cohort are shown in *black*, and nucleotides previously reported to have variant alleles in other populations are indicated by a *black dot*. Nomenclature of the SNPs indicates position in number of nucleotides upstream of the *LCT*

transcription start site (on the allele in the human golden path sequence). Note that the genome sequence (indicated by the *black bar* at the top of the figure) is in the opposite orientation from the direction of transcription of *MCM6* and *LCT* (the genomic nucleotide positions correspond to those given by the UCSC Genome Browser, March 2006 assembly)

Table 2 Variant alleles observed in the lactase enhancer in the Somali cohort

	<i>n</i>	-14010*C	-14009*G	-13915*G	-13910*T	-13907*G	-13806*G	-13779*C	-13730*G
D	42	0	2	9	1	6	0	0	2
ND	134	0	0	1	0	0	2	1	13
I	22	1	0	0	1	1	0	0	1
H ₂ NP	16	0	1	1	2	5	0	0	1
<i>P</i> -value		–	0.056	1 × 10 ⁻⁵	0.23	1 × 10 ⁻⁴	0.57	0.76	0.53

n indicates the number of chromosomes sequenced; other columns show number of alleles of each kind found in each of the phenotype categories. Two-tailed *P*-values were obtained from 2 × 2 contingency tables of D and ND phenotypes for each allele. Note that Bonferroni correction for eight tests (0.05/8) gives a significance threshold of 0.006

D digester, ND non-digester, I intermediate, H₂NP hydrogen non-producer

Table 3 Contingency table of the unambiguous digester (D) and non-digester (ND) phenotype groups showing the distribution of individuals carrying either none, one or two variant alleles within the *LCT* enhancer sequence

	No. of variant alleles (per person)			Total
	0	1	2	
D	4	14	3	21
ND	51	15	1	67
Total	55	29	4	88

Of the four people who carry two derived alleles, only one was homozygous (for -13907*G)

Association of each allele with lactose digestion capacity was also tested, by comparing allele counts in the two groups, lactose digesters and non-digesters (Table 2). The most frequent derived allele (-13730*G, rs4954492) at the 3' end of the enhancer sequence shows no hint of association, and is, unlike the other derived alleles, widespread in several populations (B. Jones and D.M. Swallow, unpublished data, and Supplementary Table 2). Both -13915*G (previously shown to associate with lactase

persistence; Ingram et al. 2007; Tishkoff et al. 2007; Imtiaz et al. 2007) and -13907*G (previously demonstrated to have function in vitro; Tishkoff et al. 2007) are individually highly significantly associated with the lactose digester phenotype in this cohort, *P* = ≤ 1 × 10⁻⁴ for a 2 × 2 contingency table (Fishers exact test), and remain significant after Bonferroni correction for eight tests (corrected significance threshold = 0.006).

The -14010*C allele, previously shown to be associated with lactase persistence in Kenyan and Tanzanian populations (Tishkoff et al. 2007) and the European -13910*T allele each occur in single individuals in whom the LTT was inconclusive. Two of the newly identified loci (-13806A>G and -13779G>C) are rare and occur only in lactose non-digesters.

The third novel allele (-14009*G) is also rare and occurs in two lactose digesters and in one hydrogen non-producer. It shows marginal significant association with lactose digestion capacity in the Somali cohort (*P* = 0.056, Fishers exact test) at the uncorrected significance threshold (0.05), but is no longer close to significance at the corrected threshold of 0.006.

The discovery of $-14010G>C$ and $-14009T>G$ led us to re-examine the Sudanese Jaali cohort in which $-13915T>G$ was originally identified (Ingram et al. 2007). $-14009*G$ was found at a frequency of 0.06 in this cohort (11/166 chromosomes). Eight of these eleven $-14009*G$ alleles occurred in lactose digesters, six of whom carried no other variant alleles within the enhancer region, see Supplementary Table 2. Association of the SNP with lactase persistence status was however not statistically significant (one tailed $P = 0.08$, Fishers exact test). Nevertheless, taken together with the Somali data, the excess of $-14009*G$ alleles in persistent individuals is noteworthy.

