European Journal of Human Genetics (1999) 7, 791–800

http://www.stockton-press.co.uk/ejhg

ARTICLE

Common polymorphism in a highly variable region upstream of the human lactase gene affects DNA-protein interactions

Edward J Hollox[1], Mark Poulter[1], Yangxi Wang[1,3], Amanda Krause[2] and Dallas M Swallow[1]

[1]*MRC Human Biochemical Genetics Unit, University College London, UK*
[2]*Department of Human Genetics, South African Institute for Medical Research and University of the Witwatersrand, Johannesburg, Republic of South Africa*

In most mammals lactase activity declines after weaning when lactose is no longer part of the diet, but in many humans lactase activity persists into adult life. The difference responsible for this phenotypic polymorphism has been shown to be *cis*-acting to the lactase gene. The causal sequence difference has not been found so far, but a number of polymorphic sites have been found within and near to the lactase gene. We have shown previously that in Europeans there are two polymorphic sites in a small region between 974 bp and 852 bp upstream from the start of transcription, which are detectable by denaturing gradient gel electrophoresis (DGGE). In this study, analysis of individuals from five other population groups by the same DGGE method reveals four new alleles resulting from three additional nucleotide changes within this very small region. Analysis of sequence in four primate species and comparison with the published pig sequence shows that the overall sequence of this highly variable human region is conserved in pigs as well as primates, and that it lies within a 1 kb region which has been shown to control lactase downregulation in pigs. Electrophoretic mobility shift assay (EMSA) studies were carried out to determine whether common variation affected protein-DNA binding and several binding activities were found using this technique. A novel two base-pair deletion that is common in most populations tested, but is not present in Europeans, caused no change in binding activity. However, a previously published C to T transition at –958 bp dramatically reduced binding activity, although the functional significance of this is not clear.

Keywords: lactase; polymorphism; nuclear protein binding; denaturing gradient gel electrophoresis; primate

Correspondence: DM Swallow, MRC Human Biochemical Genetics Unit, UCL, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. Tel: (+44) 171 504 5040; Fax: (+44) 171 387 3496; E-mail: dswallow@hgmp.mrc.ac.uk
[3]Current address: Harvard Institute of Medicine, Boston, MA, USA
Received 10 March 1999; revised 6 May 1999; accepted 17 May 1999

Introduction

The intestinal enzyme lactase is responsible for the digestion of lactose, which is the main carbohydrate in milk. In most populations lactase non-persistence is the most common phenotype, and within Europe frequencies of lactase persistence decline from north to south and from west to east (for review see Flatz[1]). The persistence or non-persistence of lactase activity is genetically determined and persistence behaves as a dominant trait in families as judged by lactose tolerance tests (for review see Swallow and Harvey[2]). The difference responsible for this phenotypic polymorphism has been shown to be *cis*-acting to the lactase gene (LCT),[3] but no causal sequence changes within the gene have been identified.

A *cis*-acting element, CE-LPH1, initially identified in the pig, 40 bp upstream from start of transcription, binds an intestine-specific factor NF-LPH.[4] Two proteins, which bind to this element, have been identified as the homeobox proteins Cdx-2 and HOXC-11[5] and transfection studies using transgenic mice showed that the 1 kb immediately upstream of pig lactase controls post-weaning downregulation.[6] The homologous region in humans is disrupted by two tail-to-tail Alu elements. Alignment of the human and pig sequences excluding the human Alu elements shows that sequence similar to that in the pig promoter region exists in humans, but sequencing of the homologous region in humans has revealed no differences which are totally associated with the lactase persistence/non-persistence phenotype (Poulter M, in preparation). However, allelic variation within this region may contribute to the phenotypic polymorphism.

Seven previously studied polymorphic sites spanning approximately 70 kb across the lactase gene form three common haplotypes: A, B and C; lactase persistence is associated with the A haplotype in Caucasians.[7–9] Two of the seven polymorphic sites are within a small region 974 bp to 852 bp upstream of the transcription start site, and these two sites are revealed as three variants when analysed by denaturing gradient gel electrophoresis (DGGE): variant 1, which is part of both the A and C haplotypes; variant 3, which is rare; and variant 4, which is part of the B haplotype.[8] In this paper we note further allelic variation in this very small area in the populations tested. Despite the high level of allelic variation, this region shows apparent conservation in the pig sequence as well as primate sequences. This highly variable region is within the 1 kb sequence which, in the pig, has been reported to control lactase downregulation. Therefore it is possible that allelic variation within this region may affect gene regulation. The highly variable region was investigated for protein binding activity using electromobility shift assays (EMSA) with a protein extract of Caco2, an intestinal cell line that expresses lactase at a low level,[10] and so contains all the *trans*-acting factors essential for basal lactase expression.

Methods

Samples

DNA samples from five groups (British of African or Afro-Caribbean ancestry, Papua New Guinean, Japanese, San bushmen and Bantu-speaking South Africans) were prepared from whole blood by standard techniques.

The primate DNA samples were prepared from cultured cell lines from one chimpanzee (*Pan troglodytes*), two gorillas (*Gorilla gorilla*), two orang-utans (*Pongo pygmaeus*) and one crab-eating macaque (or cebus monkey, *Macaca fascicularis*). The chimpanzee and gorilla are African Great Apes, the orang-utan an Asian Great Ape, and the macaque an Asian monkey.

Polymerase Chain Reaction

Polymerase chain reaction (PCR) using the oligonucleotide primers 5FS and 5FA were carried out on human DNA samples as described previously.[7] PCR was also carried out on primates, the same primers, and the same protocol. The primers PROA and PROS2 were used to amplify the region between the 5FS/5FA PCR product and the start of transcription (Figure 1) using the same reaction conditions but different cycling conditions: 95°C, 5 min, followed by 95°C 1 min, 66°C 1 min, 72°C 1 min for 35 cycles. The sequence of PROA is 5'-GACTACATGCCAAGACAGCTCC-3' (+ 35 to + 14; Genbank/EMBL M61834, nt 1060 to nt 1039) and of PROS2 is 5'-TCTTCAGACATTTTCCGGGTTC-3' (–529 to –507; Genbank/EMBL M61834, nt 497 to nt 518).

