Original Paper

Eur J Hum Genet 1995;3:27–41

Clare B. Harvey
Wendy S. Pratt
Ira Islam
David B. Whitehouse
Dallas M. Swallow

MRC Human Biochemical Genetics Unit, Galton Laboratory, University College London, UK

DNA Polymorphisms in the Lactase Gene

Linkage Disequilibrium across the 70-kb Region

Abstract

The enzyme lactase, which is responsible for the digestion of dietary lactose, is present in the intestine of some adults but not others. As a means of providing a platform to explore the molecular basis of this nutritionally relevant genetic variation we have screened for polymorphism in several regions of the lactase gene. In each case simple polymerase chain reaction-based procedures (including single-strand conformation analysis and denaturing gradient gel electrophoresis) were used, combined with silver staining as a method of detection. Allelic variation was found at 6 different sites. One previously published polymorphism was also tested. The frequencies of the alleles were determined in more than 100 unrelated individuals of the Centre d'Etude du Polymorphisme Humain (CEPH) panel, and the haplotypes were deduced. A region of linkage disequilibrium was observed, which spans the whole coding region of the lactase gene (~ 60–70 kb); there were only 3 common haplotypes in this population. When the CEPH sample was subdivided according to the population of origin (France or Utah) the haplotype frequencies were shown to be markedly different.

Key Words

Lactase gene
Single-strand conformation analysis
Denaturing gradient gel electrophoresis
Linkage disequilibrium

Introduction

Intestinal lactase phlorizin hydrolase is the brush border disaccharidase responsible for the digestion of dietary lactose. In the majority of the world's population lactase activity does not persist into adult life but declines after weaning. In some people, however, particularly in northern Europe, lactase activity persists into adulthood. The frequency of these two phenotypes varies in different populations of the world. Lactose tolerance tests in families have suggested that the polymorphism is controlled by two alleles at a single gene locus with persistence being dominant to non-persistence [1, for review 2].

Received: August 4, 1994
Revision received: November 17, 1994
Accepted: November 28, 1994

Dallas M. Swallow, PhD
MRC Human Biochemical Genetics Unit
Galton Laboratory, University College London
4 Stephenson Way
London NW1 2HE (UK)

1018–4813/95/
0031–0027$8.00/0

Studies of lactase activity in samples of adult intestine from populations of unrelated individuals show a clear trimodal distribution [3–5]. The frequencies of the individuals in these three groups are consistent with the two-allele model, in which the group of individuals with intermediate activity represent the heterozygotes and the other groups are the homozygotes. This suggested that the relevant genetic element(s) may be *cis*-acting and hence be within or close to the lactase gene (*LCT*).

Despite sequence analysis of 1 kb of the promotor region and the complete cDNA of the lactase gene in a few individuals of known phenotype, the molecular basis of this polymorphism is not yet known. Single base changes have been seen, but none of these were obviously associated with the lactase persistence/non-persistence polymorphism [6, 7]. The level at which the difference in expression of lactase is regulated has also been controversial [8–10]. Recent studies suggest that in most cases lactase non-persistent individuals show a lower level of lactase mRNA [4, 11]. These findings do not, however, distinguish between a *cis*- or a *trans*-acting mechanism.

In order to determine whether the lactase gene is directly implicated in the lactase persistence status, we have searched for polymorphisms with a view to making the *LCT* gene more informative for family and association studies, and for identifying individual lactase transcripts. We have focused on regions of the gene in which base changes had already been reported [6, 7]. We made use of polymerase chain reaction (PCR)-based techniques that are sensitive to the detection of a wide variety of small base alterations in DNA, namely single-strand conformational analysis (SSCA) [12], denaturing gradient gel electrophoresis (DGGE) [13] and simple polyacrylamide gel electrophoresis (PAGE). High resolution was obtained by using very small quantities of DNA and detection by silver staining. In addition we have determined the precise location of a previously described *Msp*I polymorphism [14].

Materials and Methods

Samples

50 large sibship families obtained from the Centre d'Etude du Polymorphisme Humain (CEPH) [15] were investigated in this study. 37 were originally from Dr. Ray White's laboratory in Utah, 10 were from France and 3 from other sources. Genomic DNA was obtained directly from CEPH or prepared from blood samples or cell lines using an ABI 340A Nucleic Acid Extractor.

Polymerase Chain Reaction

Four segments of the lactase gene were amplified: a portion 5′ to the coding region (5F); a portion spanning the second exon (F2); a region extending from exon 16 to exon 17 (LCT3); and a region of 3′ untranslated sequence spanning the polyadenylation signal (UT). The sequences of the oligonucleotide primers and the sizes of the PCR products are given in table 1. The oligonucleotide primers were synthesised using an ABI 391-PCR-MATE.

Fragments were amplified using the reaction mix recommended by the manufacturers of the Taq polymerase (Advanced Biotechnologies or Promega). The conditions of the amplification for the 5F fragment were: after initial denaturation for 5 min at 95 °C, 30 cycles of amplification consisting of denaturation for 20 s at 94 °C, annealing for 20 s at 47 °C and elongation for 40 s at 70 °C. The F2 fragment was denatured as above but the subsequent 30 cycles consisted of denaturation for 20 s at 94 °C, annealing for 20 s at 52 °C and elongation for 20 s at 70 °C. The cycles for the UT fragment were as for 5F except the elongation was only for 20 s. The LCT3 fragment was amplified using 30 cycles consisting of a denaturation stage as above, annealing for 20 s at 53 °C and elongation for 40 s at 70 °C. The F17/LCT fragment was amplified using primers LCT3A and F17S for 30 cycles of 20 s at 94, 50 and 70 °C. 10 µl of 5F PCR products were digested using 0.8 U of the restriction enzyme *Ava*II (BRL) in a final volume of 40 µl using the conditions recommended by the manufacturers. 10 µl of LCT3 or F17/LCT were digested by *Msp*I (BRL) under similar conditions.