Genetic Diversity and Differentiation

Our results show a significant difference in heterogeneity in the *LCT* enhancer locus between people of different inferred lactase persistence status. Since this might result from recent population admixture or other causes of structuring, we have examined the distribution of Y chromosome, mitochondrial and autosomal markers in the same samples. All should be sensitive to detecting population sub-structure; Y chromosome and mtDNA with their smaller effective population size, and/or high mutation rate, and the autosomal STRs because of their known power to differentiate closely related populations (Krenke et al. 2002).

Pairwise F_{ST} and Fishers exact tests of population differentiation, between the digesters and non-digesters, were calculated for HVS-1 mtDNA and Y chromosome haplotypes as well as for the *LCT* enhancer. Only the *LCT* enhancer showed a significant difference between the digesters and non-digesters ($P < 0.001$ for both F_{ST} and population differentiation). In order to compare the pattern of diversity observed at the *LCT* enhancer region with that observed for mtDNA and Y-chromosome, Nei's h (Nei 1987) was calculated for the digester and non-digester groups, and tested for significant difference (Thomas et al. 2002). In both the Somali and Jaali populations, only the *LCT* locus revealed a difference in the apportionment of genetic diversity (Table 4). Table 5 shows the results of a combined Fishers exact test of the difference in allelic distribution of 15 unlinked autosomal microsatellite markers between the digesters and non-digesters in both the Somali and the Jaali cohorts. There was no evidence of population differentiation between the phenotype groups of either cohort, although the Somali were significantly different from the Jaali.

Association of Intron 13 SNPs with Haplotype

To infer *LCT* gene haplotypes for each of the enhancer alleles, four *LCT* haplotype tagging SNPs were typed,

Table 4 Genetic diversity, Nei's h , of digesters and non-digesters of the Somali and Jaali cohorts for *LCT* enhancer (haplotypes), mtDNA (HVS-1 haplotypes) and Y-chromosome (haplotypes composed of 12 loci)

	Digester h (SE)	Non-digester h (SE)	P -value
Somali			
<i>LCT</i>	0.67 (0.03)	0.23 (0.02)	7.66×10^{-9}
mtDNA	0.99 (0.00)	0.98 (0.00)	0.76
Y	0.67 (0.07)	0.50 (0.04)	0.50
Jaali			
<i>LCT</i>	0.59 (0.03)	0.31 (0.02)	1.0×10^{-3}
mtDNA	0.99 (0.00)	0.97 (0.00)	0.32
Y	0.85 (0.02)	0.94 (0.01)	0.32

P -value taken from test-h-diff (Thomas et al. 2002)

Table 5 Outcome of contingency table comparison between the microsatellite allele distribution in the Somali and Jaali cohorts and the digesters/non-digesters

Population pair	χ^2	df	P -value
Somali and Jaali	75.77	30	<0.0001
Somali D and ND	24.39	30	0.7540
Jaali D and ND	31.50	30	0.3913

Differences in allele distribution of 15 STR markers were individually examined using Fishers exact test, and summed over all loci using Fishers combined probability test, as implemented in GenePop. There is strong statistical support for differences between but not within each cohort

allowing haplotypes to be defined according to the nomenclature published by Hollox et al. (2001). Genotype data for the intron 13 variants (see Supplementary Table 3) as well as the haplotyping SNPs in the Somali and Jaali cohorts were pooled with the same data from an additional 553 individuals from a number of geographic locations including Europe, the Middle East, and east and west Africa. The total data set used for haplotype inference included 746 individuals. Table 6 shows the frequency with which a given intron 13 lactase persistence-associated allele was observed on different *LCT* haplotypes.