Denaturing Gradient Gel Electrophoresis

Ava II restriction endonuclease digestion of the 5FS/5FA PCR product and DGGE of the resulting fragments was carried out as described previously.[7]

Cloning of PCR Products

PCR products were cloned using the TOPO-TA Cloning kit™ (Invitrogen, Netherlands) following the manufacturer's instructions. Clones were plated on to LB agar plates supplemented with X-gal (Life Technologies, Gaithersburg, Maryland, USA) and containing 50 μg/ml ampicillin (Sigma Chemical Co, St Louis, Missouri, USA). After overnight incubation, ten white or pale blue colonies were picked for analysis.

Sequencing

PCR product was sequenced using the PCR primers (5FS and 5FA on humans; 5FS, 5FA, PROS2 and PROA on primates, Figure 1) and the Thermosequenase radiolabelled terminator

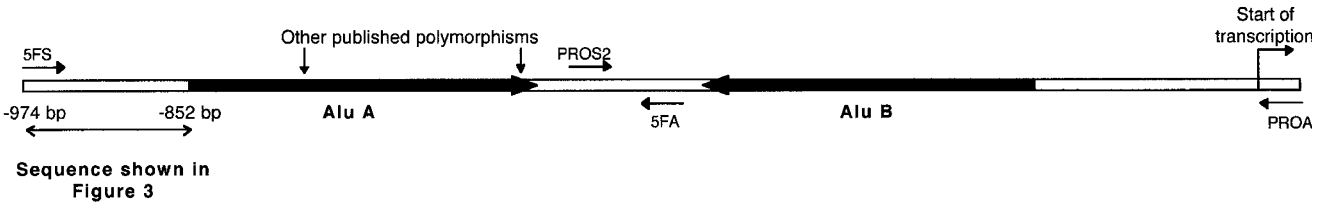

Figure 1 *Diagram showing the first 1 kb upstream from exon 1 of the human lactase gene. Positions of the PCR primers 5FA, 5FS, PROS2 and PROA are shown. The sequence shown in Figure 3 is underlined, and the other published polymorphisms are described previously.*[7]

cycle sequencing kit (Amersham Pharmacia Biotech, Amersham, Buckinghamshire, UK), according to the manufacturer's instructions.

Sequence Analysis

REPEATMASKER (available from the MRC Human Genome Mapping Project (HGMP) Resource Centre at Hinxton Hall, Cambridge, UK) was used to analyse sequence for repeat elements.

SIGNALSCAN, from the MRC HGMP, was used to analyse sequence for potential transcription factor binding sites.

The BESTFIT program, from the GCG suite (Genetics Computer Group, Wisconsin, USA), was used for most sequence alignments and identity statistics with the gap creation penalty set at 50 and the gap extension penalty set at 3. The comparison between pig sequence and the human sequence upstream of –974 bp as well as comparisons between human and rat were made using PILEUP from GCG using a gap creation penalty of 5 and extension penalty of 1. The percentage sequence identity was determined manually. The Genbank/EMBL accession number of the pig sequence and rat sequence are Y08677 and S77839 respectively.

Preparation of Nuclear Protein Extracts

Caco2 cells (passage 85) were cultured in Dulbecco's Modified Eagles Medium and 20% heat inactivated foetal calf serum as described.[11] Cells were harvested 15 days after the previous trypsination when they express maximum levels of lactase, centrifuged at 400 g for 10 min, the supernatant removed, and the pellet washed with phosphate-buffered saline solution (1 × = 0.15 M NaCl, 0.01 M NaH_2PO_4, 0.0075 M NaOH).

The pellet was resuspended in 5 volumes of 10 mM KCl, 1.5 mM $MgCl_2$, 10 mM HEPES pH 7.9, incubated on ice for 10 min and centrifuged at 400 g for 10 min. Again, the pellet was resuspended in 3 volumes of 10 mM KCl, 1.5 mM $MgCl_2$, 10 mM HEPES pH 7.9, 0.05% Nonidet P-40 and homogenised with a tight-fitting Dounce homogeniser to release the nuclei. Nuclei were pelleted by spinning at 530 g for 10 min and resuspended in 1 ml 1.5 mM $MgCl_2$, 0.2 mM EDTA, 5 mM HEPES pH 7.9, 25% (v/v) glycerol. 1 M NaCl solution was added to a final concentration of 300 mM NaCl, mixed well, and incubated on ice for 30 min. Following a spin at 25 000 g for 20 min at 4°C, the supernatant was aliquoted and snap-frozen at –70°C. Protein concentration was estimated using optical attenuance at 280 nm and 260 nm as described by Warburg and Christian (reviewed in Thorne[12]). DTT and PMSF were added to a final concentration of 0.5 mM in all solutions just before use.

