Hum Genet (2007) 120:779–788
DOI 10.1007/s00439-006-0291-1

ORIGINAL INVESTIGATION

A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence?

Catherine J. E. Ingram · Mohamed F. Elamin · Charlotte A. Mulcare · Michael E. Weale · Ayele Tarekegn · Tamiru Oljira Raga · Endashaw Bekele · Farouk M. Elamin · Mark G. Thomas · Neil Bradman · Dallas M. Swallow

Received: 20 September 2006 / Accepted: 25 October 2006 / Published online: 21 November 2006

Abstract Persistence or non-persistence of lactase expression into adult life is a polymorphic trait that has been attributed to a single nucleotide polymorphism (*C-13910T*) in an enhancer element 13.9 kb upstream of the lactase gene (*LCT*). The *-13910*T* allele occurs at very high frequency in northern Europeans as part of a very long haplotype (known as **A**), and promotes binding of the transcription factor Oct-1. However, *-13910*T* is at very low frequency in many African milk drinking pastoralist groups where lactase persistence phenotype has been reported at high frequency. We report here for the first time, a cohort study of lactose digester and non-digester Sudanese volunteers and show there is no association of *-13910*T* or the **A** haplotype with lactase persistence. We support this finding with new genotype/phenotype frequency comparisons in pastoralist groups of eastern African and Middle Eastern origin. Resequencing revealed three new SNPs in close proximity to *-13910*T*, two of which are within the Oct-1 binding site. The most frequent of these (*-13915*G*) is associated with lactose tolerance in the cohort study, providing evidence for a *cis*-acting effect. Despite its location, *-13915*G* abolishes, rather than enhances Oct-1 binding, indicating that this particular interaction is unlikely to be involved in lactase persistence. This study reveals the complexity of this phenotypic polymorphism and highlights the limitations of *C-13910T* as a diagnostic test for lactase persistence status, at least for people with non-European ancestry.

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00439-006-0291-1 and is accessible for authorized users.

C. J. E. Ingram · M. F. Elamin · C. A. Mulcare · D. M. Swallow (✉)
Department of Biology, Galton Laboratory, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK
e-mail: d.swallow@ucl.ac.uk

C. A. Mulcare · M. E. Weale · A. Tarekegn · M. G. Thomas · N. Bradman
TCGA, Department of Biology, University College London, London NW1 2HE, UK

M. E. Weale
Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

M. F. Elamin · F. M. Elamin
Elrazi College of Medical and Health Sciences, P.O. Box 4168, Khartoum, Sudan

A. Tarekegn · T. O. Raga · E. Bekele
Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia

Introduction

Lactase, the intestinal enzyme responsible for digestion of lactose in milk is down-regulated in mammals after weaning. However, in humans lactase persistence (expression of the enzyme in adulthood) is a polymorphic trait. The frequency of lactase persistence has been shown to be moderately well correlated with the cultural habit of drinking fresh milk (Holden and Mace 1997). For example, African pastoralists have been reported to show a higher incidence of lactase persistence than non-pastoralists (Bayoumi et al. 1981, 1982), and lactase persistence is more frequent in Bedouins (who are known to consume appreciable amounts of fresh milk) than in neighbouring non-Bedouin Arabs

(Cook and al-Torki 1975; Snook et al. 1976; Hijazi et al. 1983; Dissanyake et al. 1990).

In Europeans, lactase persistence is controlled by a regulatory element, *cis*-acting to the lactase gene *LCT*, located within a region of strong linkage disequilibrium that is characterised by only three common *LCT* haplotypes, designated **A**, **B** and **C** (Harvey et al. 1998; Hollox et al. 2001). One of these haplotypes, **A,** forms the core of an extended (500 kb+) haplotype, which is associated with lactase persistence and is at high frequency, probably due to the effect of selection for this trait (Poulter et al. 2003; Bersaglieri et al. 2004; Altshuler et al. 2005; Coelho et al. 2005). A single SNP 13.9 kb upstream from *LCT* (*-13910*T*), on the background of the **A** haplotype, located within intron 13 of adjacent gene (*MCM6*) is proposed to be causative of lactase persistence (Enattah et al. 2002). Promoter-reporter construct assays revealed that *-13910*T* has an enhancing effect on transcription (Olds and Sibley 2003; Troelsen et al. 2003). It has been shown more recently that the sequence in which the *-13910*T* allele is located binds the transcriptional enhancer protein Oct-1, while the sequence containing the ancestral *-13910*C* allele binds only poorly (Lewinsky et al. 2005).

The putative causative allele, *-13910*T*, however, although present in a few urban Fulani and Hausa from Cameroon, was absent in samples of several groups in sub-Saharan Africa in which the lactase persistence trait is common (Mulcare et al. 2004). This suggested that either there may be more than one cause of lactase persistence, with a different allele occurring in some parts of Africa, or that *-13910*T* is not in fact functional in relation to lactase persistence in vivo and the true causal mutation is located elsewhere on the extended **A** haplotype.

Our conclusion that *-13910*T* cannot be the worldwide cause of lactase persistence led us to claim that the use of this SNP as a diagnostic tool (Rasinpera et al. 2004) is inappropriate in non-Europeans (Mulcare et al. 2004) but this suggestion proved to be controversial (Kolho and Jarvela 2006; Swallow 2006; Weale 2006).

