Molecular and population genetic analyses of variation within and surrounding the human lactase gene

Edward John Hollox

A thesis submitted for the Doctor of Philosophy degree at the University of London

January 2000

MRC Human Biochemical Genetics Unit
The Galton Laboratory
Department of Biology
University College London

Abstract

This thesis describes an investigation of the variation within and around the human gene coding for the enzyme lactase-phlorizin hydrolase. The two major aims were to analyse sequence variation within and adjacent to the gene to study the extent and nature of diversity within and between populations, and to provide insight into the developmental and molecular bases of the lactase persistence polymorphism.

The *cis*-acting effect of lactase persistence was confirmed using a panel of 15 children and quantification of lactase mRNA from intestinal biopsies which showed variable and progressive downregulation of the lactase gene.

In order to extend sequence analysis a fosmid and cosmid contig was constructed across the lactase gene. Intron 1 was sequenced, and part of this sequence together with 1.8kb upstream and 2kb downstream of the gene was analysed in a panel of adult individuals of known lactase persistence status. The upstream region was analysed in six other species and is conserved in between humans, pigs and primates. Electromobility shift assays were done on a conserved region containing several human variants to analyse protein-binding activity of that region. Fluorescence in situ hybridisation analyses of clones from a PAC contig was used to confirm the physical map of a ~500kb region containing the lactase gene.

11 polymorphic sites were used to construct haplotypes in 596 individuals from 12 populations, and association with lactase persistence/non-persistence tested in cohorts of Finnish and Yakut individuals. A global picture of linkage disequilibrium and haplotype variation was formed which showed four common haplotypes in all non-African populations. Analysis of two sub-Saharan African populations showed these four haplotypes together with many more other haplotypes. This supports the 'Out of Africa' theory for the origin of modern humans.

Contents

Figures

Tables

Acknowledgements

I would like to thank my supervisor, Dallas Swallow, for continuous advice, support and encouragement throughout my PhD. I would also like to thank Mark Poulter for help, discussions, and teaching me practical science.

I also thank the many people, here and abroad, that have been involved in this work. Dr Vladimir Ferak and Mr Marek Svalik from Comenius University, Bratislava kindly provided me with Roma samples, and performed some of the typing of those samples. Professor Trefor Jenkins at the University of the Witwatersrand in South Africa provided useful comments on part of the manuscript, and he and Dr Amanda Krause provided the San and Bantu samples. Dr Nilmani Saha from the National University of Singapore provided all Indian samples, together with Malay and Chinese sample sets. The ever-helpful Yeti-hunting Dr Andrew Kozlov in Moscow provided all samples from the Russian Federation, and all Yakut lactose tolerance data were provided by Professor Maria Khandy and Ms Sardana Markova of the Yakutsk State University. Finnish samples were provided courtesy of Dr Riita Korpela of the University of Helsinki. The Japanese samples were provided by Dr Geoff Daniels and the Papua New Guinean samples by Dr Phil Johnson.

Within the Unit, I'd like to thank my second supervisor, Dr David Whitehouse, for valuable discussions, and Professor Yvonne Edwards for the gorilla, macaque and orangutan samples. Dr Felicity Drummond provided oligonucleotides for EMSA analysis, and, together with Dr Harry Papapetrou, provided valuable advice on this technique. Two former members of the lab, Drs Yangxi Wang and Clare Harvey, also provided advice. Drs Margaret Fox and Rosemary Ekong taught me the techniques of FISH and Strand-FISH. Daniella Estrin characterised the intron 1 polymorphisms, and Helen Bond helped construct two by two tables for allelic association analysis. Finally, I'd like to thank everyone in the Galton Laboratory, especially Dha, Eleni and Jamie, for making it a great place to work.

Abbreviations

AS-PCR	Allele-Specific Polymerase Chain Reaction
BAC	Bacterial Artificial Chromosome
cDNA	Complimentary DNA
CE-LPH	*Cis*-Element Lactase Phlorizin Hydrolase
CEPH	Centre d'Etude de Polymorphisme Humain
DGGE	Denaturing Gradient Gel Electrophoresis
DM	Myotonic Dystrophy
DNA	Deoxyribose Nucleic Acid
ECCAC	European Collection of Animal Cell Cultures
EMSA	Electromobility Shift Assay
FISH	Fluorescent In-Situ Hybridisation
HC	Haplotype Counting
LCR	Locus Control Region
LCT	Lactase
LCT*N	Lactase non-persistence allele
LCT*P	Lactase persistence allele
LD	Linkage Disequilibrium
LDC	Lactose Digestion Capacity
LPH	Lactase Phlorizin-Hydrolase
LPL	Lipoprotein Lipase
ML	Maximum Likelihood
mRNA	messenger RNA
NF-LPH	Nuclear Factor Lactase Phlorizin Hydrolase
OMIM	On-line Inheritance In Man
PAC	P1 Artificial Chromosome
PCR	Polymerase Chain Reaction
PSL	Photo-Stimulated Luminescence
RNA	Ribonucleic Acid
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SI	Sucrase Isomaltase
SNP	Single Nucleotide Polymorphism

SSCA	Single Stranded Conformation Analysis
UK-HGMP	United Kingdom Human Genome Mapping Project
YAC	Yeast Artificial Chromosome

1 Introduction

This thesis concerns analysis of variation across the human lactase gene using both molecular and population genetic approaches. In this introductory chapter, section 1.1 describes the lactase persistence/non-persistence polymorphism in humans, from clinical consequences through genetics of the trait to evolutionary scenarios. Sections 1.2 and 1.3 describe current knowledge about the biochemistry and molecular biology of lactase in humans and other animals. Sections 1.4 and 1.5 describe current molecular biological techniques for investigating gene expression and give an overview on the developmental control of two well-characterised genes. Finally section 1.6 describes how study of human polymorphisms can reveal the history and evolution of our species, and section 1.7 describes the history and milk-drinking habits of the populations studied in this thesis.

1.1 Lactase persistence polymorphism

1.1.1 Lactose intolerance

1.1.1.1 Causes of lactose intolerance

Lactose intolerance is a clinical term for the occurrence of gastrointestinal symptoms after ingestion of lactose, or food containing lactose, by individuals with a low lactose digestion capacity (LDC). Symptoms include meteorism, borborgymi, flatulence and diarrhoea, and occasionally nausea and vomiting. Low LDC is caused by low levels of the enzyme lactase located on the luminal surface of the epithelium of the small intestine. This can be due to gastrointestinal disease such as acute and chronic enteritis, tropical sprue, severe protein malnutrition, and can be recognised by the changes in intestinal morphology. It can also be due to the natural down-regulation of lactase after weaning. The symptoms are caused by undigested lactose increasing the osmotic pressure, which causes water to diffuse into the gut. The resulting increase in volume stimulates peristalsis and the unabsorbed lactose passes into the colon, where bacteria metabolise the lactose to form carbon dioxide and hydrogen (for review see Hollox and Swallow, 2000).

Not all individuals with low LDC experience symptoms of lactose intolerance after ingestion of lactose, and the severity of symptoms vary between individuals.

Additional variables include the pH and consistency of food and small intestinal passage time. It has also been found that that faecal β-galactosidase activity rises after prolonged exposure to lactose in individuals with low LDC, suggesting that bacteria adapt to higher amounts of lactose (Briet *et al.*, 1997).

When lactase deficiency was first identified, it was regarded as an abnormality and lactase activity in adulthood was regarded as normal. It was assumed that low lactase activity in adults was a rare deficiency because of the fact that most people tested were of Northern European origin, where lactase activity is usually high in adults. However it soon became clear that presence or absence of lactase activity in adulthood was polymorphic, and the two phenotypes were termed lactase persistence and lactase non-persistence (Auricchio *et al.*, 1963; Dahlqvist *et al.*, 1963).

1.1.1.2 Lactose tolerance tests

Lactose tolerance tests can be used to measure lactose digestion capacity (LDC), and hence, in individuals with no gastrointestinal disease, LDC can be used to define lactase persistence/non-persistence status. There are several types of test, but all rely on measuring a physiological trait before and after an oral challenge of 50g of lactose in water.

The most common is the blood glucose test, which measures blood glucose levels after an overnight fast but prior to ingestion of lactose, and several times after lactose ingestion (Dahlqvist, 1974). If the individual had high LDC, the lactose is hydrolysed and the resulting galactose and glucose absorbed into the blood. The rise in blood glucose level is recorded and high LDC is diagnosed if a rise of 1.4 mmol/l or more is recorded, and low LDC is diagnosed if the rise, if any, is less than 1.1mmol/l. For a rise of between 1.1mmol/l and 1.4mmol/l a repeat test, if possible, is performed.

The breath hydrogen test, as the name suggests, measures the hydrogen concentration of exhaled breath after lactose ingestion. Individuals with low LDC have high levels of hydrogen due to undigested lactose being fermented in the colon to produce hydrogen (see above). The hydrogen diffuses into the blood and then out into the lungs (Howell *et al.*, 1981; Metz *et al.*, 1976).

The urinary galactose test measures the level of galactose in the urine, and is very similar in principle to the blood glucose test (Jussila, 1969). Individuals with high LDC will have high levels of galactose as the lactase is hydrolysed and the glucose and galactose is absorbed by the gut. The liver metabolises galactose to glucose and this is inhibited by an oral administration of ethanol, so the urine galactose level represents the amount absorbed in the gut.

1.1.2 Evidence for a genetic basis for lactose intolerance

1.1.2.1 Evidence from family studies and twin studies.

Family study data are almost entirely restricted to those obtained from lactose tolerance tests. The most convincing family data are from several large Finnish families (Sahi *et al.*, 1973). These together with many small families from different populations show autosomal recessive inheritance of lactase non-persistence (low LDC) (reviewed in Swallow and Harvey, 1993). A twin study in Hungarians showed the level of concordance and discordance among the dizygotic twins agreed with expected values, confirming the genetic basis for variable LDC (Metneki *et al.*, 1984).

1.1.2.3 Evidence from studies on intestinal lactase activity

Direct examination of lactase activities in unrelated individuals provided more evidence for the genetic basis of variable lactase levels. Disaccharidase activity ratios were measured, which provided a more accurate diagnosis of lactase persistence or non-persistence. Jejunal samples from autopsies showed a trimodal distribution of sucrase/lactase ratios in Hardy-Weinberg proportions (Ho *et al.*, 1982), and this was confirmed on a different data set using biopsies and a lactase/maltase ratio (Flatz, 1984). The heterozygotes showed intermediate levels of lactase activity, consistent with a gene dosage effect.

It is now accepted that high and low LDC, and hence lactase persistence and non-persistence, are determined by two alleles, termed LCT*P and LCT*R. Because LCT*P/LCT*R heterozygotes have enough lactase to digest 50g of lactose, they are diagnosed as lactase persistenct and so the persistence allele is dominant.

1.1.3 Evolution of lactase persistence

1.1.3.1 Global distribution of lactase persistence and non-persistence

The frequency of lactase persistence varies between populations from 0% in Thais to 100% in Dutch. Allele frequencies of the populations studied in this thesis are shown in figure 6.2, and a more complete version can be found in Flatz (1987) and

Swallow and Hollox (2000). In summary, lactase non-persistence is the commonest phenotype in native populations of Australia and the Pacific islands, Eastern and Southeastern Asia, sub-Saharan Africa, and North and South America. Lactase persistence is the commonest phenotype in Central and Northern Europeans, and nomadic populations in arid areas of North and Central Africa and Arabia that depend heavily on milk. A cline in allele frequencies is observed from north to south across Europe and west to east across Asia. As well as the midpoints on these clines, intermediate allele frequencies are found in North Indians and in populations with recent mixture of high and low LDC peoples, such as American Mexicans and European-Inuit.

1.1.3.2 Distribution of the lactase phenotypes and natural selection.

The strong correlation between populations with a high frequency of lactase persistence and populations that have a tradition of production and consumption of fresh milk has interested geneticists and anthropologists. Dairying was a technique developed after the domestication of mammals which occurred around 6000 to 9000 years ago in the Fertile Crescent area of the Near East (Simoons, 1971). The development of agriculture by domesticating animals and plants resulted in significant changes for the human population, who had previously lived as hunter-gatherers (Sheratt, 1981). As well as a change in diet, exemplified by the consumption of milk and milk products, farming allowed settlements to be developed, which provided ideal conditions for rapid spread of infectious diseases. Some of these diseases, such as smallpox and tuberculosis, developed from diseases of the newly domesticated mammals that were now living in close proximity with humans. Settlement and farming allowed division of labour and large family sizes which created a rapidly expanding population interspersed with population crashes due to famine or disease (Diamond, 1998). These strong selective pressures undoubtedly are reflected in allele frequencies of human populations today, and it is thought that under these conditions, lactase persistence rose to high frequencies in certain populations.

Milk is rich in carbohydrate and protein, and is an excellent source of nutrition for individuals who can digest it. It is also an important source of water in arid regions (Cook, 1978). Camels can survive many days without water, surviving partly on the 'waters of respiration' provided by metabolising fat from the hump.

Human nomads, not having this metabolic adaptation themselves, can benefit from it by drinking camel's milk. Desert nomads also drink sheep and goat's milk, and for the Beja of Sudan, milk can be the only food for months and consumption of up to three litres a day is not unusual (Habte *et al.*, 1973). As mentioned above, a symptom of lactase non-persistence is diarrhoea, which causes water loss. Thus for these desert nomads there would be a very strong selective advantage for digesting lactose.

In Northern Europe mixed farming resulted in less dependence on milk (Ammermann and Cavalli-Sforza, 1971), and so a 'calcium absorption' hypothesis has been suggested as a selective factor to explain the high frequency of lactase persistence (Flatz and Rotthauwe, 1971). Rickets and osteomalacia were common problems in areas such as Northern Europe with low solar irradiation, and any allele reducing the risk of these diseases would be selected for. Milk is not only rich in calcium, but is rich in lactose which increases absorption of calcium in the gut, and it has been shown that individuals with high LDC absorb more calcium than than individuals with low LDC (Kocian *et al.*, 1973).

However, several populations with low levels of lactase persistence consume large amounts of milk (such as the Mongols, Herero and Dinka). This is achieved by a cultural adaptation of fermenting fresh milk into milk products such as cheese, yoghurt and kumiss. These foods have lower levels of lactose due to both lower levels of lactose because of β-galactosidase digestion by the fermenting organism, and are better tolerated because of retarded gastrointestinal passage (Kolars *et al.*, 1984; Savaiano *et al.*, 1984).

It is also clear that some low LDC individuals can adapt to a diet including a certain amount of lactose by an increase in bacterial β-galactosidase activity in the gut, as mentioned in section 1.1.1.1.

1.1.3.3 Modelling the evolution of lactase persistence

One study analysed the environmental pressures which may be responsible for the different frequencies of lactase persistence in different populations (Holden and Mace, 1997). In order to control for the effect of genetic relatedness between populations, both a genetic and linguistic tree of populations was used in a comparative method. Maximum likelihood statistics supported selection for milk drinking by increased nutrition, but not by increased water consumption or calcium consumption. The model also supported the theory that present lactase persistence allele frequencies arose after the advent of dairying.

Mathematical models have also been constructed to attempt to simulate the selective pressures and explain how modern allele frequencies were reached in 6000-9000 years of selective pressure. Bodmer and Cavalli-Sforza (1976) were the first to suggest a model, and estimated the selection coefficient required for modern frequencies of lactase persistence. The selection coefficient s is the selective advantage of an allele, so if s=0.1 then the allele in question has a 10% higher fitness than the other allele. The first domestic livestock arrived in Europe only 3500 years ago, and for the allele to reach contemporary frequencies it was estimated, using a similar model, that a selection coefficient as high as 0.07 is necessary (Flatz and Rotthauwe, 1971). Assuming a frequency of 50% for the lactase persistence allele in northern Europeans, a selective coefficient (s) of 0.04 and an initial gene frequency of 0.001% (equivalent to a new mutation) are required to lead to the present frequency in 300 generations (about 10,000 years). If the initial frequency is 1%, then the selection coefficient lowers to 0.015.

The estimate of 300 generations (10,000 years) since agriculture is probably too high, and the 6000 years used in a gene-culture co-evolution model (Aoki, 1986) may be nearer the truth. Additionally, the final frequency for lactase persistence is taken as 0.70, which reflects the high frequencies found in Scandinavia. In this model the genetic variant of lactase persistence is associated with the cultural trait of drinking milk and both are mutually dependent. The simulation suggests that, with an initial starting frequency of 5% the selection coefficient is over 0.05, but this can become lower if the number of generations since dairying is higher or the initial allele frequency is higher. A later study (Feldman and Cavalli-Sforza, 1989) found very similar results. As Aoki admits, the model takes no account of genetic or cultural migration, which is likely to have had a significant role in the introduction of agriculture into Europe. The models also assume milk drinking is entirely correlated with lactase persistence frequencies, but in several mainly non-persistent populations drinking milk (as fermented products) is common. No account is taken of random genetic drift or population growth, both factors that could alter the selection coefficient (Nei and Saitou, 1986). As both archaeological and genetic information accumulate on human prehistory, these factors may be able to be taken into account.

1.1.3.4 Origins of lactase persistence

The possible selective models mentioned above all assume that lactase non-persistence is the original phenotype in early humans. The main evidence for this comes from other mammals, which show that a characteristic decline in lactase activity after weaning is the normal trait (Plimmer, 1906).

In primates, the most convincing evidence for a post-weaning decline of lactase comes from baboons (Welsh *et al.*, 1974), but a study on macaques showed that out of ten studied, nine had high LDC as adults (Wen *et al.*, 1973). Unfortunately, the intestinal lactase activity was not measured directly. The possibility that lactase may persist in other great apes has never been ruled out, given the fact that both chimpanzees and gorillas weaning late as compared to humans (1640 days and 1583 days as compared to 720 days in hunter-gatherer humans, Harvey *et al.*, 1987). Lactase persistence may even be polymorphic in these other species.

The evolutionary models proposed also assume, for a reasonable selection coefficient, a significant starting frequency of lactase persistence before agriculture. The fact that the persistence allele is found in all populations (albeit at low frequencies) suggests that the allele originated before the peopling of the Old World, and one can speculate on the nature of the selective pressures that maintained the allele when fresh milk was not consumed by adults.

If lactase persistence was reasonably common in early humans, a hypothesis has been proposed to account for the current distribution of the lactase persistence allele (Anderson and Vullo, 1994). Milk is rich in the vitamin riboflavin, and there is evidence that in flavin-deficient erythrocytes cell multiplication of malarial parasites is inhibited. This suggests that mildly flavin deficient individuals (who drink little milk) may have increased resistance to malaria (Das *et al.*, 1988), and therefore *Falciparum* malaria could select against lactase persistence. This could explain the near absence of the allele in most African populations, although recent experimental work on the Sardinian population does not support this idea (Meloni *et al.*, 1998).

1.1.4 Medical aspects of the lactase persistence polymorphism

1.1.4.1 Disease associations

Many studies have attempted to associate lactase persistence or non-persistence with a change in risk for various diseases. In several studies, failure to allow for ethnic differences within the control and patient set have cast doubt on the interpretation of the data, especially as the frequency of lactase phenotypes varies so dramatically between populations. Most studies have been on association between lactase non-persistence and osteoporosis, which show results, even in well designed studies, that are contradictory (Honkanen *et al.*, 1997; Horowitz *et al.*, 1987; Newcomer *et al.*, 1978). An association between lactase persistence, high milk consumption and cataract formation has also been suggested (Bengtson *et al.*, 1984; Simoons, 1982). High galactose levels, formed by digestion of lactose, may lead to increased likelihood of cataract formation, but considering that the liver can rapidly convert galactose to glucose, this may only be a factor in individuals with damaged livers.

1.1.4.2 Lactase in malnourished children

Malnutrition causes disaccharidase levels in the intestine to decline. Lactase is the most sensitive of all the disaccharidases, being the first to decline due to poor diet and the last to return to normal levels following adequate nutrition (Nichols *et al.*, 1997). In children, the normal decline due to the non-persistence allele can complicate the interpretation of altered disaccharidase activities as a result of gastrointestinal disease, and so this should be taken into account by clinicians.

1.1.4.3 Consequences of variation between populations

In most populations, adult lactase non-persistence is a common condition, and this is sometimes not considered by individuals or organisations from populations where most people are lactase persistent. For example, a large amount of skimmed milk powder is supplied as aid to countries where most individuals are lactase non-persistent. The individuals may, in addition, be malnourished and suffer from gastrointestinal infections. If fed milk, children as well as adults may develop severe symptoms of lactose intolerance. Despite problems of this type in the past (Hollox and Swallow, 2000; Ransome-Kuti, 1977), modern aid agencies are more aware of the potential problems of milk.

The gene dosage effect of lactase may have clinical consequences. Lactase persistent heterozygotes have lower levels of lactase than lactase persistent homozygotes, and may be more likely to suffer symptoms of lactose intolerance as a result of gastrointestinal disease. Indeed, studies on rats have suggested that brush border transporters and hydrolases, including lactase, are not present in great excess over the amount required to digest a typical disaccharide load (O'Connor and Diamond, 1999).

1.2 *Lactase phlorizin hydrolase*

Lactase phlorizin hydrolase (LPH, lactase, EC 3.2.1.23) is present in the small intestine of most mammals and is responsible for the hydrolysis of the disaccharide lactose into its two constituent monosaccharides glucose and galactose, which can be directly absorbed by the intestine. Lactose is present in high quantities in milk, and is essential for suckling mammals.

1.2.1 Structural features of the lactase gene

1.2.1.1 Cloning of the lactase gene

The human lactase gene (LCT) was cloned in 1988 by screening an intestinal cDNA library with a degenerate oligonucleotide probe derived from the partial amino acid sequence of rabbit lactase. The amino acid sequence was deduced from the cDNA, and was found to contain a signal peptide (amino acids 1-19) as expected for a brush border membrane protein, pro-lactase (amino acids 20-866), which is cleaved during intracellular processing (see section 1.1) and the part of the polypeptide that forms mature lactase (amino acids 867-1926). Amino acid residues 1883-1901 form the only highly hydrophobic stretch of sequence, long enough to span the phospholipid bilayer in a helix, to occur in the protein. This suggests that these residues anchor lactase to the brush border membrane (Boll *et al.*, 1991).

Following characterisation of a genomic clone containing lactase, the gene was shown to consist of 17 exons spanning approximately 70kb, as determined by Southern blot analysis. S_1 nuclease mapping allowed the transcription start site to be identified, and a further 1kb 5′ to the gene was sequenced on the genomic clone (Boll *et al.*, 1991).

1.2.1.2 Polymorphism of the human lactase gene

Boll *et al.* (1991) analysed the coding sequence of lactase in six individuals and found 11 synonymous and three non-synonymous allelic changes. G666A changed valine to isoleucine, G3297A changed alanine to threonine and G4927A changed methionine to asparagine. A two-base pair deletion (TG6236/7ΔΔ) and a C to G change in the non-coding region of exon 17 were also observed.

Harvey *et al.* (1995) and Lloyd *et al.* (1992) analysed the 5′ sequence in a number of individuals and found four alleles detectable by single stand conformation analysis and denaturing gradient gel electrophoresis. The variant alleles were three single nucleotide changes (C-958T, A-875G, A-678G), one of which was in an Alu repeat element, and the other was a insertion/deletion of an A nucleotide in the Alu element poly-A tail (A_8-552/-559A_9). These polymorphic sites, together with G666A, TG6236/7ΔΔ, and C5579G were tested in DNA from families from the Centre d'Etude de Polymorphism Humain (CEPH) (Harvey *et al.*, 1995). Linkage disequilibrium was observed across the whole gene, and three common haplotypes (A, B, C) together with 6 rare haplotypes (D,E,F,G,H,I) were deduced. Haplotype A was the most common in the families, followed by B then C.

1.2.1.3 Chromosomal localisation of the human lactase gene

The lactase gene was first localised to chromosome 2 by Southern blotting a panel of rat-human chromosome hybrid cells using a lactase cDNA probe. Fluorescence in-situ hybridisation (FISH) allowed finer localisation to chromosome 2q21 (Harvey *et al.*, 1993; Kruse *et al.*, 1988).

Using an Mspl polymorphism identified as T5579C by Boll *et al.* (1991), both the NIH/CEPH chromosome 2 linkage map (1992) and the chromosome 2 integrated map (1996) localise LCT to 2q21. MCM6, a homologue of a G1/S checkpoint cell cycle gene in yeast was shown to be immediately 5′ to LCT by sequencing a genomic clone (Harvey *et al.*, 1996).

1.2.2 Structural features of the lactase enzyme

1.2.2.1 Catalytic units

Lactase phlorizin hydrolase has two activities: a β-galactosidase activity and a β-glucosidase activity hydrolysing lactose and phlorizin, a disaccharide found in roots and hips of plants of the Family Rosacaeae (Stecher, 1968) and some seaweeds (Freund *et al.*, 1997). Lactase is synthesised as a polypeptide, containing four domains (section 1.2.2.3), with an apparent molecular mass of 245,000 and is cleaved inside the cell to form the mature enzyme (domains III and IV) with an apparent molecular mass of 160,000 (Potter and Bolton, 1982; Skovbjerg *et al.*, 1981). It dimerises on the brush-border membrane to form the active enzyme. The cleaved polypeptide, containing domains I and II, has no apparent enzymatic function, but may function as a chaperone (Naim *et al.*, 1994; Oberholzer *et al.*, 1993).

It was first suggested that two glutamate residues at 1271 (domain III) and 1747 (domain IV) are responsible for lactase and phlorizin hydrolase activity respectively (Wacker *et al.*, 1992). However site directed mutagenesis of Glu1271 did not affect lactase activity, whilst site directed mutagenesis of Glu1747 abolished both enzymatic activities (Neele *et al.*, 1995). Another study suggested that neither domain II nor domain IV are sufficient, on their own, for enzyme activity (Jost *et al.*, 1997). It is possible that both domains co-operate to form one single active site, where domain III is the entry point for lactose and the substrate is channelled to the catalytic site on domain IV. Lactase has a membrane spanning domain and a small hydrophilic domain at the C-terminus, which is cytosolic (Wacker *et al.*, 1992).

1.2.2.2 Transport and glycosylation

Jacob *et al.* (1994) analysed the transport and sorting of lactase by a Madin-Darby Canine Kidney (MDCK) cell line that expressed lactase. Pulse-chase labelling experiments suggested that the kinetics of lactase transport was similar to intestinal enterocytes, and the enzyme activity was indistinguishable from that found on the brush-border membrane. The mature lactase was targeted to the apical membrane whilst the precursor form of lactase was targeted to apical and baso-lateral membranes. This suggests that signals, masked until cleavage of precursor lactase, target the enzyme to the apical membrane. Lactase is both N- and O-glycosylated, and in vitro experiments show that the N- and O- glycosylated form has a V_{max} that is four times higher that the N-glycosylated form alone (Naim and Lentze, 1992). It is possible that this O-glycosylation, which occurs in the Golgi apparatus, may be involved in post-translational control of enzyme activity. Terminal glycosylation of lactase shows variation between individuals, which is controlled by the ABH and Lewis blood group glycosyltransferases (Green *et al.*, 1988).

1.2.2.3 Evolution and comparative biochemistry

Analysis of the human, rat, rabbit genes shows that they have a four-fold internal homology (domains I, II, III and IV), suggesting evolution by gene duplication (Grabnitz *et al.*, 1991; Mantei *et al.*, 1988). Domains I and II are in the cleaved portion of the LPH polypeptide, but as mentioned above they have no enzymatic function. Since lactase activity is confined to mammals and phlorizin hydrolase activity is found in most vertebrates, the evolution of lactase active site by duplication and subsequent modification of a phlorizin hydrolase active site is likely.

Mian, 1998 used various computational techniques to compare the four domains of human, rat and rabbit LPH with other glucosidases. As expected, they showed very close similarity to each other, and the domains were also similar to *Candida wickerhamii* β-glucosidase and human KL (*klotho*) gene, which is a gene that may be involved in ageing in mice (Kuro-o *et al.*, 1997).

1.2.3 Expression patterns in the intestine

1.2.3.1 Structure and histology of the intestine

The small intestine leads from the pylorus at the bottom of the stomach to the ileocaecal valve, and includes first the duodenum, then the jejunum then the ileum. The large intestine extends from the distal end of the ileum to the anus, and is comprised of the caecum, the colon and the rectum.

The intestinal epithelium is highly invaginated creating a large surface area for the absorption of digestive products. Villi are processes that can be ridge-like in the first part of the duodenum and finger-like in the distal jejunum and ileum. The epithelium of the villi is formed from enterocytes, which are absorptive cells with an apical brush-border membrane, and mucous cells which secrete mucus. Between each villus are crypts which contain stem cells, the cells which can proliferate and migrate up the villus replenishing epithelial cells which are exfoliated from the villous tips.

1.2.3.2 Expression of lactase in the intestine

Lactase expression varies along both the crypt-villus axis and the intestine. Immunohistochemical evidence on humans shows that there is very little lactase expression in the crypt in humans. It is at its highest in the midvillus region, and declines gradually towards the apex of the villus (Skovbjerg *et al.*, 1980). The distribution of lactase enzyme and its mRNA along the crypt-villus axis has been studied in rats. Both mRNA and enzyme were low in the crypt, but at the base of the villus the enzyme was low but the mRNA was high. The midvillus region showed an opposite pattern: enzyme levels were high but mRNA was low (Rings *et al.*, 1992). Studies of varying lactase expression along the length of human intestine used both immunological and enzymatic techniques (Newcomer and McGill, 1966; Skovbjerg, 1981; Wang *et al.*, 1994). No lactase is found in the stomach or colon, and lactase activity is low in the proximal duodenum and distal ileum. Lactase activity is highest between the mid- to lower jejunum.

1.2.3 Other disaccharidases in the intestine

1.2.3.1 Sucrase isomaltase

Sucrase isomaltase hydrolyses sucrose to glucose and fructose, and also hydrolyses maltose and isomaltose to two glucose units. It is a heterodimer composed of two similar subunits with apparent molecular masses of 120000 to 160000 (Sjostrom *et al.*, 1980). A single polypeptide chain (pro-sucrase isomaltase) is synthesised which contains both sites; one site for sucrase activity and one site for isomaltase activity. This is cleaved by a pancreatic protease to form the two subunits which heterodimerise, and only one subunit of the dimer is attached to the brush border membrane by its N-terminal end (Brunner *et al.*, 1979). The heterodimers themselves dimerise to form a tetramer, which is the active catalytic unit localised in the brush-border membrane. The enzyme is glycosylated (Ghersa *et al.*, 1986), and, like lactase, carries ABH blood group antigens (Green *et al.*, 1988).

The sucrase isomaltase gene has been cloned in chicken (Uni, 1998), rabbit (Hunziker *et al.*, 1986), rat (Chandrasena *et al.*, 1994), human (Chantret *et al.*, 1992) and the house musk shrew (*Suncus murinus*, Ito *et al.*, 1998), an insectivore with sucrose absent from its diet. The human sucrase isomaltase gene has been localised to chromosome 3q25-26 (West *et al.*, 1988), and its mRNA transcript is 6kb encoding a polypeptide of 209402 molecular mass (Chantret *et al.*, 1992). The frequency of sucrase isomaltase deficiency is not clear. It may be present at polymorphic frequencies in some populations; it is relatively common in Inuit (Gudmand-Hoyer *et al.*, 1987;McNair *et al.*, 1972) and in Bantu-speaking South Africans there are lower mean sucrase activities as compared to South Africans of European origin (Veitch *et al.*, 1998). Evidence from protein studies suggests that various mutations cause the disease (Naim *et al.*, 1988b), but only one has been characterised at the DNA level: a single base-pair change substituting proline for a glutamine at amino acid 1098. This mutation prevents secretion of the polypeptide causes accumulation in the endoplasmic reticulum and *cis*-Golgi (Ouwendijk *et al.*, 1996).

Some house musk shrews have sucrase deficiency, which is caused by a 2bp deletion creating a frameshift and an immediate stop codon (Ito *et al.*, 1998). The 2bp deletion occurs after the sequence responsible for the isomaltase active site, and, as expected, isomaltase activity is unaffected.

1.2.3.2 Maltase glucoamylase

Maltase glucoamylase (MGA, EC 3.2.1.20 and EC 3.2.1.3) hydrolyses both maltose and amylose. It is synthesised as one long polypeptide chain, undergoes extensive glycosylation (Naim *et al.,* 1988a).

The maltase glucoamylase gene has been cloned in human and analysis of the sequence showed that the catalytic sites are identical to those in sucrase isomaltase, but overall the protein is only 59% identical. The gene has been tentatively mapped to chromosome 7 by sequence similarity to a known EST (expressed sequence tag) (Nichols *et al.,* 1998).

1.2.3.3 Trehalase

Trehalase (EC 3.2.1.28) hydrolyses αα-trehalose, a disaccharide that is found in insects and mushrooms. Trehalase deficiency is an autosomal recessive trait and has been reported at a frequency of 8% in Greenland (McNair *et al.,* 1972), although given the unusual distribution of the disaccharide, deficiency is unlikely to be of any nutritional consequence. The trehalase cDNA has been cloned in rabbit, rat and human (Isahara *et al.,* 1997; Ruf *et al.,* 1990; Oesterreicher *et al.,* 1998) The human mRNA transcript is 2kb long and found in kidney and liver as well as intestine. Studies using in vitro transcribed rabbit trehalase mRNA injected into *Xenopus laevis* oocytes and treatment with phospholipase C suggested that, unlike lactase, sucrase isomaltase and maltase glucoamylase, trehalase is attached to the brush-border membrane by glycosylphosphatidylinositol (Ruf *et al.,* 1990).

1.3 Control of lactase expression

1.3.1 Lactase expression in humans

1.3.1.1 Control of developmental downregulation

Lactase activity, together with sucrase isomaltase activity, appears in the foetal small intestine at the same time as villi at about 8 to 9 weeks. Lactase activity is highest in the jejunum with lower levels in the duodenum. Post-partum, lactase activity rises and this is accompanied by a rise in lactase mRNA, suggesting that control of expression is at the transcriptional level. Lactase activity remains high throughout childhood, then either declines after weaning or persists into adulthood at childhood levels. The timing of the loss of lactase activity appears to be variable, with data based on lactose tolerance tests (section 1.1.1.2) suggesting as early as 5 years for Thais (Flatz *et al.*, 1969), and as late as 20 in a Finnish study (Sahi *et al.*, 1983). There have been no studies following a cohort of individuals at various ages and measuring intestinal lactase activity. It is not known whether there could be a reasonably fast decline in lactase activity, with the age of onset of decline being variable; or a constant age of onset but a gradual decline (see figure 1.1).

The cause of this loss of activity has been hard to identify. The lactase enzyme present in adults is exactly the same protein that is present in children (Potter *et al.*, 1985; Skovbjerg *et al.*, 1981), and none of the polymorphisms identified in the lactase mRNA correlated with lactase persistence or non-persistence (Boll *et al.*, 1991). Most data suggest that the downregulation of lactase is controlled primarily by a transcriptional mechanism in humans (Fajardo *et al.*, 1994; Harvey *et al.*, 1994; Montgomery *et al.*, 1991; Rossi *et al.*, 1997), although some data has suggested a role for post-transcriptional control as well (Rossi *et al.*, 1997). Wang *et al.* showed that persistence/non-persistence was determined by a polymorphism of a *cis*-acting element by examining lactase mRNA from adults heterozygous for the exonic polymorphisms, only mRNA from one allele was present (Wang *et al.*, 1995) in eight out of nine individuals with high lactase activity. This suggested that these individuals (who also showed on average intermediate lactase activity) were heterozygous for lactase persistence, and a mutation in a *cis*-acting element was responsible for the change in developmental expression.

1.3.1.2 Characterisation of the immediate promoter

Harvey *et al.* (1995) and Lloyd *et al.* (1992) showed that no variants upstream of exon 1 associated with lactase persistence or non-persistence. Nevertheless an understanding of the *cis*-acting elements in the lactase promoter is essential for understanding lactase gene expression. Boll *et al.* (1991) sequenced the first 1kb upstream of lactase exon 1 in humans, and found a TATA box at -27 to -32. A small amount of work has been done using a human lactase promoter fused to a luciferase reporter gene and transfected into Caco2 cell line. The reporter gene was transactivated by GATA-6 expressed from a transfected plasmid, an 18bp segment at -100 corresponding to a GATA consensus motif was deleted, expression of the reporter gene was reduced (Fitzgerald *et al.*, 1998). This suggests that this GATA box is essential for expression of lactase.