Oligonucleotides for Electromobility Shift Assay

To prepare double-stranded oligonucleotides, complementary single-stranded oligonucleotides were synthesised by Perkin Elmer Biosystems (Warrington, UK), mixed in equimolar amounts and heated to 85°C for 5 min. After cooling to room temperature over a period of 30 min, double stranded oligonucleotides were adjusted to a final concentration of 1 pmol/μl using distilled water. The sequences are shown in Figure 3, except the sequence CE-LPH based on published sequence[4] which is as follows:

5′-AGTATTTTACAACCTCAGTT-3′

3′-AAAATGTTGGAGTCAACGTC-5′

and the sequence 17mer[13] which is as follows:

5′-AATTTTTTACAACACCT-3′

3′-TTAAAAAATGTTGTGGA-5′

5' End Labelling of Oligonucleotides

2 pmoles of double stranded oligonucleotide probe were labelled using 20 units of T4 Kinase (Boehringer Mannheim, Lewes, East Sussex, UK) and 30 μCi γ-^{33}P ATP >2500Ci/mmol (Amersham Pharmacia Biotech, Amersham, Buckinghamshire, UK) in a final concentration of 50 mM Tris-HCl, 10 mM $MgCl_2$, 0.1 mM EDTA, 5 mM DTT, 0.1 mM spermidine pH 8.2 at 25°C in a final volume of 20 μl, and incubated for 1 h at 37°C. After the incubation, 1 × STE buffer (1 × = 0.1 M NaCl, 0.001 M EDTA, 0.1 M Tris-HCl pH 8.0 at 25°C) was added to a final volume of 0.5 ml and the solution applied to a NAP-5™ Sephadex column (Amersham Pharmacia Biotech, Amersham, Buckinghamshire, UK). The probe was eluted in 1 ml H_2O.

EMSA Analysis

5 μl of protein extract (16 μg in Figure 5 or 40 μg in Figure 6) were incubated in binding buffer (final concentration: 20 mM HEPES pH 7.6, 1 mM EDTA, 10 mM $(NH_4)_2SO_4$, 1 mM DTT, 0.2% Tween-20, 30 mM KCl; final volume 20 μl) for 15 min on ice with 2 μg of poly(dI-dC) (Boehringer Mannheim, Lewes, East Sussex, UK) and 150 × molar excess unlabelled competitor probe. 10 fmol ^{33}P-labelled oligonucleotide probe in 5 μl was then added and the mixture incubated on ice for further 35 min. 6 × loading buffer (60% (w/v) glycerol, 0.2%

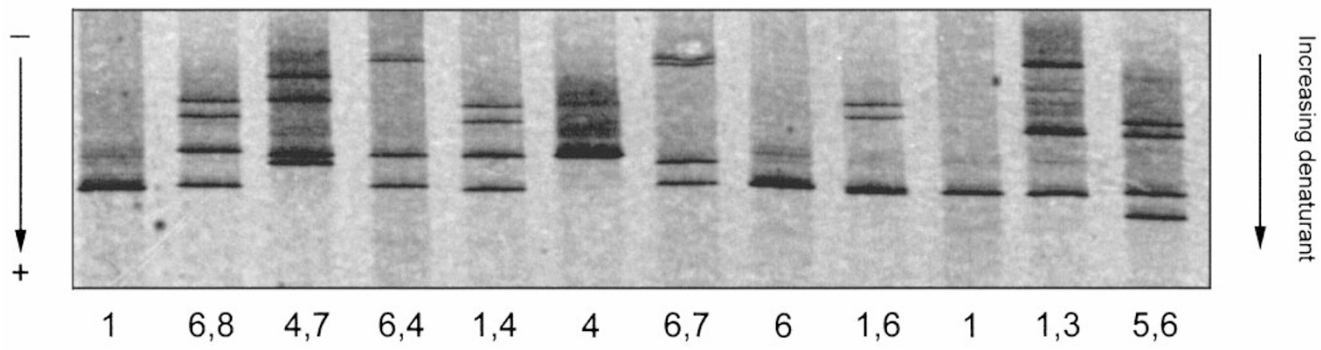

Figure 2 *The top half of a silver stained DGGE gel showing the seven variant alleles detectable in the Ava II digest fragment. The numbers assigned to the variants are below the appropriate lanes. Upper bands represent heteroduplexes. The differences between 1 and 6 could also be detected by SSCP analysis of the digested 5F amplimer as outlined previously.[7] Note that the small fragment of Ava II digestion is not shown and migrates further on the gel. The variant allele 2 is a polymorphism within Alu A detected only by SSCP analysis, and is not considered in this paper.*

(w/v) bromophenol blue, 0.25 × TBE buffer (1 × TBE = 0.09 M Tris-borate, 0.002 M EDTA) was added to a final volume of 30 µl. Routinely 10 or 20 µl was loaded on the gel.

Electrophoresis was performed using Hoefer SL600 equipment (Amersham Pharmacia Biotech, Amersham, Buckinghamshire, UK) and kept at a constant temperature of 10°C using an LKB Bromma water bath.

1.5 mm thick 5% 29:1 acrylamide:bis-acrylamide gels were pre-run for at least 1 h at 150 V. The buffer used was 0.5 × TBE. Following electrophoresis at 150 V for 2 h 30 min, the gel was dried and exposed to Kodak Biomax MR film.

Results

DGGE Analysis

A total of 157 samples from five population groups in which lactase non-persistence is the predominant phenotype were examined by DGGE. Several new variants, shown in Figure 2, were found in addition to the previously described 1, 3 and 4 alleles, and the allele frequencies in the different populations are shown in Table 1.