Here, we report the examination of more milk drinking groups from East Africa and also the Middle East, selected because of the close proximity and long history of migrations between these regions. We included three different Bedouin groups because of the known high incidence of lactase persistence in the Bedouin (Cook and al-Torki 1975; Snook et al. 1976; Hijazi et al. 1983; Dissanyake et al. 1990). Critically, we report for the first time the results of an association study in a cohort of Sudanese volunteers. We selected the Jaali because of the intermediate lactase persistence frequency (53%) published for this group (Bayoumi et al. 1981).

We show unequivocally that neither the *-13910*T* allele nor the **A** haplotype, upon which it resides, account for lactase persistence. Failing to find an association of lactase persistence phenotype and *LCT* core haplotype, we resequenced the −13.9 kb region. This revealed new SNPs in close proximity to *-13910*T*, the functional properties, association with phenotype and distribution of which we examine.

Materials and methods

Samples

DNA samples were extracted from buccal swabs from adult representatives (unrelated at the grandpaternal level) of 15 different groups living in Africa and the Middle East. Ethnic origin was self-declared and recorded on questionnaires.

Five African and three Middle Eastern groups with a present day pastoralist way of life were tested (Murdock 1967; Holden and Mace 1997; Blench 1999). These were Fulani and Shuwa Arabs from Cameroon, Afar and Somali from Ethiopia, one of the pastoralist Beja groups, the Beni Amir, from Sudan, and Bedouins from Israel/area of the Palestinian Authority (PAA), Jordan and Saudi Arabia. Some urban and non-pastoralist neighbours were also tested. These were Amhara from Ethiopia, Mambila from Cameroon, Dounglawi and Shaigi from Sudan and Druze, Israeli urban Arabs and Palestinians from Israel/area of the PAA. Milk drinking information was also collected from the Afar from Asayita, Ethiopia; Somali from Jijiga, Ethiopia; Fulani from the North and Extreme North Provinces, Cameroon and Beni Amir from Port Sudan, Sudan.

Samples were also collected from a cohort of 99 Jaali, a Sudanese group with part Arab ancestry, living in Shendi ~150 km north of Khartoum, with previously reported intermediate lactase persistence frequency (Bayoumi et al. 1981). Donors from this group who were unrelated at the grandparental level were phenotyped for lactase persistence status using the breath hydrogen method (see below) and were asked to answer questions on a milk drinking questionnaire. One Ethiopian individual of self-declared Amharic ancestry was also lactose tolerance tested using the same method.

Approvals for this study were obtained from UCLH (99/0196 and 01/0236) and from the Federal Ministry of

Health, Republic of Sudan, the Faculty of Medicine, Addis Ababa University and the Ministry of Scientific and Technical Research, Yaoundé, Cameroon.

Breath hydrogen lactose tolerance testing

Lactase persistence status was assessed by lactose tolerance testing. Fully informed consenting volunteers who fasted overnight and refrained from smoking were given 50 g lactose as a solution in 250 ml water brought to room temperature. Lactose tolerance was tested by the breath hydrogen method using a MicroH_2 meter (Peuhkuri et al. 1998) to measure breath hydrogen at 30 min intervals. Reports of any recent intestinal complaints and antibiotic treatment were recorded. To be included individuals were required to have a clearly detectable starting breath hydrogen (preferably above 2 ppm) to exclude potential hydrogen non-producers, and not above 20 ppm (to exclude individuals with potential bacterial overgrowth, or who had not fasted, or had smoked, or had eaten an excess of fermentable carbohydrate such as beans). Forty-nine individuals who showed a clear rise in breath hydrogen of over 20 ppm above baseline, during the first 2 h after the lactose load, were diagnosed as lactose non-digesters and having not declared antibiotic treatment or relevant gastrointestinal complaint were deemed most likely to be lactase non-persistent, and 45 showing no rise at all, or a rise of less than 11 ppm during the same period were diagnosed as digesters and most likely to be lactase persistent. In one case the starting breath hydrogen was 1 ppm (see above) with a transient increase to 16 ppm, and in another four cases the test gave borderline results, and it was not possible to detain the volunteers for long enough to confirm whether or not the hydrogen level would continue to rise or be maintained. These five individuals were excluded from the study.

Polymorphism typing

The *C-13910T* was typed using LAC-CM-U and LAC-CL2 primers to amplify the region containing the C/T polymorphism and *HinfI* digestion was used to discriminate between the C or T alleles as described previously (Mulcare et al. 2004). PCR-RFLP was also used to type the *A-678G, G666A* and *T5579C* polymorphisms (which define the core *LCT* haplotypes, **A**, **B** and **C**) (Hollox et al. 2001) (see supplementary data Table 1). To distinguish the **U** haplotype, *TC-942/3ΔΔ* (Hollox et al. 2001) was typed using a tetra–primer ARMS-PCR method (Ye et al. 2001) (see supplementary data, Table 2). These four SNPs were selected from the 11 described in Hollox et al. (2001), to efficiently capture the haplotype diversity detectable in other sub-Saharan Africans (Hollox et al. 2001). Note that the use of four SNPs rather than all 11 will result in grouping of some of the minor haplotypes.

All genotyping assays were conducted with positive and negative controls and typing confirmed by two investigators. Two hundred and thirty-three samples were typed both by PCR-RFLP and sequencing and in all but one case typings agreed.