Table 1. Sequence and position of the oligonucleotide primers within the lactase gene and the sizes of the PCR products

Sequence 5′–3′	Position	Primer name	Product name	Size bp	EMBL ac No.
GGA GGG TGA AGG AAT TTG CAA G	29–50	5FS	5F	534	M61834
GAG TTC AAG ACC AGC CTG G	250–268	AluS	AluS/5F	313	M61834
CAT AGG TGT GCG CCA CC	338–322	AluA	AluA/5F	310	M61834
GAC CAA CAC AAA AAC CTC AGA C	562–541	5FA	5F	534	M61834
CAG TGG TTT CCA CAG TCA GA	411–430	F2S	F2	190	M61835
CTC TCC TCA GAT GTT ACA GG	600–581	F2A	F2	190	M61835
TAC AGT GAC CCT TCT CTG CCA AG	230–252	LCT3S	LCT3	~1,400	M61849
CTG AGA ACT CAA ATC AGC GCC AG	4–26	F17S	F17/LCT	244	M61850
GGC TTC GTT GTG TTT TCC CTT GC	247–225	LCT3A	F17/LCT LCT3	244 ~1,400	M61850
CAA CTC CAT TGC ATA GAC TGC	653–673	UTS	UT	177	M61850
GAG CTC CAG ATG GCT TGT GTG AC	829–807	UTA	UT	177	M61850

The primers are shown 5′–3′ in the order of their location within the gene. It should be noted, however, that in this table the nucleotide positions are taken directly from the EMBL database, accession numbers as quoted, because some of them are located within introns. Product sizes are given in base pairs. The primers AluA and AluS and the products prepared with them were only used for sequencing purposes.

Single-Strand Conformation Analysis

Samples were either mixed 1:1 with loading buffer (consisting of 95% deionised formamide, 20 m*M* EDTA, 0.05% xylene cyanol and 0.05% bromophenol blue) in the case of digested products, or were mixed 1:1:2 (sample:water:loading buffer) in the case of neat PCR products. They were then heated to 85–95 °C for 5–10 min and snap cooled on ice. The 1-Kb ladder and φx/174 digested with *Hae*III (BRL) were used as molecular weight markers and reference points. The gel compositions used were 6% acrylamide (37.5:1, acrylamide:bis, Bio-Rad) in 0.086 *M* Tris, 1.9 m*M* EDTA and 0.09 *M* borate buffer, pH 8.4 (1 × TBE) or in 0.5 × TBE, with or without the addition of 5% glycerol. The gel size was 17 × 13 cm × 0.8 mm, and electrophoresis was performed in a BRL vertical gel tank. Electrophoresis was carried out for various times with voltage limiting at 400 V, either in a cold room, temperature 4–10 °C, the gel surface temperature remaining below 30 °C, or at room temperature (approximately 22 °C). The conditions defined for the routine analysis of the 5F PCR product were: 1 × TBE, 5% glycerol for 2 h, in a cold room, and for the F2 product were: 0.5 × TBE, 5% glycerol for 1.5 h, in a cold room.

Denaturing Gradient Gel Electrophoresis Analysis

Electrophoresis was carried out on a modified Hoefer SE 600 vertical electrophoresis apparatus [16]. The gel was submerged in 0.04 *M* Tris-acetate, 1 m*M* EDTA, pH 7.4 (1 × TAE) at 61 °C with circulation of electrolyte between anode and cathode. The gels, 14 × 18 cm × 0.75 mm, consisted of 10% acrylamide (37.5:1 acrylamide:bis, Bio-Rad) in 1 × TAE with a linear 40–50% gradient of chemical denaturant (100% denaturant being 7 *M* urea, 40% formamide). Digests of PCR products (5 µl) were mixed with an equal volume of 1 × TAE buffer and 2.5 µl of loading buffer (loading buffer composition was 20% Ficoll, 0.5% bromophenol blue, 10 m*M* Tris, 1 m*M* EDTA, pH 7.8). Electrophoresis was performed with voltage limiting at 35 V (approximately 65 mA) for 22 h.

Analysis of the UT Product by Non-Denaturing PAGE

Gels were prepared containing 6% acrylamide (19:1 acrylamide:bis, Bio-Rad) in 1 × TBE. Samples were diluted 1:19 in sterile distilled water and then mixed 1:1 with loading buffer (40% sucrose, 0.1% bromophenol blue, 0.1% xylene cyanol) such that the amount was equivalent to about 5 ng DNA (0.25 µl PCR product) and loaded onto the gel without denaturation. Electrophoresis was carried out in a cold room (4–10 °C) at 40 mA/gel and with limiting voltage of 400 V, for 2.5 h until the xylene cyanol marker had run off the gel.

Restriction Fragment Length Polymorphism (RFLP)

Two different methods were used. For Southern blot analysis of *Msp*I-digested genomic DNA from the CEPH families, the LCT3 PCR product was used as probe. The PCR product was gel purified and labelled using a Multiprime kit (Amersham). The Hybond N+ (Amersham) filters (prepared by EUROGEM) were prehybridised and hybridised as recommended by the manufacturer. Alternatively LCT3 or F17/LCT PCR products from the family members were digested with *Msp*I and the digestion products analysed by non-denaturing PAGE and silver staining.

Silver Staining

The gels were fixed in 10% ethanol and 0.5% acetic acid using two 3-min incubations, then incubated in 0.1% silver nitrate (freshly made) for 10 min. They were then washed in two changes of distilled water and incubated in staining solution (375 m*M* sodium hydroxide, 2.6 m*M* sodium borohydride and 0.148% formaldehyde) until the bands were visible (maximum 20 min). Gels were subsequently vacuum dried and stored flat.

Sequencing

The 5F and F2 PCR products were sequenced by the dideoxy chain termination method [17] using the Sequenase kit (USB). Single-stranded template was prepared by biotinylation of one strand and separation on streptavidin-coated magnetic beads. In the case of F2 the PCR products (1 µl) were reamplified using 5 pmol of the same primers, one of which was biotinylated. In the case of the 5F product the initial PCR product was gel purified and then reamplified using sense or antisense primers located in the Alu element (AluS and AluA; table 1) together with biotinylated 5FS or 5FA. Strands were separated on Dynabeads (M-280, Dynal) in 0.1 *M* NaOH and both the biotinylated strand, which was attached to the beads, and the NaOH eluate were sequenced using 2 pmol (AluS, AluA, F2S, F2A) or 5 pmol (5FS, 5FA) as primer.