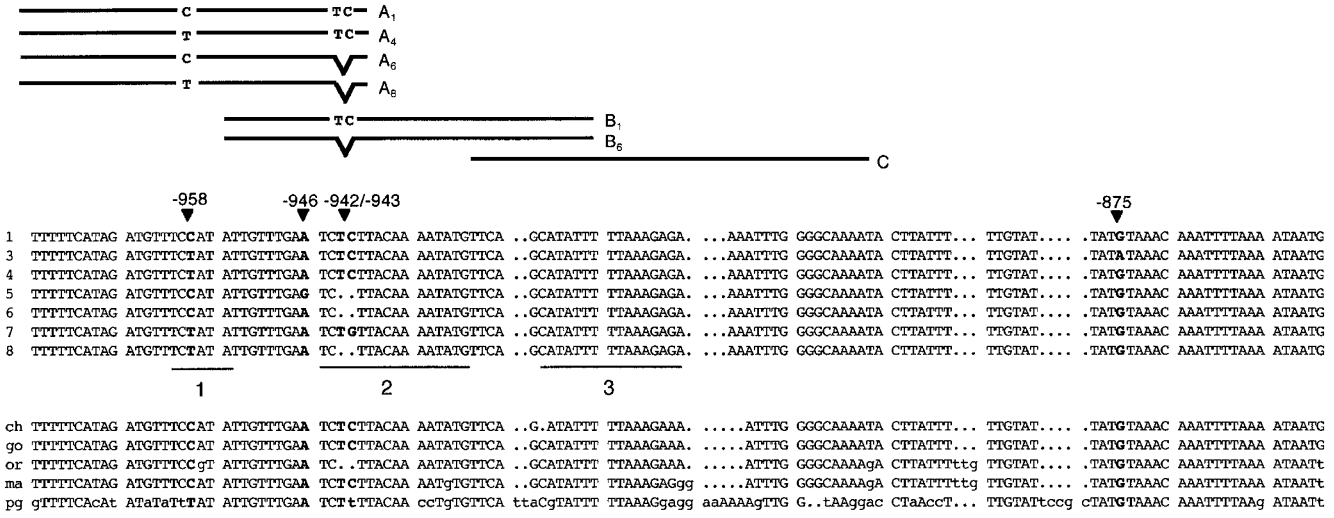

Figure 3 *A comparison of the sequence between –974 bp and –852 bp (see Figure 1) in the variants identified and in five other species. Polymorphisms are shown in bold and the position of each polymorphism is shown directly above the sequence. The numbers 1, 3, 4, 5, 6, 7, 8 on the left of the seven human sequences are the sequences of each respective variant. Non-identity between the other species and human is shown in small letters. The sequences 1, 2 and 3 are referred to in Figure 4. The double-stranded oligonucleotides used for EMSA are shown above the sequence. Any sequence differences in these oligonucleotides as compared to variant 1 are shown either by the base change or by a V indicating the two base-pair deletion. ch = chimpanzee, go = gorilla, or = orang-utan, ma = macaque, pg = pig.*

A

	1	2	3
Human	TCCATA	TCTCTTACAAAATATG	CATATTTTTAAAGAGA
Pig	**TTTATA**	TCT**TTTACA**ACCTGTG	CGTATTTTTAAAGGAG

B

Sequence 2 sense strand	TCTCTTACAAAATATG
Sequence 3 antisense strand	TCTCTTTAAAAATATG

Figure 4 *Comparison between short stretches of sequence highlighted in Figure 3.* ***A*** *Comparison of sequences 1, 2 and 3 in the human and pig. Sequence underlined shows identity between pig and human, and sequence in bold indicates a Cdx-2 consensus binding site.* ***B*** *Comparison between sequence 2 and the antisense of sequence 3 in humans. Sequence underlined shows identity between the two sequences (shown in Figure 3), and the sense strand is defined by the sense strand of the lactase gene.*

Variants 7 and 8 were each found in one individual only, although variant 7 has been observed again in two of five unrelated Chinese individuals from Taiwan (data not shown). Variant 5 was found in two individuals, one San and one Bantu-speaking South African. Variant 6 was found at varying polymorphic frequencies in all the groups tested in this study, but has not yet been found in any European data set.

Sequence Analysis

Direct sequencing of genomic PCR products of each variant (1,3,4,5,6,7,8) determined the sequence variation responsible for the gel phenotypes, and showed that all variation detected by DGGE was due to base changes within a small region between 974 bp and 852 bp upstream of the transcription start site. To confirm the haplotype across the fragment, the PCR product of each variant was cloned, several clones reamplified and the product digested with Ava II. The digests were then analysed by DGGE in comparison with the digested PCR product from the original genomic samples to confirm the identity of the cloned allele and discriminate against clones containing PCR artefacts. A cloned representative of each allele was sequenced. Homozygous individuals were used where possible, but in the case of rare variants (5 and 8) the two alleles were cloned from heterozygotes. The sequences are shown in Figure 3. Variant 6 is due to a –942/–943 two base pair deletion in comparison with the variant 1 sequence, and variant 5 has this deletion as well as an additional change (A-946G). Variant 4 is C-958T as described previously.[7] Variants 3, 7 and 8 are further mutations on the background of variant 4: variant 3 is a rare European variant with a single nucleotide substitution (A-875G), variant 7 has a change at –942 (C-942G), and variant 8 has the –942/–943 deletion.

1 kb upstream from the start of transcription was sequenced in the four primate species and compared with the human sequences. All primate individuals were heterozygous at one position at least (data not shown), so that more than one chromosome was analysed in each species. Alu elements are found at the same position in all primate species so that the complete region can be aligned and all five species showed $>93\%$ identity over the 1 kb region. Table 2 shows the percentage identity between human and each primate Alu sequence in comparison with that observed in the region –974 bp to –852 bp which is highly variable in humans. All the primates show greater sequence conservation in the region corresponding to –974 bp to –852 bp in humans than in the two Alu elements. Alignment of these primate sequences with the human region between –974 bp and –852 bp is shown in Figure 3.

The pig and rat sequences were compared with the human upstream of –852 bp. The rat showed no significant areas of identity but the pig showed 80% identity with the human region between –974 bp and –852 bp. This compared with only 36.4% identity with the human region between –1540 bp to –974 bp (Table 2).

Three interesting sequence motifs were identified within this conserved area (Figure 3 highlights them as 1, 2, 3) and the sequence homologies between human and pig are shown in Figure 4a. In pig, but not in human, sequences 1 and 2 contain consensus binding motifs for Cdx-2, an intestine-specific transcription factor. Sequence 3 is of interest because it contains an 11 bp motif which is identical in humans and pigs. In humans sequence 2 contains an inverted version of this motif in which 9 of 11 bases are identical, the two non-identical bases being found in the centre of the motif. This is shown in Figure 4b.