Sequencing

A 700 or 400 bp fragment spanning the −13.9 kb region was PCR amplified with primers MCM6i13 and MCM778 or LAC-CL2 (see supplementary data Table 3). PCR product was purified by PEG precipitation and samples were sequenced using the MCM6i13 primer and Big-Dye Terminator chemistry (Applied Biosystems, Foster City, CA, USA) on an ABI 3100 DNA analyser.

Statistical methods and haplotype inference

Haplotypes were inferred using both an expectation maximisation algorithm (Arlequin) (Excoffier et al. 2005) and a Bayesian method (PHASE) (Stephens et al. 2001) and the case/control feature of PHASE was used to compare haplotype distribution in the lactose digester and non-digester Jaali. The program GenoPheno (Mulcare et al. 2004) (available at http://www.ucl.ac.uk/tcga/software/) was used to compare the predicted lactose digester frequency based on the occurrence of candidate causative DNA variants, with the phenotypic frequencies for published matched groups. This program takes into account sampling and phenotyping error rates. Populations were accepted as 'matches' if they were the same ethnic group living in the same country or an immediately neighbouring country, and these are referenced in Tables 1 and 4, and supplementary Table 4.

Electrophoretic mobility shift assay (EMSA)

Nuclear protein enriched extracts were prepared (Hollox et al. 1999) from differentiated Caco-2 cells (as used previously by Lewinsky et al. 2005, and the only reported human cell line to express lactase). Electrophoretic mobility shift assays (EMSAs) were performed by adding approximately 8 μg nuclear extract plus 2 μg poly[d(I-C)] (with or without 100 pmols

unlabelled competitor probe) to a 10 µl total reaction volume (20 mM HEPES pH 7.6, 1 mM EDTA, 10 mM $(NH_4)SO_4$, 1 mM DTT, 0.2% (v/v) Tween-20, 30 mM KCl). Following 15 min pre-incubation on ice, 10 fmol of ^{32}P-labelled oligonucleotide was added and incubated for a further 35 min.

Samples were separated on a 5% non-denaturing polyacrylamide (TBE) gel at room temperature, 200 v for 1 h 45 min. After electrophoresis the gels were dried and exposed to Fuji Super HR-E 30 X-ray film. The following double stranded oligonucleotides (designed with 5′ overhangs) were used as probes and competitors (upper strand only shown):

13910C (ANCESTRAL):
5′-AAGATAATGTAGCCCCTGGCCTCAA-3′
13910T:
5′-AAGATAATGTAGTCCCTGGCCTCAA-3′
13915G:
5′-AAGATAAGGTAGCCCCTGGCCTCAA-3′
13913C:
5′-AAGATAATGCAGCCCCTGGCCTCAA-3′
13907G:
5′-AAGATAATGTAGCCCGTGGCCTCAA-3′
Oct-1 classic:
5′-ATGTCGAATGCAAATCACTAGAACT-3′
Non-specific:
5′-AACTCCGGTCCCCGATGTAATAGAA-3′
TAATGARAT:
5′-TCGTCGTATCTCATTACCGCCGTCG-3′

Results

*13910*T* and *LCT* core haplotype frequencies in the Middle East and Africa

The previously published restriction enzyme assay for *-13910*T* (Mulcare et al. 2004) was used to type three new African and five Middle Eastern groups for whom lactase persistence frequencies were well documented. The *-13910*T* was shown to be very rare in the East African and Middle Eastern groups (Table 1) and failed to explain reported lactase persistence in any of those with inferred lactase persistence allele frequency above 0.2 (max $P = 1.0 \times 10^{-4}$, using GenoPheno matched group comparison). In contrast we confirmed that this allele was frequent (0.39) in the pastoralist Fulani group from Cameroon.

Lactase gene core haplotype markers [*TC-942/3ΔΔ*, *A-678G*, *G666A* and *T5579C* (9)], were tested in the same groups. In the Fulani, as in Europeans, *-13910*T* is strongly associated with the **A** haplotype ($P = 1.73 \times 10^{-19}$ for a Fishers Exact test of a 2 × 2 table of allele counts), and in the two Sudanese and four Middle Eastern individuals who carry *-13910*T* it also appears to occur on an **A** haplotype background.

The core haplotypes show considerable inter group difference in frequency (Table 1). The **A** haplotype is widespread, but is not at all correlated in frequency, across these newly tested groups, with published

Table 1 Frequency of *-13910*T* and *A* haplotype frequency in eight new population groups in comparison with lactase persistence allele frequency (*LAC*P*) calculated from published lactose digester frequency, with two European groups included as controls

Population	Reported Lac*P allele frequency of matched group	*n*	-13910*T frequency	*P*-value	*n*	Haplotype frequency			
						A	B	C	Other
Beni Amir	0.641[a]	100	0.005	<0.0001	100	0.210	0.195	0.325	0.270
Saudi Bedouin	0.592[a]	56	0.000	<0.0001	46	0.109	0.239	0.554	0.098
Fulani	0.533[a]	91	0.390	0.4828	75	0.493	0.200	0.040	0.267
French CEPH[d]	0.524[b]	24	0.458	0.4126	24	0.583	0.271	0.063	0.083
Jordanian Bedouin	0.515[a]	26	0.058	<0.0001	23	0.283	0.261	0.413	0.043
Israeli Bedouin	0.515[a]	19	0.026	<0.0001	21	0.214	0.310	0.262	0.214
Jaali	0.316[c]	96	0.005	<0.0001	93	0.119	0.313	0.273	0.295
Italian-south[e]	0.187[a]	29	0.103	0.3406	29	0.367	0.333	0.217	0.083
Israeli urban Arab	0.138[a]	83	0.024	<0.0001	136	0.272	0.515	0.184	0.029
Palestinian Arab	0.138[a]	19	0.026	0.8110	19	0.316	0.368	0.211	0.105