Linkage Analysis and Determination of Haplotypes of Alleles in LCT

Lod scores were calculated from the equations described by Maynard-Smith et al. [18], using the computer program HANDLINK [J. Attwood, personal commun.]. Unambiguous haplotypes were determined for 240 chromosomes in the CEPH pedigrees by analysis of the joint segregation of the alleles detected in each fragment. This included information on the other chromosome from each of the grandparents where this was available. Chromosomes for which the information was incomplete were excluded from the analysis. In those cases where the families are known to be related the duplicated chromosomes were counted only once.

Calculation of Linkage Disequilibrium

Linkage disequilibrium (D) for pairs of sites was measured by calculating the deviation of the observed frequency of the haplotype from that expected from multiplication of the individual allele frequencies and expressed as a ratio of D_{max} (D/D_{max}). D_{max} (the maximum possible disequilibrium for a given pair of allele frequencies) was taken as the minimum value of rs or $(1 - r)(1 - s)$ where $D < 0$ or the minimum value of $r(1 - s)$ or $s(1 - r)$ where $D > 0$, where r is the frequency of the rarer allele at one site and s is the frequency of the rarer allele at the second site [19, 20]. The significance of the difference of D/D_{max} from 0 was calculated as a χ^2 with 1 d.f. using the equation $D^2N/[r(1 - r)s(1 - s)]$ [19].

Results

Several short regions of the lactase gene, including some in which base changes had previously been reported [6, 7], were amplified using the PCR technique. Three of these regions were found to show polymorphism in a preliminary screen of a test population of 8–10 individuals and were therefore studied further. One region spans an Alu element 5′ to the promotor of the lactase gene and the other two contain exon sequences. A previously reported *Msp*I polymorphism was localised and

detected, either by digestion of the appropriate PCR product with *Msp*I, or by Southern blot analysis of *Msp*I-digested DNA using the PCR product as probe.

Analysis of the 5'-Flanking Region of the Lactase Gene (5F PCR Product)

The 534-bp 5F product was digested with *Ava*II to give two fragments of 310 and 224 bp. These same digests were used for both SSCA and DGGE analysis.

SSCA

Initially both denatured and non-denatured samples were tested in order to distinguish the double- and single-stranded fragments. Under the conditions used the single-stranded fragments migrated more slowly than the double-stranded fragments (fig. 1). We noted that the single- and double-stranded DNA give slightly different colours when silver stained: the double-stranded DNA is browner and the single-stranded more orange. The single-stranded bands corresponding to each of the digestion products were distinguished by separate analysis of each of the products of the *Ava*II digestion (data not shown). The faster migrating components correspond to those produced from the smaller fragment. Initial screening of these fragments was carried out using the eight different gel conditions. Allelic variation was detected in the smaller single-stranded fragment on all four gel compositions when they were run in the cold room (sample 5, fig. 1). Variation was also detected in the larger fragment but only on gels containing glycerol (sample 2, fig. 1). All subsequent gels contained 5% glycerol, 1 × TBE and electrophoresis was conducted in the cold.

The variation detected in the smaller fragment was named '1/2' (frequencies 1, 0.93 and 2, 0.07) and that in the larger fragment was named '1/3' (frequencies 1, 0.98 and 3, 0.02). Polymorphism was also evident in the double-stranded DNA of the smaller fragment but this was clearly independent from the 1/2 allelic variation detected in the single-stranded DNA. The two alleles were named F and S (fast and slow) to describe their relative electrophoretic mobility. The individuals with 3 bands (lanes 2 and 5, fig. 1) are putative S/F heterozygotes. The extra bands can be attributed to heteroduplex formation between allelic strands which differ at a site which does not alter the mobility of the double- or single-stranded DNA under the conditions tested.

DGGE Analysis

Use of the computer programmes MELT87 and SQHTX [21] demonstrated two melting domains within each of the *Ava*II digestion products of the 5F PCR product. Since the GC-rich, higher melting domain would behave as a natural 'GC clamp' the same *Ava*II digests were analysed directly, without using primers containing a synthetic GC clamp. Initially, the digests were analysed using a gradient of 20–60% denaturant and clear evidence of genetic variation was obtained. The conditions were then optimised for the detection of these variations by the use of a narrower gradient and an experimental time course. The S/F polymorphism, observed in the double-stranded DNA on the SSCA gels, was resolved unequivocally by DGGE (lower bands, fig. 2a) as was the 1/3 polymorphism in the larger fragment (not shown). DGGE revealed additional polymorphism in the larger fragment, not seen by SSCA, which we have called 1/4 (upper bands, fig. 2a).

The allele frequencies observed in the CEPH population for the 1/4 polymorphism were 1, 0.85 and 4, 0.15, and for the S/F polymorphism S, 0.76 and F, 0.24. DGGE analysis and SSCA of the same series of samples from a single family are shown in figure 2.