Electromobility Shift Assay Analysis

EMSA was used to examine the effect of the common nucleotide sequence variation which occurs within sequences 1, 2 and 3 on protein binding. Overlapping

double-stranded oligonucleotides were designed to span these sequences (Figure 3). Two groups of double-stranded oligonucleotides were named A and B, and one double-stranded oligonucleotide was named C. The four members of group A spanned sequence 1 and part of sequence 2, and were synthesised to correspond to

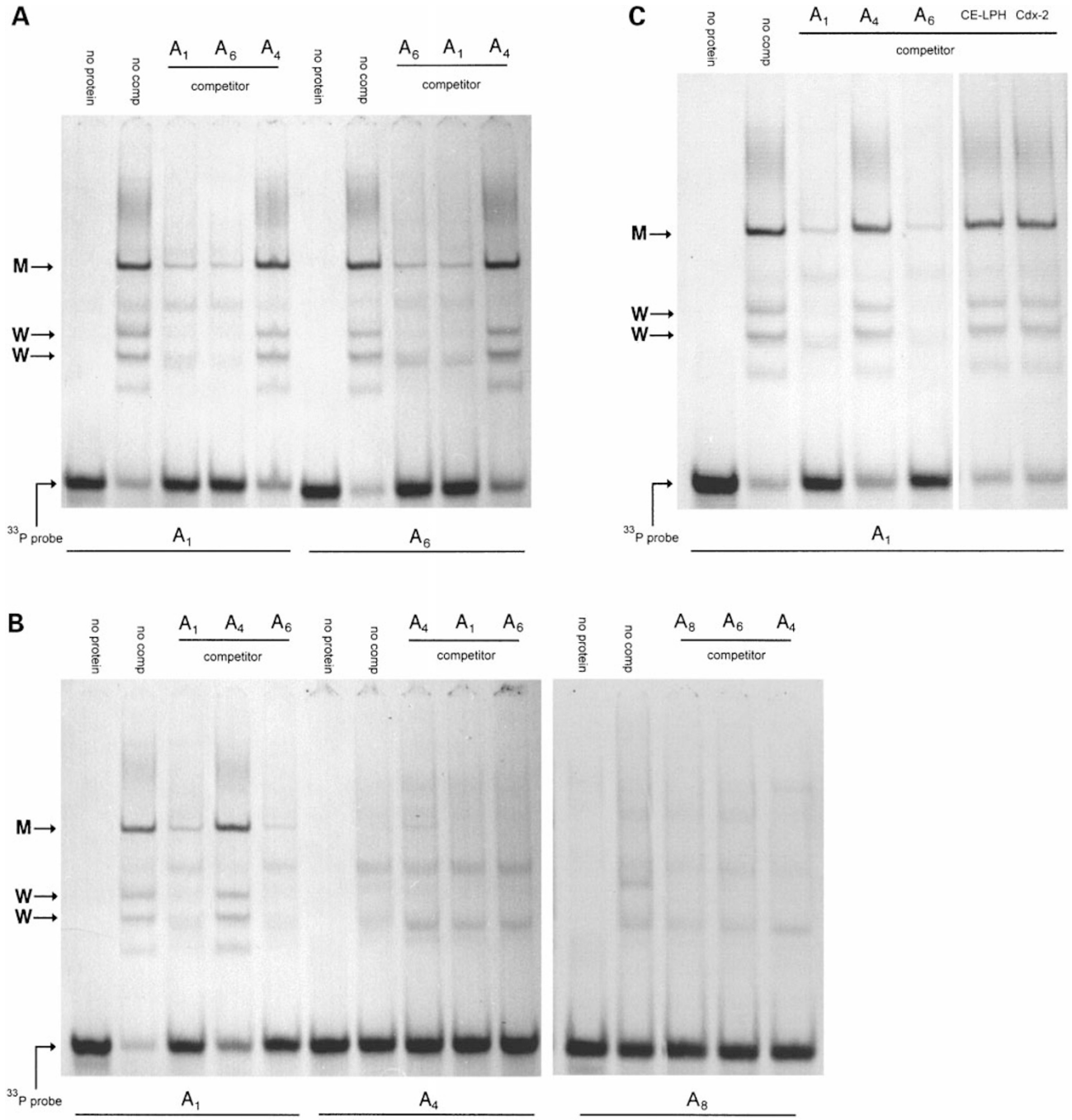

Figure 5 *EMSA using ^{33}P-labelled oligonucleotides A_1, A_4, and A_8.* no comp *indicates no unlabelled oligonucleotide competitor;* ***M*** *indicates main specific band;* ***W*** *indicates weaker specific bands. ^{33}P probe indicates unbound labelled oligonucleotide probe, and the label underneath the gel indicates the labelled oligonucleotide used as a probe. All the observations were reproducible in repeat experiments.* ***A*** *Variant 1 generates specific binding activities which variant 4 does not show. Competitor oligonucleotides are A_1, A_4, A_6 and A_8.* ***B*** *Variant 1 and variant 6 show the same binding activity using oligonucleotides A_1 and A_6. Competitor oligonucleotides are A_1, A_4, A_6, and A_8.* ***C*** *Binding activities using oligonucleotide A_1 do not involve Cdx-2 or other factors that bind to CE-LPH1. Competitor oligonucleotides are A_1, A_4, A_6, CE-LPH and Cdx-2 (which is the oligonucleotide 17mer)*

Table 1 Frequencies of variants in different groups

Variant	*Northern European*	*Southern European*	*Bantu-speaking South African*	*San*	*British (Black)*	*Papua New Guinean*	*Japanese*
1	*0.91*	*0.60*	0.72	0.40	0.60	0.56	0.58
3	*0.03*	*0.06*	0	0	0	0	0
4	*0.06*	*0.34*	0	0.03	0.08	0.40	0.12
5	*0*	*0*	0.01	0.03	0	0	0
6	*0*	*0*	0.27	0.54	0.31	0.04	0.29
7	*0*	*0*	0	0	0	0	0.01
8	*0*	*0*	0	0	0.01	0	0
N	*104*	*108*	72	30	62	72	78
LCT*P	0.78	0.26	0.12	0.03	0.09	0.05	0.10