No deviation from Hardy–Weinberg equilibrium was observed in any of the populations. The program GenoPheno was used to compare the expected lactose digester frequency, if carrying *-13910*T* was causative, with the published (observed) phenotypic frequencies for matched populations. Populations are listed in order of reported lactase persistence allele frequency. References for matched groups (superscripts) are given below

[a] Numbers taken from Holden and Mace (1997) who extracted adult data. For source references see Supplementary Table 4

[b] Cuddenec et al. (1982)

[c] Bayoumi et al. (1981)

[d] DNA kindly supplied by Centre d'Etude du Polymorphisme Humain (CEPH) http://www.cephb.fr/

[e] DNA kindly supplied by S. Auricchio, M. Rossi, samples included in Harvey et al. (1998)

lactase persistence allele frequencies. The **C** haplotype was noted to be most frequent in three of six groups where lactase persistence is reported to be high (Table 1).

Phenotyped cohort study

Of the 94 individuals for whom clear lactose tolerance test results were obtained 48% were lactose digesters, in close agreement with previously published frequencies for this ethnic group (Bayoumi et al. 1981). Lactose digester status and milk consumption behaviour were highly associated ($P = 5.84 \times 10^{-6}$, Table 2). The *-13910*T* was found in only one person in this cohort, who was a lactose digester (presumed lactase persistent). This clearly shows that it is not the sole causal variant in this group. None of the 'core' *LCT* SNPs showed significant association with lactose digestion status (min $P = 0.12$, for Fishers Exact tests of 2×2 tables of allele counts). Core haplotype distributions are shown in Fig. 1. Although the **C** haplotype is more frequent in the lactose digester group, the difference in core haplotype distribution between the two groups was also not significant ($P = 0.73$, PHASE, case-control comparison), and no one haplotype was frequent enough to carry a single allele causal of lactase persistence.

Questioning whether another nucleotide change could be located in the vicinity of *-13910*T*, we resequenced the −13.9 kb region. Several variant alleles were revealed following sequencing of the region in the Jaali cohort and in a single sample from a lactase persistent Ethiopian. Three of these were novel and very close to *-13910*T* (*-13915*G*, *-13913*C* and *-13907*G*) and two (*-13915*G*, *-13913*C*) overlap the previously reported Oct-1 binding site (Fig. 2). The *-13907*G* is just three nucleotides outside and was present in the lactase persistent Ethiopian.

The most common of these new sequence variants in the Jaali, *-13915*G*, is present at a frequency of 0.14.

Table 2 Contingency table showing numbers of lactose digester and non-digester people reporting consumption of more than 500 ml milk per day

Milk drinking	Phenotype		
	Non-digester	Digester	Total
No	41	17	58
Yes	8	28	36
			94

Five individuals were excluded from the analysis due to ambiguous lactose tolerance test results. Note that the numbers are slightly larger than in subsequent tables because individuals are included from whom no (or poor quality) DNA was obtained

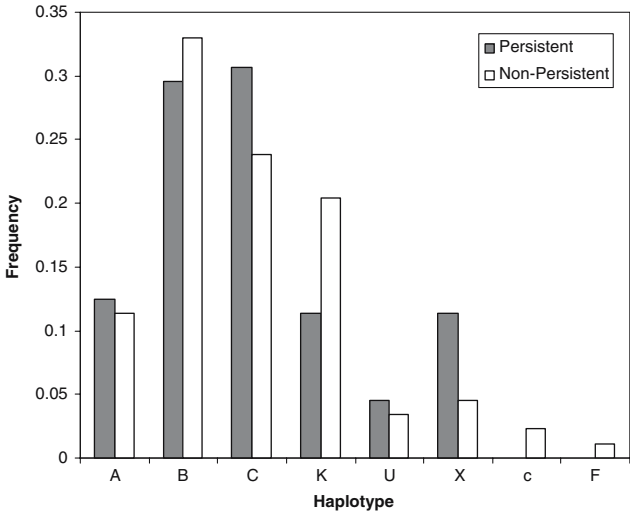

Fig. 1 Comparison of the core lactase haplotype frequencies in lactose digester (persistent) and non-digester (non-persistent) Jaali. The difference in distribution between the two groups is not statistically significant ($P = 0.73$, PHASE). Haplotype nomenclature taken from Hollox et al. (2001). The designations for these four SNP haplotypes are those of the most frequent 11 SNP haplotypes in sub-Saharan Africans. Note that lower case **c** is quite distinct from upper case **C**