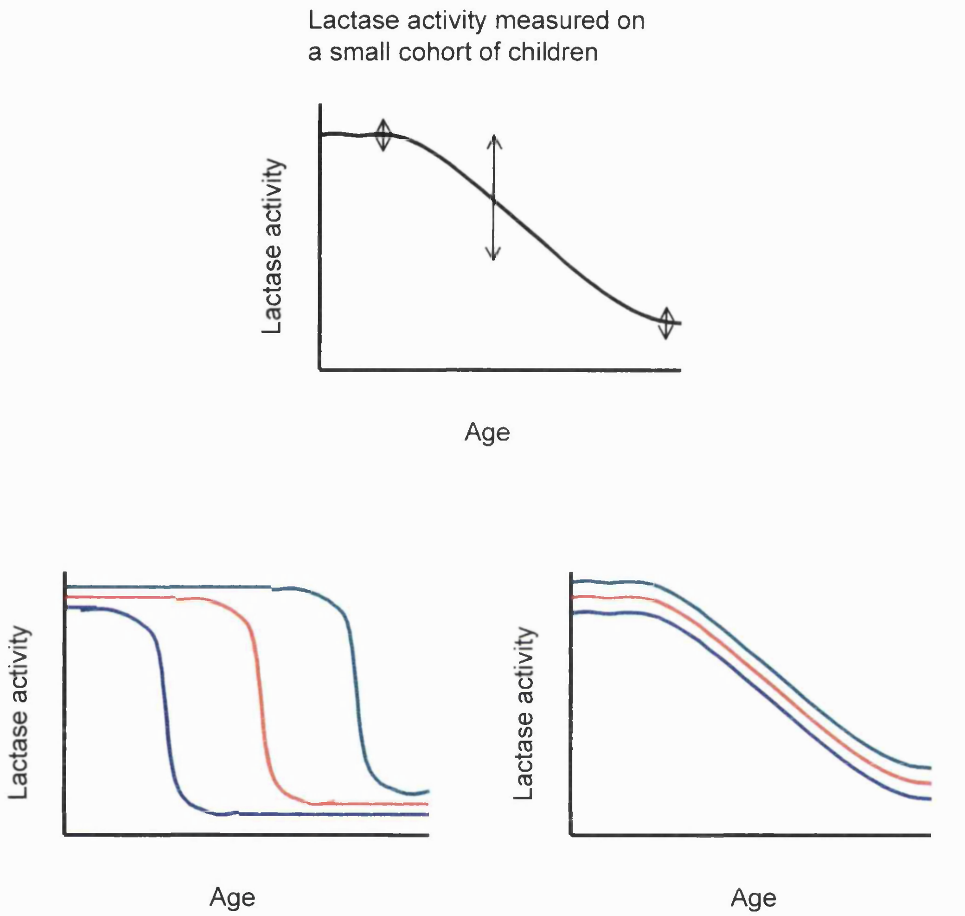
1.3.1.3 Mosaicism of lactase expression in the villus

In the jejunum of human non-persistent individuals, lactase expression was revealed by immunohistology to be patchy with patterns of expression apparently not following the clonal origin of the cell (Maiuri *et al.*, 1994). It is possible that the downregulation is inefficient and some cells escape the shutdown of expression. Therefore it is clear that lactase persistence is a mutation which allowed all the cells to escape downregulation, although it is not clear whether the functional mechanisms involved are the same. The best known case of mosiacism in tissue is position-effect variegation in *Drosophila,* which has been shown to be due to changes in chromatin structure across the locus determining the phenotypic mosiacism (Wallrath and Elgin, 1995). Mosaicism of lactase expression in enterocytes has been observed in adult rats as in humans, although in rats the pattern of mosaicism reflects the clonal origin of the cells by appearing as ribbons along the length of the villus (Maiuri *et al.*, 1992).

1.3.2 Lactase expression in the pig

1.3.2.1 In vitro experiments

Most of the studies have been conducted on pig lactase promoter transfected into human Caco2, a human colon carcinoma cell line. Analysis of mRNA and lactase activity along the neonate pig intestine showed regulation of lactase activity at the mRNA level and, in the duodenum and proximal jejunum, regulation by increased protein turnover (Torp *et al.*, 1993). DNasel footprinting and electromobility shift assays (EMSA) showed that an intestine-specific factor called NF-LPH1 bound to a sequence between -40 and -54 (called CE-LPH1). Enterocytes from neonate pigs have high levels of NF-LPH1, and also high levels of lactase, whilst adult pigs have low levels of NF-LPH1 and low levels of lactase. This suggested that NF-LPH1 levels may be requeired for the expression of lactase and play a role in its down regulation. Electromobility Shift Assays suggested NF-LPH1 includes the intestine specific homeodomain transcription factor Cdx-2, which is related to *caudal* in *Drosophila,* and is involved in the regulation of other intestinal genes such as sucrase-isomaltase and carbonic anhydrase (Drummond *et al.*, 1996; Suh *et al.*, 1994; Troelsen *et al.*, 1997). NF-LPH1 also includes the transcription factor encoded by the homeobox gene HOXC-11 (Mitchelmore *et al.*, 1998).



Lactase activity measured on a small cohort of children

Steep downregulation, variable age of onset, minimal experimental/ environmental variation

Gradual downregulation, constant age of onset, large amount of experimental/ environmental variation

Figure 1.1 Two different hypotheses on the nature of lactase downregulation
The top diagram shows hypothetical data showing the downregulation of lactase in children, and the two lower diagrams show two possible mechanisms explaining the observed effect. Blue, red and green are different individuals.

1.3.2.2 Transfection and transgenic experiments

Troelsen *et al.* (1994a) attached 1kb of the pig lactase promoter to a rabbit β-globin reporter gene, and established a mouse line transgenic for this construct. They showed by Northern blot analysis and in situ hybridisation that the transgene was expressed only in enterocytes of the small intestine of the mouse. It was also shown that the transgene was downregulated at the appropriate age after birth, which suggested that the 1kb pig promoter was sufficient for downregulation in mice.

Caco2 is the only human cell line known to express lactase. It has been used in several experiments as a model system to examine the function of fragments of the lactase promoter attached to a reporter gene. Further transfection work has attempted to characterise the *cis*-acting elements and their trans-acting factors (figure 1.2). Activation of lactase expression by HOXC-11 depends on the presence of the transcription factor HNF-1a (Mitchelmore *et al.*, 1998), which can bind to the sequence elements CE-LPH2a and CE-LPH2c (CE-LPH2b was identified by homology to the two other CE-LPH2 elements and has no known function). In addition, members of the HNF-3/forkhead subfamily of transcription factors, FREAC-2 and FREAC-3 seem to repress lactase by binding to CE-LPH3. Overexpression of C/EBP factors increased reporter gene transcription, but although CE-LPH4 has the C/EBP consensus binding site, EMSA analysis showed that the factors binding to CE-LPH4 were heat labile, in contrast to C/EBP. Additionally, a C/EBPα knockout mouse showed no change in lactase expression (Spodsberg *et al.*, 1999). Figure 1.2 summarises the current knowledge of the pig upstream elements.

1.3.3 Lactase expression in the rat

As in the pig and human, lactase mRNA levels reflect lactase protein levels which suggests that transcription is the primary level of control of lactase gene expression (Buller *et al.*, 1990; Duluc *et al.*, 1993; Krasinski *et al.*, 1994; Sebastio *et al.*, 1989). A *cis*-acting element CE-LPH1 was identified by homology with the pig promoter (see section 1.3.2) and an intestine specific factor NF-LPH1 was found to interact with the *cis*-acting element and activate transcription. However, whether or not this factor is developmentally regulated is disputed (Boukamel and Freund, 1994; Hecht *et al.*, 1997; Tanaka *et al.*, 1997).

Krasinski *et al.* (1997) created a mouse line transgenic for a rat lactase promoter/ human growth hormone transgene. They showed that the 2kb upstream

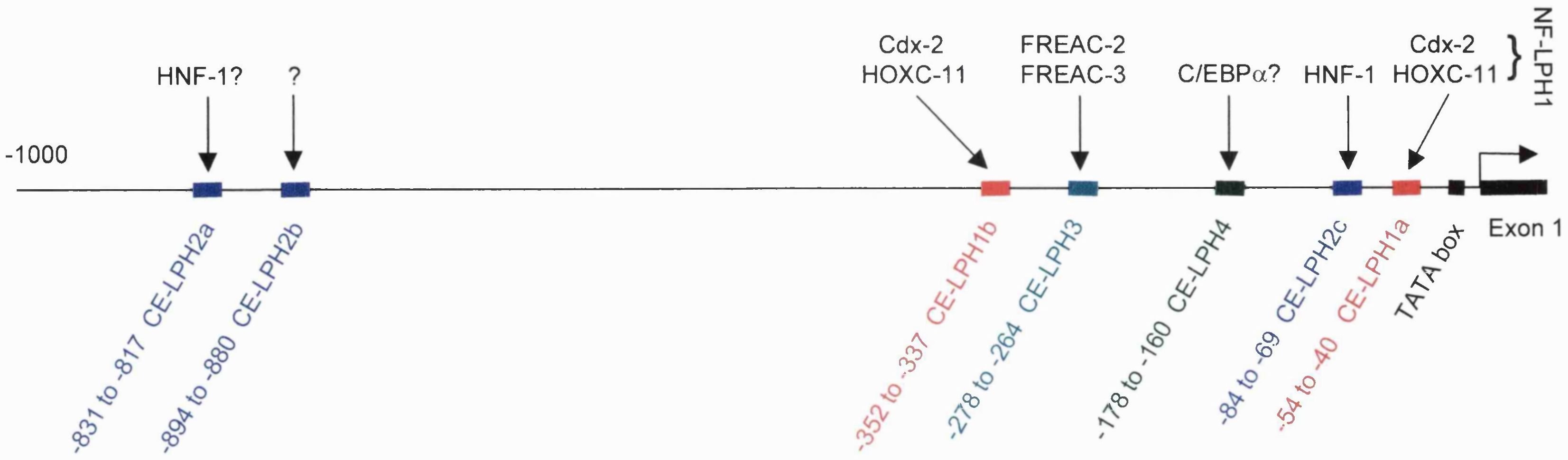


Figure 1.2 **The lactase promoter from pig showing *cis*-acting elements with *trans*-acting factors present in Caco2 cell line.**
The *cis* elements are coloured, with their position shown relative to exon 1. *Trans*-acting factors that are thought to interact with the *cis*-elements are indicated by arrows.

from exon 1 was responsible for correct tissue, cell and crypt-villus axis expression. Interestingly, mosaicism of transgene expression is observed in the transition zones of the duodenum, similar to native lactase. However the 2kb promoter does not seem to contain the elements required for appropriate expression along the intestine and appropriate downregulation.

1.3.4 Lactase expression in other animals

Rabbit lactase mRNA and lactase enzyme activity correlate strongly in the distal ileum but correlate less well in the jejunum, so like other animals, transcription appears to be the primary level of control, but post transcriptional control may also play a part. Rabbits have three lactase genes that share 94% identity at the nucleotide level and 97% identity at the amino acid level (Mantei *et al.*, 1988). The representation of each sequence in the final protein differs along the intestine, and this may reflect the different methods of regulation along the intestine.

Studies using biopsies from baboons, cats, lions, ferrets, polar bears, grey kangaroos and dogs have also shown downregulation of the lactase enzyme (Hore and Messer, 1968; Kerry, 1969; Welsh *et al.*, 1974; Welsh and Walker, 1965).

1.4 Levels of control of gene expression

There are several levels of control that allow the constant genotype encoded by the DNA in a cell to manifest itself as a selection of proteins which produce the phenotype of the cell. Each stage will be discussed briefly, with examples. Section 1.6 contains an overview of the control of genes: β-globin because it is the best studied of any developmentally controlled gene in humans, and sucrase isomaltase, because there has been several studies characterising the control of expression of this intestinal specific gene.

1.4.1 Control by transcription

For most genes, the primary level of control is gene transcription. The gene is transcribed by RNA polymerase IIcomplex, which starts at exon 1 following assembly of the transcription initiation complex (TIC) usually on TATA sequence 10bp upstream of exon 1. Assembly of the TIC depends on protein-protein interactions with trans-acting transcription factors binding to specific *cis*-acting elements upstream from exon 1 in the promoter region. For correct temporal and spatial expression of a gene, a variety of transcription factors can bind to the promoter and control transcription. The sea urchin *endo16* gene promoter, containing seven *cis*-acting elements, has been used as a model system to demonstrate that the gene responds to the combination of transcription factors in a logical way (Yuh *et al.*, 1998).

Gene transcription is affected by chromatin structure as well as binding of transcription factors. The most favoured model is one in which the cytosine in cytosine-guanine dinucleotides in genes that are inactive in a cell are selectively methylated by an unknown mechanism. These methylated cytosines recruit a protein (MeCP2) (Nan *et al.*, 1996) which in turn recruits histone deacetylases (Wade *et al.*, 1999). These deacetylate histone H3, a component of chromatin, which causes chromatin condensation to form a denser heterochromatin-like structure (for review see Razin, 1998). Because chromatin can be regarded as a 'global' mechanism of control, *cis*-elements regulating chromatin changes can be some distance from the gene they control. Loss of distant *cis*-acting elements have been implicated in some human diseases such as aniridia by analysis of recombination breakpoints in patients (Kleinjan and van Heyningen, 1998). Polymorphic expression of the TIMP1 gene from the inactive X chromosome in females has been observed, and since X chromosome inactivation is mediated by a dense chromatin structure, it is tempting to suggest that the polymorphism is in an element that controls chromatin structure around the gene (Anderson and Brown, 1999).

1.4.2 Control by splicing and nuclear export

Splicing of pre-mRNA to mRNA by removal of introns is a potential level of control. A well-studied example is control of sex differentiation in *Drosophila*, by alternatively splicing of the sex-lethal gene. In male cells exon 3, which contains a stop codon, is included and translation produces a non-functional protein. In female cells, exon 3 is skipped and a functional full-length protein is produced on translation (MacDougall *et al.*, 1995).

In humans, alternative splicing is a common method of generating multiple isoforms from one gene product. There is an example of alternative splicing caused by a polymorphism in a *cis*-acting element: the human growth hormone receptor expressed in the placenta has one allelic isoform with exon 3 and another isoform without exon 3. The causative polymorphism has not yet been identified, and the functional consequence of the missing exon has not been determined (Stallings-Mann *et al.*, 1996).

Nuclear export of mRNA is a regulated process, but little is known about the control of this process. Src protein may be involved in this process in mice, but how it regulates processing and which transcripts are regulated is not known (Neel *et al.*, 1995).

1.4.3 Control by translation

Selective translation of mRNA transcripts has been demonstrated in *Xenopus, Caenorhabditis elegans* and *Drosophila.* Recently, work on the differentiation of protomyelocytic cell line HL-60 showed that five mRNAs were specifically inhibited at that stage and five mRNAs were specifically activated. The inhibited mRNAs were all ribosomal protein genes from a 5'oligopyrimidine tract family, but the activated mRNAs were all uncharacterised. There is some evidence that the transcribed Alu-like elements may have a role in upregulating translation of their mRNA (Krichevsky *et al.*, 1999).

1.4.4 Other levels of control

1.4.4.1 Control by intracellular localisation of protein

Polarised epithelial cells maintain polarisation by sorting proteins to the correct membrane. Lactase is an example which is localised to the brush-border membrane of enterocytes (see section 1.2.1.2), and it is clear that this is an important stage in maintaining phenotype. It is probable that the cell can control localisation of proteins in response to extracellular factors.

1.4.4.2 Control by phosphorylation

Phosphorylation on serine or threonine residues is a common method of controlling the activity of intracellular proteins, especially transcription factors. Protein kinases, which are often phosphorylated themselves, phosphorylate other proteins, and phosphorylation signalling cascades are an important component of cell signalling. For example, the transcription factor CREB (cAMP response element binding protein) is active only when phosphorylated on Serine-133 (Gonzalez and Montminy, 1989).

1.4.4.3 Control by protein and mRNA degradation

The levels of many proteins are controlled by varying the rate of degradation of that specific protein. For example, the human Jab1 gene product controls levels of the cyclin-dependent kinase inhibitor p27Kip1 by altering its rate of degradation (Tomoda *et al.*, 1999). The altered protein levels created by a mutation of a protein in a degradation pathway can cause human disease (Kato 1999).

RNA degradation involves the same pathway as protein degradation (Laroia *et al.*, 1999), and is an important form of control for cytokine and proto-oncogene mRNAs.

1.5 Examples of developmentally regulated genes in humans

1.5.1 β-globin

1.5.1.1 The β-globin gene cluster

The β-globin gene cluster comprises five genes, ε-globin, γ_G-globin, γ_A-globin, δ-globin and β-globin arranged in that order 5′ to 3′. All the genes are developmentally regulated with ε-globin expressed in the embryonic yolk sac, γ-globin in the foetal liver and δ- and β- globin expressed in adult bone marrow.

1.5.1.2 The locus control region (LCR)

Expression of the globin genes is controlled by the LCR at the 5′ end of the cluster. This region spans about 10kb and has four segments that are hypersensitive to DNase 1, indicating that they have a loose chromatin conformation. All four segments can enhance reporter genes in transgenic mice, and when the LCR is deleted the globin genes are either inactive or expressed very weakly. The control of each gene depends on the interaction of the LCR with the promoter of each gene, and, unlike other enhancer elements, the orientation of the LCR with respect to the genes is crucial (Tanimoto *et al.*, 1999). The LCR is thought to mediate its effect either by binding *trans*-acting factors and looping so these factors can contact *trans*-acting factors on each promoter, or by selecting an open chromatin conformation so that *trans*-acting factors can only access the gene promoter of choice (Ellis *et al.*, 1996; Grosveld *et al.*, 1993).

The model for developmental switching of globin genes is based on the concept that there are several gene promoters competing for one LCR, and contact of promoter and LCR can either open chromatin or directly activate transcription factors, as discussed above. In the embryo, LCR contacts the ϵ-globin promoter, but in the adult the transcription factor NF-E4 is produced which may direct the LCR to associate with the β-globin promoter (Foley *et al.*, 1994).

1.5.1.3 Heriditary persistence of fetal haemoglobin (HPFH)

This condition is when the expression of fetal haemoglobin persistes into adulthood, in contrast to normally downregulating and replacement by adult haemoglobin. Expression of either the γ_G-globin or γ_A-globin is responsible, and the condition is heterogenous at the molecular level (Tuan *et al.*, 1980). A variety of small deletions and point mutations have been characterised as causing different types of HPFH, either within the gene or in regions outside the gene, which may contain possible regulatory elements (Carlson and Ross, 1986; OMIM 141749).

1.5.2 Sucrase isomaltase

Sucrase isomaltase is expressed only in the small intestinal and expression peaks at 10 weeks, concurrent with the appearance of duodenal villi, and maintains that level throughout life. The enzyme levels correlate with mRNA levels (Sebastio *et al.*, 1986) and mRNA levels are the same in fetuses and adults (Wang *et al.*, 1994), so differences in the level of transcription may be responsible for most developmental differences in expression, although differences in the level of mRNA stability cannot be ruled out.

Studies on the promoter of sucrase isomaltase by attachment to a reporter gene and transfection into Caco2 cells showed that intestine-specific regulation was controlled by a number of regulatory elements (Wu *et al.*, 1992). SIF1-BP has been shown to be related to NF-LPH1, a transcription complex identified by EMSA which binds to the lactase promoter (section 1.3.2.1) (Troelsen *et al.*, 1994b). Cdx-2 has also been shown to bind as a dimer to the sucrase-isomaltase promoter (Suh *et al.*, 1994) and may be a component of SIF1-BP.

Transgenic studies involving the 8.5kb upstream of SI linked to a reporter gene suggested that this region contained elements necessary for intestine-specific transcription. However, unlike the native gene in the trangenic mouse, the reporter gene was not solely expressed in enterocytes but in Paneth cells, goblet cells and enteroendocrine cells. This suggests that SI expression is usually repressed in these cells by a silencer that is outside the 8.5kb directly upstream from the gene (Markowitz *et al.*, 1995; Tung *et al.*, 1997).

1.6 Human molecular population genetics and human prehistory

1.6.1 Mutation and polymorphism

Individuals are unique, yet when a piece of human DNA is sequenced, it is assumed to represent the human species. The ultimate example of this is the Human Genome Project whose eventual aim is to completely sequence the DNA of a man. In fact, the sequence will be a 'composite man' because every human individual will have differences when compared with the 'composite man' sequence. Each DNA change, or mutation, can be confined to one individual, or it can be shared with most of the human population, or can be confined to any number of individuals. If a mutant allele is present at a frequency which cannot be explained by recurrent mutation (usually defined as a mutant allele frequency of >1%) then the mutation is polymorphic and is termed a polymorphism.

1.6.2 Detecting variation

Protein electrophoretic analysis of blood or other tissues from different individuals was the first technique to reveal that variation was common in humans, although previously some polymorphisms (such as ABO blood groups) were detected by serological methods. Starch gel electrophoresis and isoelectric focusing, techniques that separated proteins on basis of charge, revealed polymorphism in enzymes such as phosphoglucomutase and α-1-antitrypsin (Axelsson and Laurell, 1965; Spencer *et al.*, 1964). With the advent of DNA cloning and sequencing followed proof that these polymorphisms were, as expected, reflected in the DNA (March *et al.*, 1993). Now mutation detection techniques are almost always DNA-based, such as those used in this study (section 2.4)

1.6.3 The use of variation in understanding human diversity

Two recent studies have analysed sequence diversity in humans as compared to chimpanzees over 10kb non-coding sequence of the X chromosome (Kaessmann *et al.*, 1999a; Kaessmann *et al.*, 1999b). Human nucleotide sequence diversity (mean pairwise sequence difference per base) was found to be 0.00037, with most of that diversity in Africa (0.00031 in African populations). This figure is considerably less than that observed for chimpanzees (0.0013) suggesting at least one genetic bottleneck in human evolution. Nevertheless, the amount of diversity within the human genome has allowed many studies on allele frequencies in different populations.

Analysis of allele frequencies of phenotypic polymorphisms by protein electrophoresis has revealed a lot about human history. Cavalli-Sforza *et al.* (1994) compiled allele frequencies from various studies that used immunological and protein electrophoretic techniques. This involved a statistical technique called principal component analysis, which divides frequency data for many genes from many data sets into several principal components. In a simple case allele frequencies of two loci for several populations, each population can be plotted on a two-dimensional graph with one allele frequency as the *x*-axis and the other allele frequency as the *y*-axis. A straight line of optimum fit can be drawn on the graph. This line is the first principal component. Each population can be plotted as a point on this line perpendicular to the original point, and so the two-dimensional data can be represented as values on a one-dimensional line. When alleles from more loci are added, the graph becomes *n*-dimensional and impossible to visualise, but mathematically possible to calculate. A second principle component would be a line orthogonal to the first principal component, and only the number of loci limits the number of principal components. Each principal component can thus be plotted as a value on a map position corresponding to the place of sampling, but larger and larger principle components represent less and less of the total variance. This technique allowed Cavalli-Sforza *et al.* (1994) to infer demographic histories of populations from data from many loci.

Development of DNA based polymorphism detection techniques have allowed more studies on the nature of variation. Both analyses of sequences from many individuals and allelic frequencies in populations have contributed to the knowledge of variation. The effectiveness of a map of single nucleotide polymorphisms across the genome to identify risk loci in common genetic diseases has been debated (Kruglyak, 1999, for review see Terwilliger and Weiss, 1998), and recent studies have been conflicting on the nature of linkage disequilibrium and diversity within the genome (see section 1.6.4.1).

1.6.4 Principles of human molecular evolution

1.6.4.1 Linkage disequilibrium

When a mutation occurs, it occurs on one background chromosome, and at that moment it occurs it is completely associated with all alleles on that chromosome: it is in linkage disequilibrium with those alleles. Recombination gradually breaks down linkage disequilibrium (LD) to linkage equilibrium, in which the new allele and another allele in question associate randomly. The greater the distance between the two alleles, the more likely recombination will occur per generation and so fewer generations are required to break down disequilibrium (Ott, 1991).

Within small distances, recombination is unlikely and LD breaks down slowly. However other factors can affect linkage disequilibrium between two alleles such as relative time of mutation, genetic drift, population expansion and selection.

There are several measurements of LD each with its own benefits and drawbacks (for a review see Devlin and Risch, 1995). Commonly the normalised disequilibrium coefficient D' is used, which varies from 1 for complete LD to 0 for complete equilibrium (section 2.7.2).

1.6.4.2 General population genetic assumptions

In many of the studies below, assumptions are made about the nature of mutations or the population. For example, the coalescent analysis relies on the ability to draw trees which coalesce to a common ancestor and assumes that there is no selection acting on any allele. Whilst it may be true that an intronic polymorphism is selectively neutral, it may be associated with a polymorphism which can be selected upon. This is called selective sweep or genetic hitchhiking (Nurminsky *et al.*, 1998), and the distance over which a selected allele can 'sweep' neutral alleles depends on linkage disequilibrium, which in turn depends on recombination rate, amongst other factors.

To determine the direction of mutation (whether it was a C to T mutation or a T to C mutation) the ancestral state of polymorphism is observed by noting the sequence at the homologous site in an outgroup. Usually this is the closest living relative to humans, the common or pygmy chimpanzee (Goodman *et al.*, 1998), but sometimes human polymorphisms are also polymorphic in chimpanzees (Gyllensten and Erlich, 1989). Therefore it is preferable to also note the sequence of other apes such as gorilla or orang-utan. The consensus sequence of these species represents the ancestral human allele.

1.6.4.3 The origins of modern humans

The human fossil record suggests that bipedal hominids occupied most of the temperate Old World for between 0.5 to 1.5 million years. These early hominids are *Homo erectus* whilst the later forms are known as archaic *Homo sapiens*. There are two theories concerning the origin of *Homo sapiens* (modern humans) which can be regarded as two ends of a spectrum with many scenarios in between.

The multiregional hypothesis proposes that modern humans evolved concurrently in several parts of the world. Gene flow between populations and selection for favourable characteristics kept the evolving species homogenous (Templeton, 1997).

The 'Out of Africa' hypothesis (also called, less poetically, the replacement hypothesis, or for those with a Biblical inclination, the Noah's Ark or Garden of Eden hypothesis) proposes that an ancestral modern human population grew and moved from Africa into Eurasia and the New World. There was no contribution to the gene pool of archaic humans who already lived in those places, and the number of ancestral modern humans prior to expansion was small, perhaps several thousand adults (Rogers and Jorde, 1995).

Population characteristics should distinguish the two ideas: large amounts of genetic variation should be observed if the human population has been large during history, and small amounts of genetic variation should be observed if the ancestral population size was small or had been through a population bottleneck. So both the effective population size and the amount of genetic variation should give clues to which model is correct.

1.6.4.4 Molecular evidence on the origins of modern humans

Evidence for low diversity in non-Africans and higher diversity in sub-Saharan Africans has come from many studies. Protein polymorphism and immunological studies collated by Cavalli-Sforza *et al.* (1994) show an increase in variation in sub-Saharan Africans. Increased diversity has also been shown for sub-Saharan Africans in mitochondrial DNA (Penny *et al.*, 1995; Vigilant *et al.*, 1991), microsatellites (Bowcock *et al.*, 1994; Seielstad *et al.*, 1999), minisatellites (Armour *et al.*, 1996) and Y chromosome haplotypes (Hammer, 1995; Hammer *et al.*, 1998). Further examples of nuclear DNA diversity and their consequences for human history are discussed in more detail in section 1.6.5, 1.6.6, and 1.6.7.

Estimates of population size can be made by using a coalescent approach to draw trees of haplotypes with each node representing a mutation generating a new haplotype, and the root of the tree representing the ancestral haplotype. This analysis can only be performed on non-recombining stretches of DNA, and so has been studied using mtDNA and Y chromosome data (Harpending *et al.*, 1998). An estimate of time-depth for each node gives the characteristic shape of the tree and combined with mutation rate, an estimate of population size can be made. Unfortunately both mtDNA and Y chromosome haplotypes coalesce only a few thousand years ago which is after the postulated expansion of modern humans. However, Takahata and Satta (1997) compared nucleotide substitutions in sequences of nuclear loci between modern humans using a maximum-likelihood method. They suggest that the effective population size was 10,000 during the time when *H.erectus* was spread over the Old World. Harpending suggests that this is too small to refer to *H.erectus* spread over the Old World, and so this effective population size estimate of our direct ancestors is more likely to represent an archaic *H.sapiens* population in Africa. This figure of 10,000 was corroborated by analysis of polymorphisms in an 8kb segment of the dystrophin gene in several populations, assuming no recombination and selection (Zietkiewicz *et al.*, 1997; Zietkiewicz *et al.*, 1998).

1.6.5 Haplotype analysis of the myotonin protein kinase gene.

Myotonic dystrophy (DM, OMIM 160900) is an inherited neuromuscular disease that is at different frequencies in different populations. In Japanese, the incidence is at 5.5 in 100,000; in western Europeans between 2.2 to 5.5 in 100,000; and is much less prevalent in south-east Asians and absent in southern Africans. DM shows anticipation in families in which severity of the disease increases in successive generations. This is due to expansion of a CTG repeat in the 3′ untranslated region of the myotonin protein kinase gene (DMPK). On normal chromosomes, the repeat size of $(CTG)_n$ is n=5-38, but on DM chromosomes n>42 (Brook *et al.*, 1992). There is complete allelic association between DM causing alleles and an Alu insertion allele of an insertion/deletion polymorphism 5kb telomeric to the repeat. Haplotypes consisting of eight further polymorphic sites spanning 30kb also show complete association, at least in Europeans, and strong linkage disequilibrium has been found with polymorphisms up to 160kb away (Deka *et al.*, 1996).

Tishkoff *et al.* (1998) examined DMPK haplotypes spanning 20kb in 1235 individuals from 25 populations from around the world. The haplotypes consisted of the $(CTG)_n$ repeat and markers flanking the repeat: the Alu insertion/deletion polymorphism, a HinfI RFLP 2.5kb telomeric to the repeat and a TaqI RFLP 15 kb centromeric to the repeat. Pairwise linkage disequilibrium between the non-repeat polymorphisms is high in all populations, but sub-Saharan Africans (Bantu-speaking South Africans, San, Biaka, and Mbuti) had lower levels of disequilibrium.

The haplotype distribution reflects the patterns of linkage disequilibrium. The most common haplotypes in most populations are (+++) and (–––), and the (+++) haplotype is assumed ancestral because of the presence of Alu(+), HinfI(+) and TaqI(+). Sub-Saharan populations also have (––+), (+––), (+–+) at high frequencies, and these may be intermediates between (+++) and (–––). The (–––) haplotype is at high frequencies in Asian populations, presumably due to genetic drift. When the $(CTG)_n$ repeat is included in haplotype analyses, that the $(CTG)_5$ and $(CTG)_{>17}$ repeats were carried by (+++) haplotypes in ancestral non-African populations.

In general, there is a high level of haplotype diversity in sub-Saharan African populations and a subset of that diversity in non-African populations. This suggests an 'Out of Africa' hypothesis for the origin of modern humans (see section 1.6.4.3), and is consistent with similar studies on the CD4 locus (Tishkoff *et al.*, 1996) and the DRD2 locus (Castiglione *et al.*, 1995; Kidd *et al.*, 1998).

1.6.6 Haplotype analysis of the Lipoprotein lipase gene

Lipoprotein lipase (LPL) is an enzyme involved in lipid metabolism, and several DNA variations affecting amino acid sequence have been shown to correlate with clinical lipid profiles. It has been suggested that this is a candidate gene for susceptibility to coronary heart disease, atherosclerosis and obesity.

A different approach was used to study haplotypes of this gene from that used for the DMPK gene. The DNA sequence of part of the LPL gene (3' end of intron 3 to 5' end of intron 9, 9.7kb total) was determined in 71 individuals (Clark *et al.*, 1998; Nickerson *et al.*, 1998). Even within this small region, there was a large amount of variation: of 88 variable sites, 78 were polymorphic and the nucleotide diversity was 0.002. Using allele-specific PCR to help determine phase allowed identification of 88 haplotypes. In addition, for 64% of all the site pairs all four possible gametes were present suggesting that recombination played a major role in evolution of the haplotypes.

The 71 individuals came from three populations, two rather loosely defined: African Americans, Finns from Northern Karalia, and Americans from the Rochester Family Heart Study. Only 3 haplotypes were present in all three populations, and resampling of a subset of individuals gave haplotype frequencies far in excess of the infinite-alleles model (Hartl and Clark, 1989), again suggesting that one of the assumptions of this model - no recombination - does not hold in this case. The four-gamete test was applied to the haplotypes found in the three populations and the similarity of the estimates for past recombination events suggested that there is no variation between populations in rates of recombination.

Similar sequence-based studies of variation have been done on other candidate genes for other complex diseases. 24kb of sequence of the angiotensin converting enzyme gene (DCP1) was obtained from 11 individuals (Rieder *et al.*, 1999). 78 varying sites were found reflecting a high level of diversity reflected in the overall nucleotide diversity of 0.001. However, sequence analysis of the coding and flanking regions of 36 cardiovascular disease disorders revealed lower sequence diversity and high pairwise linkage disequilibrium, in contrast to the LPL and DCP1 studies (Cambien *et al.*, 1999). As mentioned in section 1.4.3.1 Sequence variation within 10kb of the X chromosome has also been studied using similar methods but for a different purpose (Kaessmann *et al.*, 1999a). This region is non-coding and distant from any gene and unlikely to be influenced by selection, and also shows a reduced recombination rate. Another advantage of the X chromosome is that phase can be determined directly by sequencing DNA from males. In common with studies based on allele frequencies, sub-Saharan African individuals showed considerably more diversity than non-Africans supporting the 'Out of Africa' hypothesis.

1.6.7 Phylogeny of β-globin haplotypes

The β-globin locus has been extensively studied due to the number of functional variants with clinical consequences such as the thalassemias. Allele-specific PCR was used to determine phase of polymorphisms sequenced in the 3kb across the gene, and so complete haplotypes across the 3kb could be deduced. The most recent common ancestor (MRCA) of all the sequences was estimated by the coalescent process, which assumes neutrality, no recombination and an infinite alleles model of evolution (Tavare *et al.*, 1997). The MRCA was found only in African populations and arose about 800,000 years ago. Diversity is high in both Africa and Asia, and the pattern and estimated age of one particular haplotype suggests that the ancestral human population was in Africa and Asia >200,000 years ago, and these individuals may have been archaic humans (pre-*Homo sapiens*). The effective population size was found to be 10,000, in accordance with other studies (see section 7.1.3), but no great population expansion was revealed (Harding *et al.*, 1997).

1.7 History and milk-drinking habits of the populations studied

1.7.1 Malays

Malays are Daic speakers, and although some fermented milk products exist in South East Asia, they are probably not an important part of the diet. The Malays involved in this study are from Singapore (Saha, N., personal communication).

1.7.2 Chinese and Japanese

Both Chinese and Japanese are Altaic languages, a group that dominates Central and Eastern Asia. Evidence of *H. erectus* habitation have been found in China but not in Japan, and the earliest examples of *H. sapiens* fossils have been dated to 200,000 years ago and 30,000 years ago for China and Japan respectively. Agriculture began with domestication of dogs, pigs, rice and millet around 8000 years ago (Cavalli-Sforza *et al.*, 1994). They did not domesticate any animals suitable for milking, so milk was not a part of the Chinese and Japanese diet.

The Japanese used in this study are from Osaka in Japan, and the Chinese individuals are resident in Singapore.

1.7.3 North Indians and South Indians

All individuals tested in this study from these populations were resident in Singapore for 2 or 3 generations. The Northern Indians all spoke languages from the Indo-European group and the Southern Indians were all Dravidian speakers originally from Tamil Nadu and Kerala in India and Sri Lanka (Saha, N., personal communication).

Milk products are an important part of the Indian diet - fermented milk (*dahi*), which is made into several products such as *ghee* and *lassi*, and non-fermented milk products (*khoa*) (Kanbe, 1992). The relatively high frequency of the lactase persistence allele in Northern Indians (0.42, Flatz, 1987) may be a reflection of the importance of fresh milk in the diet.

1.7.4 Bantu-speaking South Africans

This group consists of individuals living in South Africa who speak Bantu languages, part of the Niger-Kordofanian language group which is spoken over much of sub-Saharan Africa. The individuals tested in this project were all blood donors from Johannesburg, and are from a number of cheifdoms such as the Zulu, Tswana and Sotho (Jenkins, T., personal communication).

Linguistics, archaeology and genetics suggest that the Bantu originated from an area bordering what is now Cameroon and Nigeria, and expanded across east and south Africa between 3000BC and AD500 (Hiernaux, 1968). They are traditionally mixed farmers with many keeping cattle, although milk forms a small part of the diet. Indeed, in urban Bantu-speakers milk consumption appears comparable to European levels, and the frequency of the lactase persistence allele has been estimated at 0.13 (Jersky and Kinsley, 1967), although Jenkins (1982) found it to be almost absent (0.03).

1.7.5 San

San, formerly known as "bushmen", are sub-Saharan Africans living mainly in Botswana and Namibia, and are hunter-gatherers, in contrast to the Khoi (formerly known as "hottentots") who are pastoralists. Both groups speak Khoisanid languages, characterised by an abundance of vocal clicks (Ruhlen, 1987). It is thought that until the 17th century Khoisan occupied most of Southern Africa, but their territory shrank dramatically due to the Bantu expansion from the North West. More recent history shows oppression by European settlers has driven both the Khoi and San to poor land, yet the San have managed to preserve their distinctiveness in the Kalahari desert (Wilmsen, 1991). The San individuals included in this project were from the zu/wasi group, a branch of the !Kung, who live in the Tsumkwe area of north-east Namibia. During the past 20 years, some have begun herding cattle, but many still practise a traditional hunter-gatherer lifestyle (Jenkins, T., personal communication).

Evidence from the Y chromosome suggests the Khoisan are very diverse, and have a high frequency of the ancestral Y chromosomal haplotype 1A (Hammer *et al.*, 1998; Scozzari *et al.*, 1999).

1.7.6 Yakut (Sakha)

The Yakut live in Northern Siberia around the Lena valley in the Sakha autonomous republic of Russia which is also known as Yakutia. All the individuals tested in this study were from the city of Yakutsk, which is capital of the republic. They are a Turkic (Uralo-Altaic) speaking people, who were forced to move northwards from their original homeland around Lake Baikal by Buryats around 600 years ago. More recently, collectivisation had a drastic effect on the Yakut population, which suffered heavy losses between 1926 and 1959 (Wixman, 1984). Cattle breeding is an important part of their culture, as is ritual drinking of fermented mare's milk (*koumiss*) (Forde, 1934; Forster, 1767). Genetic data is virtually absent from the former Soviet Union, but Y chromosomal analysis is starting to appear. For example, analysis of the so-called Tat-C allele in European and Asian populations found it to be at highest frequencies among Altaic and Uralic speakers, including Yakut (Zerjal *et al.*, 1997).

1.7.7 Mordavians

These people are also known as Mordvinians, Mordva or Volga Finns, and are divided into two groups, Erzya and Moksha, who speak separate languages. The samples used in this study were a mixture of both Erzya and Moksha. They live across the middle Volga region of Russia, and the two languages spoken are from the Finno-Ugric branch of the Uralo-Altaic family (Wixman, 1984).

1.7.8 Russians, Northern Europeans and Southern Europeans

These three artificial groups are all Europeans speaking one of the Indo-European group of languages. The Russian language is a Slavic branch of the Indo-European family, and the individuals in this study all live in Perm, in the Ural Mountains, and are Russian by self-identification (Kozlov, A., personal communication).