Frequency of the variants in the groups tested. *N* is the number of chromosomes tested. Black denotes African, Afro-Caribbean and mixed descent. The variant allele 2 is a polymorphism within Alu A detected only by SSCP analysis, and is not considered in this paper. The frequencies of the persistence allele (LCT*P) in different groups are published[1,19], and the data in italics are from a previous paper[8].

the three common allelic variants 1, 4 and 6, which are generated by two polymorphic sites (–958 and –942/–943; Figure 3). The fourth was synthesised to correspond to the rare variant 8 which represents the remaining haplotype of these two polymorphic sites (Figure 3). Oligonucleotide A_1 corresponds to variant 1; oligonucleotide A_4 corresponds to variant 4; oligonucleotide A_6 corresponds to variant 6; and oligonucleotide A_8 corresponds to variant 8.

The two members of group B spanned sequence 2 and part of sequence 3, and were synthesised to correspond to two common allelic variants generated by the polymorphism at –943/–943 only. Oligonucleotide B_1 corresponds to variant 1 and oligonucleotide B_6 corresponds to variant 6. Oligonucleotide C spanned sequence 3. All the oligonucleotides were used in EMSA with nuclear protein extract of Caco2 cells.

Initial experiments used labelled group A oligonucleotides, which represent the common variants (A_1, A_4, A_6), as probes to investigate the effect of variation on protein binding. Figure 5a shows both A_1 and A_6 generate a specific binding activity of one strong main band (M) with two weak higher mobility bands (W). Competition assays using A_1, A_6 and A_4 as competitors revealed that both A_1 and A_6 displaced both M and W bands generated by A_1 and A_6 probes (Figure 5a). This showed that these bands represented specific protein binding activities, and that variation at –942/–943 appears not to affect binding. However, competition with unlabelled A_4 failed to displace the bands (Figure 5a) suggesting that variation at –958 affects protein binding. This was confirmed by using A_4 as a labelled probe in further EMSA experiments which showed that A_4 does not generate M or W bands (Figure 5b). Using A_8 in place of A_4 produced the same results: no protein bound when used as a labelled probe (Figure 5b) nor could it displace M or W bands produced when A_1 or A_6 were used as probes (data not shown). This again confirms that the polymorphism at –958 affects protein binding but the polymorphism at –942/–943 does not.

Table 2 Percentage identity of sequences between human and other species

Species	*% identity with human Alu elements A and B*	*% identity with human sequence –974bp to –852bp*
Chimpanzee	96.8	100
Gorilla	97.1	100
Orang-utan	96.3	97.5
Macaque	92.6	95.2
Pig	36.4*	80

A comparison between the percentage identity of human sequence and sequences from other species over the region shown in Figure 3 and over the Alu elements A and B (Figure 1). *Since the pig has no Alu elements, the percentage identity between pig and human of between –1540bp and –974bp is shown.

The M and W bands generated by A_1 and A_6 were not displaced by the oligonucleotide 17mer (Figure 5c), representing the Cdx-2 binding site 5' of the carbonic anhydrase gene,[13] nor by CE-LPH (Figure 5c), an oligonucleotide that represents CE-LPH1 sequence upstream of the human lactase gene.

Further experiments used labelled group B oligonucleotides (B_1, B_6) as well as the C oligonucleotides to investigate the significance of the 11 bp motif in sequence 2 and its inverted homologue in sequence 3. Labelled oligonucleotides B_1 and C showed apparently identical specific binding activities with two major bands H and L (Figures 6a and 6b). Band H generated by either labelled B_1 or C can be displaced by both B_1 and C indicating that this is the same activity. This is