All but one of the 127 $-13910*T$ alleles were observed on an **A** haplotype (76 from European individuals, 46 in the Cameroon Fulani and singletons in populations from the Middle East, Sudan and Ethiopia). The single non-**A** haplotype $-13910*T$ allele was found in a Fulani individual from Cameroon and was inferred to be part of the **F** haplotype, which may result from a recombination between haplotypes **B** and **A** (Hollox et al. 2001). $-13915*G$ was predominantly associated with the **C** haplotype, and the novel $-14009*G$ allele was found to associate with the **X** haplotype. $-14010*C$ was very rare and confined to the Somali ethnic group where it occurs on the **B** haplotype. $-13907*G$ shows more evidence of variation in its *LCT*

Table 6 Frequency with which a given intron 13 lactase–persistence associated allele was observed on different *LCT* haplotype backgrounds

	<i>-14010G >C</i> <i>n</i> = 3	<i>-14009T >G</i> <i>n</i> = 37	<i>-13915T >G</i> <i>n</i> = 169	<i>-13910C >T</i> <i>n</i> = 127	<i>-13907C >G</i> <i>n</i> = 58
A				0.992 (126)	0.776 (45)
B	1.000 (3)				0.017 (1)
C			0.959 (162)		
U		0.027 (1)			
X		0.973 (36)			
K			0.018(3)		0.069 (4)
F				0.008 (1)	0.138 (8)
Other			0.023 E (1), c (3)		

Number of chromosomes shown in parenthesis. The total data set included 746 individuals from different global population groups. Typing of 2 SNPs between the enhancer and the promoter (rs4954490 and rs3769005) and two (rs309180, rs4954490) upstream of the enhancer, in the phenotyped cohorts confirms the continuity of the haplotypes between *LCT* and intron 13, although it reveals some breakdown of the *-13915*G* carrying C haplotype, upstream of the enhancer (see Supplementary Table 4)

haplotype association than the other alleles. While it is observed mostly on the A haplotype (as reported by Enattah et al. 2008), we found a relatively high proportion of alleles (20%) associated with other haplotypes.

Discussion

The most striking outcome of this study is the finding that the *LCT* enhancer sequence is significantly more heterogeneous in the lactase persistent Somali than in the non-persistent members of the cohort. Reanalysis of the previously collected Jaali cohort shows the same phenomenon. Analysis of other loci (Y chromosome and autosomal STRs) demonstrates that this difference is not attributable to population stratification, and the observation of equally high diversity of mtDNA in the non-digesters as in the digesters excludes the possibility that the non-digesters are genetically more homogeneous. These findings are in dramatic contrast to the situation in Europe, where a single allele causal of lactase persistence is found, at very high frequency, in a genomic region of reduced genetic diversity, and which is a “textbook example” of the classical signal for a positive selective sweep. Here we argue that the *increased* degree of genetic diversity seen in the lactase persistent group resulting from multiple advantageous mutations is also a consequence of selection.

The clustering of the lactase persistence associated variants in a single short sequence region, the fact that they occur on different haplotype backgrounds, and the substantial degree of genetic differentiation of this region between phenotypically distinct groups, taken together support the conclusion that these changes are of functional importance, but also suggest that the enhancer region affects *LCT* expression in a complex manner. Most of the

persistence associated variants (*-13910*T*, *-13915*G*, and *-13907*G*) cluster around an Oct-1/GATA binding site (Lewinsky et al. 2005), where other rare SNPs have also been reported (*-13913T >C*/rs41456145 (Ingram et al. 2007); *-13914G >A* (Tag et al. 2007); and *-13908C >T*/rs4988236) and it now appears that a similar clustering of SNPs occurs upstream. The *-14010*C* lactase persistence associated allele is found in the centre of a run of three variable nucleotides, with *-14009T >G* and *-14011C >T* (rs4988233) on either side (Fig. 1). It is not yet known how these SNPs increase lactase expression; however, we speculate that while each of the persistence associated variants may have a different effect on protein interactions and/or chromatin structure of the enhancer region, all have the same effect of preventing the process of *LCT* down-regulation. In this sense, the lactase persistence associated SNPs can be regarded as ‘loss of function’ mutations that lead to a gain in activity. The situation with *LCT* may be directly analogous to the recently published study of the sonic hedgehog gene (*Shh*), in which a number of different point mutations in a *cis*-regulatory region located 1 Mb upstream act as ‘gain of function mutations’ activating ectopic *Shh* expression (Lettice et al. 2002, 2008), although in that case there is an associated pathology.