The Northern European individuals are from Britain, Ireland and northern France, and so speak English (Germanic branch) and French (Italic branch). The diet of these populations contains a varying amount of fresh milk depending on personal preference, although it is possible that drinking large amounts of fresh milk was more common before the Industrial Revolution. Cheese and, more recently, yoghurt are also eaten.

The Southern European individuals are mostly from Naples in Italy, with some Greeks resident in Britain. Southern Europeans from around the Mediterranean drink little fresh milk, but convert a large amount of milk into milk products. These include cheese, like Northern Europeans, yoghurt (*tiaourti* in Greece), and *trahana*, a traditional Greek food made by mixing fermented ewe's milk with wheat flour followed by drying (Kanbe, 1992).

The history of the Europeans has been studied extensively both genetically and archaeologically. The earliest humans were *Homo erectus* and archaic *H. sapiens*, who developed into modern *H.sapiens*. These modern humans developed their culture throughout the upper Paleolithic and Mesolithic, as shown in art and artefacts. The neolithic is associated in Europe with the transition from hunter-gatherer existence to agriculture, which is thought to have spread from the Fertile Crescent area of the Near East (Cavalli-Sforza *et al.*, 1994). This "Neolithic transition" was associated with population expansion (Ammermann and Cavalli-Sforza, 1971) and has been associated with the spread of Indo-European languages from their suggested homeland in Anatolia (modern Turkey) (Renfrew, 1987). This is an attractive model, suggesting slow steady movement of farmers from the Near East into Europe, carrying their language with them and merging with the existing population. However, the theory is still controversial, as some authors have cited Ukraine or even the Baltic states as the homeland of the Indo-European languages (Gimbutas, 1970).

Cavelli Sforza's first principal component shows a gradient from the Near East to North-Western Europe and mirrors the map of agricultural expansion into Europe. This south-east to north-west gradient is seen in allele frequencies of many genes (Cavalli-Sforza *et al.*, 1994).

A group of black British of African or Afro-Caribbean ancestry was also obtained, having previously been selected on their Rh blood group vs+, which is not present in native Europeans (Poulter, M., personal communication). Because of the heterogeneity of the group and the selection of individuals based on blood group, it is not treated as a population.

1.7.10 Finns

The Finnish language is classed as a Finno-Ugric branch of the Uralic-Yukaghir language group, and this suggests that Finnish have rather a different history as compared to neighbouring Scandanavians and, indeed, other Europeans. Genetic and archaelogical evidence shows that the first settlers were nomads arriving from Asia, and followed by agricultural settlers from Sweden around 2000 years ago (Kittles *et al.*, 1998). The founding population was small and did not increase until 1700. This small population often resulted in population bottlenecks and hence disease mutations were at a high frequency due to genetic drift (Sajantila *et al.*, 1996). The Finnish population has been of interest to geneticists, since these disease mutations are derived from a single common ancestor. Identity-by-descent mapping has identified many deleterious mutations of genes causing diseases such as Diastrophic Dysplasia (Hastbacka *et al.*, 1994).

Fresh milk is a significant part of the diet in Finland, despite the fact that the lactase persistence allele is at lower frequency than neighbouring Scandinavian countries (Sahi, 1974). Gastrointestinal clinics report a high frequency of lactose intolerance, and the identification of the molecular basis of lactase persistence is important for the Finns.

1.7.11 Roma ('Gypsy')

Since their arrival in Europe, the Roma have suffered repression from virtually every country they have lived in. Elizabeth I of England ordered the all Roma to leave the country or be executed, and more recently, they were victims of Nazi genocide, some 400,000 killed in extermination camps. These are not isolated examples and the Roma are still persecuted today. In Britain, 'Gypsy' is not a very pejorative term but in other languages the equivalent term is often abusive. They are therefore referred to by the name that they give to their language, Roma.

Roma belongs to the Indo-European group of languages, and is closely related to languages of North west India. They were probably a group of nomads who moved from India into Europe (Clebert, 1967). Intermarriage with non-Roma is forbidden, and so they can be thought of as a population isolate within Europe (Ferak *et al.*, 1987). Several mutations in genes leading to disease have been identified in Roma populations (Kalaydjieva *et al.*, 1999; Tournev *et al.*, 1999), and the population will continue to be interesting to geneticists both in terms of history and disease.

The Roma population studied here is a group of Slovak Roma. The frequency of the persistence allele is known only for Czech Roma (Madzarovova-Nohejlova, 1982) but it is unlikely to be significantly different from Slovak Roma. The allele frequency is considerably lower than the surrounding Czech population. Their nomadic lifestyle means that domestic animals are not kept and milk or milk products are probably not a significant part of the diet.

1.7.12 Papua New Guineans

Papua New Guinea is an independent state covering the eastern half of the island of New Guinea, north of Australia. Many tribes live on this island in both the highland and lowland regions, and speak different forms of Papuan (Indo-Pacific) languages. Language replacement has complicated the analysis of these tribal groups by linguistic methods, and genetic evidence agrees with the idea that there were several waves of settlement. The sequence and nature of these settlements is not clear (Cavalli-Sforza *et al.*, 1994).

The sample used in this work is a collection of individuals from both highland and lowland groups (Johnson, P., personal communication).

1.8 Aims of this project

The general aims of the group are to determine the molecular basis of the lactase persistence/non-persistence polymorphism, in the general context of examining common genetic variation in genes of the intestinal epithelium.

This project is part of this general aim, and its specific aims are:

- To analyse sequence upstream and downstream of the lactase gene, and in intron 1, in order to identify polymorphic sites.
- To investigate the protein-binding properties of sequence regions showing variation of possible functional significance.
- To construct a fosmid/cosmid contig across the lactase gene as part of a larger physical mapping project.
- To confirm the map by analysis of the chromosomal region around the lactase gene using fluorescence in-situ hybridisation.
- To examine the downregulation of lactase by analysis of mRNA expression on biopsies of children of different ages using polymorphic sites to identify the transcript.
- To establish the patterns of variation across the lactase gene in populations of the Old World, and analyse the nature of the haplotypes in these populations.

2 Materials and Methods

Buffer compositions are given in section 2.8.3

2.1 Sequencing

2.1.1 Preparation of PCR products

PCR products were prepared for direct sequencing by use of an enzymatic PCR product pre-sequencing kit (Amersham Pharmacia Biotech, Amersham, Bucks). 10 units of Exonuclease I and 2 units of shrimp alkaline phosphatase were added to 5μl of PCR product to make a total volume of 7μl. Following incubation at 37°C for 15 minutes and denaturing of the enzymes at 80°C for 15 minutes, 1μl was routinely used as a template for sequencing. The actual amount of template can vary according to the size and amount of the PCR product: sequencing reactions have succeeded on as little as 0.1μl of template (for small strong PCR products) to as much as 4μl (for longer/weaker PCR products). Both the manufacturer's instructions and practical experience show that, when the product is run on an agarose gel, a single clean PCR band is essential for satisfactory sequencing results.

2.1.2 Radioactive sequencing

2.1.2.1 Preparation of the sequencing reaction

Most sequencing was performed using the ThermoSequenase radiolabelled terminator cycle sequencing kit (Amersham Pharmacia Biotech) with ^{33}P labelled Redivue™ terminators. 2μl of 10x reaction buffer (10x = 260mM Tris-HCl, pH 9.5, 65mM $MgCl_2$) was combined with 2pmol oligonucleotide primer and 8 units Thermosequenase DNA polymerase (containing 0.0012 units *Thermoplasma acidophilum* inorganic pyrophosphatase) and 1μl of treated PCR product template, made up to a final volume of 20μl with H_2O. For sequencing of fosmid clone DNA, 5μl (~4μg DNA) of DNA solution was added in place of the 1μl treated PCR product. Four 4.5μl aliquots were mixed with 2.5μl of A,C,G and T termination mixes (each termination mix is 7.5μM dATP, dCTP, dGTP, dTTP with 0.075μM (0.225μCi) of the appropriate terminator [α-^{33}P]ddNTP (1500Ci/mmol)). The four resulting reactions were covered with mineral oil and cycled on a Hybaid Omnigene with tube control routinely using the following conditions: 95°C, 30 seconds, 58°C, 30 seconds, 72°C, 1 minute, 40 cycles. An annealing temperature of 58°C in the above cycling protocol was routinely used with primers of an estimated T_m of 60°C, although the temperature could be raised or lowered to suit an individual primer (see 4.1 Polymerase Chain Reaction).

Following cycling, 4µl of terminator mix (95% formamide, 20mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol FF) was added and reactions stored at -20°C prior to electrophoresis.

2.1.2.2 Electrophoresis of sequencing products

Electrophoresis was carried out using a TTE buffer system or TBE buffer system and Biorad equipment. 1xTTE buffer was found to give slightly higher band resolution as was recommended for sequencing PCR products as a 'glycerol tolerant' buffer. The denaturing 6% 19:1 acrylamide:bisacrylamide gel contained 7M urea and polymerisation was catalysed by 75µl TEMED and 150µl 25%(w/v) ammonium persulphate in a total of 100ml gel mix. Following pre-running the gel at 50W to warm the gel to 40-45°C, 4µl of each samples were loaded using a shark's tooth comb and electrophoresis at constant 45W 40-50°C carried out until the bromophenol blue was about 2cm above the bottom of the gel (length of gel is 50cm). For extended reads of sequence, electrophoresis was carried out overnight (16 hours) as above, but with a fixed voltage of 1250V. After transfer of the gel to Whatman 3MM paper and drying on a gel dryer for 1 hour at 80°C, the dried gel was routinely exposed to Kodak Biomax MR film overnight.

2.1.3 Fluorescent sequencing

2.1.3.1 Preparation of the sequencing reaction

DNA was sequenced using BigDye™ fluorescently labelled terminators and electrophoresis on an ABI Prism 377XL (Perkin Elmer Biosystems). The kit is supplied as a Terminator Ready Mix which contains AmpliTaq DNA Polymerase FS, pyrophosphatase, deoxynucleotides (dATP, dCTP, dITP, dUTP), $MgCl_2$ and Tris-HCl buffer at pH 9.0. Also included in the mix are the fluorescently labelled terminators - each containing a fluorescien donor dye linked to one of four dichlororhodamine acceptor dyes. 8µl of Terminator Ready reaction Mix were mixed with 3.2pmol primer and 1µl treated PCR product, made up to 20µl with H_2O and overlaid with

40µl of mineral oil. Cycling was performed on a Perkin Elmer DNA Thermal Cycler 480 using the following protocol: 96°C for 30 seconds, 58°C for 15 seconds, 60°C for 4 minutes, 25 cycles. Sequencing of fosmid templates was as above except 5µl of DNA solution (~4µg DNA) was used as template, 16µl of Ready Reaction Mix and 10pmol primer was used in a final volume of 40µl. Additionally, the 50°C step in the cycling protocol was for 10 seconds and the total number of cycles was 32.

After cycling and removal of the overlaying mineral oil, the sequencing products were purified by passage trough a CentriSep™ spin column (Perkin Elmer Biosystems) according to the manufacturer's instructions. A column was pre-equilabrated with H_2O, spun at 750g for 2 minutes, and the eluate discarded. The 20µl reaction was added to the column and spun again at 750g for 2 minutes. The eluate was dried using a vacuum centrifuge and the reaction stored dry at -20°C.

2.1.3.2 Electrophoresis of sequencing products

The following steps were performed as part of a laboratory sequencing service by either Ms Wendy Pratt or Ms Katie Morrison. The electrophoresis gel consists of 18g urea, 5.2ml 40%(v/v) 19:1 acrylamide SequagelXR (National Diagnostics), 27.5ml H_2O and 0.5g Amberlite MB-1A mixed-bed resin. 250µl of 10%(w/v) ammonium persulphate solution and 35µl TEMED polymerised the gel mix. Just prior to loading, the pellet was solubilised in 5µl loading buffer (5 volumes formamide, 1 volume 25mM EDTA (pH 8.0) containing 50µg/µl blue dextran) denatured for 2 minutes at 95°C and placed on ice. 1.5µl of the sample was loaded onto the gel and electrophoresis performed at a constant 50W for 7 hours with the gel surface temperature at 50°C.

Computer analysis was performed on an Apple Macintosh running ABI Prism Sequencing Analysis v.3.3 software.

2.1.4 Sequence comparison methods

All sequence analysis was performed by programs supplied as part of the GCG v.9 suite (Genetics Computer Group, Wisconsin) available at the MRC United Kingdom Human Genome Mapping Project Resource Centre (UK-HGMP).

The BESTFIT program, also from the GCG suite, was used for most sequence alignments and identity statistics. The gap creation penalty was set at 50 and the gap extension penalty set at 3. The comparison between pig and human sequence were made using PILEUP from GCG using a gap creation penalty of 5 and an extension penalty of 1. In these cases, percentage sequence identity was determined manually.

2.1.5 Identification of repeat elements

REPEATMASKER from the UK-HGMP was used to search sequences for repetitive elements (Smit and Green). It compares sequence with known human repeat sequences in the Repbase Update database compiled by the Genetic Information Research Institute (GIRI). Over 40% of an average human DNA sequence is masked by the program.

2.2 Electromobility Shift Assay

2.2.1 Caco-2 cell culture

Caco-2 cell culture was performed by Dr Yangxi Wang. Cells (passage 85) were cultured in Dulbecco's Modified Eagles Medium containing 20% heat inactivated fetal calf serum. Cells were harvested 15 days after the previous trypsination when they express maximum levels of lactase, centrifuged at 400g for 10 minutes, the supernatant removed, and the pellet washed with phosphate-buffered saline solution (1x = 0.15M NaCl, 0.01M NaH_2PO_4, 0.0075M NaOH).

2.2.3 Preparation of nuclear-enriched protein extract

The cell pellet was resuspended in 5 volumes of 10mM KCl, 1.5mM $MgCl_2$, 10mM Hepes pH 7.9, incubated on ice for 10 minutes and centrifuged at 250g for a further 10 minutes. Again, the pellet was resuspended in 3 volumes of 10mM KCl, 1.5mM $MgCl_2$, 10mM Hepes pH 7.9, 0.05%(v/v) Nonidet P-40 and homogenized with a tight-fitting Dounce homogenizer to release the nuclei.

Nuclei were pelleted by spinning at 250g for 10 minutes and resuspended in 1ml 1.5mM $MgCl_2$, 0.2mM EDTA, 5mM Hepes pH 7.9, 25%(v/v) glycerol. NaCl was added to a final concentration of 300mM, mixed well, and incubated on ice for 30 minutes. Following a spin at 25000g for 20 minutes at 4°C, the supernatant was aliquoted and snap-frozen at -70°C. Dithiothrietol (DTT) and phenyl methyl sulfonyl fluoride (PMSF) were added to a final concentration of 0.5mM in all solutions just before use.

2.2.4 5' End labelling of oligonucleotides

2 pmoles of double stranded oligonucleotide probe were labelled using 20 units of T4 Kinase (Boehringer Mannheim) and 30µCi γ-^{33}P ATP (Amersham Pharmacia Biotech) in a final concentration of 50mM Tris-HCl, 10mM $MgCl_2$, 0.1mM EDTA, 5mM DTT, 0.1mM spermidine pH 8.2 (@25°C).

After incubation for one hour at 37°C, 1xSTE buffer was added to a final volume of 0.5ml and the solution passed through a NAP-5™ Sephadex column (Amersham Pharmacia Biotech). The probe was eluted in 1ml H_2O.

2.2.5 Electrophoresis

All electrophoresis was performed using Hoefer SL600 equipment (Pharmacia Biotech) and temperature was controlled using a LKB Bromma water bath. The gel was 1.5mm thick 5%(v/v) 29:1 acrylamide:bis-acrylamide in 0.5xTBE, prerun in 0.5xTBE for at least 1 hour at 10°C 150V.

Protein extract (16µg or 40µg) with 2µg of poly(dI-dC) (Boehringer Mannheim) and unlabelled competitor probe in binding buffer (final concentration 20mM HEPES pH 7.6, 1mM EDTA, 10mM $(NH_4)_2SO_4$, 1mM DTT, 0.2%(v/v) Tween-20, 30mM KCl) was incubated for 15 minutes on ice. 10fmol ^{33}P-labelled probe was then added and the mixture incubated on ice for a further 35 minutes. 6x Loading buffer (6x=60% (w/v) glycerol, 0.2% (w/v) bromophenol blue, 0.25xTBE) was added to a final volume of 30µl and routinely 10 or 20µl loaded on the gel. Following electrophoresis at 10°C, limiting at 150V, current approximately 20mA for 2hr 45min, the gel was dried and exposed to Kodak Biomax MR film or a phosphorimager screen.

2.3 RNA analysis

2.3.1 RNA extraction from intestinal biopsies

Collection of all biopsies was performed by clinical staff and co-ordinated by Dr Clare Harvey from the Queen Elizabeth Hospital, Hackney. Biopsies from infant patients (between 1 and 203 months) were obtained using a double-port paediatric capsule positioned at the duodenal-jejunal fissure. A sample was immediately flash-frozen at -70°C.

Biopsies from adult patients were obtained from University College and Middlesex Hospitals by Crosby capsule or multiple pinch biopsy. RNA was isolated by Dr Yangxi Wang from flash frozen biopsy samples using the guanidium / ni thiocyanate-phenol chloroform method of Chomczynski and Sacchi, 1987. After drying in air, the RNA pellet was dissolved in Diethylpyrocarbonate(DEPC)-water and the concentration of RNA estimated spectrophotometrically, and the solution stored at -70°C. DNA was prepared from the residue by Dr Yangxi Wang using a standard phenol-chloroform extraction procedure (Sambrook *et al.*, 1989).

2.3.2 RT-PCR

5µl of a 0.2µg/µl total RNA solution was heated for 5 minutes at 95°C. This was combined with 5.2nmol of each dNTP, 400pmol random hexamers (Amersham Pharmacia Biotech),400 units of M-MLV reverse transcriptase (Gibco BRL), 2µl 10x reaction buffer (10x= 0.5M KCl, 15mM MgCl$_2$, 1%(v/v) Triton X-100, 0.1M Tris-HCl pH 8.8@25°C), and H$_2$O added to a final volume of 20µl. After incubation at 37°C for 1 hour, the reaction was stopped by heating for 5 minutes at 95°C, and the resulting cDNA solution diluted 1:10 with H$_2$O.

Amplification was performed using 10µl of the diluted cDNA solution from above (equivalent to 0.5µg of starting RNA) in a PCR reaction containing 18nmol of each dNTP, 50pmol of each oligonucleotide primer, 9µl of 10x reaction buffer (10x= 0.5M KCl, 15mM MgCl$_2$, 1%(v/v) Triton X-100, 0.1M Tris-HCl pH 8.8) and the final volume adjusted to 100µl with H$_2$O prior to addition of 2 units of Taq polymerase (Promega). Following addition of 50µl mineral oil, cycling was carried out in a Hybaid Omnigene for 30 seconds at 94°C, 40 seconds at 52°C and 1 minute at 70°C for 25 cycles. Amplification reactions without reverse transcriptase and without template were performed to control for genomic DNA amplification and DNA contamination respectively. This protocol represents the optimised method for semi-quantitative mRNA analysis as performed by Dr Yangxi Wang (Wang *et al.*, 1994).

2.3.3 Sequencing and phosphorimage analysis

Sequencing of exon 1, exon 2 and exon 17 PCR products of cDNA and PCR products of genomic DNA was carried out as above (section 2.1.2). After drying, the gel was exposed to a phosphorimager screen for 1 to 2 hours. The screen was then read by a Fujix BAS-1000 phosphorimager and an image of the radioactivity on the gel was produced, which was analysed using MACBAS on an Apple Macintosh. The software allows quantification of amount of radioactivity, which is expressed as a photo-stimulated luminescence (PSL) value and density of radioactivity over a fixed area, which is expressed as the PSL value divided by the area in mm^2. The PSL value has been shown to be directly proportional to radioactivity measured as disintigrations per minute for the radioisotopes ^3H, ^{14}C, ^{35}S, ^{32}P and ^{125}I, and the linear relationship holds for ^{33}P (Jeganathan, D., personal communication). The area was drawn as a box enclosing the band of interest on the screen. The level of a particular sequencing band can be described as the amount of radioactivity within the band minus the background signal. If the area of the box remains the same when comparing bands in the same sequencing lane, then density of signal is equivalent to amount of signal.

In the following example, it is arbitrarily assumed that the polymorphic site investigated has a C allele and a T allele, and the ratio of C allele to T allele is to be determined. If both alleles are present then two bands should be visible at the same position on the sequence ladder, one in the C lane and one in the T lane, and by comparing intensities of these two bands an estimate of the relative amounts of the two allelic sequences can be determined. In order to compare band intensity of the C allele with the band intensity of the T allele, the different levels of signal shown by all the bands in one lane as compared to the other lane must be taken into account. This was achieved by comparing one non-polymorphic band in one lane with another non-polymorphic band in the other lane. The ratio of these two control band densities was then used as a correction factor for the polymorphic band intensity in one lane. This also corrected for any small difference generated if the box size measuring the C lane bands was slightly different to the box size measuring the T lane bands. A formula of the above approach is shown below.

$$\%C \text{ of total sequence} = \frac{x(C_{pm} - C_{background})}{(T_{pm} - T_{background}) + x(C_{pm} - C_{background})}$$

$$\text{where the control factor } x = \frac{(T_{control} - T_{background})}{(C_{control} - C_{background})}$$

which can be simplified to :

$$\%C = \frac{(T_{control} - T_{background})(C_{pm} - C_{background})}{(T_{control} - T_{background})(C_{pm} - C_{background})+(C_{control} - C_{background})(T_{pm} - T_{background})}$$

This formula was used to estimate the percentage of each allele at a polymorphic site, and the percentage was then expressed as a fraction of 1. This represents the fraction of total transcript/genomic sequence due to the sequence of the particular allele.

2.4 Polymorphism detection

2.4.1 Polymerase chain reaction (PCR)

See also 2.4.6 Allele-specific PCR.

PCR was usually carried out in a final aqueous volume of 50µl containing reaction buffer (10x=0.75M Tris-HCl pH 8.8@25°C, 0.2M $(NH_4)_2SO_4$, 0.1%(v/v) Tween, 25mM $MgCl_2$), dNTP mix (10x=2mM dATP, 2mM dCTP, 2mM dGTP, 2mM dTTP, Amersham Pharmacia Biotech), 1.25 units Taq DNA polymerase (Advanced Biotechnologies), 25pmol of each oligonucleotide primer, and made up to 49µl with H_2O. After gentle mixing, 1µl of genomic DNA (~100ng/µl) is added, and the reaction mix overlaid with 50µl of mineral oil (Sigma).

Non-specific bands occasionally appeared in the product after analysis by agarose gel electrophoresis, and this can be remedied by repeating the PCR reaction using a technique called "Hotstart". This was carried out as above, except the Taq DNA polymerase was not added until the reaction mix containing the DNA has been heated for 5 minutes at 95°C, and subsequently cooled to 80°C. This removes non-specific extension products generated as the reaction mix heats gradually up to 95°C for the first cycle.

Exact cycle conditions varied according to the primers used and length of product. The annealing temperature of the primer depend on its T_m, the temperature at which 50% of the oligonucleotide would be bound to the complementary DNA at a specific salt concentration. Usually, the T_m of each oligonucleotide primer in a pair was designed to be 60°C by counting the number of A/T and C/G nucleotides, and counting 4°C for every C or G, and 2°C for every A or T. Ideally, the ratio of GC to AT should be 1:1. A generalised cycling program for a pair of primers with a T_m of 60°C is 95°C for 30 seconds, 58°C (annealing temperature) for 30 seconds and 72°C for 1 minute, for 32 cycles with a final extension of 72°C for 5 minutes. Appendix A lists the sequence of all primers used in the work described by this thesis.

Cycling was routinely performed on an MJ Research DNA Engine on calculation mode, which allows it to estimate mathematically the temperature inside the tube. In some cases, a Hybaid Omnigene was used in tube mode, which measures the temperature inside a tube filled with oil using a thermocouple; and in other cases a Techne Phoenix machine was used in block mode, which simply measures the temperature of the heating block. The machine used and the mode of temperature measurement was usually not critical.

PCR on bacterial DNA was occasionally performed as above, except 1µl bacterial cell suspension replaced the genomic DNA. This was pre-heated to 99°C for 10 minutes to lyse the cells.

2.4.2 Agarose gel electrophoresis

PCR products and restriction enzyme digests of PCR products were routinely analysed on 1%-2% agarose gels in 1xTBE, as described in Sambrook *et al.*, 1989. Ethidium bromide was added to gel and buffer to a final concentration of 0.5µg/ml, prior to pouring, and a mixture of 8µl of PCR product/restriction enzyme digest and 2µl loading buffer (15%(w/v) Ficoll (Type 400, Amersham Pharmacia Biotech), 0.1%(w/v) bromophenol blue, 0.1%(w/v) xylene cyanol FF in H_2O) was loaded in each lane. A marker was run in one lane, usually ϕX174 HaeIII digest (Gibco BRL) but occasionally 1kb ladder (Gibco BRL) or Superladder (Advanced Biotechnologies). Gels were stained in ethidium bromide (0.5µg/ml) and visualised under ultra-violet light.

2.4.2 Denaturing Gradient Gel Electrophoresis (DGGE)

Denaturing gradient gel electrophoresis is a method of detecting mutations based on the fact that DNA base changes will alter the physical properties of the DNA duplex; in this case the amount of denaturant required to dissociate (or melt) the duplex (Fischer and Lerman 83). The technique was developed with the addition of a GC-rich sequence (known as a GC-clamp) on one end by positional cloning (Myers 85) or, when used on PCR products, a modified oligonucleotide. This ensured that denaturation occurred from only one end, that the whole PCR product was analysed, and if the DNA duplex was electrophoresed though a denaturing gradient gel, the point at which it melted and arrested in the gel varied. In some cases one end of the duplex is naturally GC-rich so acting as a natural GC-clamp and ensuring melting from the other end.

Conditions for the analysis of the 5F AvaII digested fragments by DGGE were determined by Dr Clare Harvey (Harvey, 1994). Briefly, the sequences were analysed by MELT87 and SQHTX (Lerman and Silverstein, 1987) which predicted the melting profile of the DNA, and allowed the gel conditions to be estimated.

Electrophoresis was performed on a suitably modified Hoefer SE 600 electrophoresis apparatus, with circulation of 1xTAE at 61°C. 0.75mm thick gels were composed of 10%(v/v) acrylamide (37.5:1) in 1xTAE with a linear 40% to 50% gradient of denaturant (100% denaturant = 7M urea, 40%(v/v) deionised formamide). 5F Ava II digests (1µl to 5µl, depending on concentration of PCR product estimated by agarose gel electrophoresis) was diluted to a final volume of 5µl with H_2O and mixed with an equal volume of loading buffer (20%(w/v) Ficoll, 0.5%(w/v) bromophenol blue, 10mM Tris-HCl pH 7.8, 1mM EDTA). 5µl was applied to preheated gels which were placed in preheated buffer. Electrophoresis was for 22 hours with voltage limiting at 35V.

2.4.3 Single strand conformation analysis (SSCA)

Single strand conformation analysis (SSCA) allows detection a single nucleotide changes and small deletions in DNA. The technique commonly uses PCR products visualised on a non-denaturing acrylamide gel by silver staining (Harvey *et al.*, 1995; Yip *et al.*, 1999). The technique relies on the fact that DNA, after denaturation by heating and renaturation by rapid cooling, will form both double stranded DNA and self-annealed single stranded DNA. This self-annealed single stranded DNA will form different conformations depending on the base composition of the sequence, and this is often to just one DNA base change.

Electrophoresis was performed on a Hoefer SE 600 apparatus attached to a circulating water bath (LKB Bromma or Grant). 5FA/5FS Ava II digests (1µl to 5µl, depending on concentration of PCR product estimated by agarose gel electrophoresis) was diluted to a final volume of 5µl with H_2O, mixed with an equal volume of loading buffer (95%(v/v) formamide, 0.5%(w/v) bromophenol blue, 0.5%(w/v) xylene cyanol FF) and heated on a heat block to 85°C for 3 minutes. The gel was 6%(37.5:1) acrylamide, 5.5%(w/v) glycerol, 1xTBE; and polymerised using 50µl TEMED and 250µl ammonium persulphate (25%(w/v) solution) per 30ml gel solution.

After rapid cooling on ice, 3µl was loaded carefully in the bottom of the well. Electrophoresis was at 25°C with voltage limiting at 400V for 2 hours.

F2A/F2S PCR products were electrophoresed as above except the gel was 6%(37.5:1) acrylamide, 6.75%(w/v) glycerol, 0.5xTBE; and electrophoresis was at 20°C for 1¾ hours.

2.4.4 Simple acrylamide gel electrophoresis (SAGE)

2µl UTA/UTS PCR product was mixed with 4µl agarose gel loading buffer, 14µl H_2O and 2µl of the mix loaded carefully at the bottom of the well on a 6%(37.5:1) acrylamide gel with 1xTBE as electrophoresis buffer. Electrophoresis was at 25°C with voltage limiting at 400V until the xylene cyanol dye was 4cm from the bottom of the gel.

2.4.5 Silver staining of acrylamide gels

This method was used on all SAGE, SSCA and DGGE gels. All solutions were made with Milli-Q reverse osmosis water. Gels were fixed in 10% ethanol 0.5% glacial acetic acid using two three-minute incubations, and stained in 0.1% silver nitrate for 10 minutes. Following a brief rinse in H_2O, the gels were incubated in developing solution (375mM NaOH, 2.6mM sodium borohydride, 0.148%(v/v) folmaldehyde) for a few minutes until bands were visible. After a final rinse in H_2O, the gels were drained for 10 minutes, transferred to Whatman 3MM paper and dried on a gel dryer for 2 hours at 80°C.

2.4.6 Allele Specific PCR

Allele specific PCR (AS-PCR, also known as ARMS, Amplification Refractory Mutation System) is a method of selective amplification of one allele. PCR was as above, except oligonucleotide primers were designed with the 3′ base complementary to the polymorphic base. For each polymorphism, two oligonucleotide primers were designed, each corresponding to an allele. Allelic phase and allelic status of each polymorphism (haplotype) can be determined by using double AS-PCR: a pair of allele specific primers in PCR amplification allows amplification only when certain haplotypes are present in the DNA amplified (figure 2.1). Table 2.1 shows the annealing temperatures which were determined for each oligonucleotide pair to amplify only the specified pair of alleles.

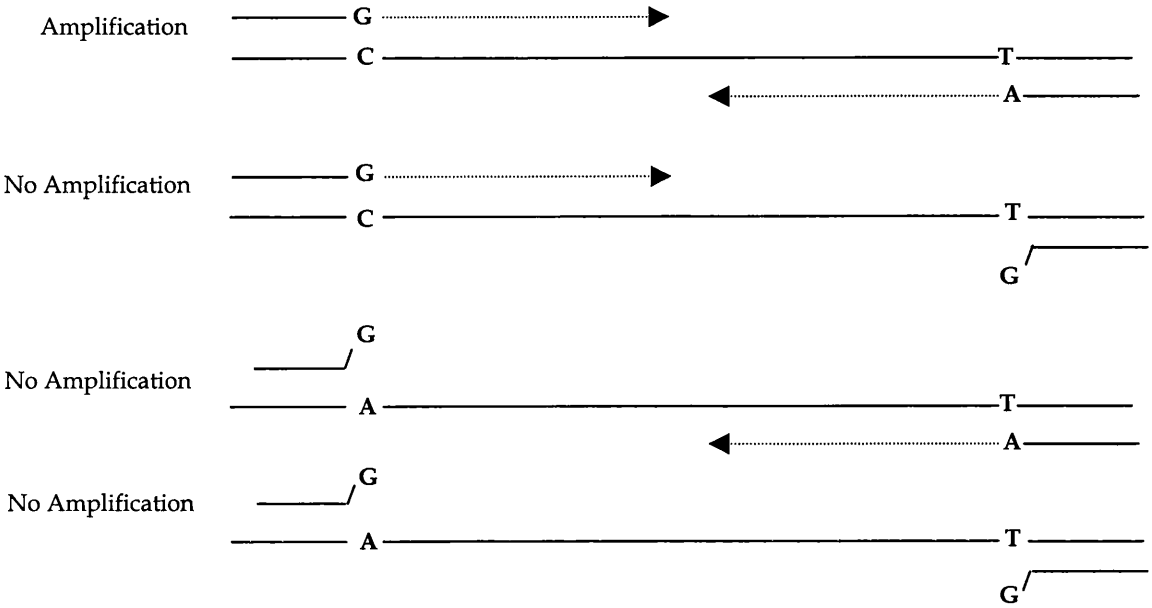


Figure 2.1 The principle of double AS-PCR
Amplification occurs only when both primers match the genomic sequence at their corresponding polymorphic sites.

Polymorphic sites	Allelic combination	Primers	Annealing temperature °C
T5579C TG6236/7ΔΔ	C/TG	X17ARMS-S X17ARMSIA	63 (5 cycles) 62 (28 cycles)
T5579C TG6236/7ΔΔ	C/ΔΔ	X17ARMS-S X17ARMSDA2	58
T5579C TG6236/7ΔΔ	T/TG	X17ARMS+S X17ARMSIA	63 (5 cycles) 62 (28 cycles)
T5579C TG6236/7ΔΔ	T/ΔΔ	X17ARMS+S X17ARMSDA2	63 (5 cycles) 62 (28 cycles)
TG6236/7ΔΔ CATT+225ΔΔΔΔ	TG/CATT	+10ARMSSI +271ARMSAI	60
TG6236/7ΔΔ CATT+225ΔΔΔΔ	TG/ΔΔΔΔ	+10ARMSSI +271ARMSAD	60
TG6236/7ΔΔ CATT+225ΔΔΔΔ	ΔΔ/CATT	+10ARMSSD +271ARMSAI	60
TG6236/7ΔΔ CATT+225ΔΔΔΔ	ΔΔ/ΔΔΔΔ	+10ARMSSD +271ARMSAD	60
CATT+225ΔΔΔΔ C+658T	CATT/C	+271ARMSSI +704ARMSAC	61
CATT+225ΔΔΔΔ C+658T	CATT/T	+271ARMSSI +704ARMSAT	60
CATT+225ΔΔΔΔ C+658T	ΔΔΔΔ/C	+271ARMSSD +704ARMSAC	60
CATT+225ΔΔΔΔ C+658T	ΔΔΔΔ/T	+271ARMSSD +704ARMSAT	60

Table 2.1 Primers and annealing temperatures used for AS-PCR

2.5 *Microbiological methods*

2.5.1 Cloning PCR products

PCR products were cloned using the TOPO-TA Cloning kit^{TM} (Invitrogen) according to the manufacturer's instructions. 1μl of a strong PCR product (~25ng), as analysed by agarose gel electrophoresis, was incubated with 10ng pCR2.1-TOPO plasmid vector in 10%(w/v) glycerol, 10mM Tris-HCl pH7.4 at 25°C, 0.2mM EDTA, 0.2mM DTT, 0.02%(v/v) Triton X-100, 20μg/ml BSA for 5 minutes at 25°C. The plasmid is activated with a topoisomerase present in the vector mix which allows ligation to occur. 2μl of this mixture was then added to OneShot TOP10^{TM} *E.coli* competent cells (genotype F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^{R}) *end*A1 *nup*G) previously treated with 1μmol β-mercaptoethanol, and incubated on ice for 15 minutes.

The cells were heat-shocked at 42°C for 30 seconds, then added to 250μl SOC medium (2% Tryptone, 0.5% yeast extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10mM $MgSO_4$, 20mM glucose), and placed in a shaking incubator (225rpm) for 30 minutes at 37°C. 50-100μl of the cells were then spread on a pre-warmed plate of LB-agar coated with 40μl of 40mg/ml 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-Gal) and containing 50μg/ml ampicillin, and incubated at 37°C overnight. Twelve white or pale blue colonies were picked and grown overnight in 5ml LB broth with 50mg/ml ampicillin. 1μl of the cells were then boiled and analysed by PCR amplification.

2.5.2 Preparation of clone DNA from fosmid vector

Clones were grown up from glycerol stock by plating 1μl of glycerol stock on and LB agar plate containing 25μg/ml chloramphenicol, and following incubation overnight, one colony incubated in 10ml of LB broth for a further night. 20μl of this culture was then transferred to another 10ml of LB broth and incubated during the day. 500μl of this culture was the/added to 500ml of LB broth and incubated, with n shaking, overnight. All LB broth was treated with chloramphenicol to a final concentration of 25μg/ml. Cells were harvested by centrifugation at 6000g for 15 minutes at 4°C and removal of supernatant. Fosmid DNA was extracted using the Maxi-kit from Qiagen (Crawley, Sussex) using the very low-copy protocol. The buffers used are as follows:

P1	50mM Tris-Cl, pH 8.0; 10mM EDTA; 100µg/ml RNase A
P2	200mM NaOH; 1%(w/v) SDS.
P3	3.0M potassium acetate, pH 5.5
QBT	750mM NaCl; 50mM MOPS, pH 7.0; 15% isopropanol; 0.15%(v/v) Triton X-100
QC	1.0M NaCl; 50mM MOPS, pH 7.0; 15% isopropanol
QF	1.25M NaCl; 50mM MOPS, pH 7.0; 15% isopropanol
QN	1.6M NaCl; 50mM MOPS, pH 7.0; 15% isopropanol
STE	100mM NaCl; 10mM Tris-Cl, pH 8.0; 1mM EDTA
TE	10mM Tris-Cl, pH 8.0; 1mM EDTA

The bacterial pellet was resuspended thoroughly in 20ml P1, and then 20ml P2 added, mixed gently by inversion and incubated at room temperature for 5 minutes. 20ml chilled P3 was then added to the viscous lysate, mixed gently and incubated, with occasional mixing, on ice for 30 minutes. Following centrifugation at 20000g for 30 minutes at 4°C, the supernatant was removed and recentrifuged, again at 20000g for 30 minutes at 4°C. This clears the lysate of all precipitated material.