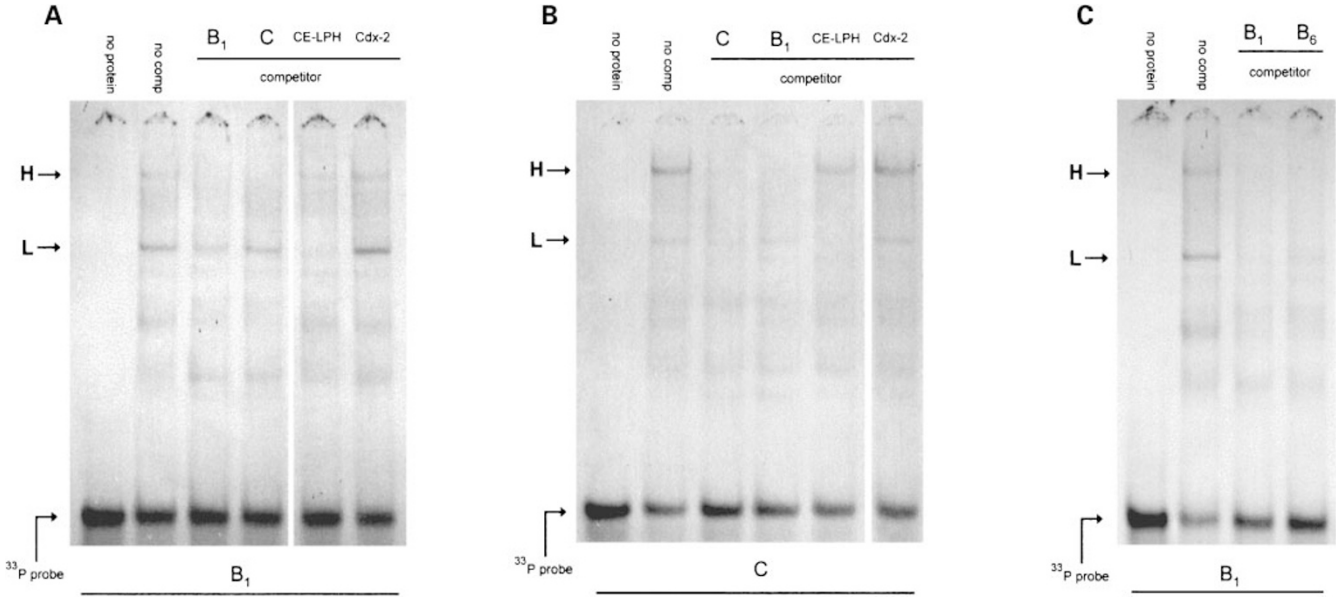

Figure 6 *EMSA using ^{33}P-labelled oligonucleotide B_1 or C.* no comp *indicates no unlabelled oligonucleotide competitor;* H *indicates the higher specific band,* L *indicates the lower specific band. ^{33}P probe indicates unbound labelled oligonucleotide probe, and the label underneath the gel indicates the labelled oligonucleotide used as a probe. All the observations were reproducible in three independent experiments, but the intensity of the L band varied.* ***A*** *Binding activities using oligonucleotide B do not involve Cdx-2, but one involves a factor that binds to CE-LPH1. Competition oligonucleotides are B_1, C, CE-LPH, and Cdx-2 (17mer).* ***B*** *Binding activities using oligonucleotide C do not involve Cdx-2, but one involves a factor that binds to CE-LPH1. Competition oligonucleotides are B_1, C, CE-LPH, and Cdx-2 (17mer).* ***C*** *Variant 1 and variant 6 both displace the binding activities shown by oligonucleotide B_1. Competition oligonucleotides are B_1 and B_6.*

due to a shared motif either as a result of the 11 bp inverted repeat (Figure 4b) or the overlap between oligonucleotides (Figure 3).

The band L generated by labelled B_1 can be displaced, albeit ineffectively, by B_1 but not by C. Conversely, band L generated by labelled C can be displaced by C but not by B_1. This shows that the activities generating L are different. Both activities generating L are displaced by CE-LPH, suggesting that both activities are likely to be due to a protein or proteins that bind to CE-LPH1. Cdx-2 does not displace either band generated by either oligonucleotide (Figures 6a and 6b).

Using B_1 and B_6 as competitors, both H and L bands generated by labelled oligonucleotide B_1 are displaced (Figure 6c). Conversely, B_1 and B_6 can displace both bands generated when B_6 is the labelled probe (data not shown). This again illustrates that the polymorphism at –942/–943 has no effect on protein binding.

Discussion

The region –974 bp to –852 bp of human lactase is an unusually variable stretch of DNA sequence with marked allele frequency differences in different populations. The previously described variant 4 is found at polymorphic levels in southern Europeans, Japanese, and New Guineas, but rare in the black British cohort, San, and northern Europeans, and absent in Bantu-speaking South Africans. The newly described variant 6 (the deletion at –942/–943) was present at polymorphic levels in all the new population groups tested, yet was not observed in the Europeans tested using the same detection method.

The population differences in allele frequencies may be due to genetic drift, but it is perhaps more likely that it is a result of past natural selection. It is unlikely that selection operated on these sequence differences directly, but it is more probable that their frequency reflects selection for the haplotype carrying lactase persistence. This would result in an increase of the neutral alleles which were present on the same haplotype, in this case the A haplotype which carries persistence in virtually all Caucasians. Analysis of haplotypes in other populations will help to reveal the role of 'selective sweep' or 'genetic hitch-hiking' effects on diversity within the lactase gene.

Comparison of the human sequence with those of the primates suggests that variant 1 is the ancestral primate

variant (Figure 3). However both orang-utans were homozygous for –942/–943 deletion characteristic of human variant 6. This could be explained by a *de novo* mutation in the orang-utan species or maintenance of a polymorphism in human and orang-utan lineages. In chimpanzees, which are more closely related to humans, there are examples of the same sites being polymorphic as humans, most notably at the HLA locus.[14,15] It is difficult to envisage polymorphism being maintained in both human and orang-utan lineages given the 14 million years since the most recent common ancestor,[16] suggesting that the mutation may have arisen independently on both lineages.

Sequencing of the seven human variants shows that they represent combinations of five base changes, and that DGGE can reveal haplotypes across this region. Despite the highly variable nature of this region, it shows evidence of greater conservation between species than the surrounding sequence. This suggested that protein-binding sites important for regulation might be within this region, and that polymorphism might have an effect on their function.

Analysis of the human sequence revealed an inverted repeat between –945 bp and –909 bp which contains the polymorphic deletion at –942/–943 (variant 6). EMSA experiments using two oligonucleotides (B_1 and C) covering both repeated units showed two binding activities: one is the L band generated by protein(s) that also binds to CE-LPH1. Indeed others have identified part of one repeat as CE-LPH1b by comparison with CE-LPH1.[17] The transcription factor Cdx-2 has been shown to bind to CE-LPH1, but in both repeat units the Cdx-2 consensus binding motif TTTA$^{C}/_{T}$A has been disrupted. We show that the protein that generates the L band is not displaced by another Cdx-2 binding oligonucleotide, which suggests that it is not Cdx-2. Other homeobox proteins may bind to this region, such as HOXC11 that binds to the *cis*-element CE-LPH1.[5] There are fewer clues as to the identity of the protein that binds to oligonucleotide A_1 since no candidate recognition sequence was identified by searching transcription factor databases.

Analysis of the effect of the polymorphic deletion at –942/–943 (variant 6) on protein binding revealed no differences in EMSA experiments with either the A or B group of oligonucleotides.