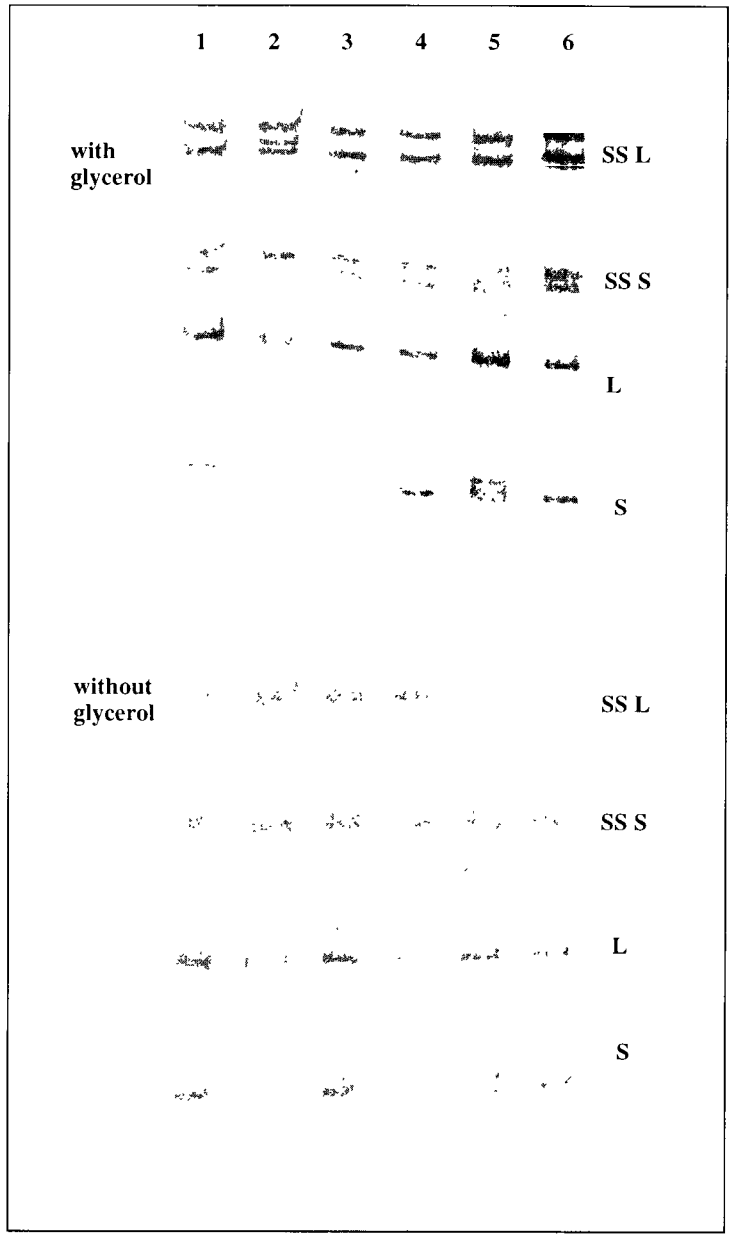


Fig. 1. Photographs of SSCA of the same series of samples under two different electrophoretic conditions (with glycerol and without glycerol, 1 × TBE in the cold). The 5F fragment was digested with AvaII and the positions of the small (S) and large (L) fragments are indicated. The single-stranded components are indicated by SS. The bands were visualised after silver staining.

Sequence Analysis of the 5F PCR Product

Sequencing of the 5F PCR products from 7 individuals of different phenotypes identified the base changes responsible for the alterations in mobility of the fragments detected by SSCA and DGGE and confirmed the existence of four different polymorphic sites. The data are summarized in table 2. The site responsible for the 1/3 polymorphism was identified by the analysis of four different hetero-

zygous individuals, because this allele has so far not been found in homozygous form. In all cases the allele was shown to carry a T at nucleotide position −957 (like the 4 allele) in addition to the substitution at position −874.

SSCA of Exon 2

The F2 PCR product which spans exon 2 was analysed under a variety of conditions with the aim of revealing the sequence polymorphism described previously at nucleotide 666 [6]. This polymorphism was revealed by use of 0.5 × TBE and glycerol in the gels and electrophoresis in the cold. An example of this polymorphism is shown in figure 3. Each homozygote shows a pattern of three single-stranded bands. The heterozygote phenotype appears to be a straightforward combination of the homozygote patterns. The observation of three bands corresponding to each of the alleles suggests that one of the strands can form two equally stable conformers. Sequence analysis confirmed the nucleotide substitution at position 666 (table 2) and showed that allele A corresponds to the presence of a G and allele B an A (frequencies A, 0.83 and B, 0.17).

Analysis of an MspI *RFLP in Exon 17*

Examination of the distribution of *Msp*I sites and the base changes observed in the published sequences [6, 22] suggested that the previously reported *Msp*I RFLP [14] might be due to variation at an *Msp*I site in exon 17.