The DNA was precipitated by addition of 0.7 volumes of room-temperature isopropanol and centrifugation at 20000g for 30 minutes at 15°C. The DNA pellet was then dissolved in 500µl 1xTE and 5ml buffer QBT added. This solution is then added to a QIAGEN-tip pre-equilibrated with buffer QBT, and allowed to enter the resin by gravity flow. The QIAGEN-tip was then washed with 2x30ml buffer QC and eluted with 15ml buffer QF pre-heated to 65°C. DNA was precipitated from the eluate by adding 0.7 volumes of isopropanol at room temperature and centrifuged at 15000g for 30 minutes at 4°C. After decanting the supernatant, the pellet was washed with ~4 ml 70% ethanol at room temperature, followed by centrifugation at 15000g for 10 minutes. The supernatant was discarded, the pellet dried in air and then redissolved in 250µl of 1xTE. The typical concentration of DNA obtained was ~0.5µg/ml

2.6 *Physical mapping*

2.6.1 Random priming radioactive labelling of PCR product

PCR product was treated by Exonuclease 1 and Shrimp alkaline phosphatase as for preparation for sequencing, and labelled using the Multiprime™ labelling kit (Amersham Pharmacia Biotech) essentially as the manufacturer's instructions. The equivalent of 1μl of PCR product (~25ng) was heated for 10 minutes at 99°C, then added to 5μl (50μCi) [α-^{32}P]dCTP (3000Ci/mmol) in a buffer containing random hexanucleotides, deoxynucleotides, and BSA. 2 units of DNA polymerase 1 "Klenow fragment" (Amersham Pharmacia Biotech) were added and incubated at 37°C for 1 hour. The labelled probe was then purified from the unincorporated label by spinning through a Sephadex column.

2.6.2 Screening of chromosome 2 libraries

The human chromosome 2 genomic library used in this work was constructed at the Human Genome Center, Biology and Biotechnology Research Program. L-452, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA under the auspices of the National Laboratory Gene Library Project sponsored by the US Department of Energy. The library filters were obtained from the UK-HGMP. These were incubated in 250ml pre-hybridisation solution (6xSSC, 5xDenhardt's solution, 0.5%SDS) for 2 hours at 65°C. 65μg Herring sperm DNA was added to the ^{32}P-labelled probe was boiled for 5 minutes and 50μl added to the pre-hybridisation solution containing the filters, and incubated for 16 hours at 65°C. Positive clones were ordered from the UK-HGMP.

The filters were then washed with four dilutions of SSC starting at 2x and finishing with 0.2x containing 0.1% SDS for 15 minutes each; 2xSSC at 25°C, 2xSSC at 65°C, 0.5xSSC at 65°C and 0.2xSSC at 65°C, 500ml each. The filters were then exposed to Fuji Super-HRE30 film at -70°C.

2.6.3 Probe labelling

2.6.3.1 Biotin labelling

DNA was labelled with biotin using the Bionick Labeling Kit (Gibco BRL). 1μg DNA was incubated with 2.5 units DNA polymerase I, 0.0375 units DNase I, 0.02mM dCTP, 0.02mM dGTP, 0.02mM dTTP, 0.01mM dATP, 0.01mM biotin-14-dATP, 1μg BSA in Tris-HCl (pH 7.8@25°C), 5mM $MgCl_2$ in a final volume of 50μl at 16°C for at least 1 hour. 5μl 500mM EDTA pH 8.0 was added to stop the reaction.

The reaction products were then purified by ethanol precipitation. 200ng of probe (equilvalent to 11μl of the reaction mix described above) together with 10μl human Cot-1 DNA (1μg/μl) and 2μg herring sperm DNA (1μg/μl), 0.1V 3M sodium acetate (pH5.2@25°C) and 2V ice cold absolute ethanol. The solution was mixed then incubated at -70°C for 30-45 minutes. Following a spin in a microfuge for 5 minutes at 15000g, the supernatant was discarded and the pellet dried in air.

2.6.3.2 Digoxygenin labelling

DNA was labelled with digoxygenin (DIG) using the DIG-Nick Translation Mix (Boehringer Mannheim). 1 μg DNA was incubated with DNA polymerase I, DNase I, 0.05mM dATP, 0.05mM dCTP, 0.05mM dGTP, 0.034 mM dTTP, 0.016mM DIG-11-dUTP in the buffer supplied in a final volume of 20μl at 16°C for at least 1 hour. 1μl 500mM EDTA pH 8.0 was added to stop the reaction, and the reaction products purified by ethanol precipitation, as in section 2.6.3.1 above.

2.6.4 Metaphase FISH

2.6.4.1 Culture of lymphoblastoid cell lines

Lymphoblastoid cell lines were established from blood by ECCAC, and stored in liquid nitrogen. Cells were grown in 1xRPMI (Gibco BRL) containing 10% foetal calf serum and a final concentration of 2mM glutamate, 60μg/ml streptomycin and 100μg/ml penicillin. Incubation was at 37°C in a controlled moist atmosphere containing 5% CO_2.

2.6.4.2 Culture of fresh blood

A culture was set up using 1ml of fresh blood, 16ml Iscoves modification of Dulbecco's medium (or alternatively 1xRPMI), 2ml Fetal calf serum, 0.4ml phytohaemagglutunin (Sigma) and incubate for 72 hours at 37°C. After incubation, 6mg thymidine was added and incubated at 37°C for a further 17 hours. The blood culture was then divided between two tubes of warm Iscoves/10% fetal calf serum, and centrifuged at 180g for 5 minutes. The supernatant was discarded and the pellet resuspended in 5ml Iscoves/10% fetal calf serum. Again the suspension was centrifuged at 180g for 5 minutes and the supernatant discarded, and the pellet resuspended in 5ml Iscoves/10% Fetal calf serum. The suspension was incubated for 4 hours 35 minutes at 37°C. 15 minutes before harvesting the cells, 0.5μg of colcemid (Gibco BRL) was added to the culture.

2.6.4.3 Cell harvest and preparation of metaphase spread slides

The 5ml of cell suspension (from blood or lymphoblastoid cell culture) was centrifuged at 180g for 5 minutes, and supernatant removed. 5ml of freshly prepared fixative (3:1 methanol:glacial acetic acid) was added and left for 30 minutes at 4°C. After centrifuging at 180g for 5 minutes, the fixative was discarded and replaced with fresh fixative. The suspensions were incubated for 16 hours at 4°C. Slides were cleaned in methanol to which a few drops of HCl had been added. The cell suspensions were centrifuged, the supernatant discarded as above and replaced with enough fresh fixative to create a cloudy suspension. A drop of suspension was dropped from about 12 inches above the slide, which is held at a slight angle. As soon as the slide dried, it was briefly flooded with 1ml of 70% acetic acid in H_2O, and dried in air. The slide was then dehydrated in an ascending ethanol series (70%, 90%, 100%) in Coplin jars, dried in air then stored at 4°C.

2.6.4.4 Slide pre-treatment and hybridisation

Slides were treated with 100µg/ml RNase A (Sigma) in 2xSSC and incubated for 1 hour at 37°C in a humidified box. After rinsing in 2xSSC at room temperature, slides were dehydrated by placing in an ascending ethanol series (70%, 90% and 100%) for 3 minutes each, and dried in air. Slides were then washed in proteinase K buffer (1x = 0.02mM Tris HCl, 0.002M $CaCl_2$, pH 7.4) at 37°C for 10 minutes, then treated with 50ng/ml proteinase K (Boehringer Mannheim) in proteinase K buffer at 37°C for 7 minutes. Slides were then rinsed in 1xPBS at room temperature for 5 minutes, and transferred to 1% formaldehyde solution (50ml 1xPBS, 0.5g $MgCl_2.6H_2O$, 1.3ml 37% formaldehyde which has been saturated with $NaHCO_3$) and incubated at room temperature for 10 minutes. After rinsing in 1xPBS, slides were dried using the ascending ethanol series as before.

100µl of 70%(v/v) formamide in 2xSSPE was then pipetted onto each slide, covered with a cover slip and placed in a fan oven at 75°C for 3.5 minutes. The cover slips are then removed and the slides are then placed in ice cold 70% ethanol for 2 minutes, then dehydrated further though cooled 90% and 100% ethanol for 2 minutes each. After drying in air, the probe was resuspended in in hybridisation solution (2xSSPE, 1%(w/v) dextran sulphate, 50%(v/v) formamide, pH7.2@25°C), denatured at 75°C for 5 minutes and preannealed at 37°C for 30 minutes. 5µl of each probe was carefully pipetted on to a cover slip, which was then placed on the slide. The edges of the cover slip were sealed with Cow Gum (Cow Proofings Ltd, Slough, Berkshire) and incubated overnight at 37°C in a humidified box.

2.6.4.5 Signal detection

After removal of the Cow Gum, the slide was placed in a Coplin jar and washed in three changes of 50ml 50% formamide in 2xSSC at 42°C. The slide was then washed in two changes of 2xSSC for 2.5 minutes each at 42°C, and finally two changes of 0.1xSSC for 2.5 minutes each at 42°C.

The slide was then transferred to 50ml 4xSSC containing 0.05%(v/v) Tween 20 and incubated, with shaking, for 5 minutes at room temperature. It was then incubated in 50ml 4xSSC 2.5%(w/v) non-fat milk powder (Marvel brand) for 20 minutes at room temperature.

After draining, 100µl of fluoresceinisothiocyanate(FITC)-avidin solution (5µg/ml FITC-avidin (Vector Laboratories) in 0.1M Tris-HCl, 0.15M NaCl pH 7.5, 0.5%(w/v) blocking agent (Boehringer Mannheim)) was placed on the slide which was covered with a coverslip, and incubated at 37°C for 20 minutes in the dark. This and subsequent steps were performed protected from light since fluorescent dyes are sensitive to light. The slide was then washed in three changes of 50ml 4xSSC containing 0.05%(v/v) Tween 20 for 5 minutes each at room temperature, then drained briefly. 100µl of biotinylated anti-avidin solution (5µg/ml biotinylated anti-avidin (Vector Laboratories) in 0.1M Tris-HCl, 0.15M NaCl pH 7.5, 0.5%(w/v) blocking agent (Boehringer Mannheim)), the slide covered with a coverslip and incubated at 37°C for 20 minutes. The slide was again washed in three changes of 50ml 4xSSC containing 0.05%(v/v) Tween 20 for 5 minutes each at room temperature, then drained briefly. 100µl of FITC-avidin/anti-DIG rhodamine (5µg/ml FITC-avidin (Vector Laboratories) and 10µg/ml anti-DIG rhodamine (Boehringer Mannheim) in 0.1M Tris-HCl, 0.15M NaCl pH 7.5, 0.5%(w/v) blocking agent (Boehringer Mannheim)), the slide covered with a coverslip and incubated at 37°C for 20 minutes. The slide was washed in two changes of 50ml 4xSSC containing 0.05%(v/v) Tween 20 for 5 minutes each at room temperature, then rinsed in 1xPBS for 5 minutes. The slide was then dehydrated for 3 minutes each in an ascending ethanol series (70%, 90%, 100%) in Coplin jars, and dried in air. After mounting in 20µl Vectashield antifade with DAPI (Vector Laboratories), the slide was covered with a coverslip and kept at 4°C until microscopy.

2.6.5 Strand FISH

Strand FISH, also known as fiber-FISH, involves combing genomic DNA on a specially prepared cover slip so that long strands of DNA extend along the slide. Fluorescently labelled clone DNA is used as a probe to hybridise with the genomic DNA. The technique has been used to analyse deletions in the TSC1 gene (Michalet *et al.*, 1997).

2.6.5.1 Preparation of agarose blocks

A pellet of cultured lymphoblastoid cells was resuspended in 500µl of 1xPBS and placed on ice. A 1% dilution of the suspension in 1xPBS was placed on a haemocytometer, and the cell density estimated by eye using the dye trypan blue. The total cell suspension was then diluted to a final concentration of 2x10^{7} cells/ml. After incubation of the suspension for 5 minutes at 37°C, an equal volume of molten 1% low melting point NuSieve GTG™ agarose (FMC Bioproducts, Rockland, Maine, USA) was added, the solution swirled gently and poured into 100µl well Teflon block formers. These were placed at 4°C for 10 minutes to set.

The blocks were then placed into pre-digested 1xESP (1x=2mg/ml Proteinase K, 1% Sarcosyl, 0.5mM EDTA pH 8.0) in a volume equivalent to at least 250µl solution for each block, and incubated at 50°C for 48 hours. After incubation the solution is removed and the blocks rinsed twice with 0.5M EDTA pH 8.0. They are then washed three times at 50°C for 30 minutes each in 1xTE containing 40µg of PMSF (from a 40µg/µl solution dissolved in isopropanol). The blocks, each containing 10^{6} cells, (6.6µg DNA), are then stored at 4°C in 0.5M EDTA pH 8.0.

2.6.5.2 Dynamic molecular combing of genomic DNA

Two blocks were washed three times for two hours each wash in 1xTE at room temperature. They were then placed in a 2ml round-bottomed tube, 20µl diluted YOYO stain (0.1mM YOYO-1 (Molecular Probes, Eugene, Oregon, USA), 40mM Tris-HCL 2mM EDTA) added, the volume made to 200µl using 40mM Tris-HCL 2mM EDTA and incubated at room temperature for 1 hour to stain. After staining, the supernatant was discarded, 1ml 1xTE added, and the agarose blocks melted by incubation for 30 minutes at 68°C. 120µl of 10xAgarase-1 buffer (1x=10mM bis-Tris-HCl, 1mM EDTA, pH6.5@25°C) was added and the solution equilibrated to

40°C. After addition of 4 units of β-agarase-1, the solution was incubated overnight at 40°C.

600μl of 0.5M 2-[N-Morpholino]ethanesulfonic acid (MES) pH5.50@25°C was added and the DNA mixed very gently by inverting the tube three times. The solution was then gently poured into a 15ml round-bottomed tube and the volume made up to 4ml with 50mM MES pH 5.50. After heating at 75°C for 30 minutes to mix the DNA, the solution is allowed to equilibrate to room temperature and transferred to the Teflon reservoir.

Silane coated coverslips (22mm^2, from Institut Pasteur) were then combed using a mechanised lift (specially constructed at the National Institute of Medical Research workshop at Mill Hill to move at 300μm/s). The coverslip was then glued, DNA side upwards, to a slide and the combing quality of each coverslip checked using a fluorescence microscope. Following baking of the slides overnight at 60°C, they were dehydrated through an ethanol series (70%, 90%, 100%) and stored at 4°C.

2.6.5.3 FISH analysis of combed DNA slides

Precipitation of labelled probe DNA was by the same method as 2.6.3 except that 1μg of probe DNA is used. The probe pellet is resuspended in hybridisation solution (as 2.6.4.4) and denatured at 100°C for 5 minutes. The slides were equilibrated to room temperature and dehydrated through the ethanol series as described previously. The slides were then denatured in a Coplin jar containing 70% formamide 2xSSC for 2 minutes at 75°C, then placed in ice-cold ethanol for 3 minutes, and then placed through the ethanol series as above. After drying in air, the slides were covered with a coverslip with the probe on it, sealed with Cow Gum, and incubated overnight at 37°C.

The slides were washed three times (5 minutes each) in a Coplin jar containing 2xSSC/50%(v/v) formamide at 37°C, then washed three times (5 minutes each) in a Coplin jar containing 2xSSC at room temperature. 50μl of 1.5% blocking solution (0.375g blocking agent, 25ml 4xSSC/0.05%(v/v) Tween-20 heated to 68°C and stored at -20°C) was added to each slide and incubated in a humidified box for 30 minutes at 37°C with coverslip.

All antibodies involved in the detection were diluted using a mix of 1.5% blocking solution and 4xSSC/0.05%(v/v) Tween-20 in the ratio of 2:1 and a final volume of 10μl applied to each slide at each step. The slides were washed using

4xSSC/0.05% (v/v) Tween-20 at room temperature for three five minute washes. All incubations were in a humidified box at 37°C with coverslip. The first step is a 30 minute incubation with 1/50 dilution of avidin-Texas red (Vector Laboratories) and 1/50 dilution anti-Dig mouse-FITC (Interchim) followed by the washes. The second step is a 30 minute incubation with a 1/100 dilution of biotinylated anti-avidin (Vector Laboratories) and a 1/50 dilution of donkey anti-mouse-FITC (Interchim), followed by washes. The third step is a 30 minute incubation with a 1/50 dilution of avidin-Texas red and a 1/50 dilution of mouse anti-rabbit-FITC (Interchim) , followed by washes. The fourth step is a 20 minute incubation with a 1/100 dilution of biotinylated anti-avidin and a 1/200 dilution of anti-FITC (Cambio), followed by washes. The final antibody step is a 20 minute incubation with a 1/50 dilution of avidin-Texas red and a 1/100 dilution of anti-[anti-FITC]-FITC (Cambio), followed by washes.

The slides were then washed in 1xPBS at room temperature for five minutes, and then a drop of Vectashield (with no DAPI) placed on a coverslip, which is then placed on the slide. The edges were then sealed with nail polish, and stored at 4°C in the dark.

2.7 Population genetic methods and statistics

2.7.1 Fisher's exact test

This test involved constructing a two-by-two contingency table, implemented by the program 2BY2, which gave the one-sided p-value which is the probability of obtaining a table as extreme or more extreme than the one observed at random (Ott, 1991).

2.7.2 Linkage disequilibrium and ASSOCIATE

Ott's ASSOCIATE program (Terwilliger and Ott, 1994) analyses population data from two biallelic loci as three-by-three genotype tables and generates the component of χ-squared that is due to allelic association and the corresponding p-value testing the significance of any allelic association. It also generates maximum-likelihood estimates of the frequencies of the four possible haplotypes and the deviation of these frequencies from haplotype frequencies assuming no allelic association, which allowed a value for pairwise linkage disequilibrium between the two loci to be estimated. This was expressed as D' which has a value between 1, which indicates complete linkage disequilibrium, and 0, which indicates complete linkage equilibrium. D′ is calculated as follows:

$$D'=D/D_{max}$$

where D is the positive deviation of the observed haplotype frequencies from the expected haplotype frequencies assuming no association, and D_{max} is the minimum value of *r*(1-*s*) or *s*(1-*r*) where *r* and *s* are the frequencies of the rarer alleles at the two polymorphic loci analysed (Harvey *et al.*, 1995).

2.7.3 Maximum-likelihood estimation of haplotypes using EH

ASSOCIATE estimates the frequency of two-loci haplotypes from phenotype frequencies. EH (Terwilliger and Ott, 1994) is an extension of this method, estimating multi-loci haplotypes from phenotype frequencies, and can be regarded as a more complex form of linkage disequilibrium mapping. Five sites (either single loci or multiple loci where the phase is determined by another method) were used in haplotype analyses.

2.7.4 Estimation of haplotypes by haplotype counting

The extensive linkage disequilibrium across the gene, at least in most populations, signifies that recombinations are rare and haplotypes assumed in the population from the individuals homozygous for all sites. Haplotype counting involved noting the haplotypes of all the individuals homozygous for all sites, or heterozygous for only one. Then in ambiguous individuals heterozygous for more than one site, all possible haplotypes were considered and then compared with known haplotypes in homozygous and individuals heterozygous for one site. The two possible haplotypes at the highest frequencies in the known haplotype individuals were assumed present in the ambiguous individual.

2.7.5 Other statistical tests

Gene diversity or heterozygosity (*H*) is a test implemented in the program HET (Ott) and corresponds to the probability that an individual is heterozygous for a given allele, or in this case haplotype, if the population is in Hardy-Weinberg equilibrium. It is defined as

$$H=\sum_{i\neq j}^{n} p_i p_j$$

for a locus with n alleles, where p_i is the frequency of the i-th allele, and p_j is the frequency of the j-th allele.

2.8 Miscellaneous

2.8.1 Suppliers

Unless stated in the text, the following companies were suppliers of laboratory consumables for this thesis:

New England Biolabs, Beverly, Massuchusetts, USA (restriction enzymes)
Fisher Scientific, Loughborough, Leicestershire (General)
Fisons Scientific Equipment, Loughborough, Leicestershire (General)
Sigma-Aldrich, St Louis, Missouri, USA (General)
Merck BDH chemicals, Poole, Dorset (General)
Biorad Laboratories, Hercules, California, USA (DNA Sequencing equipment)
National Diagnostics, Atlanta, Georgia, USA (Acrylamide solutions)
Advanced Biotechnologies, Epsom, Surrey (Taq polymerase)
Amersham Pharmacia Biotech, Amersham, Bucks (^{33}P labelled nucleotides)
ICN Pharmaceuticals, Costa Mesa, California, USA (^{32}P labelled nucleotides)
Perkin Elemer Biosystems, Warrington, Cheshire (oligonucleotides)

2.8.2 Units

All values measured in this thesis use SI units except radioactive energy, which is measured in Curies (Ci). The SI unit is Becquerel (Bq), with 1Ci = 3.7x10^{10}Bq. All measurements of radioactivity refer to the calculated value at the reference date.

2.8.3 Buffers

All pH values at 25°C.

1xTTE =0.03M Tris-HCl, 0.03M Taurine (2-aminoethane sulphonic acid), 0.5mM Na(EDTA).2H_2O pH 9.2.
1xTBE=0.09M Tris-borate pH 8.0, 2mM EDTA, pH 8.3
1xSTE=0.1M NaCl, 1mM EDTA, 0.01M Tris-HCl pH 8.0
1xTE=1mM EDTA, 0.01M Tris-HCl pH 8.0
1xSSC=0.15M NaCl, 0.015M Sodium citrate pH 7.0
1xRPMI=RPMI 1640 (Gibco BRL)

1xPBS=0.01M phosphate buffer, 2.7mM KCl, 0.137M NaCl pH 7.4

1xSSPE=0.01M phosphate buffer, 0.15M NaCl, 1mM EDTA pH 7.4

2.8.4 Human samples

All human samples were collected with the ethical permission of the appropriate authority of the country of origin. In Russia, where no such authority exists, samples were collected with the fully informed consent of the individuals participating.

3 Sequence analysis and physical mapping of the lactase gene

3.1 Sequence of upstream and intron 1

At the start of this project, only the exonic sequences of lactase and 1kb of upstream sequence from a clone (pLC1) was published. Partial sequence of the 1kb upstream region from individuals of haplotypes A,B,C and D was also published. To extend this knowledge, 2kb upstream of exon 1, intron 1 and 2kb downstream of exon 17 was sequenced from clones. A panel of five individuals was used to search for polymorphisms by sequencing of PCR products of the upstream and downstream sequence, and parts of intron 1. These five individuals have been defined at the mRNA level for lactase persistence/non-persistence genotype by analysis of the levels of expression of mRNA transcripts from intestinal biopsies, and their allelic status is known at all known polymorphisms across the lactase locus (table 3.1). The DNA used is from five cell lines derived from the five patients, so there is an unlimited supply of DNA from these individuals.

3.1.1 Upstream sequencing strategy

Sequence was derived from the clone pLC1 by using the primer 5seqA (figure 3.1). PCR product was generated using the primers shown (figure 3.1) and then sequenced as shown. Further upstream has been sequenced using the same procedure by Mark Poulter.

3.1.2 Analysis of upstream sequence

The sequencing of PCR products from the panel of five individuals confirmed sites of known variation but no new polymorphisms were found.

3.1.3 Intron 1 sequencing strategy

Part of the intron 1 sequence was derived from the clone pLC1, which extended 1.5kb into the intron. A combination of genomic PCR products and the mini-gene clone pDB2 were used to sequence the remaining part of the intron. Using this approach, the remaining sequence was difficult to amplify and sequence.

Since clones from a contig spanning the lactase gene have become available (section 3.4), the complete sequence of intron 1 was generated from a single fosmid clone (fosmid 9, for code see appendix 2). Sequencing was conducted both using primers covering the fragment of intron 1 that had been sequenced and using a sequence walking approach from F2A and INT1S to complete the sequence of the intron. In one case, a primer was designed on repetitive Alu sequence, but to minimise the chances of this primer annealing to another element, the primer was compared with other Alu elements and a consensus Alu sequence to maximise mismatches with these sequences. The primers used to generate all intron 1 sequence and the length of read from each primer are shown in figure 3.2.

To confirm that the clone had no deletions or rearrangements, a variety of cosmid and fosmid clones, together with a YAC clone, were amplified using primers INT1S (designed on a previously sequenced fragment of intron 1) and F2A. This amplification also indicated that the distance between these two primers was around 2kb. Amplification of genomic DNA generated a PCR product of the same size (figure 3.3).

3.1.4 Analysis of intron 1 sequence

Repeat elements were identified by REPEATMASKER from the new sequence and previously published sequence upstream and downstream and are shown in figure 3.2. Analysis of sequence shows that there are 8 Alu elements, of which three are Alu Y; a fragment of an L1 LINE element, and two $(TA)_n$ dinucleotide repeats. Alu Y is thought to be the most recent of all Alu subfamilies, and is primate specific (Batzer *et al.*, 1996; Kapitonov and Jurka, 1996)

Analysis of base composition using the computer program COMPOSITION shows that there are fewer CG nucleotide doublets than other doublets (table 3.2), and far fewer than expected given the frequency of C and G nucleotides (χ^2=110.7, $p \leq 1x10^{-25}$ 1d.f.). This is probably due to the higher mutation rate of methylated CG doublets (Cooper and Krawczak, 1989). Intron 1 reveals two interesting $(TA)_n$ polymorphisms (table 3.1) which were analysed further by Daniella Estrin. The P4f/P4r PCR product was analysed on a panel of 20 European individuals and found two sizes of repeat, the larger of which (n=23) occurs on the A haplotype and the smaller (n=10) occurs on the B and C haplotypes. Analysis of a small number of Japanese and Bantu-speaking South Africans showed more alleles.

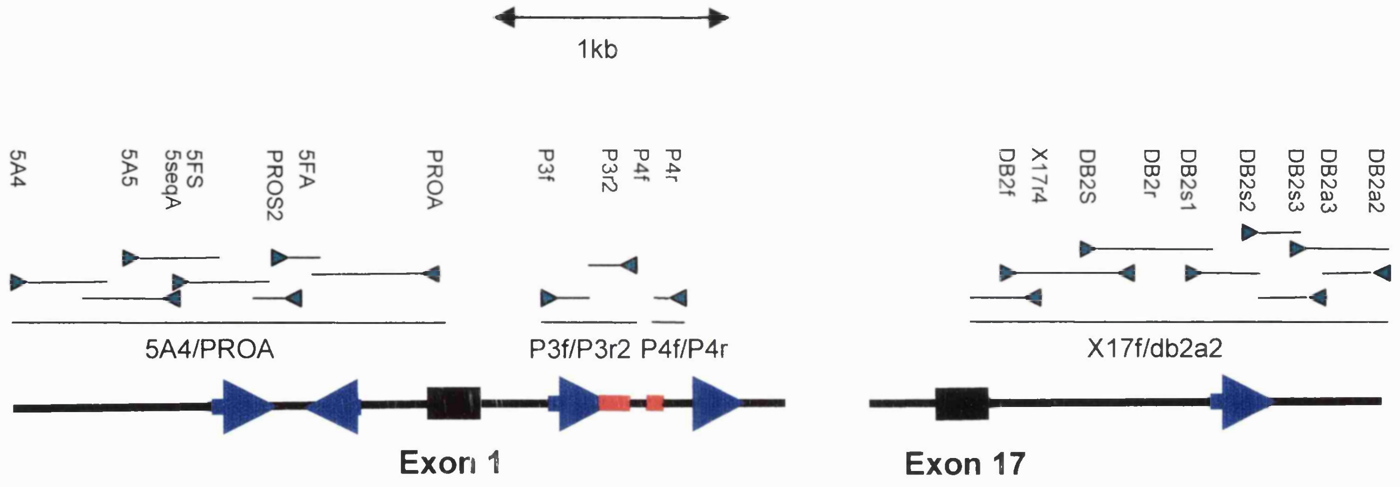


Figure 3.1 **Regions amplified and sequenced in the panel of five individuals**
Alu elements are shown in blue, with the arrowhead indicating direction (point at tail), and $(TA)_n$ repeats are shown in red. PCR products are shown as single lines, and primers are shown as green triangles withy the length of the sequencing read from that primer shown as a line extending from the triangle.

region	position	182 (AB)	187 (BC)	198 (AB)	209 (AB)	210 (AB)
	Persistence /non-persistence	PP	NN	PN	NN	PN
5′	-958	CT	CT	CT	CT	CT
5′	-875	G	AG	G	G	G
5′	-552/-559	A_8A_9	A_8	A_8A_9	A_8A_9	A_8A_9
Int 1	*i*509/*i*538	$(AT)_{12}TT(AT)_2$ /$(AT)_{12}$	$(AT)_{12}TT(AT)_2$ /$(AT)_{21}$	$(AT)_{12}TT(AT)_2$ /$(AT)_{12}$	$(AT)_{12}TT(AT)_2$ /$(AT)_{12}$	$(AT)_{12}TT(AT)_2$ /$(AT)_{12}$
Int 1*	*i*980/*i*999	$(AT)_{23}$/$(AT)_{10}$	$(AT)_{10}$/$(AT)_{10}$	$(AT)_{23}$/$(AT)_{10}$	$(AT)_{23}$/$(AT)_{10}$	$(AT)_{23}$/$(AT)_{10}$
3′	+225	CATT/ΔΔΔΔ	CATT/ΔΔΔΔ	CATT/ΔΔΔΔ	CATT/ΔΔΔΔ	CATT/ΔΔΔΔ
3′	+226	CT	CT	CT	CT	CT
3′	+658	CT	T	CT	CT	CT
3′	+988	AG	AG	AG	AG	AG
3′	+1214	AG	G	AG	AG	AG
3′	+1651	T	T	T	CT	T

Table 3.1 Panel of five lymphoblastoid cell lines from unrelated individuals of known genotype

The genotype of each individual at each polymorphic site is shown, together with the position of the polymorphic site. A negative number indicates downstream from exon 1, a positive number indicates downstream from exon 17, and an *i* number indicates sequence from the start of intron 1. * typings assumed from clone sequence.

P3f/P3r PCR product were also analysed on different samples, and two alleles were found based on size. The large allele (n=21) occurs on the C haplotype and the smaller allele (n=12) occurs on the A and B haplotypes.

3.1.5 Downstream sequencing strategy

Using the clone pDB2, 2kb downstream of exon 17 was sequenced using primers DB2f, DB2s, DB2s1 and DB2s2 (figure 3.1).

3.1.6 Analysis of downstream sequence

Several polymorphisms were discovered in the sequence 3′ to exon 17 by amplification and resequencing on the panel of individuals as shown in figure 3.4. Table 3.2 shows the nature and localisation of these polymorphisms. Two of these loci were later typed using allele-specific PCR in a variety of samples, as described in section 2.4.6. Figure 3.4 shows an example of a sequence difference revealed by this technique.

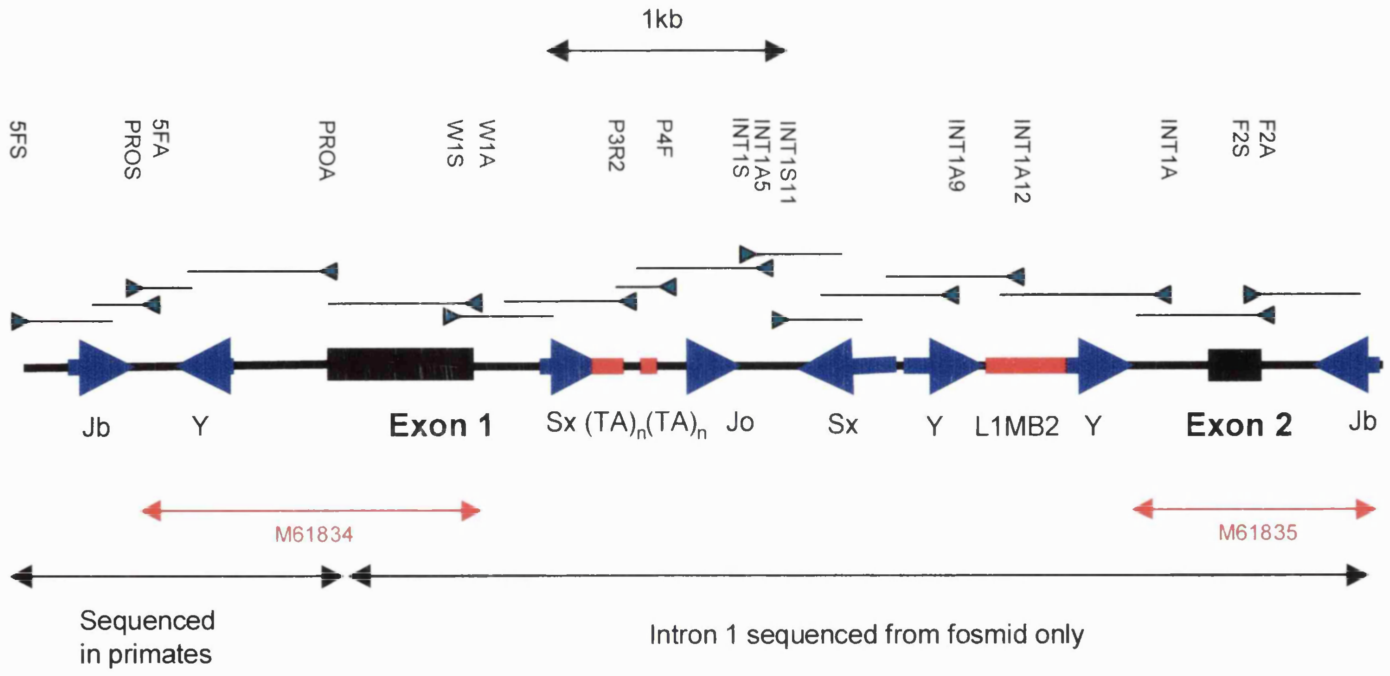


Figure 3.2 **Sequencing strategy of intron 1 on fosmid clone and of primate sequences**
Alu elements are shown in blue, with the arrowhead indicating direction (point at tail), and LINE elements and $(TA)_n$ repeats are shown in red. The subfamily of each Alu element is shown underneath. Accession numbers for published sequences are shown in red. Sequencing primers are shown as green triangles and sequencing readings from each primer are shown as lines extending from each triangle.

Second nucleotide

First nucleotide	A	C	G	T
A	328	191	262	286
C	264	220	83	260
G	235	195	215	159
T	240	221	243	315
Total	1067	827	803	1020

Table 3.2 Analysis of doublet composition of intron 1
Sequence analysed between W1S and F2A primers

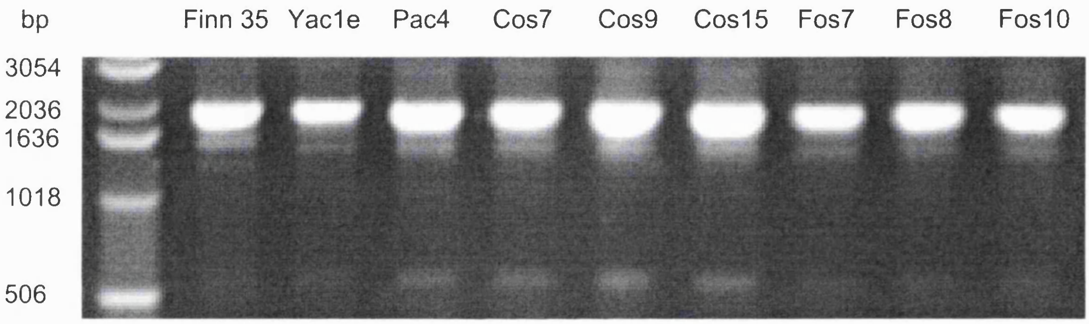


Figure 3.3 PCR product using INT1S/F2A primers on various clones and one genomic sample.
The size of the band is approximately 2kb, and according to the sequence is 1996bp.

3.2 Genetic and phylogenetic analysis of upstream region

3.2.1 Analysis of 1kb upstream of exon 1 in primates

PCR products were generated using primers 5FS/5FA and PROS2/PROA spanning the 1kb region upstream from exon 1 (figure 3.1) in one chimpanzee (Buttons), two gorillas, two orangutans and one crab-eating macaque. PCR products were generated from five other chimpanzees (Harv, Tank, Carl, Colin, Kasey) using 5FA and 5FS primers only. These PCR products were sequenced using the PCR primers, and aligned using PILEUP. The sequence is shown compared to human sequence in figure 3.5 together with primer positions and polymorphisms detected.

Several sites were polymorphic in non-human primates, and these are highlighted in figure 3.5. The human sequence was compared with the sequence of each species in turn using BESTFIT, and percentage identity determined across the total sequence, and across different regions within the sequence. The results are shown in figure 3.6. Several interesting observations can be made about these comparisons. Over the total sequence, both the chimpanzee and the gorilla differ from humans by the same amount, followed by orangutan which shares 96.8% of the sequence and then macaque which shares 93.9% of the sequence. This follows the phylogenetic relationship between the animals reasonably well, with chimpanzees being the closest living relatives of humans, followed by gorillas then orangutans. Macaques are, of the group of four primates analysed, the most distantly related (Goodman *et al.*, 1998). Comparison of the combined Alu sequences between each species and the human showed a lower percentage identity, consistent with their lack of function. The gorilla shares slightly more sequence with the human than the chimpanzee does, but this difference is small (97.1% to 96.8%). The comparison of the region between -974 to -874 shows higher identity with human sequence, suggesting that parts of the sequence may be functionally conserved, and this is expanded on in section 4. Again, the gorilla shares more sequence with the human than does the chimpanzee. Finally, the comparison of the immediate promoter and 13 bases of exon 1 (-153 to 13) shows the highest levels of identity, as expected for a region which is known to contain functional elements (see section 1.3). Both the chimpanzee and the orangutan share the most sequence identity with the human (both 99.4%) and the gorilla and macaque share slightly less.