In contrast EMSA shows that a T nucleotide at the polymorphic site C-958T (variant 4 and variant 8) polymorphism drastically affects the ability of a protein to bind to oligonucleotide A_1. This suggests that the C at –958 is a critical part of the protein binding site. The pig sequence has a T at this position and a different base compared with the human at –959, suggesting that this protein will have a low affinity for the pig sequence at this point. The variant 4 (–958T) occurs in Europeans as part of the B haplotype and most LCT B haplotype chromosomes show low LCT expression in adults, but high expression of the B haplotype has been observed in two adults who were interpreted as being homozygous for lactase persistence. Heterozygous foetuses show equal but very low expression of both transcripts using marker polymorphisms characteristic of A and B haplotypes, and four very young children (aged 2 months to 8 months) showed equal high expression of both transcripts.[9] Lactase non-persistence can occur on A and C haplotypes, which both contain variant 1 and so carry the C nucleotide at –958. These observations exclude this change from causing the phenotypic polymorphism, but we cannot exclude an effect on the timing of downregulation in young children, which appears to be variable. Furthermore, spatial regulation of lactase along the length of the intestine could possibly be affected by the C-958T polymorphism, resulting in asymmetric expression of C and T carrying alleles in certain regions of the intestine. Analysis of nuclear extracts isolated from enterocytes, from different regions of the intestine and different developmental stages, together with transfection studies using Caco-2 cells, may help to clarify the functional significance of this binding.

Several homeobox proteins are expressed in the small intestine,[18] and one or more of these could act to ensure correct spatial and temporal patterning of lactase expression.

Finer localisation of these *cis*-elements and the proteins that bind to them will help our understanding of lactase regulation, although the *cis*-element controlling lactase persistence or non-persistence is probably distant from the immediate promoter.

Acknowledgements

We thank Professor Yvonne Edwards and Dr Phil Johnson for DNA samples, Dr Felicity Drummond for Cdx2 consensus oligonucleotides, and Dr Nikolaj Spodsberg for helpful comments on the manuscript.

References

1 Flatz G: Genetics of lactose digestion in humans. *Adv Hum Genet* 1987; **16**: 1–77.

2 Swallow DM, Harvey CB: Genetics of adult-type hypolactasia. *Dyn Nutr Res* 1993; **3**: 1–7.
3 Wang Y, Harvey CB, Pratt WS *et al*: The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Hum Mol Genet* 1995; **4**: 657–662.
4 Troelsen J, Olsen J, Noren O, Sjöstrom H: A novel intestinal trans factor (NF-LPH1) interacts with the lactase phlorizin hydrolase promotor and co-varies with the enzymic activity. *J Biol Chem* 1992; **267**: 20407–20411.
5 Mitchelmore C, Troelsen JT, Sjöstrom H, Noren O: The HOXC11 homeodomain protein interacts with the lactase-phlorizin hydrolase promoter and stimulates HNF1α-dependent transcription. *J Biol Chem* 1998; **273**(21): 13297–13306.
6 Troelsen JT, Mehlum A, Olsen J *et al*: 1 kb of the lactase-phlorizin hydrolase promoter directs post-weaning decline and small intestinal-specific expression in transgenic mice. *FEBS Lett* 1994; **342**: 291–196.
7 Harvey CB, Pratt W, Islam I, Whitehouse DB, Swallow DM: DNA polymorphisms in the lactase gene: linkage disequilibrium across the 70 kb region. *Eur J Hum Genet* 1995; **3**: 27–41.
8 Harvey CB, Hollox EJ, Poulter M *et al*: Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. *Ann Hum Genet* 1998; **62**: 215–223.
9 Wang Y, Harvey CB, Hollox EJ *et al*: The genetically programmed down-regulation of lactase in children. *Gastroenterology* 1998; **114**: 1230–1236.
10 Wang Y, Harvey C, Swallow DM: Towards an understanding of the genetic basis of the lactase persistence/non-persistence polymorphism in man. In: Lentze MJ, Naim HY, Grand RJ (eds). *Mammalian Brush Border Membrane Proteins II*. Theime Medical Publishers: New York, 1994.
11 Chantret I, Rodolosse A, Barbat A *et al*: Differential expression of sucrase-isomaltase in clones isolated from early and late passages of the cell line Caco-2: evidence for glucose-dependent negative regulation. *J Cell Sci* 1994; **107**: 213–225.
12 Thorne CJR: *Techniques in Protein and Enzyme Biochemistry*, Part 1. Elsevier: North Holland, 1978.
13 Drummond F, Sowden J, Morrison K, Edwards YH: The caudal-type homeobox protein Cdx-2 binds to the colon promoter of the carbonic anhydrase 1 gene. *Eur J Biochem* 1996; **236**: 670–681.
14 Fan MW, Kasahara M, Gutknecht J *et al*: Shared class II polymorphisms between humans and chimpanzees. *Hum Immunol* 1989; **26**(2): 107–121.
15 Gyllensten UB, Erlich HA: Ancient roots for polymorphism at the HLA-DQ alpha locus in primates. *Proc Natl Acad Sci USA* 1989; **86**(24): 9986–9990.
16 Goodman M, Porter CA, Czelusniak J *et al*: Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* 1998; **9**(3): 585–598.
17 Spodsberg N, Troelsen JT, Carlsson P, Enerback S, Sjöstrom H, Noren O: Transcriptional regulation of pig lactase-phlorizin hydrolase. Involvement of HNF-1 and FREACs. *Gastroenterology* 1999; **116**: 842–854.
18 Walters JRF, Howard A, Rumble HEE, Prathalingam SR, Shaw-Smith CJ, Legon S: Differences in expression of homeobox transcription factors in proximal and distal human small intestine. *Gastroenterology* 1997; **113**: 472–477.
19 Iqbal TH, Wood GM, Lewis KO *et al*: Prevalence of primary lactase deficiency in adult residents of West Birmingham. *Br Med J* 1995; **306**: 1303.