Genotypic data for this SNP are shown in Table 3. The *-13915*G* allele shows a significant association with lactose digester status ($P = 6.05 \times 10^{-3}$, for a Fishers Exact test of a 2×2 table of allele counts). However, the association was not 100%, and there were discrepancies in both directions. Five lactose non-digester individuals carried the *-13915*G* allele, and this included one homozygote. None of these five reported any recent gastro-intestinal complaints when questioned, and four of five reported drinking less than 500 ml milk per day. Furthermore, the *-13915*G* allele was not frequent enough to explain all the lactase persistence observed in this population; 23/39 lactose digesters did not carry the allele (Table 3). The *-13910*T*, *-13913*C* and *-13907*G* were each present in a single lactose tolerant individual, leaving 20 individuals with the ancestral allele in this region. Twelve of these 20 reported drinking 0.5 l or more of fresh milk per day.

```
ANCESTRAL    TGGCAATACAGATAAGATAATGTAGCCCCTGGCCTCAAAGGAACTCTCC
-13915*G     TGGCAATACAGATAAGATAAGGTAGCCCCTGGCCTCAAAGGAACTCTCC
-13913*C     TGGCAATACAGATAAGATAATGCAGCCCCTGGCCTCAAAGGAACTCTCC
-13910*T     TGGCAATACAGATAAGATAATGTAGTCCCTGGCCTCAAAGGAACTCTCC
-13907*G     TGGCAATACAGATAAGATAATGTAGCCCGTGGCCTCAAAGGAACTCTCC
```

Fig. 2 Sequence comparisons of the ancestral and variant sequences within *MCM6* intron 13 at −13.9 kb upstream of *LCT*. The Oct-1 binding site (determined by TRANSFAC analysis, Lewinsky et al. 2005) is *shaded*, and the variant alleles are shown in *bold underline*

Table 3 Contingency table showing numbers of each *-13915*G* genotype in the lactase digester and lactose non-digester categories

Genotype	Phenotype		
	Non-digester	Digester	Total
GG	1	1	2
TG	4	15	19
TT	39	23	62
			83

Data obtained by sequencing. Five genotyped individuals were excluded from the analysis due to ambiguous lactose tolerance test results

Population distribution and haplotype background of the new variant alleles

We sequenced the region around *-13910*T* in a further 434 individuals from various Middle Eastern and African groups, including pastoralists and non-pastoralists from each location, in order to better understand the distribution of the new alleles (Table 4). Of the populations tested, the *-13915*G* allele was found to be fairly widespread in eastern Africa and the Middle East. It was most common in the Saudi Bedouins and appeared to be more common in the milk-drinking pastoralists, though the frequencies are in most cases significantly lower than those published for lactose tolerance. This is also the case if it is assumed that any variant allele at this locus is causative of lactase persistence (Table 4). Only one individual was found to carry two different variant alleles in the −13.9 kb region. The majority of the Beni Amir, Afar and Somali tested reported drinking more than 500 ml fresh milk per day and there was no association of milk drinking with carrier status for one, or any one, of the variant alleles (data not shown).

The newly observed *-13915*G* allele was found to occur exclusively on a **C**-haplotype background in the Jaali, and only 10 of 131 *-13915*G* alleles were found on non-**C** haplotypes in the other populations tested; three on an **E** (an **A**/**C** recombinant) haplotype background, four on a **K** haplotype, and three on different, rare haplotypes. These non **C** *-13915*G* carrying chromosomes were found in Beni Amir ($n = 3$), Saudi Arabian Bedouin ($n = 2$), Israeli urban ($n = 1$) and Bedouin Arabs ($n = 2$) and Jordanian Bedouin ($n = 2$) groups. The other new SNPs (*-13907*G* and *-13913*C*) are not associated with the **C** haplotype, and were rare in the populations genotyped thus far, but from the

Table 4 Allele frequencies of new *MCM6* intron 13 polymorphisms

Country/Area	Population group	Pastoralists[c]	*n*	Allele frequency					Published *LAC*P* allele frequency	*P*-value
				*-13915*G*	*-13913*C*	*-13910*T*	*-13907*G*	Any allele		
Israel/PAA	Urban Arabs	N	81	0.049	0.000	0.025	0.000	0.074	0.13[a]	**0.0030**
Israel/PAA	Druze	N	14	**0.107**	0.000	0.036	0.000	**0.143**	–	–
Israel/PAA	Bedouin	Y	19	**0.132**	0.000	0.026	0.000	**0.158**	0.51[a]	**0.0001**
Israel/PAA	Palestinians	N	18	0.000	0.000	0.028	0.000	0.028	0.13[a]	0.2791
Saudi Arabia	Bedouin	Y	46	**0.489**	0.011	0.000	0.000	**0.500**	0.59[a]	0.2507
Jordan	Bedouin	Y	23	**0.391**	0.000	0.065	0.000	**0.457**	0.51[a]	0.4073
Sudan	Beni Amir[d]	Y	82	**0.244**	0.000	0.006	0.061	**0.311**	0.64[a]	**0.0001**
Sudan	Shaigi	N	9	0.056	0.000	0.000	0.000	0.056	0.21[b]	0.2805
Sudan	Dounglawi	N	6	0.000	0.000	0.000	0.083	0.083	0.10[b]	0.7670
Sudan	Jaali[d]	N	88	**0.142**	0.006	0.006	0.006	**0.159**	0.31[b]	**0.0032**
Ethiopia	Amharic	N	19	**0.132**	0.000	0.000	0.053	**0.184**	–	–
Ethiopia	Afar[d]	Y	10	**0.150**	0.000	0.000	**0.200**	**0.350**	–	–
Ethiopia	Somali[d]	Y	9	0.056	0.056	0.000	0.056	**0.167**	–	–
Cameroon	Mambila	N	38	0.000	0.000	0.000	0.000	0.000	–	–
Cameroon	Fulani[d]	Y	63	0.000	0.024	**0.389**	0.000	**0.413**	0.53[a]	0.6237
Cameroon	Shuwa Arabs	Y	16	0.063	0.000	0.000	0.000	0.063	–	–