3.2.2 Trees of sequence relationships

Multiple alignment of consensus primate sequences using PILEUP produces a dendrogram depicting the distance between the sequences. These were used to analyse the sequence in a similar way to that described in section 3.1. Instead of comparing each sequence in turn to human sequence, this analysis generates pairwise alignments between every possible pair of sequences. These pairwise alignments are scored, and the resulting distance matrix used to generate a multiple alignment and

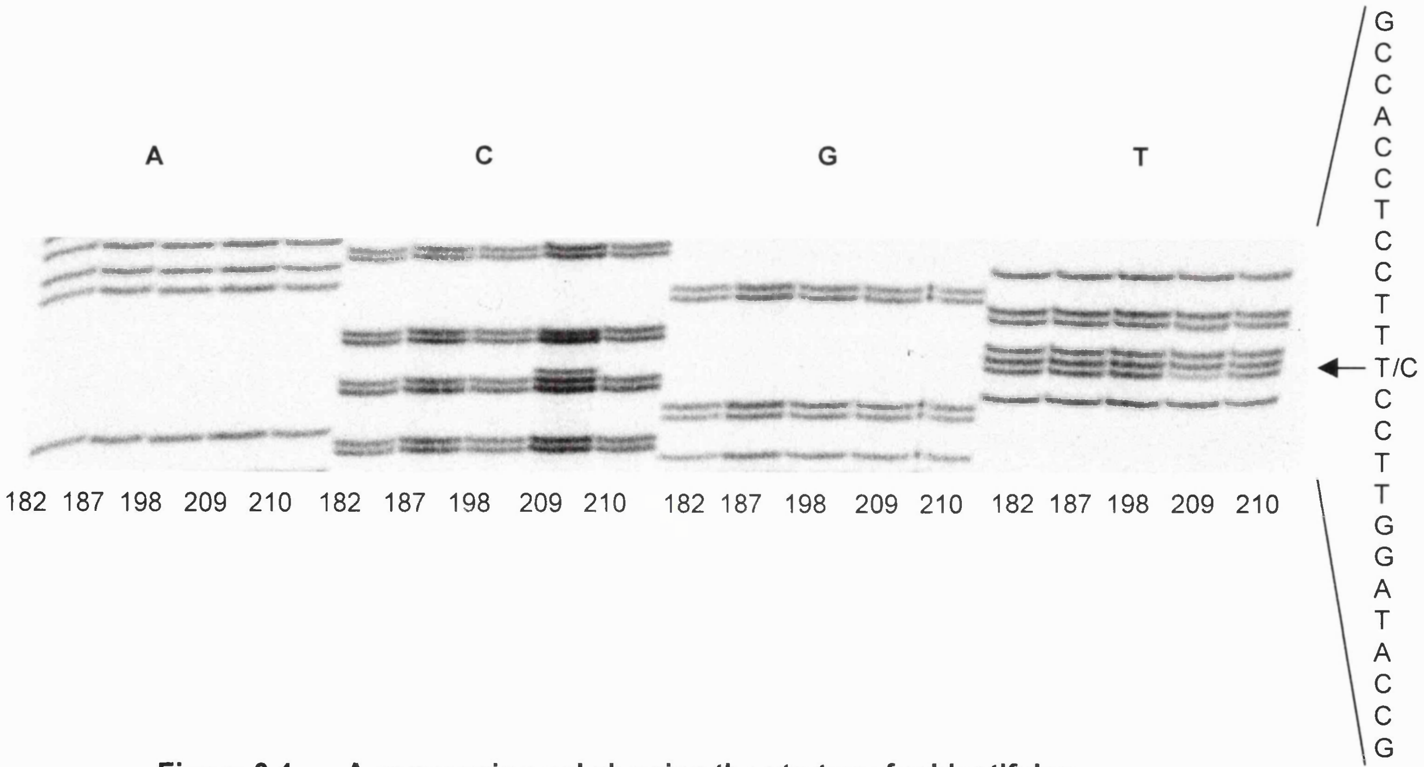


Figure 3.4 **A sequencing gel showing the strategy for identifying polymorphisms**

The sequence change shown is in WW209 which is heterozygous for a T to C change at postion +1651 (1651 bases after exon 17).

a dendrogram. This clustering strategy uses the unweighted pair-group method using arithmetic averages (UPGMA) (Li, 1997).Analyses of the total sequence, the sequence excluding Alu elements and Alu elements only were performed. The dendrograms for each set of aligned sequences (see figure 3.7) show the clustering relationships used to determine the order of each sequential pairwise alignment that PILEUP performs in generating the multiple alignment. The vertical length of each branch is proportional to the difference between the sequences.

Analysis of the total sequence and the sequence without Alu elements gives essentially the same dendrogram, the chimpanzee being placed closer to the human when the Alu elements are removed. The sequence relationships show that the orangutan is closer to the chimpanzee/human clade than the gorilla, which is in contrast to standard phylogenies that place the gorilla nearest to the human/chimp clade. This emphasises the danger in extrapolating phylogenetic influences from simple short sequence comparisons. Analysis of the Alu element sequence shows that the positions of the orangutan and chimpanzee have swapped in comparison with the two other dendrograms. This occurs even when the Alu element tails are removed, and shows that PILEUP finds the smallest distance matrix between human and orangutan, in contrast to the results from BESTFIT identity matches shown in figure 3.6. The BESTFIT identity scores consistently show chimpanzee and gorilla to share more sequence with humans as compared to orangutan, even under different values for the gap creation and gap extension penalty. Either result may be an artefact of the different methods of aligning sequences employed in each program.

```
            -975              ▽ C>T    A>G ▽    ▽-TC          -926
human       tttttcatag atgtttccat attgtttgaa tctcttacaa aatatgttca
chimpanzee  tttttcatag atgtttccat attgtttgaa tctcttacaa aatatgttca
gorilla     tttttcatag atgtttccat attgtttgaa tctcttacaa aatatgttca
orangutan   tttttcatag atgtttccgt attgtttgaa tc..ttacaa aatatgttca
macaque     tttttcatag atgtttccat attgttcgaa tctcttacaa aatgtgttca

            -925                                              -871
human       gcatattttt aaaga..gaa aatttggggc aaaatactta ttt...ttgt
chimpanzee  ...tattttt aaaga..gaa aatttggggc aaaatactta ttt...ttgt
gorilla     gcatattttt aaa....gaa aatttggggc aaaatactta ttt...ttgt
orangutan   gcatattttt aaa....gaa aatttggggc aaaagactta tttttgttgt
macaque     gcatattttt aaagagggaa aatttggggc aaaggactta tttttgttgt

               ▽ A>G                                          -821
human       attatgtaaa caaattttaa aataatgtgt ggctgggtgc gctggctcac
chimpanzee  attatgtaaa caaattttaa aataatgtgt ggctgggtgc gctggctcac
gorilla     attatgtaaa caaattttaa aataatgtgt gtctgggtgc gctggctcac
orangutan   attatgtaaa caaattttaa aataatttgt ggctgggtgc actggctcac
macaque     attatgtaaa caaatttta aataatttgt ggctgggtgt ggtggctcac

             ▽ C>G              ▽ T>A                    T>C ▽-771
human       acctgtaatc ccaacacttt aggaggctga ggcaagagga ttgcttgagc
chimpanzee  acctgtaatc ccaacacttt gggaggctga ggcaagagga ttgcttgagc
gorilla     acctgtaatc ccaacacttt gggaggctga ggcaagagga ttgcttgagc
orangutan   acctgtaagc ccaacacttt gggaggctga ggcaagagga ttgcttgagc
macaque     acctgtaatc ccaacacttt gggaggctga ggcaagagga tcgcttgagc

            -770                                              -721
human       ccaggagttc aagaccagcc tgggtgacat ggcaaaactc catctctact
chimpanzee  gcaggagttc aagaccagcc tgggtgacat ggcaaaactc catctctact
gorilla     ccaggagttc aagaccagcc tgggtgacat ggcaaaactc catctctact
orangutan   ccaggagttc aagaccagcc tgggtgacat ggcaaaactc catctctact
macaque     ccaggagttc gagaccagca t.ggtgacat tgcaaaactc catctctact

            -720                          ▽ C>T           -672
human       aaaaatacaa aaaattagcc agtcgtggtg gcg.cacacc tatggtccca
chimpanzee  aaaaatacaa aaaattagcc agtcgtggtg gcg.cacacc tatggtccca
gorilla     aaaaatacaa aaaattagcc agtcatggtg gcg.cacacc tatggtccca
orangutan   aaaaatacaa aaaattagcc agttgtggtg gcgccacacc tatggtccca
macaque     aaaagtacaa aaaattagcc agtcgtgatg gca.cacacc tatggtccca

               ▽ A>G                                          -622
human       cctacccagg atgctgagat gggaggatca cttgagccca ggaagtcaag
chimpanzee  cctacccagg atgctgagat gggaggatca cttgagccca ggaagtcaag
gorilla     cctaccaagg atgctgagat gggaggatca cttgagccca ggaagtcaag
orangutan   cctacccagg atgctgagat gggaggatca cttgagccca gggagtcaag
macaque     cctacccagt atgctgagat gggaggatcg cttaagccca ggaagtcaag

            -621                                              -572
human       gctgcaggaa gctgtgatcg caccactgca ctcccacctg ggcaacagag
chimpanzee  gctgcaggaa gctgtgatcg caccactgca ctcccacctg ggcaacagag
gorilla     gctgcaggga gctgtgatcg caccactgca ctcccacctg gtcaacagag
orangutan   gctgcaggga gctgtgatcg caccactgca ctctcacctg ggcaacagag
macaque     tctgcagtga gctgtgatcg caccactgca ctcccacctg ggcaacagag

            -571                  -A ▽▽ C>A               -522
human       tgagacccgg tcaccaaaaa acaaaaaaaa caaaaaaaat tggtaatcgt
chimpanzee  tgagacccgg tcacca.... ..aaaaaaac aaaaaaaaat tggtaatcat
gorilla     tgagaccctg tcaccaa... ..aaaaaaaa aaaaaaaaat tgataattgt
orangutan   tgagaccctg tcacc..... .......aaa aaaaaaaaat tgataattgt
macaque     tgagaccctg tcaccaaaa. ..taaataaa taaataaaac ttgatattgt

            -521                    ▽C>A   ▽ A>T          -472
human       tttcttcaga cattttccgg gttcctctgc ttaacttgta taggaagtct
chimpanzee  tttcttcata cattttccgg gttcctctgc ttaaattgta taggaagtct
gorilla     tttcttcaga cattttccgg gttcctctgc ttaaattgta taggaagtct
orangutan   tttcttcaga cattttccgg gttcctccgc ttaaattgaa taggaagtct
macaque     tttcttcaga cattttccag gttcctctgc ttaaattgta taggaagttt

            -471                                            A>T ▽
human       gaggttttttg tgttggtctt tacctttttt tttttttttt tttttttaa
chimpanzee  gaggtttttg tgttggtctt tacc...... .......... tttttttaa
gorilla     gaggtttttg tgttggtctt taccttttttt tttttttttt ttaagatgga
orangutan   gaggtttttg tgttggtctt tacc...... .ttttttttt ttttttttaa
macaque     gaggtttttg tgttggtctt ta...ttttt ttatttttta ttttttttga
```

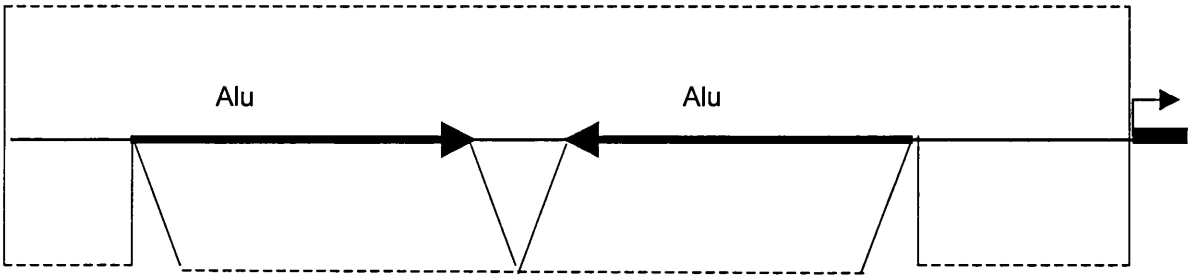
```
               -421   ▽ T>G      ▽ -T   ▽C>G                     -383
human          gatggagtct cattctgtt. gcccaggctg gagtgcagtg gcatgatctt
chimpanzee     gatggagtct cattctgtt. gcccaggctg gagtgcagtg gcatgatctt
gorilla        gctcatttta ggctct.... gcccaggctg gagtgcagtg gcatgatctt
orangutan      gatggagtct cattctgttc gcccaggctg gagtgcagtg gcatgatctt
macaque        gatggagtct cattctgtt. gcccaggctg gagtgcagtg gcatgatctc

               -372               ▽ C>T                          -334
human          ggctcct.gc aacctccgcc tcctgggttc aagtgattct cctgcctcag
chimpanzee     ggctcctcac ctccgcctcc tcctgggttc aagtgattct c.tgcc.cag
gorilla        ggctcct.gc aacctccgcc tcctgggttc aagtgttctc ctgccctcag
orangutan      ggctcct.gc aacctccgcc tcctgggttc aagtgattct cctgcctcag
macaque        ggctcactgc aacctccg.c tcccgggttc aggtgattct cctgcctcag

               -333                                              -284
human          cctcctgagt agccgggact acaggcgcat gccacgatgc ctggctaatt
chimpanzee     cctcctgagt agctgggact acagcgcact gccacgatgc ctggctaatt
gorilla        cctcctgagt agccgggact acaggcg.at gccacgatgc ctggctaatt
orangutan      cctcctgagt agccgggact acaggcgcat gccacgatgc ccagctaatt
macaque        cctcctgagt agccgggact acagacgcat gccactatgc ctggctaatt

               -283                                              -234
human          ttttgtattt ttagtagaga tggggtttca ccatgttagc taggacggtc
chimpanzee     ttttgtattt ttagtagaga tggggtttca ccatgttagc caggacggtc
gorilla        ttttgtattt ttagtagaga tgg.gtttca ccatgttagc caggacgat.
orangutan      ttttgtattt ttagtagaga cggagtttca ccgtgttagc caggacagtc
macaque        ttttgtattt ttagtagaga cag.gtttca ccgtgttagc caggatggtc

               -233                                       A>G ▽  -184
human          tcgatctcct gacctcgtga tccgcccacc tcggcctccc aaagtgctgg
chimpanzee     tcgatctcct gacctcgtga tccgcccgcc tcggcctccc aaagtgctgg
gorilla        tcgatctcct gacctcgtga tccgcccg.c tcggcctccc aaagtgctgg
orangutan      tcgatctcct gacctcgtga tccgcccgcc tcagcctccc aaagtgctgg
macaque        ttgatctcct gacctcgtga tctgccc.cc tcggcctccc aaagtgctgg

               -183   ▽ -G                                       -134
human          aattacaggt gtgagccacc acgcccggcc ctgatcttta catttttaaa
chimpanzee     aattacacgt gtgagccacc acgcccggcc ctgatcttta catttttaaa
gorilla        aattaca.gt gtgagccacc acgcccggcc ctga.cttta catttttaaa
orangutan      aattacaggt gtgagccacc acgcccggcc ctgatcttta catttttaaa
macaque        aattacaagt gcaagccacc gcacccggcc ctgatcttta catttttaaa

               -133                                               -84
human          tattgcatta gtgaaccgtg tactgatttt gtgatcatag ataacccagt
chimpanzee     tattgcatta gtgaaccgtg tactgatttt gtgatcatag ataacccagt
gorilla        tattgcatta gtgaaccgtg tactgatttt gtgatcatag ataacccggt
orangutan      tattgcatta gtgaaccgtg tactgatttt gtgatcatag ataacccagt
macaque        tattgcatta gtgaaccatg tactgatttt gtgatcatag ataacccagt

               -83                                                -34
human          taaatattaa gtcttaatta tcacttagta ttttacaacc tcagttgcag
chimpanzee     taaatattaa gtcttaatta tcacttagta ttttacaacc tcagttgtag
gorilla        taa.tattaa gtcttaatta tcacttagta ttttacaacc tcagttgtag
orangutan      taaatattaa gtcttaatta tcacttagta ttttacaacc tcagttgtag
macaque        taattattaa gtcttaatta tcacttagta ttttacaacc tcagttgttg

               -33                                    ▽ -C      13
human          ttataaagta agggttccac atacctccta acagtt.cct agaaaat
chimpanzee     ttataaagta agggttccac atacctccta acagtt.cct agaaaat
gorilla        ttataaagta aggg.tccac atactcccta .cagtt.cct agaaaat
orangutan      ttataaagta agggttccac atacctccta acagttccct agaaaat
macaque        ttataaagta agggttccac ataccttcta acagtt.cct agaaaat
```

Figure 3.5 Sequence alignment of lactase upstream sequence of five species.

Polymorphisms are shown highlighted in red, and Alu elements are shown highlighted in blue. Note that chimpanzee polymorphisms were detected by sequencing six individuals between 5FS and 5FA, and only one individual between PROS2 and PROA.

Total sequence

Species	% identity with human
chimpanzee	97.7
gorilla	97.6
orang-utan	96.8
macaque	93.9



-974 to –852

Species	% identity with human
chimpanzee	97.5
gorilla	98.4
orang-utan	97.5
macaque	95.2

Alu elements

Species	% identity with human
chimpanzee	96.8
gorilla	97.1
orang-utan	96.3
macaque	92.6

Promoter (-153 to 13)

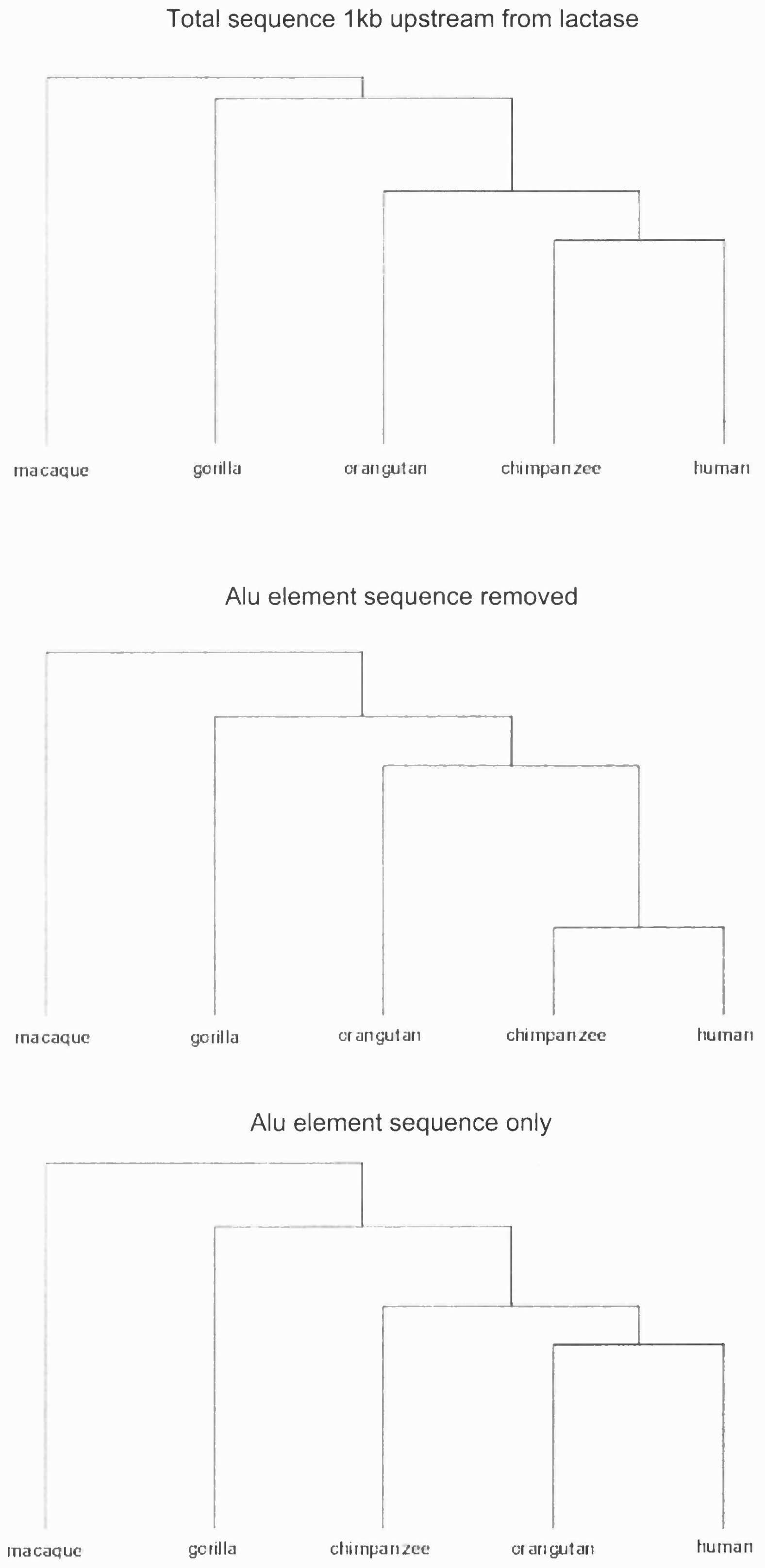
Species	% identity with human
chimpanzee	99.4
gorilla	97.5
orang-utan	99.4
macaque	97.0

Figure 3.6 Percentage identity of sequence between human and other species

3.3 Construction of a contig across the lactase locus

3.3.1 Libraries used

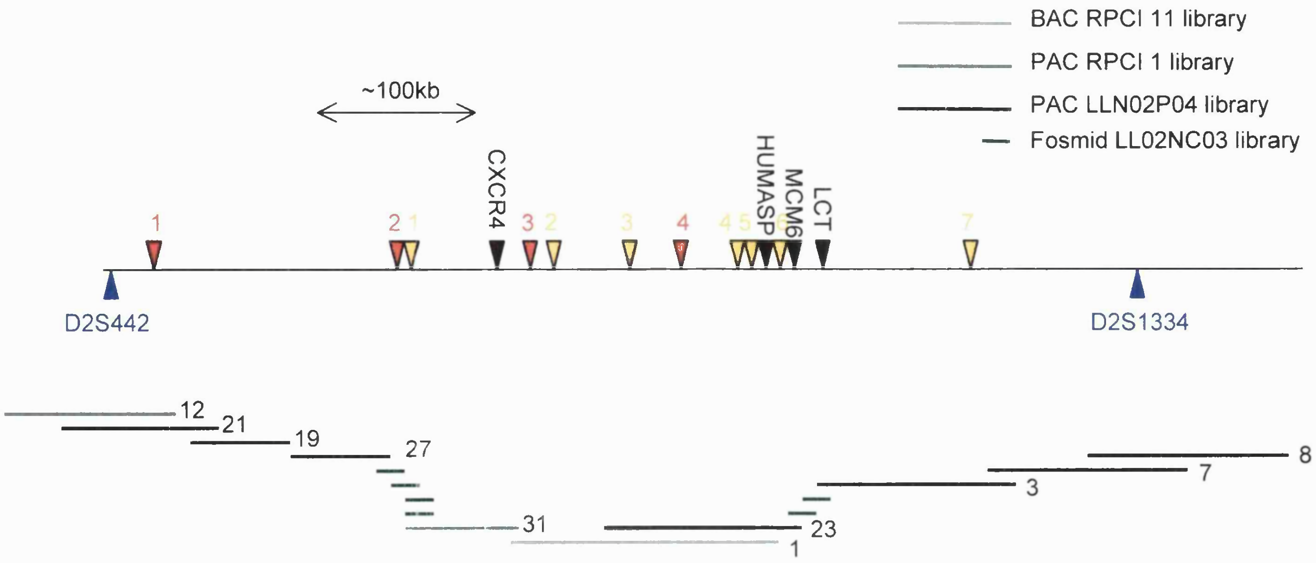
Several types of clone can be used to create a physical contig of overlapping clones spanning a certain region of the genome. Yeast artificial chromosomes (YACs) contain around 1Mb of cloned DNA, but are prone to deletion and rearrangement which may not make them representative of the genomic sequence they are intended to cover. Plasmids are smaller and typically contain inserts of between 2 and 10kb, too small to cover substantial physical distance. Bacterial artificial chromosomes (BACs) have insert sizes of 70-100kb, similar to P1 artificial chromosomes (Gingrich *et al.*, 1996). Fosmids and cosmids have insert sizes of around 40kb, and are more suitable for finer scale mapping, and were more suitable for constructing a high-density map spanning the lactase gene.


Total sequence 1kb upstream from lactase

Alu element sequence removed

Alu element sequence only

Figure 3.7 Dendrograms showing relationships of sequences in the upstream region of lactase between different species



Expressed Sequence Tags
1 EM:S74678 93% identity with hnRNP complex K
2 EM:AI184702
3 EM:HSXT02304
4 EM:1163295
5 EM:AI589561
6 EM:HSA69273 homology to endo α-D-mannosidase
7 EM:HSD684

Genomic Survey (October 1999)
1 CIT-HSP-2166NS.TR
2 CIT-HSP-2197E9.MF
3 RPCI 11 43A12.TK
4 CIT-HSP-2377N1.TF

Figure 3.8 Contig and physical map between D2S442 and D2S1334

Several libraries of cloned DNA dried on filters were used to isolate clones. The cosmid, fosmid, and PAC libraries were constructed at the Lawrence Livermore National Laboratory and obtained through the UK-HGMP. These libraries were constructed from flow-sorted DNA so that only chromosome 2 clones were represented (Gingrich *et al.*, 1996). The RPCI-1 PAC library is, by contrast, a whole genome library, constructed at the Roswell Park Cancer Institute. The RPCI-11 library is a whole genome BAC library available from Research Genetics, and is the basis of the whole human genome 'shotgun' sequencing project undertaken by Celera Inc. and several university laboratories.

3.3.2 A large contig using BACs and PACs

3.3.2.1 Construction of contig

A large contig between D2S442 and D2S1334 was constructed using a BAC-end sequencing (or sequence-tagged connector) and chromosome walking approach by Mark Poulter in the group. Initially, PCR products containing the appropriate microsatellites were used to isolate two groups of clones containing the PCR product. The clones were then orientated and the appropriate end sequenced, then this sequence was used to generate PCR primers to create another probe. At one stage, no clones were found to make the next stage in the walk, so the fosmid library was probed to isolate several fosmids which bridged the gap in the BAC/PAC contig.

3.3.2.2 Analysis of contig

Figure 3.8 shows a minimum tilepath of clones and genes, expressed sequence tags (ESTs) and sequence tagged sites (STSs) identified in the end sequences by Blast database searches. Several genes that had been cloned and mapped to this region of the chromosome by recombination breakpoint maps were physically mapped. MCM6 was previously identified by sequencing outward from the lactase gene (Harvey *et al.*, 1996) and is the human homologue to a yeast cell-cycle regulatory gene. HUMASP is the aspartyl tRNA synthetase gene, whose protein product catalyses the attachment of aspartate residues to their cognate tRNAs prior to incorporation into polypeptides by the ribosome. CXCR4 is a chemokine receptor gene, which is involved in Human Immunodeficiency Virus entry and AIDS progression.

Several potentially new genes, previously identified as ESTs, have been localised. One shows 93% identity with part of the poly(C)-binding heterogenous nuclear ribonucleoprotein complex K protein, which plays a role in pre-mRNA processing (Matunis *et al.,* 1992). Another EST shows homology to endo α-D-mannosidase, which is an enzyme involved in early N-linked oligosaccharide processing (Spiro *et al.,* 1997).

3.3.3 A fosmid and cosmid contig across the lactase gene

Isolation of a high resolution fosmid and cosmid contig across the lactase gene used a different approach to that of the larger BAC/PAC contig, since more was known about the sequence within the area to be covered by the contig. PCR products from 5′ to (F1t7f/F1t7r), exon 1 (W1S/W1A), exon 2 (F2S/F2A), exon 6 (X6f/X6r), and 3′ to exon 17 (P3P1f/P3P1r) were labelled and used to probe the filters in one experiment. From the identity of the dots that hybridised these probes, clones were ordered and liquid cultures grown from the colonies supplied on plates.

The extent of each clone was determined by PCR analysis of the liquid culture. Several pairs of PCR primers spanning the gene were used as shown in figure 3.9. Presence or absence of the appropriately sized band in each clone was used to determine the approximate position of the each clone relative to the gene. A large degree of coverage was observed, and the clones comprising the contig were stored as glycerol stocks for use in further projects, such as the sequencing of intron 1 discussed in section 3.1.

3.4 Fluorescence in-situ hybridisation (FISH)

3.4.1 Analysis of metaphase and interphase cells

FISH was used to localise clones and to check for large scale deletions or rearrangements upstream of the lactase gene. The same panel of five lymphoblastoid cell lines was used, derived from individuals of known lactase persistence genotype, to check if any large scale change corresponded to the lactase persistence/non-persistence polymorphism (table 3.1). Metaphase chromosomes derived from lymphoblastoid cell lines are known to appear more compact and fuzzier than those derived from whole blood, but the advantage of a plentiful supply of chromosomes from patients of known lactase persistence status was more important than fine localisation of the signal on the chromosome. LCT had already been mapped to 2q21 using FISH (Harvey *et al.,* 1993).

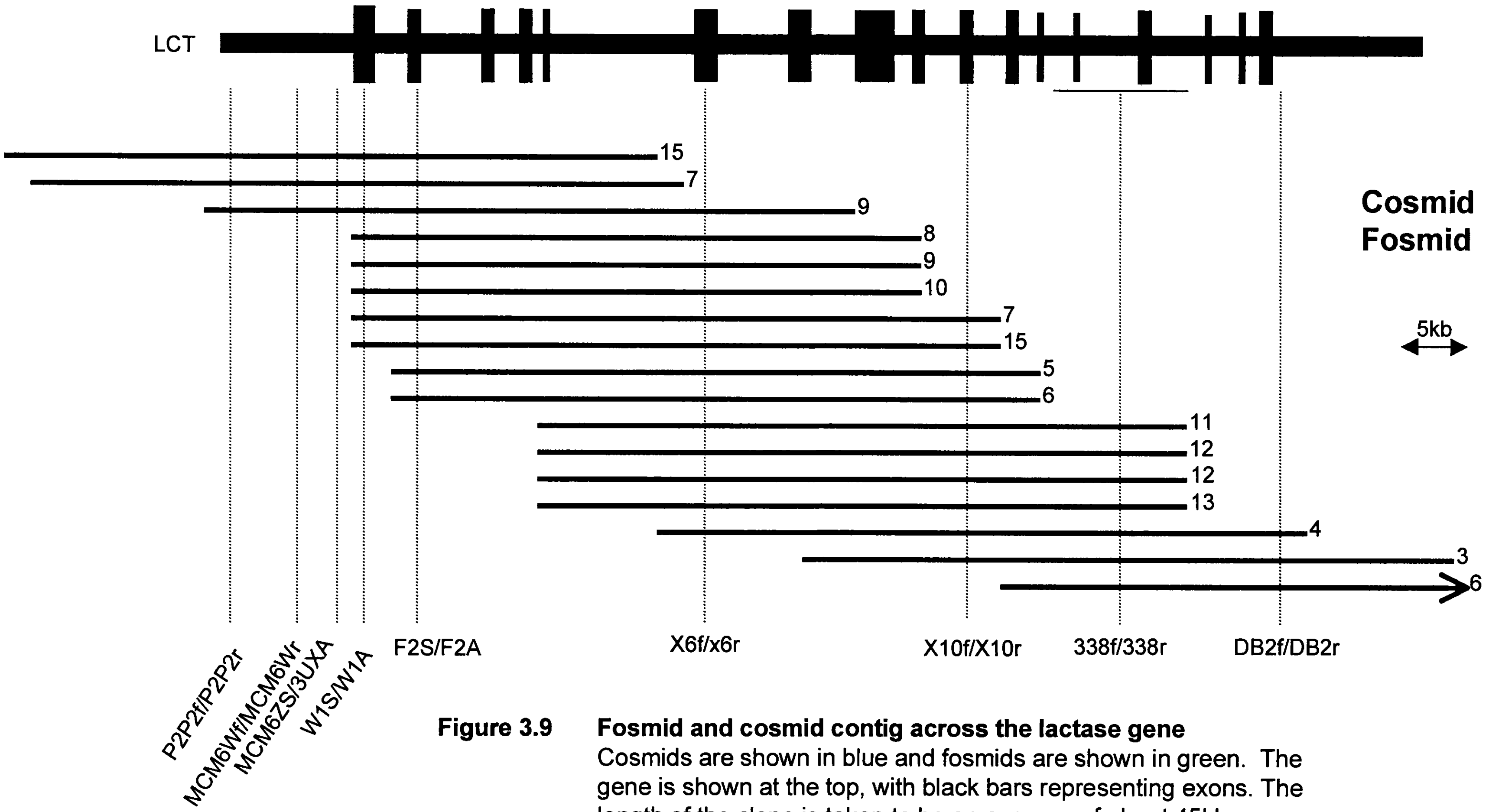


Figure 3.9 **Fosmid and cosmid contig across the lactase gene**
Cosmids are shown in blue and fosmids are shown in green. The gene is shown at the top, with black bars representing exons. The length of the clone is taken to be an average of about 45kb.

The probes used were all PACs from the chromosome 2 library, and were selected at intervals across not only the contig, but further upstream as well. The full clone addresses can be found in appendix 2. PAC 1 and PAC 15, at the extremes of the area under study, were first used in two-colour FISH analysis of the five cell lines. Figure 3.11 shows two representative images of the homozygous persistent cell line WW182 and the homozygous non-persistent cell line WW209, with PAC 1 in green and PAC 15 in red. In both cases both probes label the chromosome, but the distance between the probes is very small, too small to reliably infer orientation on the chromosome. Interphase nuclei show separation of the two probe signals because of the decondensation of the chromatin. Figure 3.12 shows PAC 12 (green) and PAC 13 (red) in WW182 and WW209. Again, in all cell lines both probes labelled the chromosome but as expected, the distance is too small to infer orientation on the metaphase chromosomes. Interphase nuclei show separation of the signals in some cases, but not all.

Figure 3.13 shows two typical images of PAC 1 (green) and PAC 12 (yellow) hybridised to WW198 cell line chromosomes (heterozygous persistent). The distance between these two probes is known to be approximately 500kb, and the two signals cannot be separated in metaphase chromosomes nor in virtually all interphase nuclei.

3.4.2 Analysis of dynamically molecular combed DNA (strand-FISH)

Dynamic molecular combing is a method of aligning thousands of copies of a genome on a slide, which can then be probed by fluorescently labelled clones in a similar manner to normal FISH. It can be used to directly measure clone sizes and distances between clones, and there should be sufficient repeat images on one slide that variability in the estimates can be assessed. The technique has been used to identify deletions in the TSC2 gene (Michalet *et al.*, 1997).

Agarose blocks containing a high concentration of DNA are melted in a well and a specially coated slide lowered into the DNA and slowly withdrawn at a constant speed using a specially designed machine. This process allows DNA molecules to attach by surface tension and be stretched along the slide as it is withdrawn. The slide is then dried and ready for normal hybridisation procedures (for full details see section 2.6.5).

The analysis was performed on DNA extracted from the WW182 cell line. PACs 1 and 4, which lie at either end of the lactase gene, were used as the fluorescently labelled probes. Several contiguous red and green lines were observed, from which length estimates were made. There were only two instances of a red and green line together, both of which are shown in figure 3.14. However, there were five instances of a red line alone and five instances of a green line alone, which allows an accurate estimate of the size of the clones to be made. Using the program CARTOGRAPHIX LITE, the length of contiguous line can be measured and distance in kilobases derived from that length. The results are shown in table 3.3, together with the mean and standard deviation, which are also shown on figure 3.14.

The length of PAC 1 is shown to be 41.1±5.1kb and the length of PAC 4 is 48.5±3.3kb. The end of PAC 1 is almost at the 5′ end of the lactase gene, and the beginning of PAC 4 is a few kilobases after exon 17, so the distance should approximate to the length of the lactase gene. The gap, derived from only two independent measurements, is suggested to be 53.3±4.7kb, which is slightly smaller than the expected length of the lactase gene as measured by Southern blot analysis (Boll *et al.*, 1991).

3.4.3 Discussion

The combination of the shortened chromosomes from lymphoblastoid cell lines and the small distances between each probe made orientating the probes difficult. However, all cell lines showed the expected localisation for the probes, so that there were no large translocations or deletions involving sequence detected by the probes used. Further planned work will include more strand FISH experiments to size clones in the contig, and using pulsed-field gel electrophoresis to confirm these sizes.

PAC 1 (green)		PAC 4 (red)		Gap	
Length (µm)	Length (kb)	Length (µm)	Length (kb)	Length (µm)	Length (kb)
22.7	45.4	21.5	43	28.5	57
14.9	29.8	25.4	50.8	25	50
21.6	43.2	24.7	49.4	-	-
21.4	42.8	24	48	-	-
21	42	26.5	53	-	-
21.6	43.2	24.8	49.6	-	-
20.6	41.2	22.9	45.8	-	-
mean	41.1	mean	48.5	mean	53.3
s.d.	5.1	s.d.	3.3	s.d.	4.7

Table 3.3 Analysis of PAC1 and PAC4 sizes using strand FISH

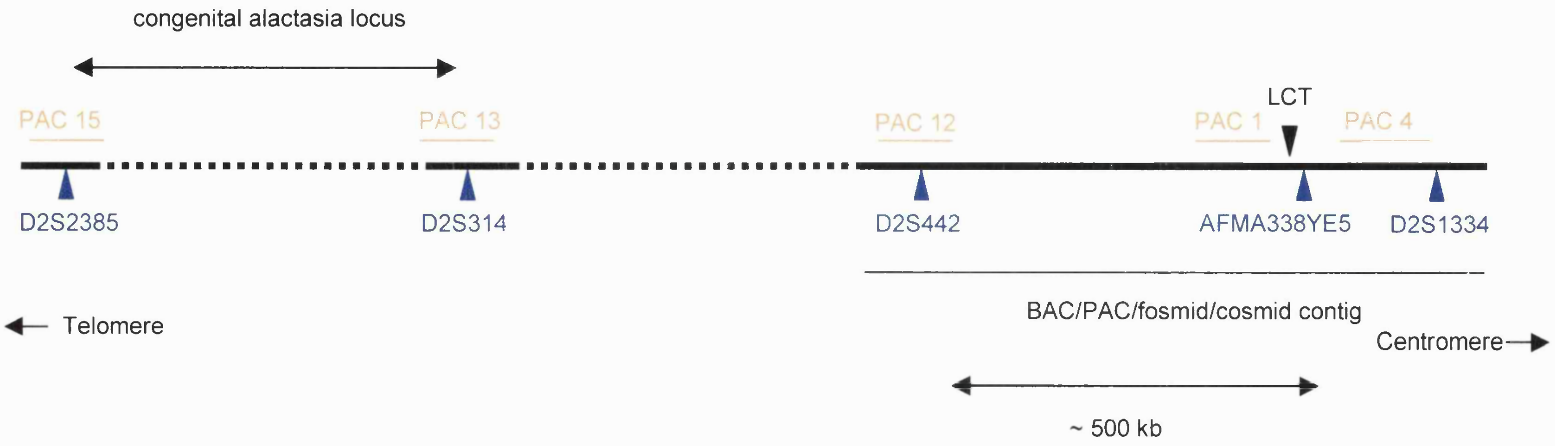


Figure 3.10 **PAC clones used for FISH analysis**
PACs are from the RPCI1 library (clone addresses shown in appendix 2). Microsatellites are shown in blue, and regions not in a contig are shown as dashed lines. The region suggested to contain the congenital alactasia locus (see section 8) is also shown.