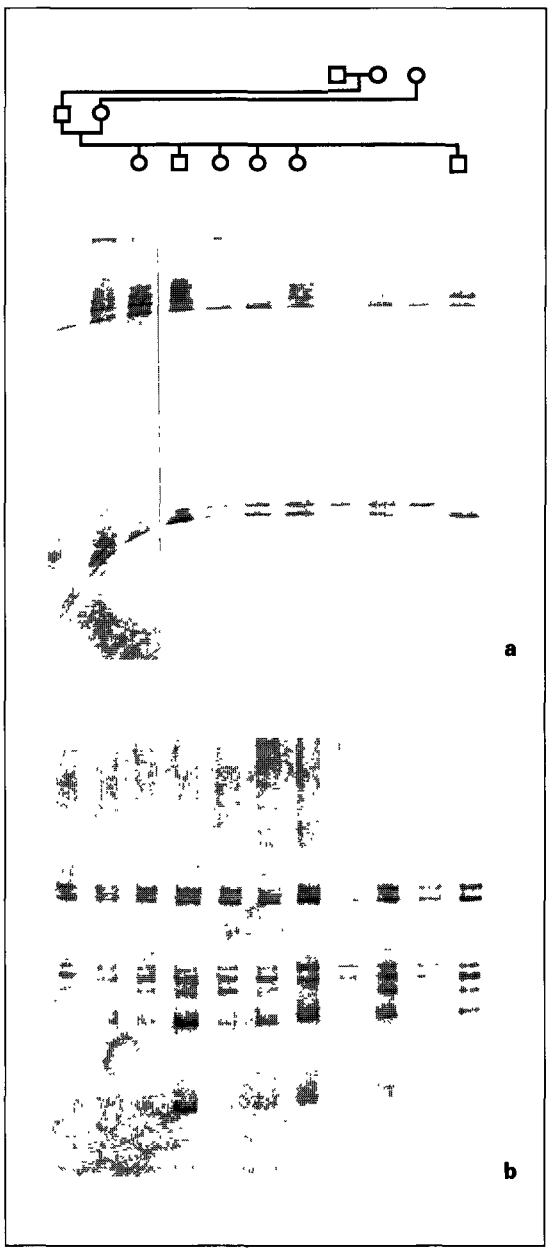


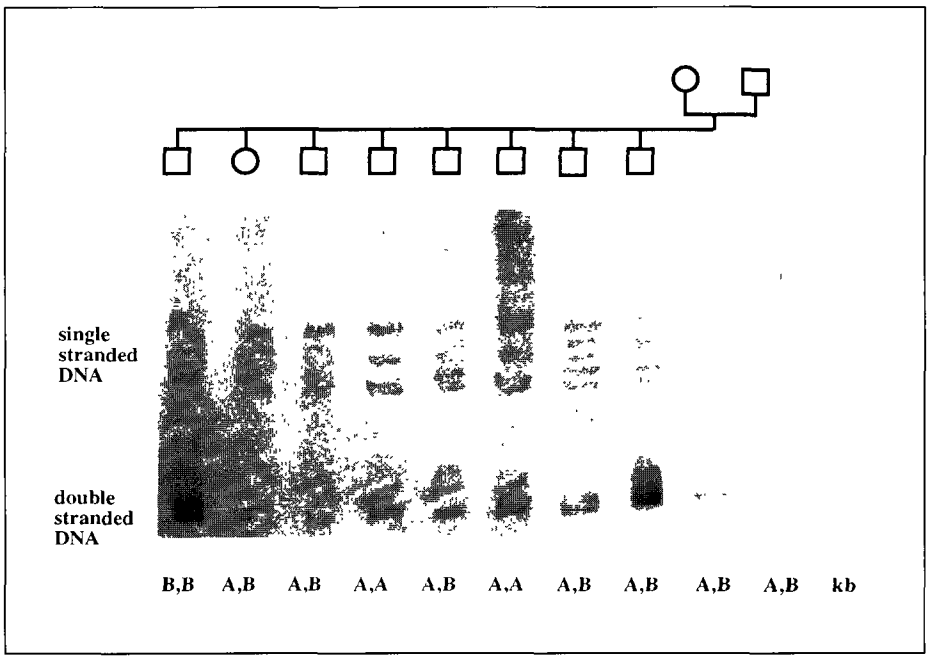
Fig. 2. Photographs of the analysis of *Ava*II digests of the 5F PCR product by DGGE (**a**) and SSCA (**b**) of the CEPH family 1447. The relationship between the people is shown in the pedigree above the gels. The tracks below a symbol in the pedigree correspond to DNA from that individual. The phenotypes assigned by DGGE analysis (**a**) were: father 1, S/F; mother 1/4, S/F; mother's mother 1, S; the last child 1/4, F (representative individuals). The phenotypes for the variation detected by SSCA (**b**) were: father 1/2, and mother 1. In individuals heterozygous for the 1/4 or S/F polymorphisms the fainter pair of bands of lower mobility may be explained as heteroduplexes between the two alleles. These were not seen in all cases since the samples were not treated to promote heteroduplex formation. The bands were visualised after silver staining.

Table 2. Nucleotide differences responsible for the polymorphisms in the *LCT* gene

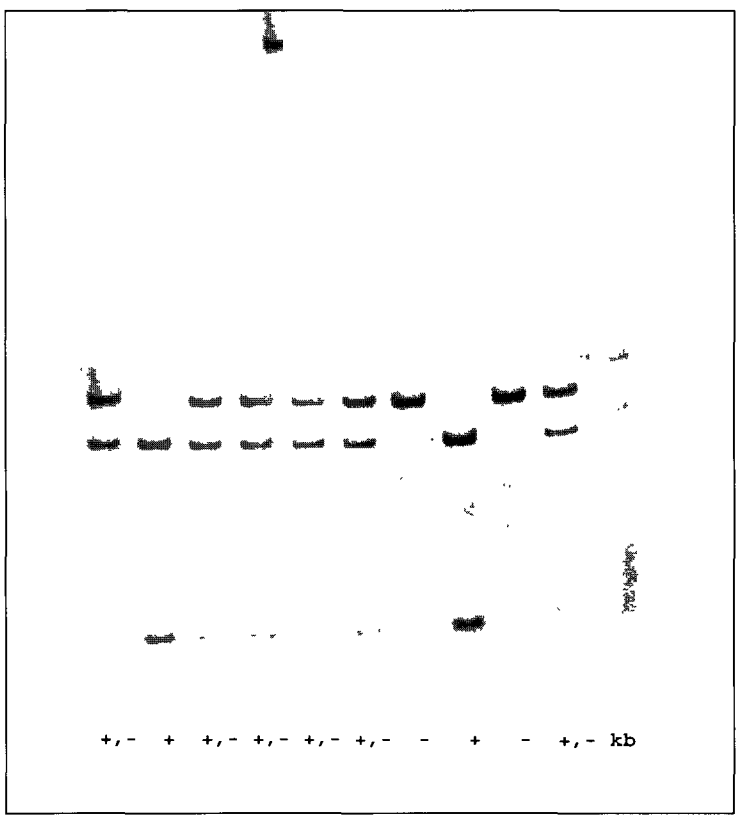
Gene position	PCR product	Allele name	nt position	Nucleotide	Detection method
5′ flanking	5F large	1	−957	C	D, S
		3		T	D, S
		4		T	D
5′ flanking	5F large	1	−875	G	D, S
		3		A	D, S
		4		G	D
5′ flanking	5F small	1	−678	A	S
		2		G	S
5′ flanking	5F small	S	−552 to	AA	D, S
		F	−559	A	D, S
Exon 2	F2	A	666	G	S
		B		A	S
Exon 17	LCT3 and F17/LCT	+	5579	C	R
		−		T	R
Exon 17	UT	Δ	6236/7	ΔΔ	A
		I		GT	A

The nucleotide positions given for each site are taken from the cDNA sequence, or in the case of the negative numbers indicate the position upstream from the start of transcription. The detection methods used to analyse each change are identified using the code: S = SSCA; D = DGGE; R = RFLP, and A = non-denaturing acrylamide gel electrophoresis. The nucleotide differences in the 5F and F2 PCR products were determined by sequencing as described. The substitution causing the *Msp*I RFLP polymorphism (+/−) was deduced from the recognition site of *Msp*I, and the polymorphism in the UT fragment was assumed to be that described previously.

This hypothesis was tested by *Msp*I digestion of PCR products (LCT3 and F17/LCT) spanning the relevant region. In each case, two alleles were observed, one where the PCR product was not digested and the other in which it was digested. The PCR product LCT3 generates digestion fragments of 184 and approximately 1,250 bp, whereas the F17/LCT product gives fragments of 184 and 60 bp (fig. 4). The LCT3 PCR product was also used as a probe on Southern blots of *Msp*I-digested genomic DNA. Two bands of approximately 5–6 kb were distinguished which differed in size by approximately 200–300 bp, consistent with the previously reported polymorphism. The CEPH samples were, in most cases, tested by Southern blot analysis of *Msp*I-digested genomic DNA and probing with the undigested PCR product. Some samples were analysed by digestion of the LCT3 PCR products. The results obtained were in complete agreement with the *Msp*I polymorphism data already on the CEPH data base [Kruse et al, unpublished]. In the



3



4

Fig. 3. Photograph of a gel showing SSCA of the F2 PCR product in samples from CEPH family 17. Bands corresponding to the double- and the single-stranded DNA are indicated. The genotypes shown are deduced from the family structure. The track labelled 1 kb contains the kilobase ladder molecular weight markers (BRL). The bands were visualised after silver staining.