It is possible that the *MCM6* gene region is susceptible to mutations due to having an open chromatin state because of expression in gametogenesis (Swiech et al. 2007). However, if this is the case and the enhancer alleles are selectively neutral, a similar level of nucleotide diversity should be observed in both lactose digesters and non-digesters.

In this study, we also report the unexpected finding that lactose digestion capacity is not necessarily correlated with milk consumption. This is contrary to our findings that individuals adapted their milk intake to reflect digester

status in the Sudanese Jaali cohort (Ingram et al. 2007). The frequency of lactose digesters in this cohort is lower than might have been expected (24%), but agrees well with a large previously published study of the same ethnic group (Flatz 1987). It is possible that in this population, adaptation of the gut flora has occurred, allowing non-persistent individuals to consume lactose without symptoms. We did observe a general trend of lower starting breath hydrogen readings, and a large number of hydrogen non-producers in the Somali cohort, and this may signify increased colonic acidity (possibly due to dietary factors) which can prohibit colonization by hydrogen producing bacteria (Perman et al. 1981; Vogelsang et al. 1988). Whatever the nature of the adaptation that allows lactose non-digesters to consume milk in large quantities, the observation has implications for interpretation of the genetic pattern observed. This and the low lactase persistence frequency implies that the selective pressure has not acted to drive one particular lactase persistence allele to high frequency, and may indicate that selection has been either weak, or has fluctuated over time, which is possible if lactase persistence is more advantageous during periods of famine and drought. The underlying selective advantage of lactase persistence is still not clearly defined and further work is required in order to understand the circumstances under which selection for the phenotype increases, for example by studying two or three generations of a famine-exposed population.

Each of the alleles described here also occurs in other populations, and the distribution patterns of each allele suggest quite different origins. It is possible that -13907^*G arose in Ethiopia, being most frequent in the nomadic camel milking Afar (Supplementary Table 3), but the other alleles probably did not (Ingram et al. 2007; Tishkoff et al. 2007; Imtiaz et al. 2007; Enattah et al. 2008). Therefore, the occurrence of -13910^*T , -13915^*G , -13907^*G and -14010^*C together in a single ethnic group may signify past contact between migratory milk-drinking peoples through shared cultural practices.

Haplotype associations of the intron 13 alleles show that each of them is primarily restricted to a single haplotype background. However, it is interesting to note the relatively high proportion of -13907^*G alleles that are observed on haplotypes that differ from **A** (the predominant haplotype, and assumed to be the haplotype upon which -13907^*G originally arose). Although our data are consistent with the allele arising on **A**, as reported by Enattah et al. (2008), who found evidence for extended **A** haplotypes carrying -13907^*G in Middle Eastern populations, our data in the Ethiopian populations show more evidence of disruption of this haplotype, most likely by recombination, and may reflect a longer presence of -13907^*G in Africa.

In this study, the breath hydrogen lactose tolerance testing was the most thoroughly conducted recent survey of

lactose digestion within an African population, with full 3-h breath hydrogen readings obtained for nearly all participants. All were requested to observe an overnight fast and not smoke, and individuals with elevated starting breath hydrogen who might have deviated from these requirements were excluded. We also took care to exclude from the analysis individuals with ambiguous results, though the raw data are presented here. Despite these strict procedures, four lactose digesters carried no variants in the entire sequenced region. These observations suggest the presence of additional genetic changes outside the enhancer region, or modifying factors in the lactase persistence phenotype. Our studies of the Senegalese Wolof population seem to support these findings, as despite having a calculated persistence allele frequency of 0.30 (published lactose digestion frequency = 0.51, representing $2pq + q^2$; Arnold et al. 1980), no intron 13 variation has been identified in this group (Supplementary Table 3).