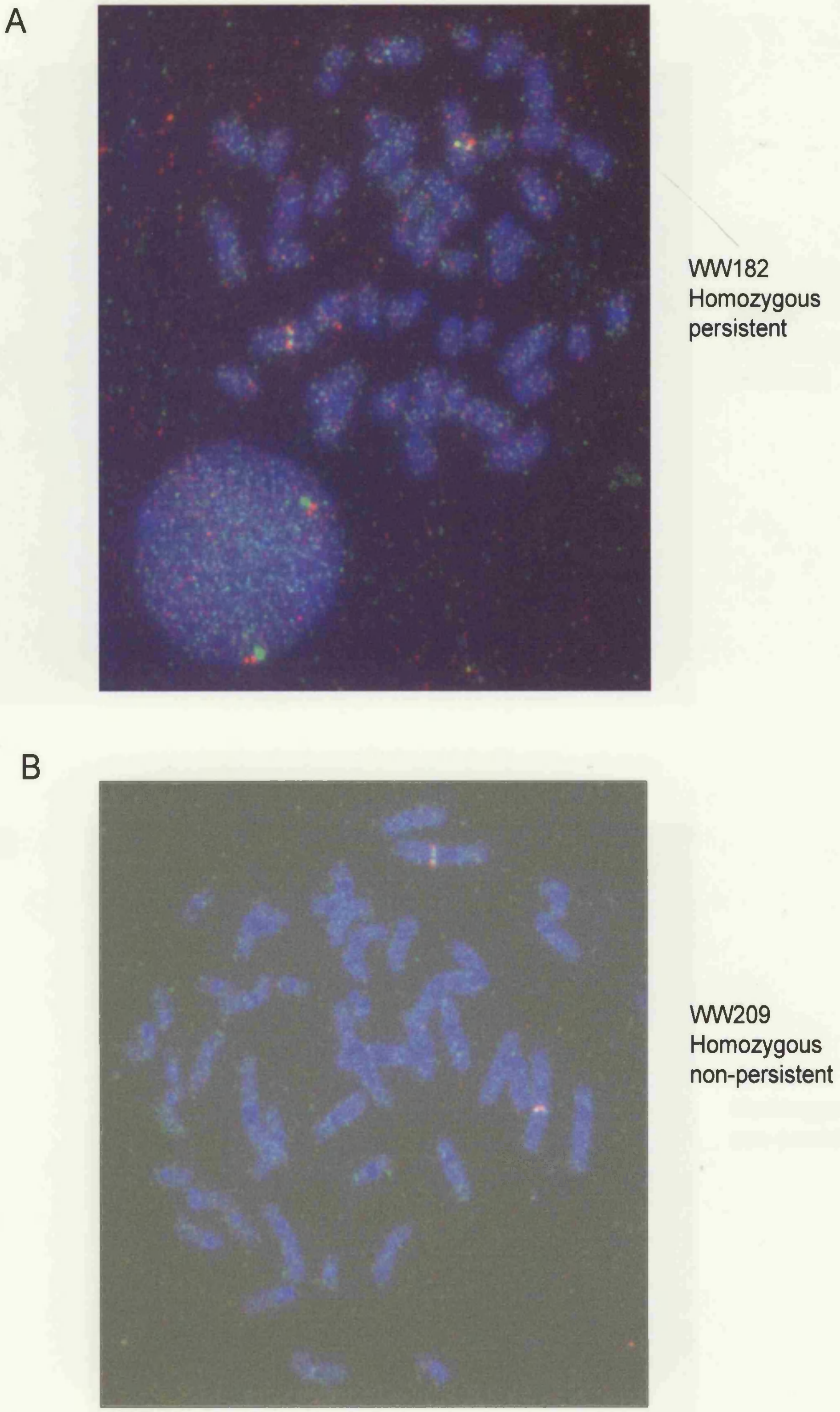


Figure 3.11 **FISH analysis of metaphase spreads of lymphocyte cell lines**

PAC 1 - biotin-label green

PAC 15 - DIG-label red

A shows that the two probes can be distinguished on interphase cells.

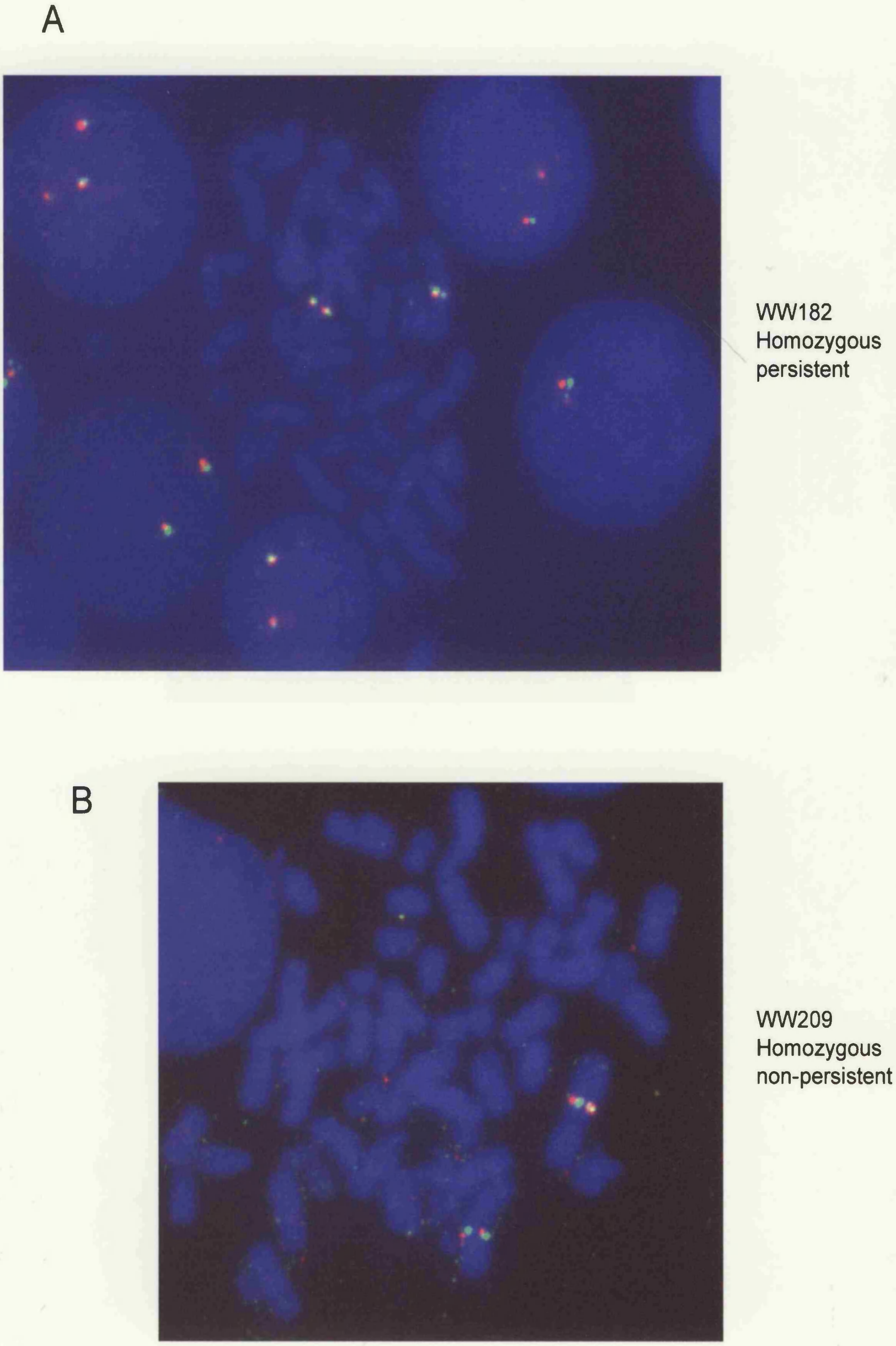


Figure 3.12 FISH analysis of metaphase spreads of lymphocyte cell lines

PAC 12 - biotin-label green
PAC 13 - DIG-label red

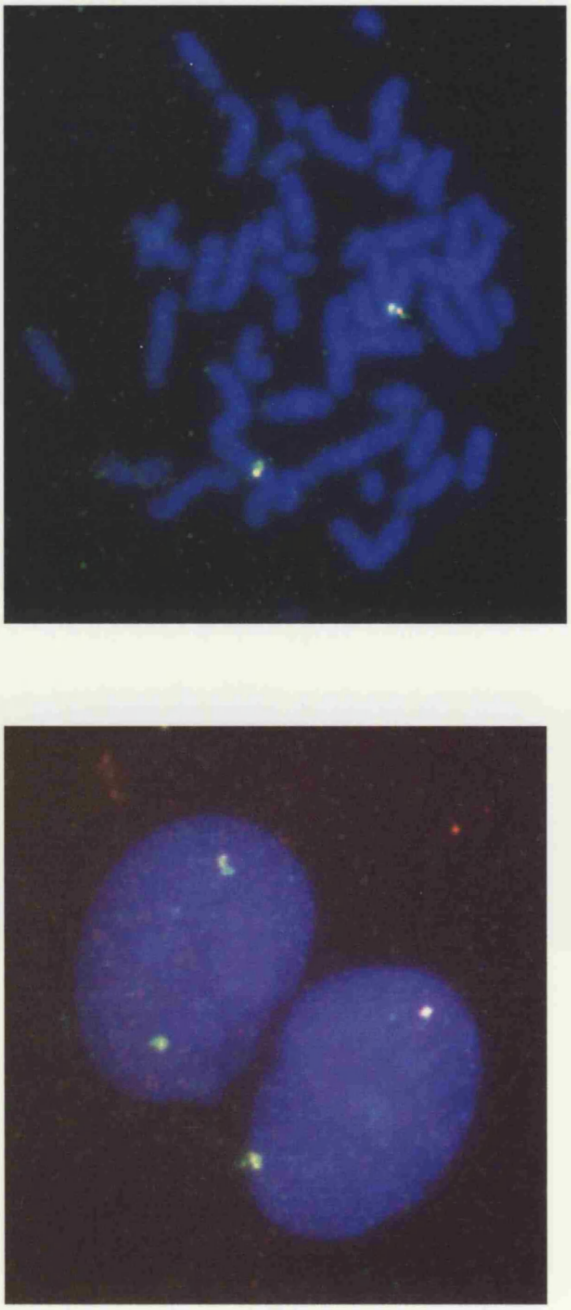


Figure 3.13 FISH analysis of metaphase spreads of WW198 cell line
PAC 1 biotin label green
PAC 12 biotin and DIG label yellow

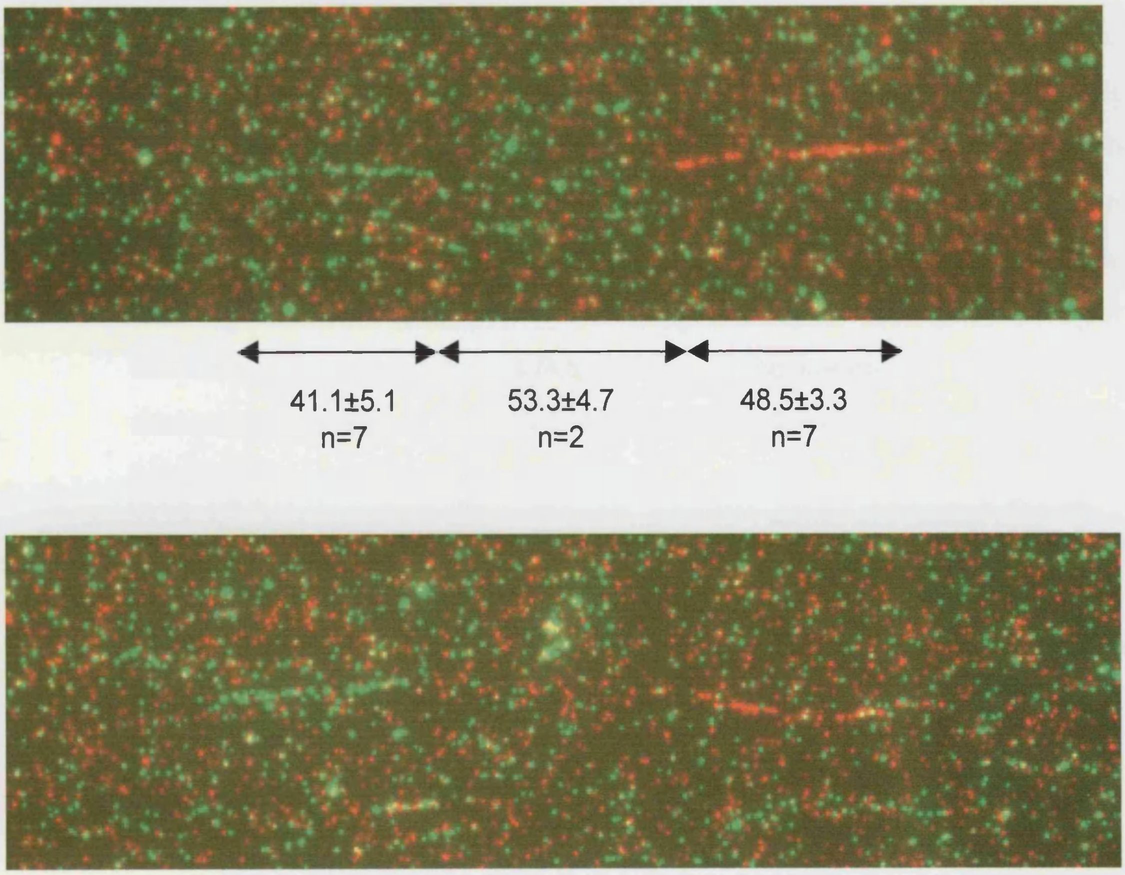


Figure 3.14 Strand FISH using DNA from WW182 cell line
PAC 1 biotin label green
PAC 4 DIG label red

4 Biochemical analysis of the region upstream of lactase

4.1 Comparison of promoter regions between human, pig and rat

4.1.1 Sequence analysis of the promoter regions of the three species

In order to identify *cis*-elements that have previously been characterised in the pig, the published sequences of rat (EMBL acc. no. S77839) and pig (EMBL acc. no. Y08677) were compared with human sequence using the program BESTFIT. Repeat elements were identified using the program REPEATMASKER. Since it had been suggested that downregulation of expression of the promoter construct in mice transgenic for the pig promoter is controlled by1kb of the upstream pig sequence (Troelsen *et al.*, 1994a), I have concentrated on this region. The equivalent region in humans spans approximately 1.5kb, because of the insertion of two ~250bp Alu elements (see figure 4.1). In the case of the rat, all the known sequence upstream from lactase (1.2kb) was used in the comparisons.

Several other repeat elements were found by REPEATMASKER in the pig and rat sequences, as well as the Alu elements present in the promoter region of the human lactase gene (figure 4.1). The pig has two LTR39 repeats, and the rat has a series of B-repeats and an Alu-like short interspersed nuclear element.

4.1.2 Comparison of the functional elements identified in the pig

Figure 4.1 shows the position of potential *cis*-elements (CE-LPH) identified in pig (Spodsberg *et al.*, 1999; Troelsen *et al.*, 1994b), and homologous sequences found in rat and human. Homologues of all the CE-LPH elements were found in rat and human, except CE-LPH4whichcannot be found in the rat sequence. Figure 4.2 shows the sequence of these *cis*-elements (CE-LPH1 to CE-LPH4) in the three species. The percentage identity of each element between species varies which suggests that certain elements may not be functional in certain species. For example, homologues had been found to all pig CE-LPH sites in the human, except CE-LPH2a and CE-LPH2b (Spodsberg *et al.*, 1999). However, on further analysis possible homologues of these two elements can be found, although they are less than 29% identical between pig and human. The order of two elements is different in rat, where the CE-LPH1b

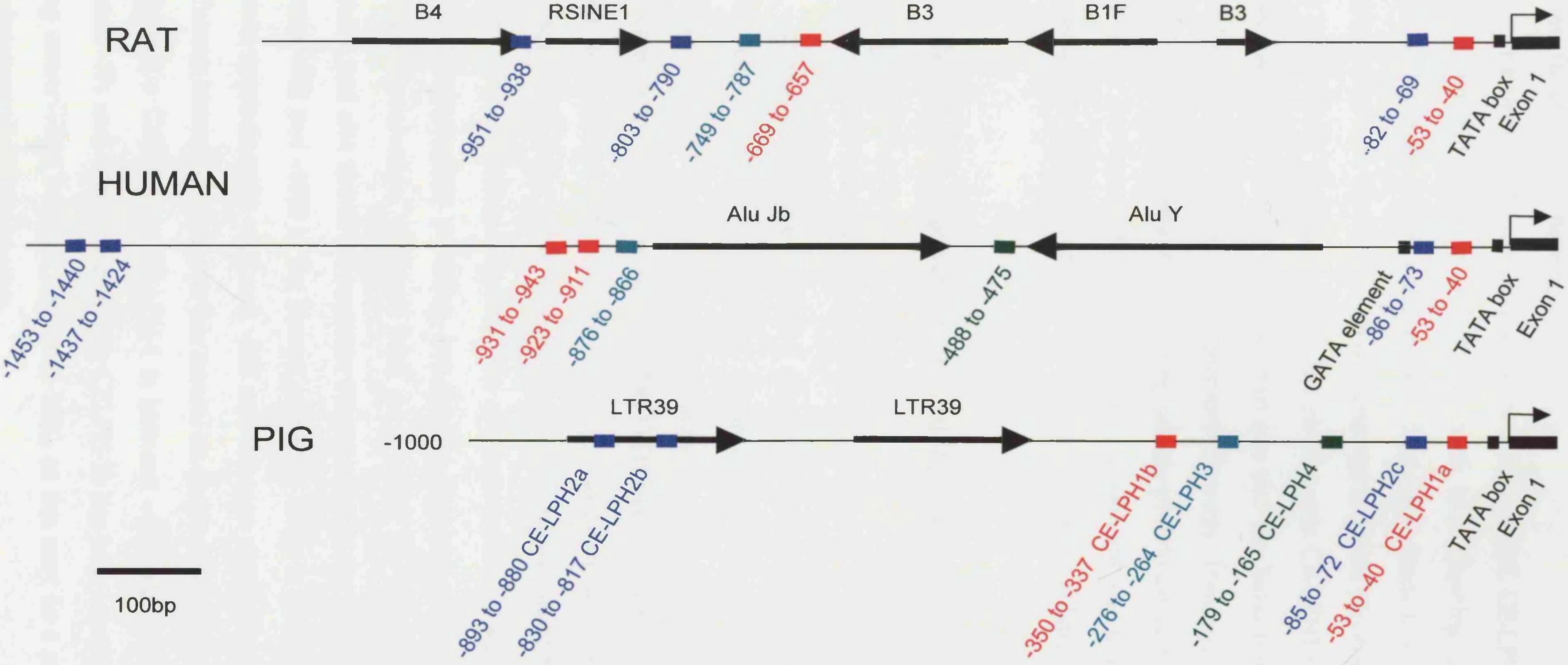


Figure 4.1 **Comparison of human, rat and pig sequences upstream of lactase**
CE-LPH elements are indicated in the pig, together with sequence positions relative to exon 1. Homologous sites in the human and rat are similarly identified, and all numbers are inclusive Repetitive elements are shown as black arrows with the arrowhead representing the tail of the element.

homologue is closer to the gene and the CE-LPH3 suggest that these elements have a different function or perhaps none at all.

In general, both order and sequence are conserved. CE-LPH1a and CE-LPH1b are 77% and 62% identical respectively, with both showing conservation of the consensus motif for Cdx-2, a transcription factor that binds to CE-LPH1a in rat and pig (shown in bold, figure 4.2) and activates transcription (Troelsen *et al.*, 1997). It is clear that in both humans and rats the *cis*-elements CE-LPH1b, 2a, 2b, and 3 are further away from the transcription start site and the basal transcription complex than in pig because of insertion of repetitive elements. Transcription factors binding to these elements mediate their effect by protein-protein interactions, and, in the rat and human, the increased distance between *cis*-element and transcription start site may make this less effective. Other factors such as distal enhancer elements may be required.

4.1.3 Identification of possible functional regions

4.1.3.1 Analysis of conservation of promoter sequence

Promoter sequence containing elements essential for gene expression is likely to be conserved through evolution. In order to find regions containing further possible functional elements, dot plots comparing human to rat and pig sequences were constructed using the program DOTTER from the UK-HGMP. The dot plot shown in figure 4.3 compares the 1kb of pig sequence with the 1.5kb of human sequence, and the dot plot shown in figure 4.4 compares the 1.2kb of rat sequence with the 1.5kb of human sequence. The rat sequence shows sequence conservation only at the immediate promoter, but the pig sequence shows several regions of sequence conservation. It shows both gaps where the human sequence has Alu elements, and also shows the sequence conservation in the immediate promoter and between -900 and -1000 in the human sequence. There is also weaker conservation between approximately -480 and -500 and between approximately -830 and -870. The human immediate promoter has several *cis*-elements identified already: CE- LPH 1a, 2c and a GATA box. CE-LPH-4 is between -488 and -475, and CE-LPH3 is between -876 and -866. However only CELPH-1b has been identified to the region of strong conservation between -900 and -1000, so this may be a possible area which contains more *cis*-acting elements.

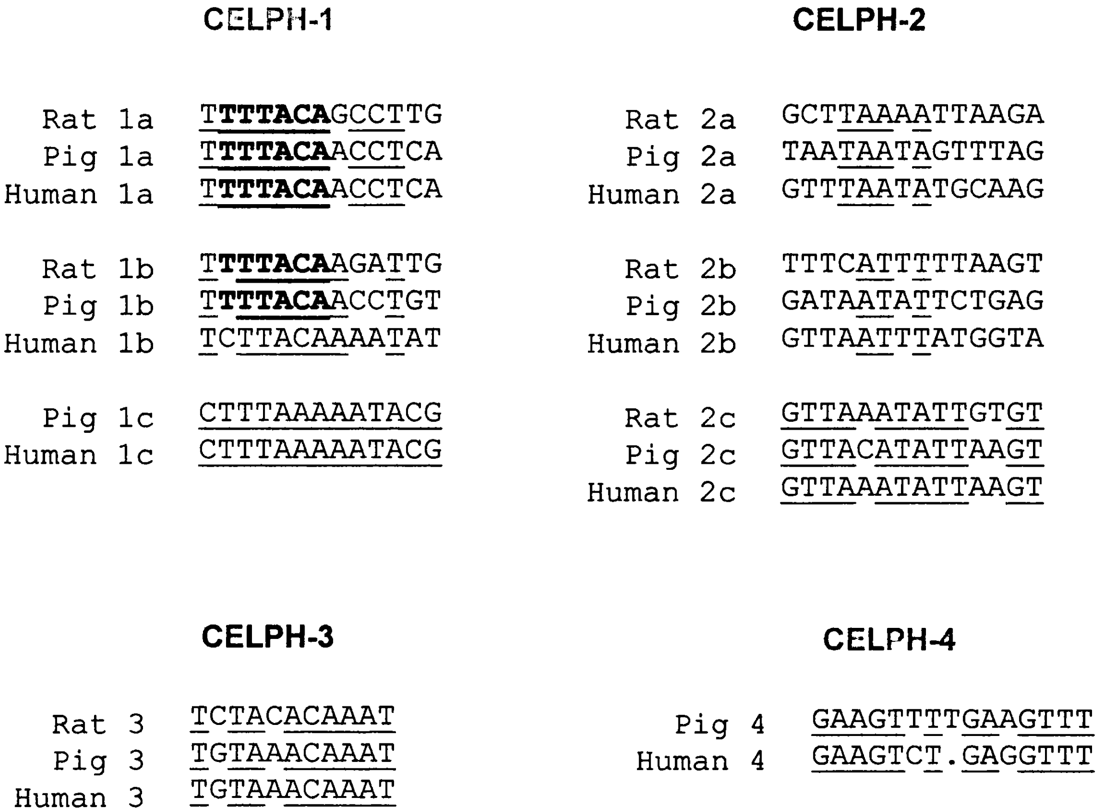


Figure 4.2 **Sequence comparison of human, rat and pig CE-LPH elements**
Underlined nucleotides are shared between all three species, and nucleotides in bold correspond to the Cdx-2 consensus binding motif.

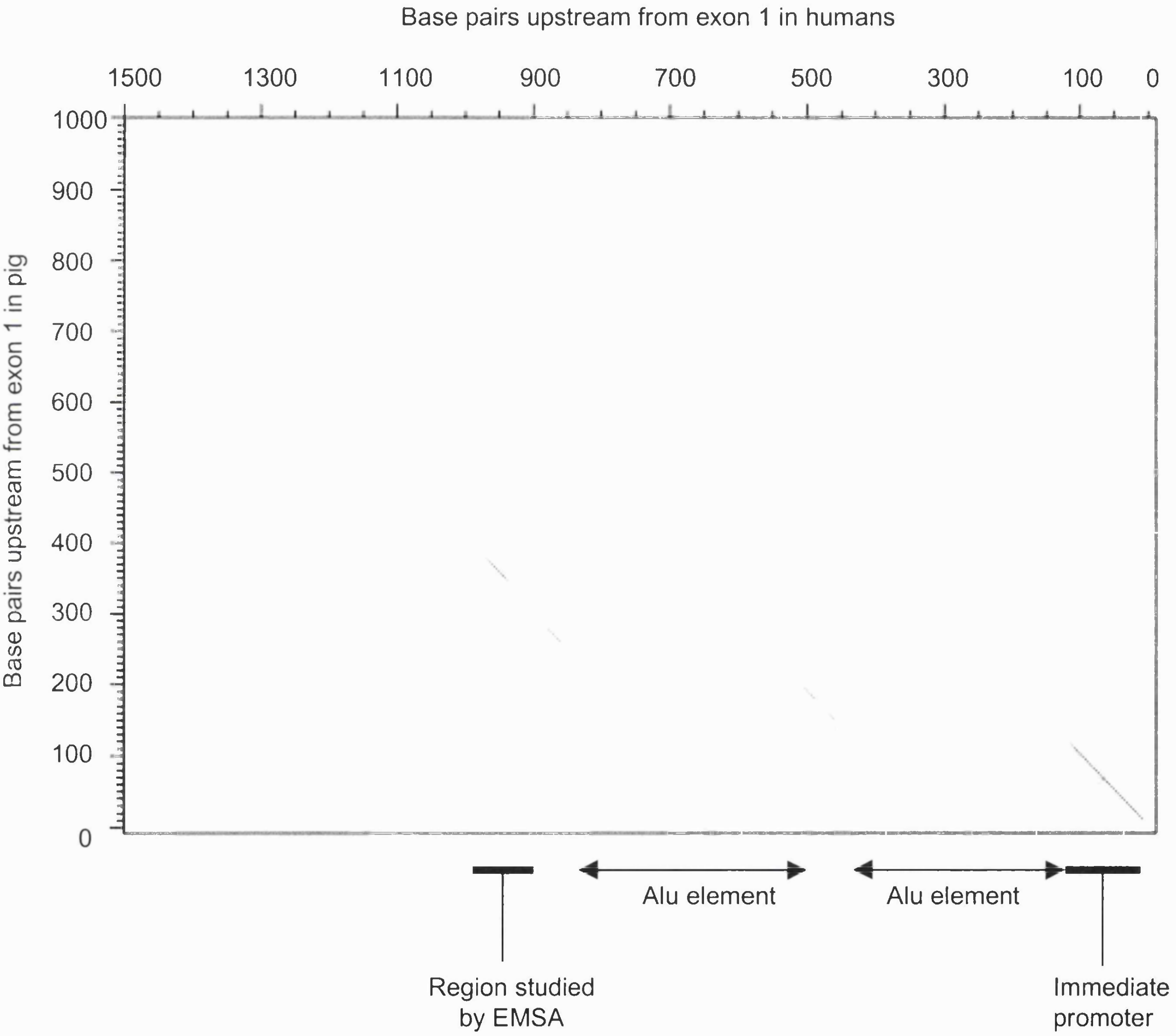


Figure 4.3 **Dotplot comparison between human and pig sequence upstream from lactase.**
Window size is 30, minimum cutoff value is 30, saturation value is 70.
Lines indiacate areas of homology.

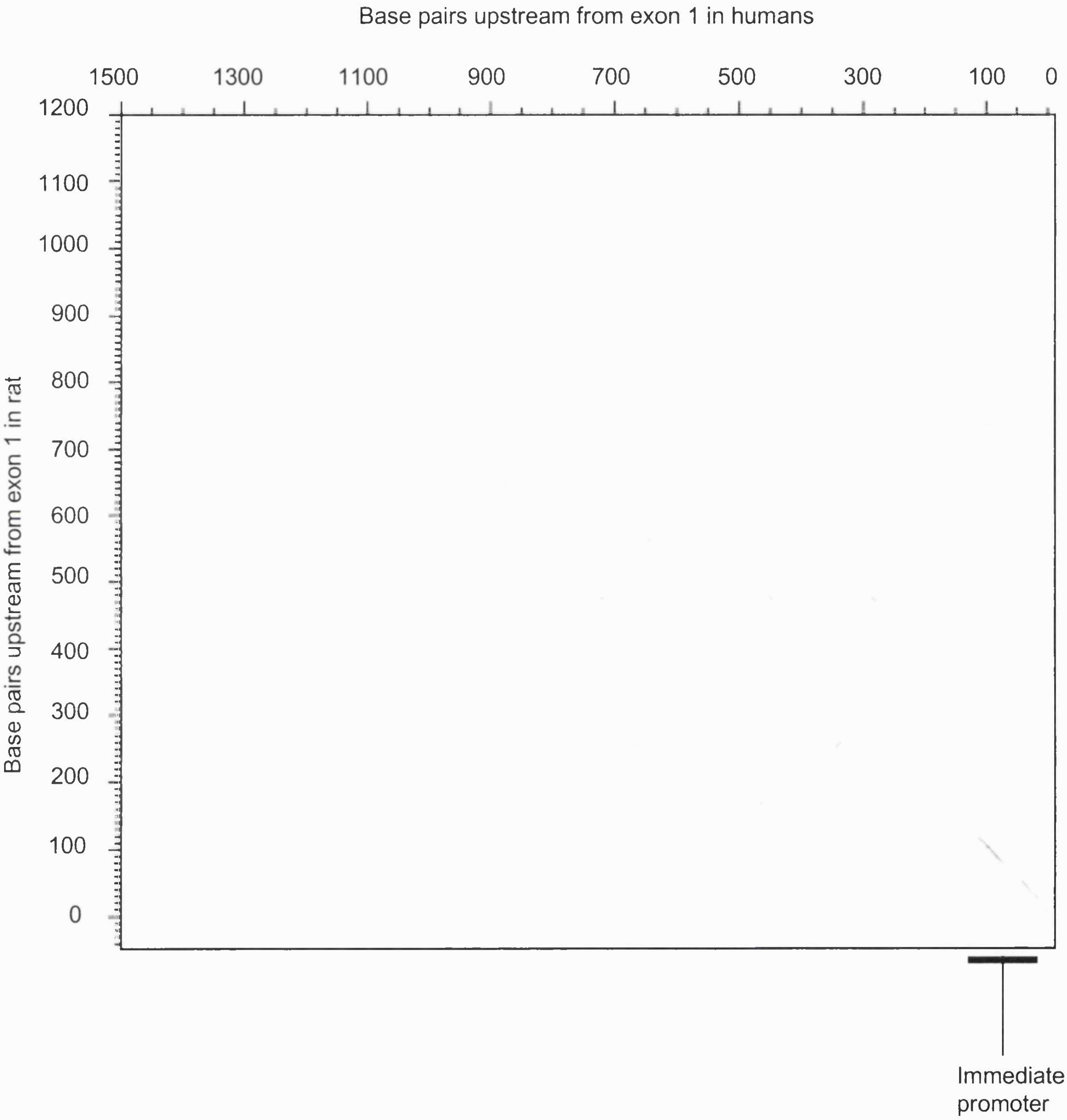


Figure 4.4 **Dotplot comparison between human and rat sequence upstream from lactase.**
Window size is 30, minimum cutoff value is 30, saturation value is 70. Lines indicate areas of homology.

4.1.3.2 Analysis of sequence of a variable region

The region between -800 and -1000 in the human sequence has also been found to contain polymorphic sites at -958, -946, -942 and -875 (Harvey *et al.*, 1995, section 6.1), and if any of these variant sites occurred within a protein binding motif, protein binding may be affected by the allele at that position. This possibility was tested by electromobility shift assays (section 4.2)

Figure 4.5 shows the human sequence of the -850 to -1030 region compared with the pig sequence together with the pig *cis*-acting elements and the human polymorphic sites. Further sequence analysis of this region shows a possible conserved *cis*-element related to CE-LPH1b, which is called CE-LPH1c (figure 4.5, table 4.1). This is on the opposite strand to the other CE-LPH1 elements, and it is 12/13 identical between human and pig. However pig CE-LPH1c is more closely related to human CE-LPH1c than pig CE-LPH1b (figure 4.6). The level of conservation suggests that CE-LPH1c may be a functional element binding the same *trans*-acting factors in both species. CE-LPH1b in comparison is less conserved, so, even if functional, the same proteins may not bind in pig and human.

Other consensus motifs within this region were identified (figure 4.5). An H4TF1 motif and a c-myc motif is showed by both human and pig sequences, and an Oct1/Oct2 site is present in the human sequence. All threebindknown transcription factors involved in controlling many genes.

4.2 Electromobility Shift Assay

4.2.1 Experimental design

In order to investigate the region between -976 and -879 in humans, three groups of double-stranded overlapping oligonucleotides were designed, each 36 base pairs long, and termed group A, group B and group C (figure 4.5). Each group consisted of several oligonucleotides corresponding to an allele in that sequence, and

sense strand containing CE-LPH1b	human	-945 TCTCTTACAAAATATG -930
antisense strand containing CE-LPH1c		-911 TCTTTAAAAATATGCT -926
sense strand containing CE-LPH1b	pig	-354 TCTTTTACAACCTGTG -339
antisense strand containing CE-LPH1c		-317 CCTTTAAAAATACGTA -332

Table 4.1 Comparison of CE-LPH1b and CE-LPH1c elements

were designed to test if allelic differences affected any protein binding observed. Group A oligonucleotides consisted of A_1, A_4, A_6, and A_8 which corresponded to (-958C, -942/-943TC), (-958T, -942/-943TC), (-958C, -942/-943ΔΔ), and (-958T, -942/-943ΔΔ) respectively. The numbers refer to the DGGE gel phenotype of each allelic combination (section 6.1.2). Group B oligonucleotides consisted of B_1 and B_6 which corresponded to -942/-943TC and -942/-943ΔΔ respectively. Only one oligonucleotide formed group C, since the oligonucleotide covered no polymorphic sites. Oligonucleotides were not designed corresponding to A-946G and C-942G since alleles at these sites are comparatively rare (table 6.2). The polymorphism at -875 was not tested since this is also a relatively rare variant (table 6.2).

In the EMSA experiments, nuclear-enriched extracts of the Caco2 cell line were used as a source of proteins. Caco2 is an intestinal cell line that is commonly used for studies of enterocyte gene expression, and is regarded as a valuable cell culture model. Lactase is present, but at quite low levels. This does suggest that all the *trans*-acting factors necessary for expression of lactase are present in the cell line. Thus these *trans*-acting factors would be present in the nuclear-enriched protein extract of Caco2 used in the EMSA experiments described above.

4.2.2 Results using oligonucleotide A

Figure 4.7 shows an EMSA using labelled oligonucleotide A_1 for the first five lanes, and labelled oligonucleotide A_6 for the last five lanes. In both cases no retarded bands are formed in the absence of protein (lanes 1 and 6). When protein is added, in both cases three low mobility bands are observed: the lowest mobility band is the strongest and is named M; the three bands with higher mobility are weaker and are named W (lanes 2 and 7). When 150x molar excess of unlabelled probe is added as a competitor, the bands become very faint indicating that these bands are specific protein-DNA complexes (lanes 3 and 8). Unlabelled A_6 can displace the binding activity shown by A_1 (lane 4), and vice versa (lane 9), showing that the TC-942/-943ΔΔ polymorphism has no effect on binding activity. Labelled A_4 (-958T) cannot displace the binding activity shown by A_1 (lane 5), nor that shown by A_6 (lane 10), which suggests that the C-958T polymorphism affects the binding of protein to the DNA. To confirm this observation the experiment was repeated using A_1 (lanes 1 to 5)

Oligonucleotide name	Sequence
CELPH	AGTATTTTACAACCTCAGTT AAAATGTTGGAGTCAACGTC
17mer (Cdx2)	AATTTTTACAACACCT TTAAAAATGTTGTGGA

Table 4.2 Sequence of competitor oligonucleotides for EMSA

and A_4 (lanes 6 to 10) (Figure 4.8) as labelled probes. As expected, A_1 showed the behaviour mentioned above, and A_4 did not show any protein binding activity (lane 7). As further confirmation, A_8 was used as a labelled probe to confirm that C-958T was the only polymorphism affecting protein binding, and the -942/-943ΔΔ allele did not rescue protein binding activity. The absence of any protein-DNA complex bands showed that -958T abolished protein binding to the sequence of group A oligonucleotides.