Allele frequencies above 0.10 are shown in bold. No deviation from Hardy–Weinberg equilibrium was observed in any of the populations. Lactase persistence allele frequency (*LAC*P*) calculated from published lactose digester frequencies, is shown for matched groups for which published data are available. References (superscripts) are given below. The program GenoPheno was used to compare the expected lactose digester frequency, if carrying any one of the non-ancestral variants at −13.9 kb was causative, with the published (observed) phenotypic frequencies for matched populations. In four cases there were significantly fewer expected digesters than observed in matched groups in the literature (bold)

[a] Numbers taken from Holden and Mace (1997), for source references see Supplementary Table 4

[b] Bayoumi et al. (1981)

[c] Pastoralist taken from Blench (1999), Murdock (1967) and Holden and Mace (1997)

[d] Milk drinking data available

preliminary data it appears that *-13913*C* is associated with **B** haplotype and *-13907*G* with **A**.

Oct-1 binding affinity for the new sequence variants

Gel shift analysis confirmed binding of the *-13910*T* oligonucleotide to the same protein as bound by the Oct-1 classic and TAATGARAT probes (Fig. 3), which was inferred from the study of Lewinsky et al. (2005) to be Oct-1. Specificity of the interactions was demonstrated by adding competitor probes of either unspecific or Oct-1 classic unlabelled probes. Binding was not affected by addition of unspecific probe, but addition of unlabelled Oct-1 classic probe completely displaced binding of the labelled *-13910*T* probe. Binding of the *-13913*C* and *-13915*G* sequence variants to the protein was not detected and the ancestral and *-13907*G* variants showed only very faint binding after long exposures (not shown). In reciprocal experiments the *-13910*T* oligonucleotide was shown to be the only variant to displace the Oct-1 oligonucleotide (data not shown).

Discussion

This study provides new insight into the origins of lactase persistence in human populations. We provide direct evidence that *-13910*T* is not the worldwide cause of lactase persistence. In a cohort study, 45 volunteers were phenotyped as lactose tolerant but only one of these people carried the *-13910*T* allele, demonstrating it cannot account for lactase persistence in this group. We also found no overrepresentation of the **A** haplotype in the persistent members of the cohort, ruling out the hypothesis that a haplotype similar to the extended European **A** haplotype but lacking the marker *-13910*T*, carries a common cause of lactase persistence. We also show that neither the *-13910*T* allele nor the **A** haplotype are frequent enough to account for lactase persistence in other East African and Middle Eastern milk drinking groups. Where it was present, *-13910*T* was most often (as in Europe) on an **A** haplotype background suggesting a common origin for this chromosomal segment.

There was also no statistically significant difference in core *LCT* haplotype distribution between the lactose tolerant and intolerant Jaali, although we note that the **C** haplotype was slightly more frequent in the digester than the lactose non-digester group. These data pointed to another cause for lactase persistence in Africa and the Middle East, and even to the possibility that this may be trans-acting to *LCT*. However, the lack of association between lactose digester status and the *LCT* core haplotypes in the cohort study could in part reflect the shorter haplotype blocks seen in Africans or the somewhat reduced power resulting from errors in phenotyping, as discussed in detail below.

The discovery of new SNPs in the vicinity of the Oct-1 binding site, one of which is significantly associated with lactase persistence, does suggest that lactose tolerance is at least influenced by *cis*-acting variants in these populations, and pointed to a potential functional role, at least for *-13915*G*. However, despite the location of this SNP in the Oct-1 binding site, the agreement with phenotype was not as tight as one would have expected for a causal SNP.

It is well known that lactose tolerance testing has an error rate by whichever method is used, and for breath hydrogen testing (which is considered most accurate) this has been reported as about 5% false positive (i.e. people found to be lactose intolerant by this, but not by other methods) and 8% false negative (Mulcare et al. 2004). The identification of 23/39 individuals in the lactose digester group who do not carry *-13915*G* (and 20/39 who carry no variant allele) suggests an implausible false negative error rate of about 50%, if *-13915*G* is causal. There is also a high-false positive rate (5/44). However, it is just possible that some or all of the five individuals who carry *-13915*G*, but are lactose non-digesters, were suffering from undiagnosed or undisclosed intestinal illness, which caused them to have secondary lactose intolerance. In the context of a field study it was not possible to do more than question the