The findings reported here provide an important example of multiple rare variants being responsible for a common phenotype. It is interesting to speculate on the reasons why very different patterns of genetic diversity in *LCT* are found in Europe. In Europe, selection on *LCT* appears to have been very strong but the reasons for this are still unclear. The calcium assimilation hypothesis, which proposed that the calcium in milk is more advantageous at high latitudes due to reduced incident sunlight and reduced vitamin D synthesis (Flatz and Rotthauwe 1973), was not supported in a recent study (Itan et al. 2009). Itan et al. (2009) propose that the lactase persistence associated allele homogeneity in Europe may be a consequence of underlying demographic processes in addition to strong selection, an interpretation consistent with that of others (Hollox et al. 2001; Gerbault et al. 2009). Demographic constraints are likely to have been different in Africa. Here, human settlement was longer standing, thus preventing population expansion in the same way as took place in Europe.

We propose that the contrasting pattern in Africa may be an example of a ‘soft’ selective sweep, as described by Pennings and Hermisson (2006a, b). Such sweeps can result from high mutation or migration rate, or large effective population size. The migration rate may be of particular importance in pastoralist groups where close regular contact with multiple other communities is to be expected. Pennings and Hermisson (2006a, b) describe soft sweeps involving a low but constant coefficient of selection. Another factor which may also result in allelic diversity is variable selection in time and/or space. The phenomenon of soft sweeps is poorly recognized in human genetics, but is of potential relevance to disease association studies in which independent mutations at the same locus may be involved. However, soft sweeps are more difficult to detect than hard sweeps given the statistical tools

currently at our disposal (Pennings and Hermisson 2006b). Our findings illustrate the clinical value of phenotype/genotype research in multiple different groups across Africa and also the need to develop suitable statistical methods that would recognize multiple causative mutations located in close proximity to each other and having a similar effect, which will be increasingly important with the advent of high-throughput whole genome sequencing.

Acknowledgments We thank all the sample donors who participated in this study. We are grateful to H. Babiker, Elizabeth Caldwell, Matthew Forka, Dominic Gomis, M. Hawary, Steve Jones, Tudor Parfitt, Pat Smith and David Zeitlyn who all contributed to sample collection. We wish to thank Ranji Araseratnam for laboratory assistance, Bryony Jones for allowing us to quote unpublished data and Naser Ansari Pour, Sue Povey and Pascale Gerbault for helpful discussion of the manuscript. We also wish to thank two anonymous reviewers for their thoughtful comments. C.J.E. Ingram was funded by a BBSRC CASE studentship.

Conflicts of interest N.B. is Chairman of The Centre for Genetic Anthropology (TCGA) and an Honorary Lecturer in the research department of Genetics, Evolution and Environment at University College London. He is also joint chairman of the London and City Group of Companies and has extensive business and financial interests including involvement in biotechnology ventures and educational material used by researchers in biomedicine and the life sciences. Nevertheless, he does not have any specific commercial interest in the subject matter of this study. The research has been funded in part by a charitable trust of which N.B. is a trustee. The charitable trust has no intellectual property or other rights whatsoever with respect to the research which forms the subject matter of the paper. All other authors have no conflict of interest.