To investigate the nature of the proteins binding activity observed, an EMSA experiment using oligonucleotides containing the CE-LPH1a motif and a Cdx-2 consensus binding motif (table 4.2) in competition with labelled oligonucleotide A_1 was performed (figure 4.9). When either oligonucleotide was added in excess there was no dilution of the bands suggesting that these oligonucleotides did not displace the protein complex and so no part of the complex was Cdx2 nor any other protein that binds to CE-LPH1. Analysis of the sequence around C-958T shows that in the pig the sequence TTTATA forms a Cdx2 binding motif, but this is disrupted by two C

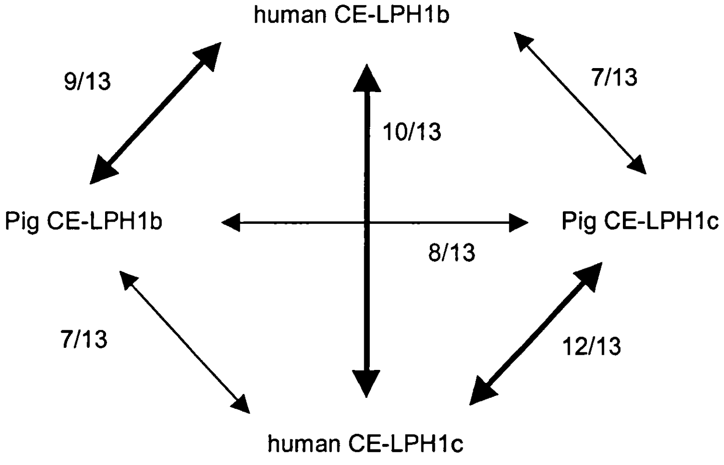


Figure 4.5 Sequence identity of CE-LPH1b and CE-LPH1 compared within and between species
Over 2/3 identity shown by thick line

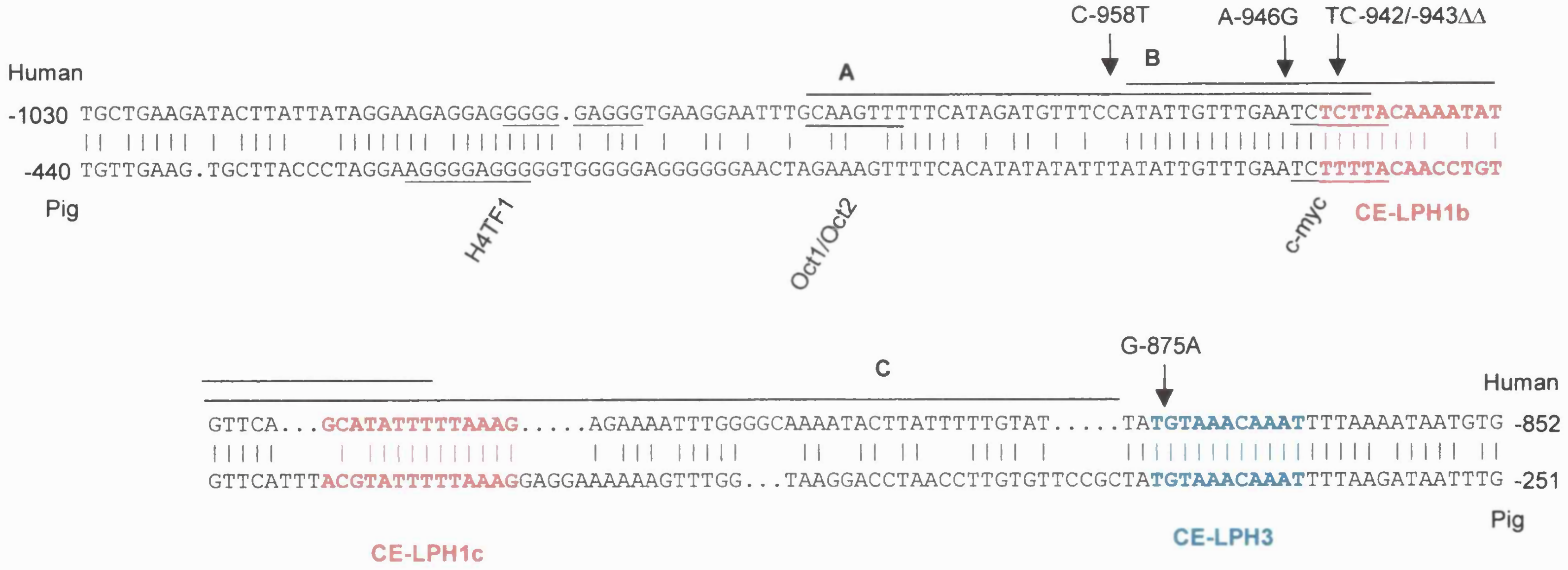

C-958T
A-946G
TC-942/-943ΔΔ
B
Human
A
-1030 TGCTGAAGATACTTATTATAGGAAGAGGAGGGGG.GAGGGTGAAGGAATTTGCAAGTTTTTCATAGATGTTTCCATATTGTTTGAATCTCTTACAAAATAT
-440 TGTTGAAG.TGCTTACCCTAGGAAGGGGAGGGGGTGGGGGAGGGGGGAACTAGAAAGTTTTCACATATATATTTATATTGTTTGAATCTTTTACAACCTGT
Pig
H4TF1
Oct1/Oct2
c-myc
CE-LPH1b
G-875A
C
Human
GTTCA...GCATATTTTTAAAG.....AGAAAATTTGGGGCAAAATACTTATTTTTGTAT.....TATGTAAACAAATTTTAAAATAATGTG -852
GTTCATTTACGTATTTTTAAAGGAGGAAAAAAGTTTGG...TAAGGACCTAACCTTGTGTTCCGCTATGTAAACAAATTTTAAGATAATTTG -251
Pig
CE-LPH3
CE-LPH1c

Figure 4.6 **Comparison of human sequence between -1030 and -852 and pig sequence between -440 and -251 upstream of lactase.**

Pig CE-LPH elements are highlighted together with their human homologues. Other cis-element consensus binding motifs identified by eye and by SIGNALSCAN are also shown. Human polymorphic sites are shown by the vertical arrows. The three groups of oligonucleotides (A, B, and C) designed on human sequence and used for EMSA are shown.

nucleotides in the human sequence. Furthermore, the C to T change abolishes protein binding, whereas if the protein were Cdx2 it would be expected to increase the binding affinity since -958T causes the sequence to become more similar to the Cdx2 consensus binding site. Because the corresponding nucleotide at human position -958 is a T in the pig, it is unlikely that the same protein-DNA complex would be formed using pig sequence. Therefore this is a potential *cis*-acting element which is present in humans but not in pigs.

4.2.3 Results using oligonucleotides group B and C

The sequence covered by oligonucleotides group B and C is shown in figure 4.6, and these oligonucleotides were used in further EMSA experiments. Figure 4.10a shows EMSA analysis using labelled oligonucleotide B_1 as a probe. No retarded bands are present when no protein is added (lane 1), but two bands are formed when protein is added: a slow mobility high band named H and a faster mobility lower band named L (lane 2). Both can be displaced by 150x molar excess of the unlabelled B_1 (lane 3), but a faint trace of band L can be seen. Band H is displaced when oligonucleotide C is a competitor, but band L is only slightly displaced (lane 4).

Using oligonucleotide C as the labelled probe in figure 4.10b, both H and L bands appear (lane 2), but they are displaced equally effectively by oligonucleotide C and B1 (lanes 3 and 4). Comparison of the band intensities of H and L shows variation between oligonucleotide B_1 probe and oligonucleotide C probe. Since the band intensity did vary in repeat experiments, this is probably not a real effect.

Both oligonucleotides B_1 and C span candidate *cis*-elements CE-LPH1b and CE-LPH1c. To test whether the oligonucleotides spanned an element that bound to proteins that bind to CE-LPH1a, the ability of a CE-LPH1 consensus oligonucleotide and a Cdx2 consensus oligonucleotide to displace the bands was tested. Figure 4.10a lane 5 and figure 4.10b lane 5 show that the CELPH oligonucleotide displaced the L band but not the H band which suggests that the protein forming the protein-DNA complex in the L band can bind to CE-LPH1. Cdx2 is a protein that can bind to CELPH1 but the Cdx2 oligonucleotide cannot displace either H or L band (Figure 4.10a and figure 4.10b lane 6).

Figure 4.11 shows an EMSA experiment which tests whether the polymorphism at -942/-943 affects binding of the proteins to the oligonucleotide B sequence. B_6 and B_1 can displace both H and L bands generated by oligonucleotide B_1

(lanes 3 and 4), and both H and L bands generated by B_6 (lanes 7 and 8). Interestingly, the H and L bands generated by B_6 were not competed out as effectively by either B_1 or B_6 as the H and L bands generated by B_1.

Together with the evidence described in section 4.2.2, it seems that the TG-942/-943ΔΔ polymorphism does not affect the binding of proteins to oligonucleotide B sequence.

4.3 Interpretation of results

4.3.1 Binding activities shown by group A oligonucleotides

The identification of a protein binding activity affected by a polymorphism raises questions about its significance with respect to the lactase persistence/non-persistence polymorphism. Persistence occurs virtually exclusively in association with the -958C allele, but non-persistence can occur on both alleles (section 6.3, Harvey *et al.*, 1998) so this change is not responsible for the lactase persistence polymorphism. Assuming that the protein binding activity does have some functional consequence for the expression of the gene, it could have another role and hence may affect some aspect of gene expression. The timing of downregulation of lactase in children is variable (section 5) and this polymorphism may be responsible for that variation.

The sequence around the C-958T polymorphism has no recognised consensus binding motifs, so there is no hint at the possible function.

4.3.2 Binding activities shown by group B and C oligonucleotides

Both B_1 and C oligonucleotides appear to show the same protein binding activity. This could be due to one of two reasons: the binding site for the protein is in the 12bp overlap shared by the two oligonucleotides, or both oligonucleotides share some other sequence. As discussed in section 4.1.3.2, the oligonucleotides span an inverted repeat which contains two related CE-LPH elements, 1b and 1c. It is possible that the two H and L binding activities generated by both oligonucleotides bind to CE-LPH1b and CE-LPH1c, and this is supported by the fact that the protein forming the L band binds to CE-LPH1a. Competition experiments suggest that the L band generated by the the oligonucleotide B_1 is different from that generated by oligonucleotide C, because it is not displaced by unlabelled C. It is possible that they represent two different but related proteins that bind to CE-LPH1a, but it may be an artefact caused by the different intensities of the two L bands.

The protein forming the H band does not bind to CE-LPH1a and could possibly be a CE-LPH1b/CE-LPH1c specific binding protein. It could also bind to a motif in the overlap between oligonucleotides B_1 and C.

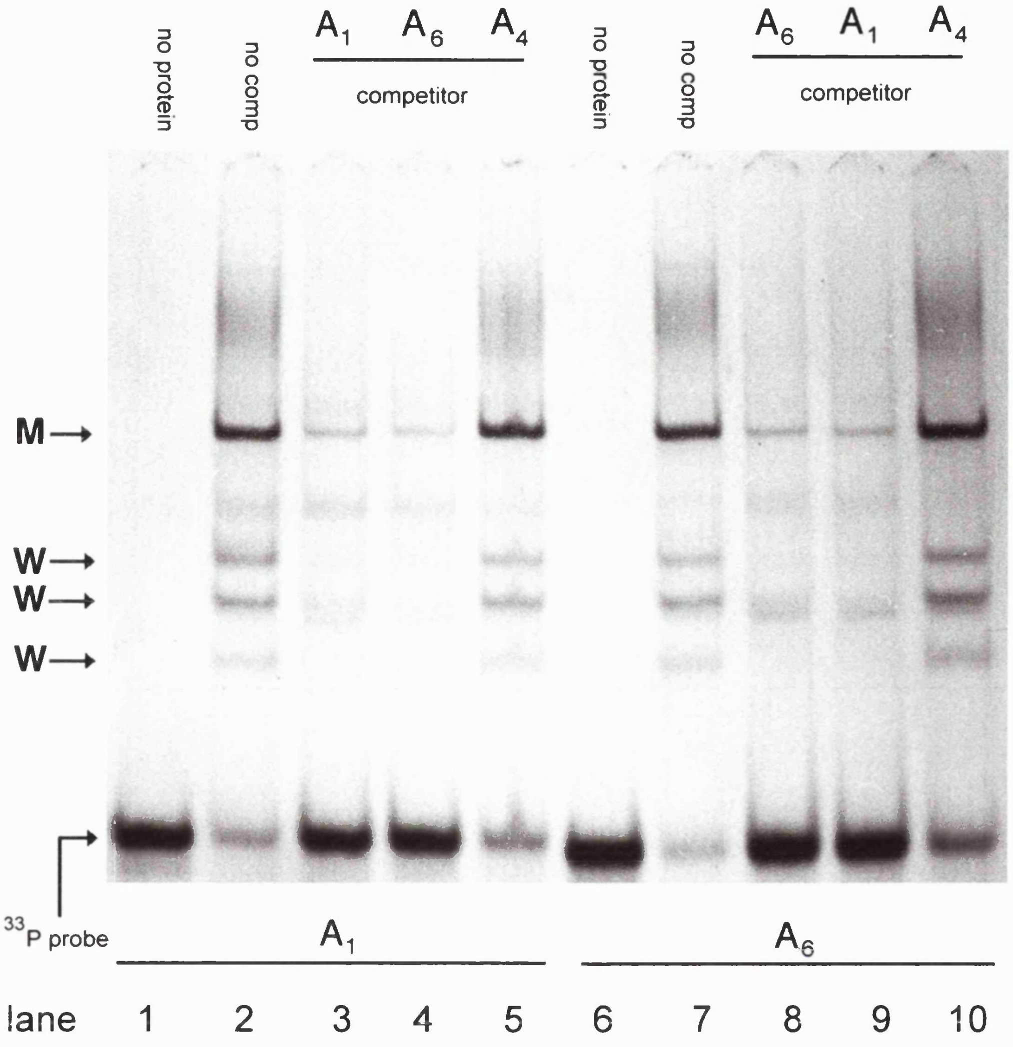


Figure 4.7 **EMSA using ^{33}P labelled oligonucleotides A_1 and A_6**
no comp indicates that no unlabelled oligonucleotide competitor was used. M indicates the main low mobility band and W indicates the three weaker higher mobility bands.

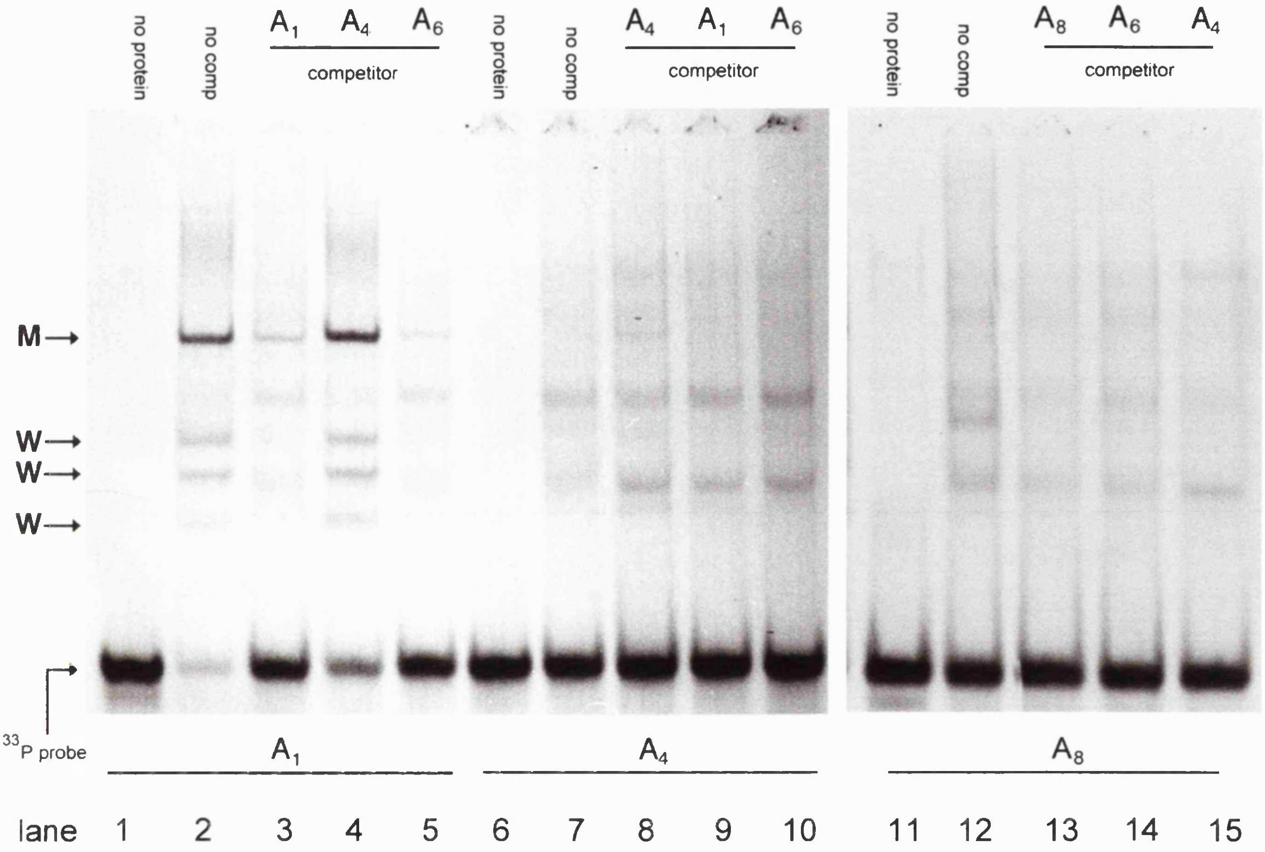


Figure 4.8 **EMSA using ^{33}P labelled oligonucleotides A_1 , A_4, and A_8.**
no comp indicates that no unlabelled oligonucleotide competitor was used. M indicates the main low mobility band and W indicates the three weaker higher mobility bands.

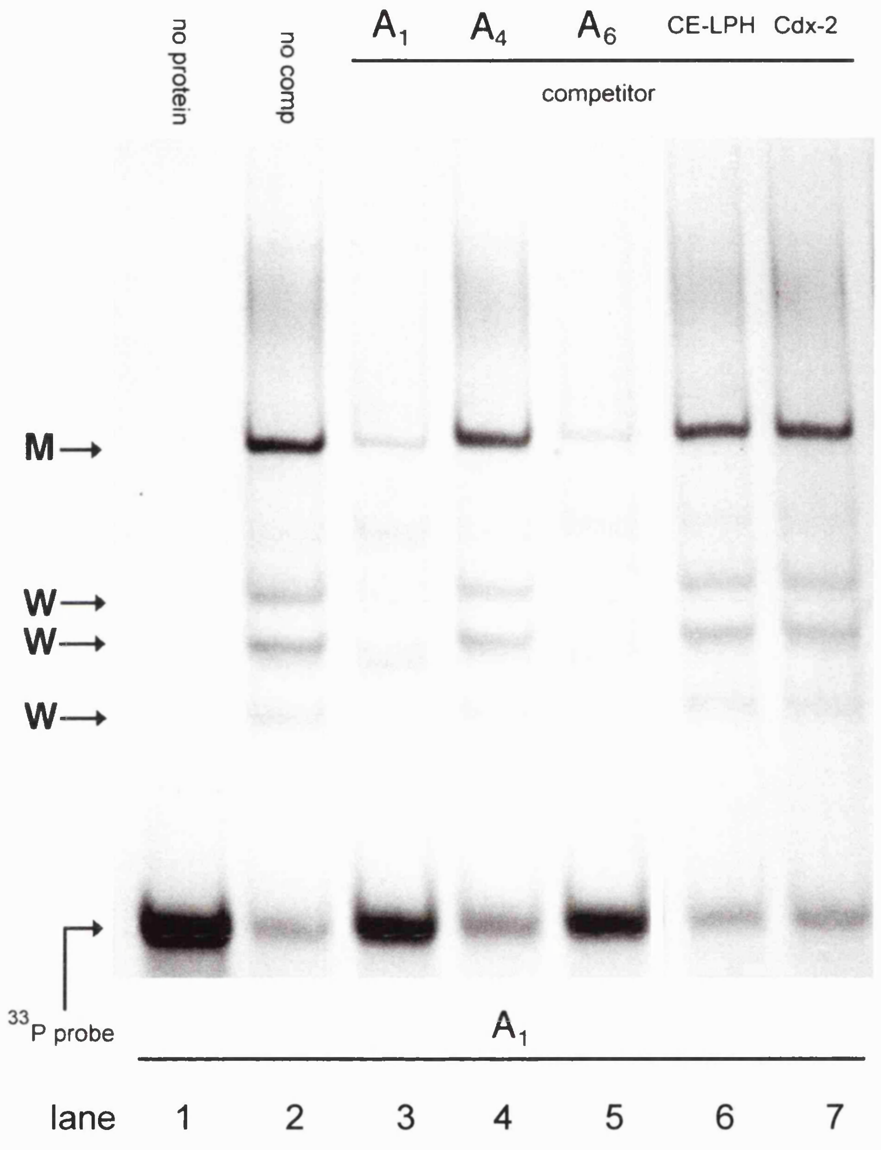


Figure 4.9 **EMSA using ^{33}P labelled oligonucleotide A_1 showing that the binding activity does not involve Cdx-2 or other factors that bind to CE-LPH1a.**
no comp indicates that no unlabelled oligonucleotide competitor was used. M indicates the main low mobility band and W indicates the three weaker higher mobility bands.

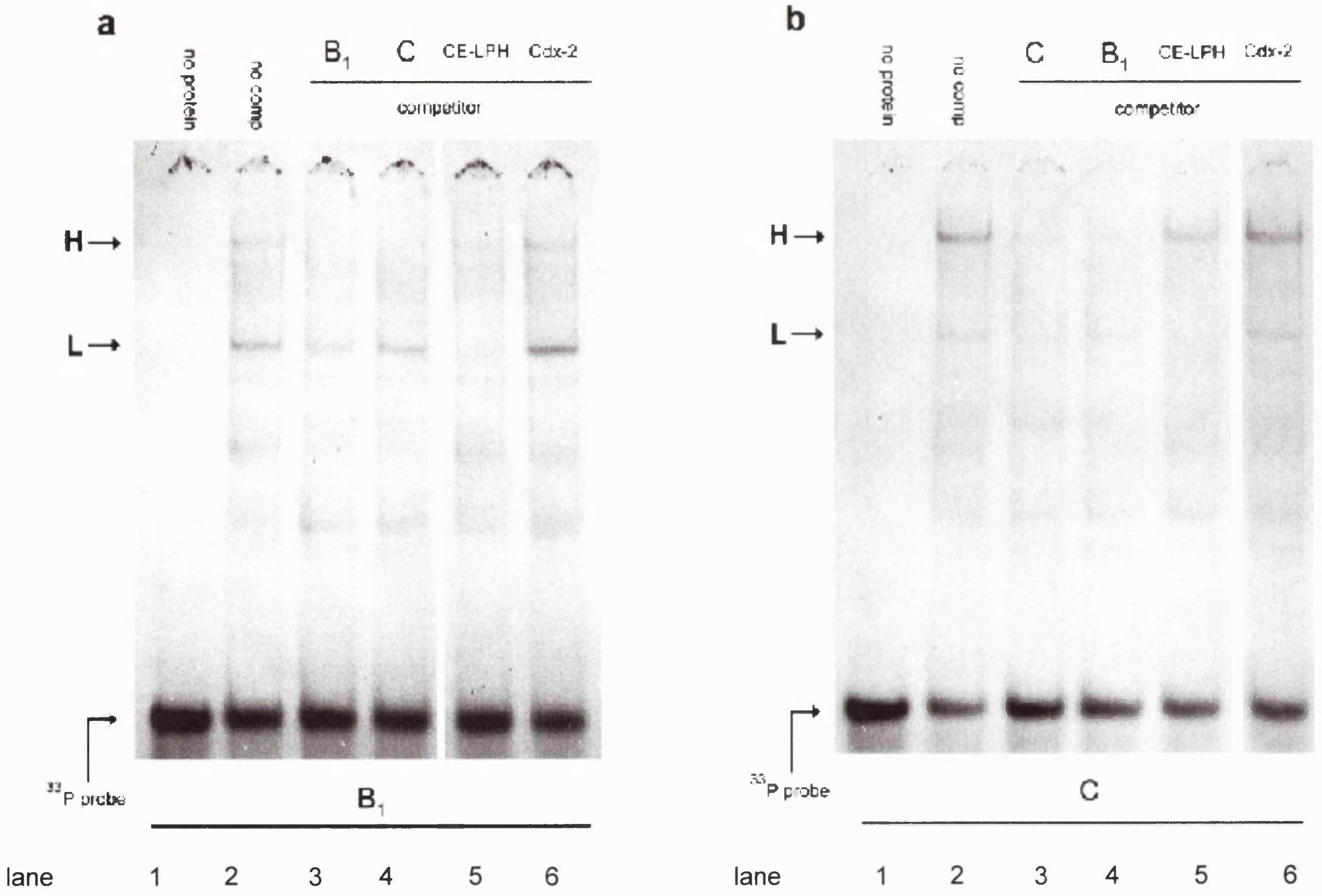


Figure 4.10 **EMSA using ^{33}P labelled oligonucleotides B_1 or C.**
no comp indicates that no unlabelled oligonucleotide competitor was used. H indicates the upper band and L indicates the lower band.

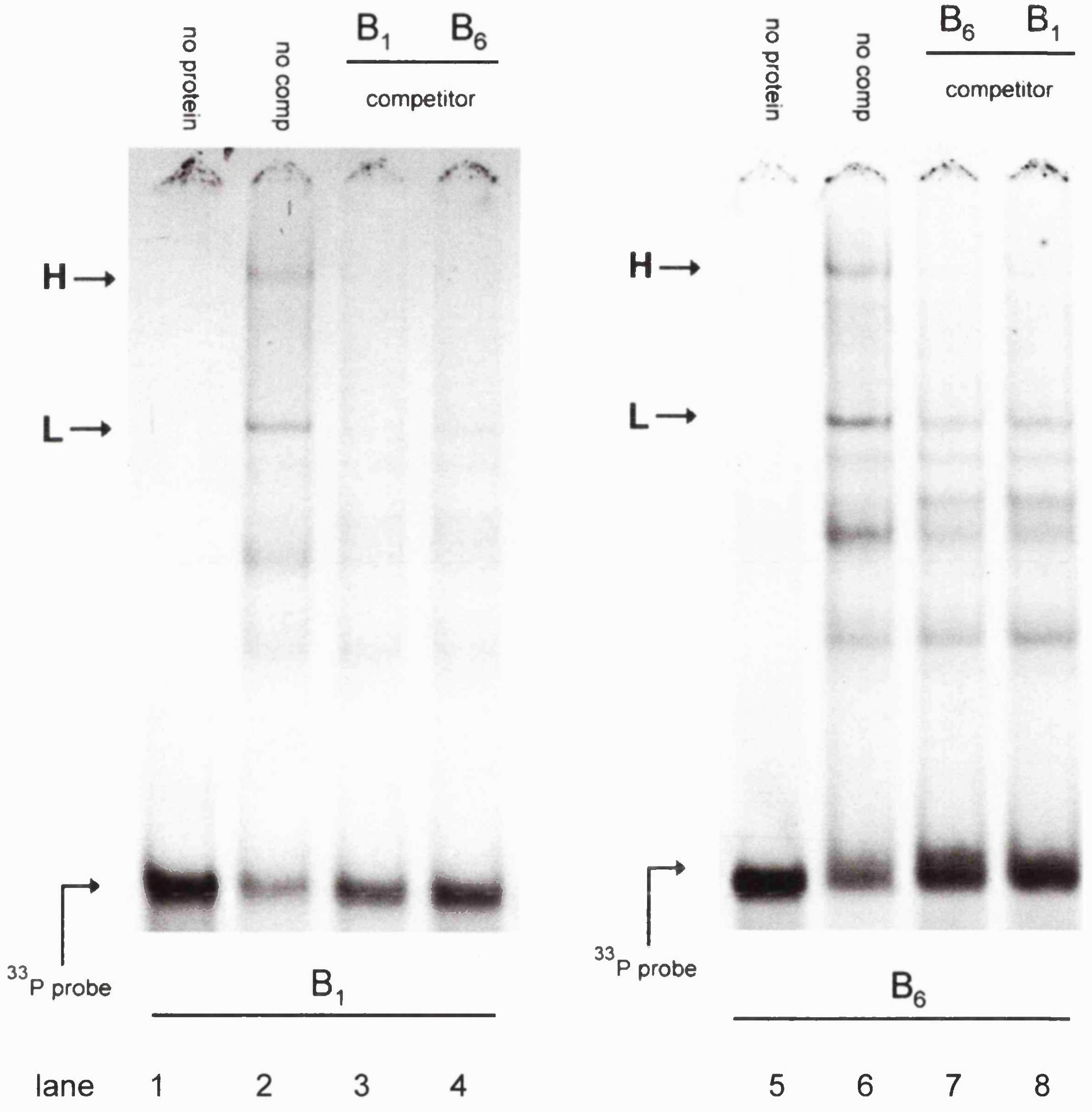


Figure 4.11 EMSA using ^{33}P labelled oligonucleotides B_1 and B_6 showing that the binding activity is not affected by the -942/-943 polymorphism.

no comp indicates that no unlabelled oligonucleotide competitor was used. H indicates the upper band and L indicates the lower band.

5 Analysis of mRNA in intestinal biopsies

5.1 *Analysis of lactase mRNA*

5.1.1 Biopsy

Duodenal biopsies had previously been collected from 32 children of varied ethnic origin at the Queen Elizabeth Hospital for Children, Hackney, London. They had been referred to the gastroenterology clinic for a variety of reasons, including malbsorption, failure to thrive and suspected coeliac disease. In addition, 32 fetal autopsies were collected from the MRC Tissue Bank. Adult biopsies were those previously described in Harvey *et al.* (1994).

5.1.2 Amplification of cDNA

Semi-quantitative reverse transcriptase polymerase chain (RT-PCR) reactions had previously been performed so that the levels of mRNA in each biopsy could be assayed (Wang *et al.*, 1994; Wang *et al.*, 1995). The fetuses and lactase non-persistent adults (defined by having high sucrase:lactase activity ratios) had low levels of lactase mRNA, and children and lactase persistent adults (defined by having low sucrase:lactase activity ratios) had high levels of lactase mRNA. In an extension to these studies, RT-PCR was used to create cDNA PCR products from the mRNA extracted from these samples, and to amplify two regions of the lactase gene (exons 1 to exon 3, and exon 17) using primers XIPS/LCT3A and LCT3S/X17PA respectively. These two RT-PCR products each contained two polymorphisms (C593T, G666A, and T5579C, C5845G).

The alleles at all four polymorphic loci were determined on genomic DNA by PCR using W1S/W1A for exon 1, F2S/F2A for exon 2, and X17f/X17r3 for exon 17. If the individual was heterozygous at one or more loci, the alleles present in the mRNA transcript were determined by direct sequencing of the RT-PCR products. By observing the alleles in the cDNA PCR product as compared to the alleles in the genomic DNA PCR product, it was possible to determine whether one or both alleles were expressed.

5.1.3 Phosphorimaging analysis of sequence

Previous studies had used single stranded conformation analysis or restriction enzyme digest analyses of the PCR products. The level of expression of the two transcripts can be seen by inspecting the band strengths of the two alleles in a sequencing gel (figure 5.1). To quantify the alleles, the amount of radioactivity in each band was measured using phosphorimaging.

Analysis of the 'strength' of the polymorphic bands must take into account the background values and the variation in strength between sequencing lanes. To do this, for each polymorphic band analysed, a control (non-polymorphic) band and the background was measured. Analysis of band strengths along a sequence showed small variations between bands, so it appeared that, ideally, three or more control band values should be measured and the average taken. However, analysis of multiple samples showed that this pattern of variation along the sequence did not vary significantly between samples, apart from a general raising or lowering of values due to sequence lane intensity. One control band, preferably immediately below the polymorphic band, which showed the least sample-to-sample variation, was chosen for each polymorphic locus analysed, and used in every sample measurement (figure 5.2).

For every sample, six readings are taken, and the percentage of cDNA comprised of a specific allele calculated, as described in section 2.3.3. Analysis of homozygous sites on the cDNA ranged from 96% to 103%, or -4% to 4%, depending on which allele was homozygous. Four genomic heterozygotes for exon 1 were also analysed, which would be expected to produce a %C value of 50%. Phosphorimaging produced a result of %C=51.25 (standard error of the mean ±1.29). The technique has since been developed to examine allelic imbalance in the TSC1 gene (Jeganathan *et al.*, 2000).

5.2 Analysis of adult biopsies

To examine more quantitatively, RT-PCR products of five of the original series of adult biopsies were sequenced in three persistent heterozygotes to observe the extent of downregulation of the lactase non-persistent allele, and in two persistent homozygotes to examine the relative levels of expression of both alleles. The results are shown in table 5.1, with two examples of sequencing gels shown in figure 5.1.

			Exon 1 (C593T)		Exon 2 (G666A)	
	hapl	hapl up	Genomic	cDNA (%C)	Genomic	cDNA (%A)
I11	AD	A	CT	7	AG	5
I18	AB	AB	CT	36	AG	41
I26	AB	A	CT	3	AG	-2
P26(182)	AB	AB	CT	56	AG	53
P39	AB	A	CT	3	AG	0

Table 5.1 Allelic imbalance in adult lactase cDNA PCR products
hapl indicates the haplotype of the individual, and *hapl up* indicates the haplotype that is not downregulated.

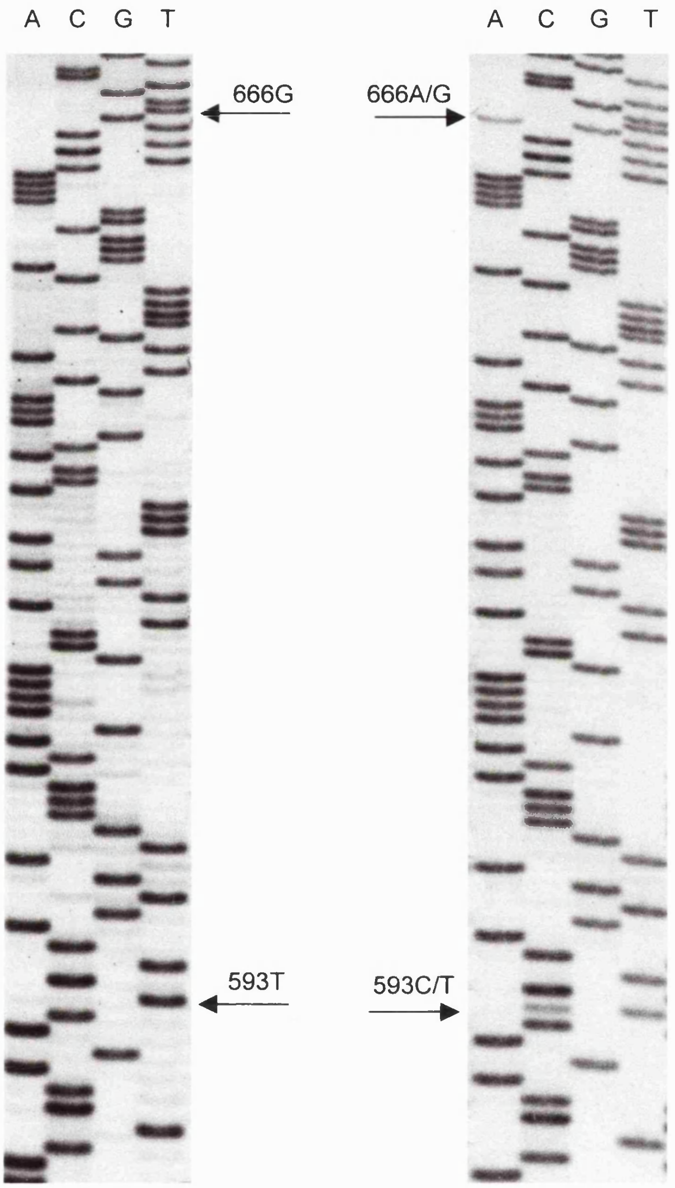
The three persistent heterozygotes show almost complete absence of one allele, and the persistent homozygotes show expression of both alleles. Interestingly, the Italian sample I18 shows slight downregulation of the B haplotype, although the P26 sample (Northern European) shows approximately equal expression. The significance of this slight down regulation is not clear and it may be a 'partial rescue' of a persistent B haplotype.

5.3 Analysis of fetal and infant biopsies

5.3.1 Results

Table 5.2 shows the results of mRNA analysis on the fetal and child biopsy samples. An *h* indicates that the sample was homozygous and therefore not informative at that site, and a dash indicates that the sample was not tested at that site.

All the values are consistent over all polymorphic sites, which discounts any form of splicing regulation between exon 2 and exon 17, where perhaps the first half of the mRNA is in non-persistent individuals but not the second half of the mRNA. The fetal samples all show expression of both alleles (mean expression of one allele 49.0%, standard deviation 4.3%), although the mRNA levels in fetal samples is low (Wang *et al.*, 1998). In children, the lactase mRNA rises (Wang *et al.*, 1998), and in young children (≥16 months) both alleles are equally expressed although the 14 month old child, and even the 7 month old child, do show some evidence of downregulation of one allele. Figure 5.3 shows examples of cDNA PCR products showing different levels of downregulation.



Sample I11

Patient 26
(cell line 182)

Figure 5.1 Sequencing analysis of cDNA PCR products
Part of the sequence of an X1PS/LCT3A PCR product is shown, with the two polymorphic sites shown by arrows. Genomic DNA PCR products from both samples show both sites to be heterozygous.