Fig. 4. Photograph of a gel showing the *Msp*I polymorphism in the F17/LCT fragment. The bands were visualised after silver staining.

Fig. 5. Photograph of a gel showing the three phenotypes detectable for the ΔI polymorphism in the UT PCR product from a selection of unrelated individuals. I indicates the presence of the insertion and Δ the deletion. Approximately 5 ng of PCR product was loaded in each track. The bands were visualised after silver staining.

few cases where samples were temporarily unavailable to us, the results already on the data base were used to complete the haplotypes. The allele frequencies observed were 0.78 for *Msp*I+ and 0.22 for *Msp*I–.

Analysis of the UT Product (Exon 17)

The previously reported deletion (Δ)/insertion (I) of two base pairs at nt 6,236/7 [6] was found to be distinguishable by simple nondenaturing polyacrylamide electrophoresis, provided that the amount of DNA loaded was reduced to approximately 5 ng. It is noteworthy that heteroduplexes can be detected in the heterozygotes. An example of this analysis is shown in figure 5. The allele frequencies observed were 0.83 for I and 0.17 for Δ.

Haplotype Determination

Analysis of each of these polymorphisms in the CEPH families confirmed that they are inherited in a Mendelian fashion. All seven polymorphisms were linked, showing a high lod score and no recombination. In the case of the 5F polymorphisms and the *Msp*I polymorphism, which are located at opposite ends of the gene, all individuals in the informative families were tested and the lod score was 32 at $\theta = 0$. Subsequent analysis of the haplotypes was conducted assuming no recombination. Analysis of the variation in the 5F fragment revealed that the 4 allele and the 2 allele always occurred on different chromosomes (fig. 2) but that these each showed complete association with the F allele at the fourth site in this fragment (S/F), thus generating three haplotypes (table 3a). The rarer 3 allele was shown to be related to the 4 allele by carrying an additional nucleotide substitution in the same DNA fragment and generated a fourth haplotype. Clear patterns of associations were also found with and between each of the other individual sites (table 3a, b). In particular it is noteworthy that the S/F polymorphism (in the 5′-flanking region) shows the highest level of association with the *Msp*I polymorphism (exon 17), while the F2 polymorphism (exon 2) shows the highest level of association with the UT polymorphism (exon 17). Although this analysis revealed a number of rare haplotypes, there were only three common haplotypes in the CEPH population, with the haplotype that carries the 3 allele being the fourth most frequent. All the haplotypes observed are shown in table 4 together with the ten haplotypes that might have been expected but were not observed. It can be seen that the frequencies of these haplotypes deviate significantly from those expected by random association of the alleles. Calculation of the link-

Table 3. Associations between the alleles located at different sites in the *LCT* gene

a Association of the 2, 3 and 4 alleles at three of the polymorphic sites in the 5F fragment with the F and S alleles at the fourth site in this fragment

S/F	5F haplotype			
	1	2	3	4
S	182			
F		17	4	37

In each case the chromosomal origins were deduced from the family data. This table is presented to show the four haplotypes in the 5F fragment (see text). The chromosomes assigned the haplotype 1 do not carry any one of the 2, 3 or 4 alleles.

b Pairwise comparisons of the alleles at the 6 most polymorphic sites

F2	5F 1/2	
	1	2
A	182	17
B	41	

*Msp*I	5F S/F	
	S	F
+	181	6
–	1	52

I/Δ	5F 1/4	
	1	4
I	198	2
Δ	1	39

*Msp*I	5F 1/2	
	1	2
+	184	3
–	39	14

I/Δ	5F S/F	
	S	F
I	181	19
Δ	1	39

*Msp*I	Exon 2	
	A	B
+	184	3
–	15	38

I/Δ	5F 1/2	
	1	2
I	183	17
Δ	40	

F2	5F 1/4	
	1	4
A	199	
B		41

I/Δ	Exon 2	
	A	B
I	198	2
Δ	1	39

F2	5F S/F	
	S	F
A	182	17
B		41

*Msp*I	5F 1/4	
	1	4
+	184	3
–	15	38

I/Δ	Exon 17 MspI	
	+	–
I	185	15
Δ	2	38

In each case the chromosomal origins were deduced from the family data. The 3 allele which also carries the '4' substitution (–957C/T) is considered together with the 4 allele for this analysis. See table 2 and figure 6 for further information on the location of these polymorphisms.

age disequilibrium parameter (D/D_{max}) for each pair of sites is depicted in figure 6 and reveals that there is a high level of disequilibrium across the region with no hint of any correlation with distance.

The CEPH families which share the common characteristic of large sibships come from various sources. They comprise a group from France, a large group collected by Dr. Ray White's laboratory in Utah and a few assorted

Table 4. Frequencies of the haplotypes observed in the CEPH population in comparison with those expected from random assortment of the alleles

PCR product				Haplotype code	Observed number	Expected number
5F	F2	LCT3	UT			
1 1 1 S	A	+	I	A	181	74.5
4 1 1 F	B	–	Δ	B	33	<0.1
1 1 2 F	A	–	I	C	14	0.6
4 3 1 F	B	–	Δ	D	4	≪0.1
1 1 2 F	A	+	I	E	3	2.2
4 1 1 F	B	+	Δ	F	2	0.2
4 1 1 F	B	–	I	G	1	0.3
1 1 1 S	A	–	Δ	H	1	4.3
4 1 1 F	B	+	I	I	1	1.0
1 1 1 F	A	+	I		0	23.5
1 1 1 S	A	–	I		0	21.1
1 1 1 S	B	+	I		0	15.4
1 1 1 S	A	+	Δ		0	15.4
4 1 1 S	A	+	I		0	14.6
1 1 2 S	A	+	I		0	7.0
1 1 1 F	A	–	I		0	6.7
1 1 1 S	B	–	I		0	4.3
4 1 1 S	A	–	I		0	4.1
1 3 1 S	A	+	I		0	1.9

The numbers expected and observed of the nine haplotypes detected in the CEPH population are shown. The 10 other haplotypes that might have been expected if there was random assortment of the alleles are also shown in the lower part of the table. The haplotypes are described using the allele symbol notations and these are listed within the relevant PCR products in the order in which they are found 5′ to 3′ in the gene.

others. It was very noticeable that the frequencies of the two commonest haplotypes A and B differ significantly ($p < 0.001$, by Fisher's exact test) in the two major groups (table 5).