References

- Andres AM, Clark AG, Shimmin L, Boerwinkle E, Sing CF, Hixson JE (2007) Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet Epidemiol* 31:659–671
- Arnold J, Diop M, Kodjovi M, Rozier J (1980) Lactose intolerance in adults in Senegal. *C R Seances Soc Biol Fil* 174:983–992
- 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 R (1999) Why are there so many pastoral groups in eastern Africa? In: Azarya V, Breedveld A, De Bruijn M, Van Dijk H (eds) *Pastoralists under pressure? Fulbe societies confronting change in west Africa*. Brill Press, Boston, pp 29–49
- 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
- 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, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, Natah A, Ali A, Natah S, Comas D, Mehdi SQ, Groop L, Vestergaard EM, Imtiaz F, Rashed MS, Meyer B, Troelsen J, Peltonen L (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
- Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491
- Excoffier L, Laval LG, Schneider L (2005) Arlequin ver.3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online* 1:47–50
- Flatz G (1987) Genetics of lactose digestion in humans. *Adv Hum Genet* 16:1–77
- Flatz G, Rotthauwe HW (1973) Lactose nutrition and natural selection. *Lancet* 2:76–77
- Freeman B, Smith N, Curtis C, Hockett L, Mill J, Craig IW (2003) DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behav Genet* 33:67–72
- 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
- Gilat T, Ben HH, Gelman-Malachi E, Terdiman R, Peled Y (1978) Alterations of the colonic flora and their effect on the hydrogen breath test. *Gut* 19:602–605
- 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
- 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
- Imtiaz F, Savilahti E, Sarnesto A, Trabzuni D, Al-Kahtani K, Kagevi I, Rashed MS, Meyer BF, Jarvela I (2007) The T/G 13915 variant upstream of the lactase gene (LCT) is the founder allele of lactase persistence in an urban Saudi population. *J Med Genet* 44:e89
- Ingram CJE, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N, Swallow DM (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 (2009) Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet* 124:579–591
- Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The origins of lactase persistence in Europe. *PLoS Comput Biol* 5:e1000491
- Krenke BE, Tereba A, Anderson SJ, Buel E, Culhane S, Finis CJ, Tomsey CS, Zachetti JM, Masibay A, Rabbach DR, Amiott EA, Sprecher CJ (2002) Validation of a 16-locus fluorescent multiplex system. *J Forensic Sci* 47:773–785
- Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci USA* 99:7548–7553
- Lettice LA, Hill AE, Devenney PS, Hill RE (2008) Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Hum Mol Genet* 17:978–985
- 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

- Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, Thomas MG (2004) The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet* 74:1102–1110
- Nei M (1987) *Molecular evolutionary genetics*. Columbia University Press, New York
- Pennings PS, Hermisson J (2006a) Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol* 23:1076–1084
- Pennings PS, Hermisson J (2006b) Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet* 2:e186
- Perman JA, Modler S, Olson AC (1981) Role of pH in production of hydrogen from carbohydrates by colonic bacterial flora. *Studies in vivo and in vitro*. *J Clin Invest* 67:643–650
- 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
- Raymond M, Rousset F (1995a) An exact test for population differentiation. *Evolution* 49:1280–1283
- Raymond M, Rousset F (1995b) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Heredity* 86:248–249
- Sokal RR, Rohlf FJ (1994) *Biometry: the principles and practice of statistics in biological research*, 3rd edn. WH Freeman and Company, New York
- Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978–989
- Swiech L, Kisiel K, Czolowska R, Zientarski M, Borsuk E (2007) Accumulation and dynamics of proteins of the MCM family during mouse oogenesis and the first embryonic cell cycle. *Int J Dev Biol* 51:283–295
- Tag CG, Schifflers MC, Mohnen M, Gressner AM, Weiskirchen R (2007) A novel proximal -13914G>A base replacement in the vicinity of the common-13910T/C lactase gene variation results in an atypical LightCycler melting curve in testing with the MutaREAL lactase test. *Clin Chem* 53:146–148
- Thomas MG, Bradman N, Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum Genet* 105:577–581
- Thomas MG, Weale ME, Jones AL, Richards M, Smith A, Redhead N, Torroni A, Scozzari R, Gratrix F, Tarekegn A, Wilson JF, Capelli C, Bradman N, Goldstein DB (2002) Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am J Hum Genet* 70:1411–1420
- 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 (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
- Vogelsang H, Ferenci P, Frotz S, Meryn S, Gangl A (1988) Acidic colonic microclimate—possible reason for false negative hydrogen breath tests. *Gut* 29:21–26
- 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* 4:657–662