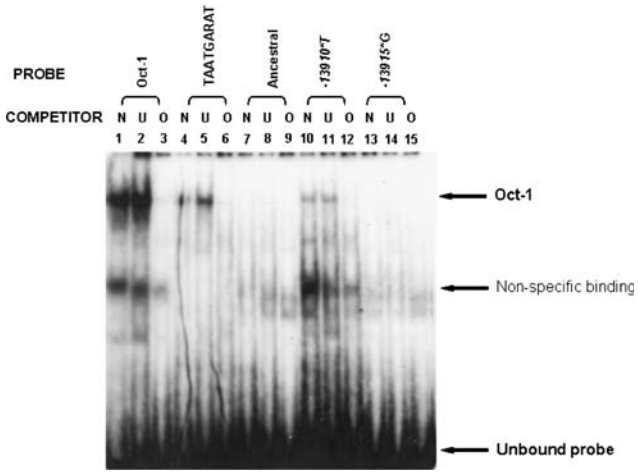

Fig. 3 Electrophoretic Mobility Shift Assay (*EMSA*) of −13.9 kb sequence variants ^{32}P-labelled probes were incubated with nuclear protein extract. Unlabelled competitor probes were pre-incubated with the protein extract to demonstrate binding specificity. Competitor is indicated above each lane; none (*N*), unspecific (*U*) and Oct-1 (*O*). N.B. *Vertical lines* of radioactivity in lanes 4 and 6 are artefacts of the gel drying process

volunteers, and further lactose tolerance tests and other examinations were not feasible. Consequently, we cannot exclude the possibility that *-13915*G* is one of several causes of lactase persistence in this group of people, along with other variants that are yet to be identified. With this in mind it was important to ask whether this allele, like *-13910*T*, promotes binding to Oct-1.

Analysis of Oct-1 binding, however, shows that in vitro neither *-13915*G* nor any of the other alleles resemble *-13910*T* in their effect on strength of binding to the protein we infer to be Oct-1. It is relevant to note that we found the affinity of this protein for the *-13910*T* containing sequence (TAATGTAG**T**) was similar to that seen with the TAATGARAT probe and significantly weaker than for the classic Oct-1 motif (ATGCAAAT), and from sequence inspection it seems possible that Oct-1 binds in its (OCTA-)TA ATGARAT conformation (Cleary and Herr 1995). If this is the case, the SNP (*-13915*G*) that associates with lactase persistence would *be expected* to disrupt the Oct-1 binding site, since it converts a critical TAA**T** to TAA**G** of the Pou_H binding domain (Verrijzer et al. 1992). Thus, the identification of these new SNPs in the immediate vicinity of *-13910*T* casts serious doubt on the physiological role of this Oct-1 binding site in the expression of lactase, although this site may be functional in relation to lactase persistence in some other way.

If on the other hand the −13.9 kb SNPs that are associated with phenotype (*-13910*T* and *-13915*G)* are not functional, one possibility is that they reside in a mutation hotspot (suggested by the occurrence of four SNPs in such close proximity) and that in different populations *-13910*T* and *-13915*G* happen to be associated with the true causal mutations. Interestingly, a fifth SNP immediately adjacent to *-13907*G* has been identified in the NIH Polymorphism Discovery Resource Panel (rs4988236, *C-13908*T*) (Entrez SNP database, NCBI, http://www.ncbi.nih.gov/).

The *-13915*G* allele has not been reported so far in Europeans (Enattah et al. 2002; Mulcare et al. 2004) (C. J. E. Ingram, unpublished data), and has not been reported in any of the public databases. Our preliminary studies on its distribution suggest that this allele may have originated in the Middle East, where it is seen at highest frequency in Bedouin groups. It is frequent in East Africa but hardly found in West Africa. It is possible that it was introduced in the last 1,400 years as a result of the Arab expansion, which accompanied the spread of Islam. This is consistent with the presence of the allele in the Beni Amir and Jaali of eastern Sudan, as well as the Shuwa Arabs of Cameroon, all of whom are Muslim and claim some Arab ancestry (Robinson 1927; Warburg 1978; Levy and Holl 2002; Vanhove 2006). However, if it is a marker of the principle cause of lactase persistence in Sudan, a longer history seems more likely since there is linguistic and archaeological evidence for herding of cows in the Nile valley and eastern Africa at least 4–5,000 years ago (Ehret 1979; Smith 1992). It is noteworthy that the frequency of *-13915*G* is low in the Israeli urban Arabs and Palestinians. The sharp differences in allele frequency in the Middle Eastern groups may reflect genetic drift magnified by the endogamous nature of these communities, or selection for lactase persistence in the nomadic groups.

The *-13907*G* is widespread in Sudan and Ethiopia, being at highest frequency in the Afar (Table 4 and C. J. E. Ingram, unpublished data). It was not detected in the Middle East, suggesting a different origin for this allele. It would be of interest to estimate a date of origin for both of these SNPs using microsatellite markers, as has been done for *-13910*T* (Coelho et al. 2005). The rarest of the new alleles, *-13913*C*, is seen at highest frequency (in the populations tested thus far) in the Fulani, who also carry the putative causative SNP *-13910*T* at high frequency.

This study demonstrates the value of resequencing, which allowed discovery of new, possibly recent, SNPs that are frequent in East Africans and even some West Africans, but which have not been reported in the public databases, and which enabled us to show an association with phenotype.

The results presented here provide an important starting point to better define the haplotypes which carry the *-13915*G* allele and the other new variants as well as the haplotype blocks that associate with phenotype in this cohort and in other African groups. Work is in progress to attempt to demonstrate directly that the regulation of expression is in fact *cis*-acting in Africans. Because of the strong possibility of multiple genetic and even epigenetic (Maiuri et al. 1991) causes of lactase persistence and the inherent error rate in the phenotypic test it will be important in the collection and testing of further cohorts to pay particular attention to the size of the study group to obtain adequate power, and to consider combining putative causative alleles (or their absence) in the analysis.