Discussion

Seven different polymorphisms in the lactase gene were analysed in this study. Their relative positions in the gene can be seen in figure 6. Four sites were in the 534-base pair region (5F) located between −997 and −464, upstream from the start of transcription. Previous sequence analysis had revealed two sites within this region that showed sequence differences in 1 of 10 chromosomes examined [7]. This study shows that these substitutions (−957C/T and −552/9AA/A) are indeed responsible for the variation we detected and correspond to the 1/4 and S/F polymorphisms, respectively. The 1/3 and 1/2 polymorphisms, on the other hand, represent new sites (−875A/G and −678A/G, respectively). Sequence analysis of exon 2 by Boll et al. [6] previously demonstrated a polymorphic site at nucleotide position 666 in the cDNA, 7 chromosomes possessing G and 4 having A at this site. This polymorphism results in a valine to isoleucine change in the pre-pro-protein (V/I219). The

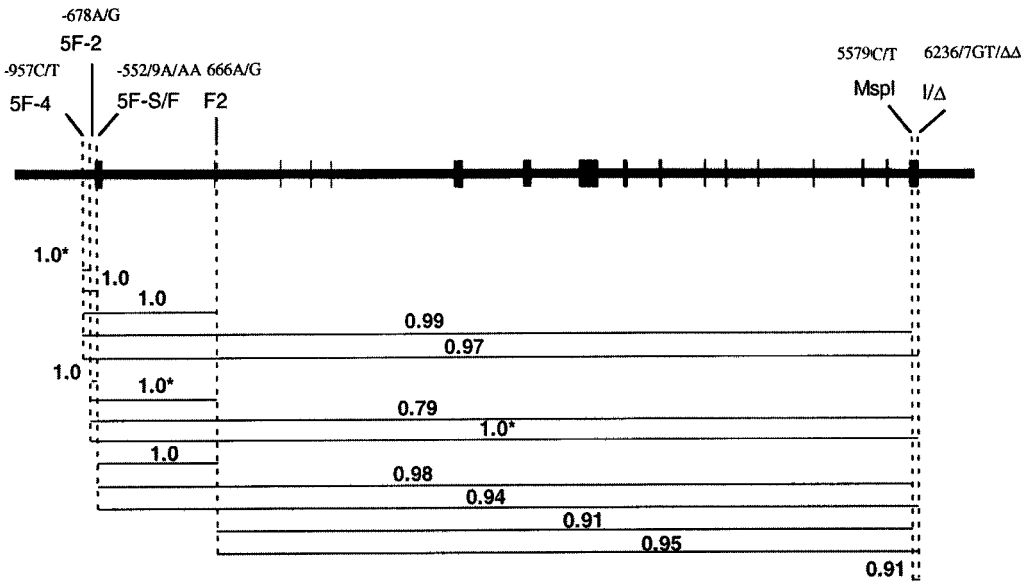


Fig. 6. Diagrammatic representation of the linkage disequilibrium (expressed as D/D_{max}) across the lactase gene. At the top is a schematic representation of the lactase gene showing the gene structure and the positions of the polymorphic sites. Shown below are the values obtained for D/D_{max} for the pairs of sites. The 3 allele is excluded from this analysis because the sample size is too small. * Not significantly different from 0 ($p > 0.01$).

SSCA conditions described here allow the simple detection of this polymorphism, allele A corresponding to the presence of a G at this site and allele B corresponding to an A. Boll et al. [6] also observed a GT duplication 7 nucleotides upstream of the putative polyadenylation signal in 7 chromosomes of the 11 sequenced [6]. The simple PAGE conditions described here allow the detection of this variation as a mobility difference. We also located the previously reported *Msp*I polymorphism to exon 17, which provides another positioned marker in the lactase gene.

Analysis of these 7 polymorphisms has allowed the frequencies of these alleles and the haplotypes to be determined using unrelated individuals of the CEPH panel. In the 240 chromosomes analysed only 9 of the possible 128 (2^7) haplotypes were observed and only 3 of these were common. 10 other haplotypes predicted to occur at high frequency were not seen. The region of linkage disequilibrium extends across the whole gene (60–70 kb).

The particularly high level of association between alleles in the F2 (666A/G) and UT (6,236/7GT/ΔΔ) fragments and between S/F in the 5F fragment (−552/9A/AA) and the *Msp*I site (5,579C/T) is interesting, since these two regions overlap and each span 50–60 kb. These overlapping associations mean that it is

Table 5. Comparison of the frequencies of the haplotypes observed in the two major subpopulations of the CEPH series

Haplotype notation	Code	French n = 48	French freq	Utah n = 176	Utah freq	Others n = 16	Total n = 240	Total freq
111S A + I	A	28	0.583	142	0.807	11	181	0.754
411F B – Δ	B	13	0.27	17	0.097	3	33	0.138
112F A – I	C	3	0.063	9	0.051	2	14	0.058
431F B – Δ	D	2	0.042	2	0.011	–	4	0.017
112F A + I	E	1	0.021	2	0.011	–	3	0.013
411F B + Δ	F	1	0.021	1	0.006	–	2	0.008
411F B – I	G	–	0	1	0.006	–	1	0.004
111S A – Δ	H	–	0	1	0.006	–	1	0.004
411F B + I	I	–	0	1	0.006	–	1	0.004

'Utah' signifies the families collected by Dr. Ray White's laboratory which were mostly, but not all, collected locally in Utah (see text). 'Others' comprises a family from Venezuela and one Amish family. The haplotype notation used is the same as in table 4.

not easy to hypothesise the possible evolutionary phylogeny of the three common haplotypes. There is no suggestion that reciprocal recombination was involved, and it is tempting to implicate gene conversion. However, the D haplotype presumably arose from the B haplotype due to an additional, more recent, point mutation at nt –875.

It was noteworthy that the B haplotype differs markedly in frequency between the French and Utah families. It was thus also of some interest that the distribution of the B haplotype among the Utah families was uneven, 5 of the 17 chromosomes being found in one family. Most of the Utah families come from the local community. This relatively recent population which originated largely from deliberate colonisation by the Church of the Latter Day Saints (Mormons) during the second half of the last century came mainly from eastern USA, UK and Scandinavia. The pronatalist policy of the Mormon Church encourages large sibships and has made available many suitable families for genetic study. The studies of McLellan et al. [23], based on the analysis of multiple polymorphic enzyme markers and blood groups, have indicated that the Utah population is representative of northern Europe. It was thus of some considerable interest to discover from Dr. R. White and Dr. M. Leppert that the family which carries 5 B haplotype chromosomes is one of a few families that was not collected in Utah and is probably of central Europe origin.

The existence of a large region of linkage disequilibrium means that if the sequence which determines the lactase persistence polymorphism is indeed cis-acting it is possible that lactase persistence will show some association with the DNA polymorphisms even if the relevant sequence is located at some distance from the gene itself. Association studies are therefore underway, as well as other genetic studies, to locate the polymorphism which determines lactase persistence status. It will be of interest to determine the haplotypes in other populations and also in higher primates, since this may help towards an understanding of the evolutionary history of the lactase gene and the lactase persistence polymorphism.

Acknowledgements

We would like to thank S. Jeremiah for operating the DNA extractor and Dr. P. Johnson for help in setting up the DGGE analysis. We also thank Dr. N. Mantei for helpful discussions and supplying us with clones and control DNA, Dr. H. Cann and Prof. J. Dausset for the CEPH DNA samples, Dr. T.A. Kruse for unpublished data, Drs M. Leppert and R. White for the information about the origin of family 1427, and Prof. E.B. Robson for support and encouragement. C.B.H. was supported by the MRC on a Human Genome Mapping Project studentship, and we received additional support from EUROGEM.

References

1 Sahi T: The inheritance of selective adult-type lactose malabsorption. Scand J Gastroenterol 1974;9:1–73.
2 Swallow DM, Harvey CB: Genetics of adult-type hypolactasia. Dyn Nutr Res 1993;3:1–7.
3 Flatz G: Gene dosage effect on intestinal lactase activity demonstrated in vivo. Am J Hum Genet 1984;36: 306–310.
4 Harvey CB, Wang Y, Hughes LA, Swallow DM, Thurrell WP, Sams VR, Barton R, Lanzon-Miller S, Sarner M: Studies on the expression of intestinal lactase in different individuals. Gut 1995;36:28–33.
5 Ho M-W, Povey S, Swallow D: Lactase polymorphism in adult British natives: Estimating allele frequencies by enzyme assays in autopsy samples. Am J Hum Genet 1982;34: 650–657.
6 Boll W, Wagner P, Mantei N: Structure of the chromosomal gene and cDNAs coding for lactase-phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. Am J Hum Genet 1991;48: 889–902.
7 Lloyd M, Mevissen G, Fischer M, Olsen W, Goodspeed D, Genini M, Boll W, Semenza G, Mantei N: Regulation of intestinal lactase in adult hypolactasia. J Clin Invest 1992;89: 524–529.
8 Sebastio G, Guzzetta V, De Vizia B, Ballbaio A, Boll W, Mantei N, Semenza G, Auricchio S: Genetic study of human adult-type hypolactasia by analysis of RFLPs of the lactase gene. Pediatr Res 1990;27:532.
9 Escher JC, de Koning ND, Van Engen CGJ, Arora S, Buller HA, Montgomery RK, Grand RJ: Molecular basis of lactase levels in adult humans. J Clin Invest 1992;89:480–483.
10 Witte J, Lloyd M, Lorenzsonn V, Korsmo H, Olsen W: The biosynthetic basis of adult lactase deficiency. J Clin Invest 1990;86:1338–1342.
11 Fajardo O, Naim HY, Lacey SW: The polymorphic expression of lactase in adults is regulated at the mRNA level. Gastroenterology 1994;106:1233–1241.
12 Orita M, Suzuki Y, Sekiya T, Hayashi K: Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. Genomics 1989;5: 874–879.
13 Myers RM, Fischer SG, Lerman LS, Maniatis T: Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis. Nucleic Acids Res 1985;13:3131–3145.
14 Kruse TA, Bolund L, Byskov A, Sjostrom H, Noren O, Mantei N, Semenza G: Mapping of the human lactase-phlorizin hydrolase gene to chromosome 2. Cytogenet Cell Genet 1989;51:1026.
15 Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J-M, White R: Centre d'étude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics 1990;6:575–577.
16 Myers RM, Maniatis T, Lerman LS: Detection and localization of single base changes by denaturing gradient gel electrophoresis. Methods Enzymol 1987;155:501–527.
17 Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain termination inhibitors. Proc Natl Acad Sci USA 1977;74:5463–5467.
18 Maynard-Smith S, Penrose LS, Smith CAB: Mathematical Tables for Research Workers in Human Genetics. London, Churchill, 1961.
19 Elbein SC: Linkage disequilibrium among RFLPs at the insulin-receptor locus despite intervening Alu repeat sequences. Am J Hum Genet 1992;51:1103–1110.
20 Thompson EA, Deeb S, Walker D, Motulsky AG: The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. Am J Hum Genet 1988;42:113–124.
21 Lerman LS, Silverstein K: Computational simulation of DNA melting and its application to denaturing gradient gel electrophoresis. Methods Enzymol 1987;155:482–501.
22 Mantei N, Villa M, Enzler T, Wacker H, Boll W, James P, Hunziker W, Semenza G: Complete primary structure of human and rabbit lactase-phlorizin hydrolase: Implications for biosynthesis, membrane anchoring and evolution of the enzyme. EMBO J 1988;7:2705–2713.
23 McLellan T, Jorde LB, Skolnick MH: Genetic distance between the Utah Mormons and related populations. Am J Hum Genet 1984;36: 836–857.