It is tempting to speculate that a phenotype that may have had a great selective advantage is likely to have multiple causes. However, intuitively it seems more likely that several different mutations would result in disruption, rather than gain of function, as attributed to *-13910*T*.

The work described here demonstrates directly that using *C-13910T* as a diagnostic test for lactase persistence status is inappropriate for people of East African or Arabian ancestry.

Acknowledgments We thank Steve Jones, Tudor Parfitt, H. Babiker, Pat Smith, David Zeitlyn, Matthew Forka and Elizabeth Caldwell for help with collection of samples, and Ranji Araseratnam, Abigail Jones and many undergraduate students, in particular Naser Ansari Pour, Fiona Pring and Rhonda Sturley for preparing the DNA and/or testing for *LCT* markers. C. J. E. Ingram was funded by a BBSRC CASE studentship.

References

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320

Bayoumi RA, Flatz SD, Kuhnau W, Flatz G (1982) Beja and Nilotes: nomadic pastoralist groups in the Sudan with opposite distributions of the adult lactase phenotypes. Am J Phys Anthropol 58:173–178

Bayoumi RA, Saha N, Salih AS, Bakkar AE, Flatz G (1981) Distribution of the lactase phenotypes in the population of the Democratic Republic of the Sudan. Hum Genet 57:279–281

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, MA, USA

Cleary MA, Herr W (1995) Mechanisms for flexibility in DNA sequence recognition and VP16-induced complex formation by the Oct-1 POU domain. Mol Cell Biol 15:2090–2100

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

Cook GC, al-Torki MT (1975) High intestinal lactase concentrations in adult Arabs in Saudi Arabia. Br Med J 3:135–136

Cuddenec Y, Delbruck H, Flatz G (1982) Distribution of the adult lactase phenotypes—lactose absorber and malabsorber—in a group of 131 army recruits. Gastroenterol Clin Biol 6:776–779

Dissanyake AS, El-Munshid HA, Al-Qurain A (1990) Prevalence of primary adult lactose malabsorption in the eastern provine of Saudi Arabia. Ann Saudi Med 10:598–601

Ehret C (1979) Antiquity of Agriculture in Ethiopia. J Afr Hist 20:161–177

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

Excoffier, Laval LG, Schneider L (2005) Arlequin Version 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

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

Hijazi SS, Abulaban A, Ammarin Z, Flatz G (1983) Distribution of adult lactase phenotypes in Bedouins and in urban and agricultural populations of Jordan. Trop Geogr Med 35:157–161

Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. Hum Biol 69:605–628

Hollox EJ, Poulter M, Wang Y, Krause A, Swallow DM (1999) Common polymorphism in a highly variable region upstream of the human lactase gene affects DNA-protein interactions. Eur J Hum Genet 7:791–800

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

Kolho KL, Jarvela I (2006) DNA test for hypolactasia premature: authors' reply. Gut 55:131–132

Levy TE, Holl AFC (2002) Migrations, ethnogenesis, and settlement dynamics: Israelites in iron age Canaan and Shuwa-Arabs in the Chad Basin. J Anthropol Archaeol 21:83–118

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

Maiuri L, Raia V, Potter J, Swallow D, Ho MW, Fiocca R, Finzi G, Cornaggia M, Capella C, Quaroni A, Auricchio S (1991) Mosaic pattern of lactase expression by villous enterocytes in human adult-type hypolactasia. Gastroenterology 100:359–369

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

Murdock G (1967) Ethnographic atlas. University of Pittsburg Press, Pittsburgh, PA, USA

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

Peuhkuri K, Poussa T, Korpela R (1998) Comparison of a portable breath hydrogen analyser (Micro H2) with a Quintron MicroLyzer in measuring lactose maldigestion, and the evaluation of a Micro H2 for diagnosing hypolactasia. Scand J Clin Lab Invest 58:217–224

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

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

Robinson AE (1927) Notes on the gamuia tribe, Sudan. J R Afr Soc 26:138–144

Smith AB (1992) Origins and spread of pastoralism in Africa. Ann Rev Anthropol 21:125–141

Snook CR, Mahmoud JN, Chang WP (1976) Lactose tolerance in adult Jordanian Arabs. Trop Geogr Med 28:333–335

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Swallow DM (2006) DNA test for hypolactasia premature. Gut 55:131–132

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

Vanhove M (2006) The Beja language today in Sudan: the state of the art in linguistics. In: Proceedings of the 7th International Sudan Studies Conference, Bergen Norway, CD Rom, University of Bergen, Bergen

Verrijzer CP, Alkema MJ, van Weperen WW, Van Leeuwen HC, Strating MJ, van der Vliet (1992) The DNA binding specificity of the bipartite POU domain and its subdomains. EMBO J 11:4993–5003

Warburg GR (1978) Islam, nationalism and communism in a traditional society: the case of Sudan. Frank Cass & Co, London, UK

Weale ME (2006) DNA test for hypolactasia premature. Gut 55:131–132

Ye S, Dhillon S, Ke X, Collins AR, Day IN (2001) An efficient procedure for genotyping single nucleotide polymorphisms. Nucleic Acids Res 29:E88