			Exon 1 (593)		Exon 2 (666)		Exon 17 (5579)		Exon 17 (5845)	
	hapl	hapl up	Genomic	cDNA (%C)	Genomic	cDNA (%A)	Genomic	cDNA (%T)	Genomic	cDNA (%G)
10.1	AB	-	CT	40	AG	46	CT	45	CG	-
11.2	AB	-	CT	45	AG	47	CT	43	-	-
12.0	BC	-	C	h	AG	46	T	h	C	h
12.6	AC	-	CT	-	G	h	CT	47	C	-
12.8	AB	-	CT	50	AG	42	CT	52	-	-
13.1	?B	-	C	h	A	h	CT	47	G	h
13.1	BC	-	C	h	AG	52	T	h	CG	57
13.2	AC	-	CT	53	G	h	CT	52	C	-
13.3	AB	-	CT	53	AG	51	CT	51	-	-
13.7	BC	-	C	h	AG	57	T	h	G	h
13.8	AB	-	CT	46	AG	53	CT	54	-	-
14.8	AB	-	CT	52	AG	47	CT	54	CG	56
2	AB	-	CT	49	AG	45	CT	48	CG	45
3	AB	-	CT	49	AG	49	CT	52	CG	51
5	BC	-	C	h	AG	50	T	-	CG	55
7	AC	C	CT	59	G	h	CT	57	C	h
8	BC	-	C	h	AG	51	T	h	CG	54
14	AB	A	CT	40	AG	42	CT	40	CG	43
16	AB	A	CT	47	AG	40	CT	46	CG	49
16	CE	-	C	h	G	h	CT	51	C	-
22	AB	A	CT	13	AG	15	CT	13	CG	17
42	AC	A	CT	27	G	h	CT	27	C	-
49	AA	A	CT	37	G	h	C	-	C	-
50	AC	A	CT	20	G	h	CT	20	C	-
54	AC	A	CT	40	G	h	CT	33	C	-
98	AB	A	CT	14	AG	16	CT	20	CG	22
132	AB	A	CT	6	AG	0	CT	2	CG	0

Table 5.2 Allelic imbalance in child and fetal lactase cDNA PCR products
The numbers in the first column are ages of the sample: in italics are fetal samples with the age in weeks since last menstruation, and in non-italics are child samples with ages in months. *hapl* indicates the haplotype of the individual (where it can be determined), and *hapl up* indicates the haplotype that is not downregulated.

It would be interesting, if it were possible, to follow expression of a cohort of individuals through their development from birth to adulthood.

5.3.2 Haplotype analysis of the individuals

The C593T and C5845G polymorphisms do not form part of the studied lactase haplotype (section 7) although it appears that 593T associates with the A haplotype, and may possibly show better association with lactase persistence than the the lactase A haplotype alone (Fisher's exact tests shown in table 5.3). However, both Boll *et al.* (1991) and the panel of five individuals of differing persistence status (table 3.1), show that this change is not causative. In the older children, one allele is expressed at progressively lower levels until, as in the 132 month old child, there is virtually no expression of one allele. This child has a lactase expression pattern of an adult heterozygous for lactase persistence. Although, in general, the allelic imbalance increased with the age of the individual, there was some variation. The 22-month old is almost completely downregulated on one allele, yet the 54-month old still has considerable expression of the allele which presumably will be fully downregulated when an adult. It appears that the time of downregulation is variable, although, as discussed in section 1.3.1.1, a larger study would be required to fully describe the downregulation of lactase.

Determining the haplotype of the polymorphic sites across the lactase gene is fully described in section 7.2.2. However, in section 7.3.7 the association of lactase persistence with haplotype A (of which 666G and 5579C are component alleles) is discussed. The haplotype can be deduced directly from the cDNA sequencing data if it is assumed that allelic imbalance of one allele (e.g. low C at C593T) is associated with the low alleles at the other polymorphic loci (e.g. low A at G666A, low T at T5579C and low G at C5845G). In these cases, any allelic imbalance (±8% from 50%) observed at any site was assumed to indicate downregulation of one lactase mRNA. This can be observed in ten children, and in nine children the mRNA that is not downregulated is haplotype A. The significance of this association between haplotype A and lactase persistence can be tested using Fisher's exact test on a two-by-two table (table 5.3), which produces a value for p of less than 0.003. However there is a possibility that this association is due to admixture from a non-persistent population which has a low frequency of haplotype A due to drift.

	Downregulated (non-persistent)	Not downregulated (persistent)
Haplotype A	2	9
Not haplotype A	8	1

p<0.003

	593T	593C
Haplotype A	10	1
Not haplotype A	0	9

p<0.0001

	Downregulated (non-persistent)	Not downregulated (persistent)
593T	1	9
593C	9	1

p<0.001

Table 5.3 Two-by-two contingency tables using data from table 5.2
The one-sided p-value, obtained by Fisher's exact test, is shown underneath each table.

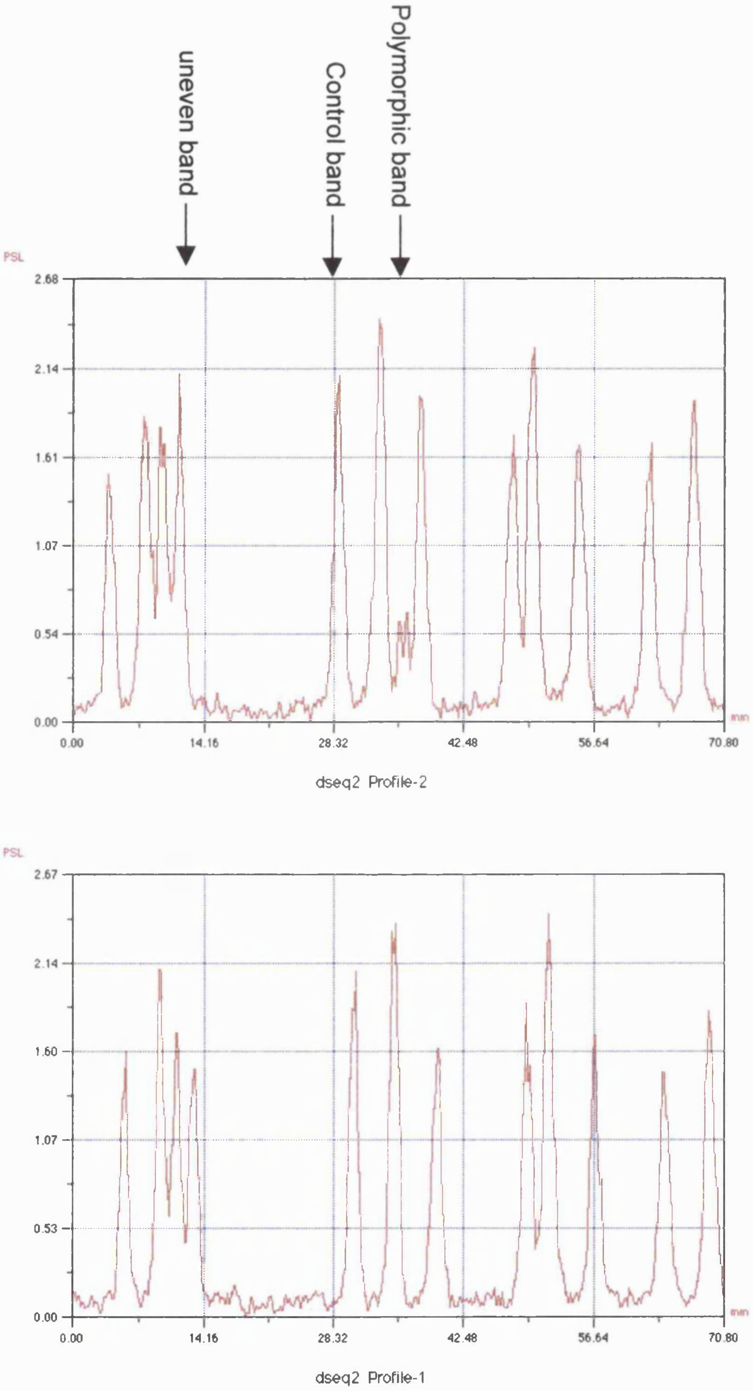


Figure 5.2 **Profile of sequencing lanes from two different samples**
The peaks represent bands on the gel, with bottom to top on the gel represented by left to right on the plot. The top sample shows the polymorphic band as a small peak. The control peak (band on the gel) in both samples has a similar intensity (PSL value), in contrast to the peak labelled *uneven band*.

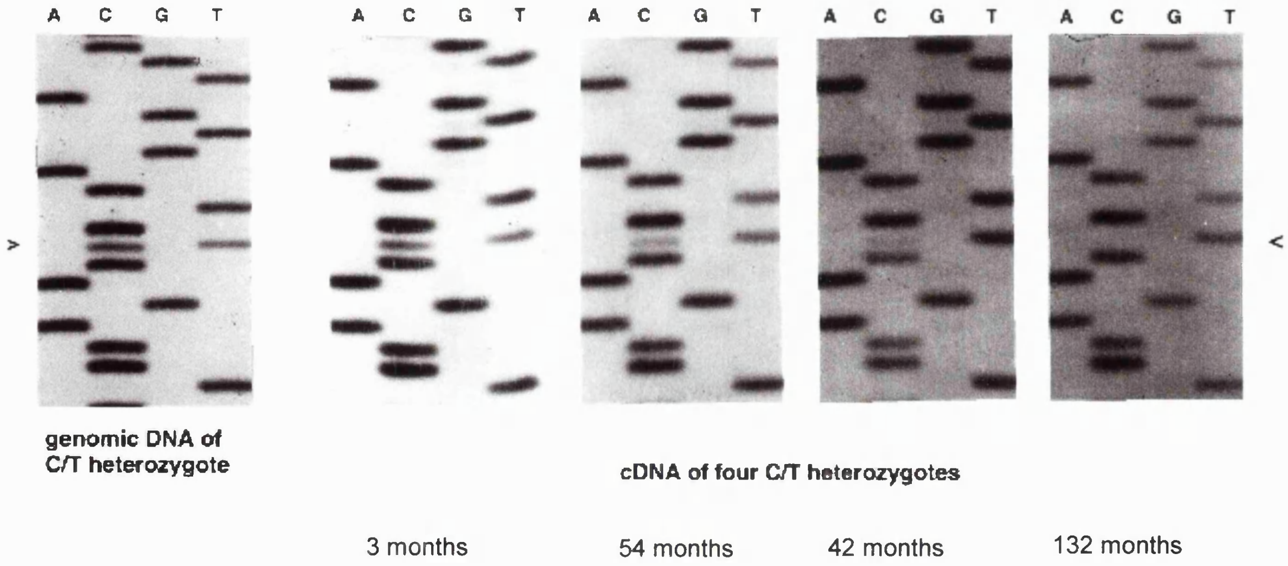


Figure 5.3 **Various stages of downregulation shown in samples of different ages**
A representative genomic heterozygote is shown on the left. The C593T polymorphism is shown.

6 Analysis of polymorphism in the lactase gene in different populations

6.1 *Techniques of polymorphism detection*

6.1.1 Polymorphisms analysed across the lactase gene

Harvey *et al.* (1995) previously analysed seven polymorphic sites across the lactase gene in four different PCR products, which are shown in figure 6.1. The same polymorphic sites were analysed in the work described here in samples from different populations and using essentially the same techniques, as described below. In several cases, extra polymorphic sites were detected; bringing the total number of polymorphic sites analysed to eleven. Several rare alleles were also detected.

6.1.2 Denaturing gradient gel electrophoresis

Harvey *et al.* (1995) developed a DGGE technique to analyse the AvaII digest of the 5F PCR product in the upstream region of LCT (figure 6.2). Analysis of the sequence by two computer programs (MELT87 and SQHTX) showed that, after AvaII digestion, the GC-rich Alu part of each fragment would melt last (Harvey, 1994). Therefore the sequence acted as a natural GC-clamp. Three variants were detected in the large fragment by this technique: variant 1 (-958C -875A), variant 3 (-958T -875G) and variant 4 (-958T -875A). Two variants, S and F were described in the smaller fragment, which corresponded to variation in the length of the poly A tail of the Alu element.

In addition to these alleles, nine more variants were found in the course of this work, four of which, termed 5,6,7 and 8, were found in the large fragment of the Ava II digest. A representative gel showing all the large fragment variants is shown in figure 6.3, and an enlargement showing distiction of the 6 variant from the 1 variant in figure 6.4. To confirm the sequence of the variants that were present as heterozygotes, a number of PCR products from heterozygous individuals were cloned, the 5F PCR and AvaII digestion repeated on ten clones, and each clone analysed on a DGGE gel with the original heterozygote PCR product. A PCR product prepared using the primers 5FS and 5A10 (figure 6.2) from a clone whose identity had

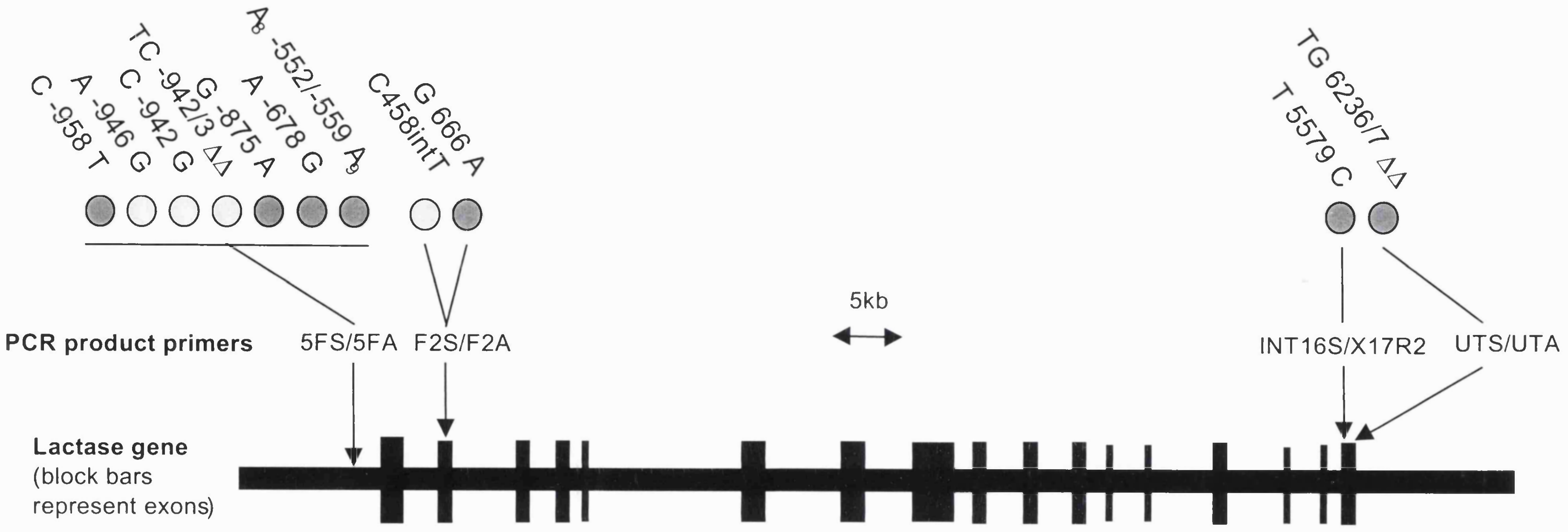


Figure 6.1 **Polymorphic sites analysed in different populations**
Polymorphic sites are shown as dots together with the PCR product used to analyse the loci and the position within the gene. Dark grey dots show polymorphisms analysed by Harvey *et. al.* (1995) and light grey dots show additional polymorphisms found in this study. The numbering is as Boll *et al.* (1991), with numbers corresponding to the transcriptional start site of lactase cDNA, except C458intT which is a polymorphism in intron 1 at base 458 of the genomic sequence for exon 2, Genbank/EMBL accession number M61835. The two exon 17 polymorphisms were also analysed using different primers by allele-specific PCR.

been confirmed by DGGE was then sequenced to confirm the identity of the nucleotide changes of each variant (figure 6.5). Five variants were found in the small fragment, and were named R1,2,3,4 (for the alleles with slower mobility than S) and L for the allele with higher mobility than F. The sequence of each allele is shown in figure 6.6. Three of these new variants are in the poly-A tail of the Alu element, suggesting that this region may be highly variable.

6.1.3 Single strand conformation analysis (SSCA)

SSCA was used to analyse polymorphism in the AvaII digest of 5F PCR product, and polymorphism in the F2 PCR product which spanned exon 2. In both cases the SSCA conditions previously developed were used, with two minor differences described in the section 2.4.3: a slightly higher concentration of glycerol and temperature control by a circulating water bath which made the separations more reliable. Figure 6.7 shows a typical SSCA gel of the 5F PCR product AvaII digest showing the 1,2 (G-875A) polymorphism and the 1,6 polymorphism (TC-942/-943ΔΔ). The resolution of the 1 and 6 alleles can be seen clearly in figure 6.8.

Figure 6.9 shows the location of the F2 primers which span exon 2. A typical SSCA gel of this PCR product is shown in figure 6.10a, which shows the AB polymorphism and a further allele called C. Figure 6.10b shows the distinction of AC and BC heterozygotes from other phenotypes (A, AB, and B). Sequence analysis confirmed that the AB polymorphism corresponds to A666G (Boll *et al.*, 1991; Harvey *et al.*, 1995) and the C variant is 458intT associated with an A at position 666 (figure 6.11). This allele has not been seen associated with 666G but would presumably have been detected by SSCA.

6.1.2 Restriction enzyme digests

Primers were designed spanning the known polymorphism at T5579C, which is within exon 17 (figure 6.12). The 5579C allele creates an Msp1 site, so the allele can be tested conveniently by restriction enzyme digest and analysis on agarose gels. The PCR product was designed to contain another Msp1 site to act as a control for complete enzyme digestion. However, in two individuals variation at this other Msp1 site was found (figure 6.13), but this was very rare and could be detected by different sized bands on the gel.

6.1.3 Simple acrylamide electrophoresis

Analysis of the polymorphism TG6236/7ΔΔ (Boll *et al.*, 1991; Harvey *et al.*, 1995) by simple acrylamide electrophoresis of the UT PCR product was conducted as before. An example of a gel is shown in figure 6.14, with the I allele corresponding to 6236/7TG and the D allele corresponding to 6236/7ΔΔ. As can be seen, the heterozygote is easily identified by its two heteroduplex bands, but distinguishing homozygote D from homozygote I may be difficult. Because of potential ambiguity in some gels, an allele-specific PCR technique was developed to analyse this polymorphism.

6.1.4 Allele-specific polymerase chain reaction (AS-PCR)

AS-PCR is a convenient method of analysing two polymorphic sites and determining phase between alleles at these sites. The polymorphisms at 5579 and 6236/7 were analysed by AS-PCR for three reasons:

a. Both allelic variants have changes convenient for discrimination by oligonucleotides. The 5579 polymorphism is a C to G change, and a G-G and C-C base pairings are not very stable thus aiding specificity (Huang *et al.*, 1992). The 6236/7 polymorphism is a two-base pair deletion.

b. The two allelic sites are relatively close (657 bp) which is not too large for PCR using Taq polymerase to generate a reasonable quantity of product.

c. By knowing the phase of the two alleles the two loci can be treated as one locus with four potential alleles (+I,+D,-I and –D), reducing potential errors when haplotypes are constructed (see section 7.2.2).

Specific oligonucleotides were designed (see section 2.4.6) and the optimum annealing temperature determined by annealing at 58°C and rising a degree at a time if allele-specificity was not observed in that experiment (see figure 6.15). The allele specificity was defined as absence or near absence of product from the incorrect genotype DNA, and the appropriate annealing temperature was determined using all primer pairs (figure 6.16).

Allele-specific oligonucleotides were also used in PCR reactions to type CATT+225ΔΔΔΔ and T+658C in selected samples (figure 6.12) and a similar procedure was used to determine the optimum annealing temperature for each combination of oligonucleotide primers.

6.2 Patterns of allele frequencies

6.2.1 Determining the ancestral allele by analysis of chimpanzee sequence

In most cases the chimpanzee allele was determined from the sequence amplified from five unrelated chimpanzees using the same PCR conditions as human samples. The chimpanzees were all male common chimpanzees (*Pan troglodytes*) from West Africa. The chimpanzee allele at each site is shown in table 6.1. Interestingly, Carl is heterozygous for TG6236/7ΔΔ, which suggests either that it is a very old polymorphism predating the separation of *Homo* and *Pan*, or that it has occurred by a recurrent mutation. Several other polymorphisms were identified in the chimpanzees by sequencing the PCR products from the five chimpanzees. These are shown in figure 6.17 together with an example of sequence of a chimpanzee polymorphism.

Polymorphic site	Tank	Harv	Colin	Carl	Kasey	Buttons
C-958T	-	-	-	-	-	C
TC-942/-943ΔΔ	TC	TC	TC	TC	TC	TC
G-875A	G	G	G	G	G	G
G-678A	A	A	A	A	A	A
A_8-552/-559A_9	n/a	n/a	n/a	n/a	n/a	n/a
Cint458T	C	C	C	C	C	-
G666A	G	G	G	G	G	-
C1430T	C	C	C	C	-	-
C4617T	C	C	C	C	C	-
T5579C	T	T	T	T	T	-
TG6236/7ΔΔ	TG	TG	TG/ΔΔ	TG	TG	-
CATT+225ΔΔΔΔ*	CATT	CATT	CATT	CATT	CATT	-
T+658C*	T	T	T	T	T	-

Table 6.1 Chimpanzee sequence at sites which are polymorphic in humans
* chimpanzee sequence assumed from allele-specific PCR

6.2.2 Variation between populations

Table 6.2 shows the allele frequencies of every locus tested in each population. The standard error σ is calculated assuming that allele frequencies are binomially distributed using the equation

$$\sigma=\sqrt{pq/2n}$$

where n is the sample size and p and q the frequency of each allele.

	N.Europe		S.Europe		N.Indian		S. Indian		Malay		Chinese		Japanese		N.Guinean		Roma		Mordavian		Yakut		Russian		Bantu		San	
statistic	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H	p	H
-958T	0.08 ±0.03	0.15	0.40 ±0.05	0.48	0.28 ±0.04	0.40	0.23 ±0.05	0.35	0.17 ±0.03	0.28	0.20 ±0.04	0.32	0.11 ±0.03	0.20	0.42 ±0.06	0.49	0.35 ±0.04	0.46	0.28 ±0.05	0.40	0.10 ±0.03	0.18	0.27 ±0.05	0.50	0	-	0.03 ±0.03	0.06
-946G	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0.01 ±0.01	0.02	0.03 ±0.03	0.06
-942G	0	-	0	-	0	-	0	-	0.01 ±0.01	0.02	0.03 ±0.02	0.06	0.01 ±0.01	0.02	0	-	0	-	0	-	0	-	0	-	0	-	0	-
-942/3ΔΔ	0	-	0	-	0	-	0	-	0.10 ±0.02	0.18	0.20 ±0.04	0.32	0.28 ±0.05	0.40	0.04 ±0.02	0.08	0	-	0.04 ±0.02	0.08	0.05 ±0.02	0.01	0.03 ±0.02	0.06	0.28 ±0.05	0.40	0.57 ±0.09	0.49
-875A	0.03 ±0.02	0.06	0.06 ±0.02	0.11	0	-	0	-	0	-	0	-	0	-	0	0	0.01 ±0.01	0.02	0.01 ±0.01	0.02	0	-	0.01 ±0.01	0.02	0	-	0	-
-678G	0.04 ±0.02	0.08	0.12 ±0.03	0.21	0.25 ±0.03	0.38	0.32 ±0.05	0.44	0.23 ±0.03	0.35	0.12 ±0.03	0.21	0.19 ±0.04	0.31	0.17 ±0.04	0.28	0.13 ±0.03	0.23	0.13 ±0.04	0.23	0.37 ±0.05	0.47	0.20 ±0.04	0.32	0.42 ±0.06	0.49	0.03 ±0.03	0.06
-552 to –559 A_9	0.89 ±0.03	0.20	0.37 ±0.05	0.47	0.44 ±0.04	0.49	0.45 ±0.05	0.50	0.61 ±0.03	0.48	0.68 ±0.05	0.44	0.68 ±0.05	0.44	0.42 ±0.06	0.49	0.49 ±0.04	0.50	0.59 ±0.06	0.48	0.53 ±0.06	0.50	0.50 ±0.06	0.50	0.43 ±0.05	0.49	0.70 ±0.08	0.42
458intT	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0	-	0.03 ±0.02	0.06	0.13 ±0.06	0.23
666A	0.08 ±0.03	0.15	0.39 ±0.05	0.48	0.29 ±0.04	0.41	0.21 ±0.04	0.33	0.27 ±0.03	0.39	0.37 ±0.05	0.47	0.41 ±0.05	0.48	0.46 ±0.06	0.50	0.34 ±0.04	0.45	0.33 ±0.06	0.44	0.18 ±0.04	0.30	0.30 ±0.05	0.42	0.36 ±0.06	0.46	0.80 ±0.07	0.32
5579C	0.90 ±0.03	0.18	0.47 ±0.05	0.50	0.47 ±0.04	0.50	0.44 ±0.05	0.49	0.52 ±0.04	0.50	0.50 ±0.05	0.50	0.40 ±0.05	0.48	0.29 ±0.05	0.41	0.53 ±0.04	0.50	0.56 ±0.06	0.49	0.45 ±0.06	0.50	0.50 ±0.06	0.50	0.19 ±0.05	0.31	0.23 ±0.08	0.35
6236/7ΔΔ	0.08 ±0.03	0.15	0.38 ±0.05	0.47	0.25 ±0.03	0.38	0.19 ±0.04	0.31	0.17 ±0.03	0.28	0.15 ±0.04	0.26	0.16 ±0.04	0.27	0.36 ±0.06	0.46	0.35 ±0.04	0.46	0.25 ±0.05	0.38	0.10 ±0.03	0.18	0.25 ±0.05	0.38	0.19 ±0.05	0.31	0.23 ±0.08	0.35
P(≥1 het)		0.65		0.97		0.97		0.96		0.96		0.96		0.96		0.97		0.97		0.96		0.94		0.97		0.96		0.95
P(≥2 het)		0.59		0.95		0.94		0.93		0.94		0.94		0.96		0.95		0.94		0.94		0.92		0.94		0.96		0.94

Table 6.2 **Allele frequencies in different populations.** p indicates frequency of non-ancestral allele, and H indicates heterozygosity. P(≥1 het) is the probability of one or more heterozygotes in 11 sites, P(≥2 het) is probability of two or more heterozygotes in 11 sites.

	N.Europe		S.Europe		N.Indian		S. Indian		Malay		Chinese		Japanese		N.Guinean		Roma		Mordavian		Yakut		Russian		Bantu		San	
n	52		52		78		41		100		51		41		36		85		34		39		40		36		15	
	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E
1	43	43.4	34	32.3	40	40.2	26	25.0	50	53.3	20	20.1	15	14.6	9	10.6	40	36.2	18	17.7	32	31.4	22	21.0	21	18.8	3	2.4
14	9	8.2	14	17.3	32	31.6	12	14.0	28	24.8	13	11.9	8	6.0	20	16.3	32	28.5	13	13.7	6	7.2	14	15.9	0	0	0	0.4
4	0	0.4	4	2.3	6	6.2	3	2.0	2	2.9	0	1.8	0	0.6	4	6.3	13	10.2	3	2.7	1	0.4	4	3.0	0	0	0	0.0
16	0	0	0	0	0	0	0	0	18	14.6	11	11.9	11	13.7	1	1.6	0	0	0	0	0	0	0	0	10	14.4	6	6.8
46	0	0	0	0	0	0	0	0	2	3.4	6	3.5	2	2.8	2	1.3	0	0	0	0	0	0	0	0	0	0	1	0.6
6	0	0	0	0	0	0	0	0	0	1.0	1	1.8	5	3.2	0	0.1	0	0	0	0	0	0	0	0	5	2.8	5	4.8
1	47	47.1	19	17.9	47	42.4	20	19.8	58	60.1	20	20.1	27	26.6	26	25.0	66	64.4	26	25.6	14	15.4	26	25.6	15	17.5	13	13.1
12	5	4.8	23	25.2	21	30.2	17	17.4	39	34.9	24	23.8	12	12.9	8	10.0	16	19.2	7	7.8	21	18.2	12	12.8	12	12.3	2	1.9
2	0	0.1	10	8.9	10	5.4	4	3.8	3	5.1	7	7.1	2	1.6	2	1.0	3	1.4	1	0.6	4	5.4	2	1.6	9	6.2	0	0.1
S	38	38.9	4	6.6	19	16.2	9	8.3	33	37.2	23	23.3	18	19.1	6	6.2	40	36.2	14	11.8	8	10.8	8	10.0	11	6.7	8	7.3
SF	14	12.1	29	23.8	33	38.7	19	20.3	56	47.6	23	22.3	20	17.8	18	17.5	31	38.5	12	16.5	25	19.4	24	20.0	9	17.7	5	6.3
F	0	0.9	19	21.6	26	23.2	13	12.3	11	15.2	5	5.3	3	4.1	12	12.2	14	10.2	8	5.8	6	8.8	8	10.0	16	11.7	2	1.3
A	43	43.4	19	18.5	48	43.9	27	25.8	49	53.3	41	39.7	15	14.0	9	10.6	24	19.8	16	16.2	32	30.6	21	19.6	17	16.0	0	2.0
AB	9	8.2	24	25.0	21	29.2	11	13.5	48	39.4	8	10.6	18	19.9	21	17.9	34	42.4	15	14.5	6	8.8	14	16.8	9	11.3	5	3.7
B	0	0.4	9	8.5	9	4.9	3	1.8	3	7.3	2	0.7	8	7.0	6	7.6	27	22.8	3	3.2	2	0.6	5	3.6	3	2.0	1	1.7
AC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4.7	6	3.3
BC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1.7	3	3.0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0	1.4
+	40	40.7	13	15.1	21	17.1	9	7.9	22	26.5	15	13.3	8	6.6	4	3.1	27	22.8	14	11.8	8	11.9	10	10.0	2	1.4	0	0.8
+-	12	10.6	30	25.8	31	38.8	18	20.2	59	50.0	22	25.5	17	19.7	13	14.9	34	42.4	12	16.5	27	19.3	20	20.0	10	11.3	7	5.4
-	0	0.7	9	11.1	26	22.1	14	12.9	19	23.5	14	12.3	16	14.6	19	18.1	24	19.8	8	5.8	4	7.9	10	10.0	24	23.4	8	8.8
I	43	43.4	19	19.1	44	43.9	29	26.6	69	68.9	36	37.1	28	28.1	13	14.7	42	38.2	21	19.1	33	31.4	24	22.5	24	23.4	9	8.8
ID	9	8.2	25	24.8	29	29.2	8	12.9	28	28.2	15	12.8	13	10.9	20	16.6	30	37.6	9	12.8	6	7.2	12	15.0	10	11.3	5	5.4
D	0	0.4	8	8.1	5	4.9	4	1.6	3	2.9	0	1.1	0	1.1	3	4.7	13	9.2	4	2.1	0	0.4	4	2.5	2	1.4	1	0.8

Table 6.3 Observed and expected gel phenotype frequencies in different populations. Shading in blue indicates departure from Hardy Weinberg frequencies ($p<0.05$). p-values are calculated from χ-squared values using one degree of freedom, except when three alleles are observed where three degrees of freedom are used. n is number of individuals tested.

The heterozygosity value for each locus in each population, calculated as shown in section 2.7.5) is shown next to the allele frequency. Also shown at the bottom of each column is the probability that an individual from the population is heterozygous for one site or more, and the probability that the individual is heterozygous for two sites or more, both assuming no association. If an individual is heterozygous for two or more sites then the two haplotypes of that individual cannot be deduced with certainty, so this value has implications for haplotype analysis, as discussed in the next chapter.

Table 6.3 shows observed numbers of gel phenotypes with expected numbers of genotypes calculated from the allele frequency assuming Hardy-Weinberg equilibrium. The program ASSOCIATE tests the observed data for significant deviation from Hardy-Weinberg equilibrium, and any deviation significant at the 5% level ($p \leq 0.05$) is shaded in green. Several variants detected by the techniques described and sequenced above are not shown in tables 6.2 and 6.3. They are rare, or possibly unique mutations, and are shown, together with the population in which they occurred, in table 6.4.

6.2.3 Departures from Hardy-Weinberg Equilibrium

Table 6.3 shows, by shading in blue, that the observed gel phenotypes differ is several populations from expected proportions calculated using Hardy-Weinberg expectations. The apparent deviations from Hardy-Weinberg equilibrium are probably due to small sample sizes, or the result of testing the significance of multiple observations. However, the AB (G666A) polymorphism is within the coding sequence and, in the Malay and North Indians, is not in Hardy-Weinberg equilibrium. A heterozygote advantage or some other form of selection may skew genotype distributions, and polymorphism within the coding region of a protein expressed on epithelial cells would be a good candidate for selection. Indeed, the AB polymorphism encodes a change from valine to isoleucine (Boll *et al.*, 1991). However, the AB polymorphism is in exon 2, which codes for the part of the lactase polypeptide which is cleaved before transport to the brush border membrane (section 1.2.1.1). The T5579C polymorphism in exon 17 does not alter amino acid sequence, and so is an unlikely candidate for selection.

Allele	PCR product	Base change	Number	Population/cohort
8	5F (large fragment)	-958T -942/3ΔΔ	1	Black British
L	5F (small fragment)	-551A	3	San (probably polymorphic)
R1	5F (small fragment)	-543/-550A$_9$	2	Finns (probably polymorphic)
R2	5F (small fragment)	-552/-559A$_{10}$	1	Papua New Guinean
R3	5F (small fragment)	-528T	1	Unknown (Child biopsy series)
R4	5F (small fragment)	-533G	1	Malay
Msp control cut	INT16S/X17r2	5788A	4	only black British tested (probably polymorphic)

Table 6.4 Details of rare variants detected

6.3 Association with persistence/non-persistence in Finns

6.3.1 Previous work

Previous work showed that in an Italian cohort, the -552/-559A$_9$ allele and 5579C allele showed strongest association with lactase persistence (Harvey *et al.*, 1998). A Finnish population, which was less likely to contain population stratification, was studied to see if the same result was obtained.

6.3.2 Method of lactose tolerance testing

Lactose tolerance testing was conducted in Helsinki by Dr Riitta Korpela using a urinary galactose and breath hydrogen protocol to determine low or high LDC (see section 1.1.1.2), and from this to deduce lactase persistence or non-persistence. The cohort of individuals were all female, and selected so that approximately half had low LDC. Because the cohort is not randomly chosen, it is not regarded as representative of the population and is not treated as a population sample.

6.3.3 Two by two test

Two by Three tables showing presumed genotype and lactase persistence status derived from the lactose tolerance tests were constructed for each of the six informative loci (table 6.5). Two by two tables were derived from the two by three tables by converting the three presumed genotypes (for example CC, CT, TT) to two presumed allele frequencies (for example C, T).

Fisher's exact two by two test, described in section 2.7.1, was used to test association of alleles with lactase persistence or lactase non-persistence status. Significant association was found with both -552/-559A_9 (p=0.03) upstream of exon 1 and 5579C (p=0.01) in exon 17, which gives no clue to the location of the persistence polymorphism but suggests that these alleles are of similar age. Analysis of the haplotype background of these alleles is discussed in the next chapter.

6.4 Association with persistence/non-persistence in Yakut

6.4.1 Method of lactose tolerance testing

Lactose tolerance testing was conducted by Ms Sardana Markova and Dr Andrew Kozlov in Yakutsk using the blood glucose assay described in section 1.1.1.2 that involved a single blood glucose determination 40 minutes after ingestion of lactose, and from these LDC tests lactase persistence/non-persistence is deduced. The raw data is shown in appendix 3.

6.4.2 Two by Two test

Two by three tables shown in table 6.5 were created and two by two tables deduced as in the Finns above. Association with lactase persistence/non-persistence phenotype was determined using Fisher's two by two exact test, and no significant association with any allele was found. This was unexpected, and contrary to other data. There could be several explanations:

1. The lactose tolerance tests were probably less accurate because they measure lactase levels indirectly, and ideally more than one measuremenf of blood glucose levels should be taken after the ingestion of the lactose.
2. A different mutation could be causing lactase persistence in the Yakut, one which could be *trans*-acting. Measurement of lactase mRNA from intestinal biopsies of heterozygotes would be informative.
3. It is known that selection increases the distance of linkage disequilibrium within a genomic region. If selection for lactase persistence occurred in Europeans then certain alleles quite far from the causative mutation would be in disequilibrium and show allelic association. If selection did not occur in

Yakut than linkage disequilibrium between the loci tested and the causative mutation may break down and allelic association would not be observed.

4. Two individuals showed unusually high levels of glucose after fasting overnight (appendix 3). Although none was diagnosed as diabetic (Kozlov, A., personal communication) it is possible that the individuals did not understand instructions about overnight fasting or there is another metabolic trait which clouds interpretation of the lactose tolerance tests.

Finnish

C-958T	CC	CT	TT
Persistent	7	2	0
Non-persistent	5	6	0

p=0.19

A-875G	AA	AG	GG
Persistent	6	3	0
Non-persistent	4	6	1

p=0.15

A_8-552/-559A_9	A_9 A_9	A_9A_8	A_8 A_8
Persistent	4	5	0
Non-persistent	1	6	4

p=0.03

G666A	GG	GA	AA
Persistent	7	2	0
Non-persistent	5	6	0

p=0.19

T5579C	CC	CT	TT
Persistent	5	4	0
Non-persistent	1	6	4

p=0.01

TG6236/7ΔΔ	TGTG	TG/ΔΔ	ΔΔ/ΔΔ
Persistent	7	2	0
Non-persistent	5	6	0

p=0.19

Yakut

C-958T	CC	CT	TT
Persistent	14	4	1
Non-persistent	16	2	0

p=0.15

A-875G	AA	AG	GG
Persistent	7	7	4
Non-persistent	7	11	0

p=0.23

A_8-552/-559A_9	A_9 A_9	A_9A_8	A_8 A_8
Persistent	3	11	4
Non-persistent	5	12	1

p=0.17

G666A	GG	GA	AA
Persistent	12	6	1
Non-persistent	12	6	0

p=0.43

T5579C	CC	CT	TT
Persistent	1	11	6
Non-persistent	1	13	2

p=0.26

TG6236/7ΔΔ	TGTG	TG/ΔΔ	ΔΔ/ΔΔ
Persistent	14	4	1
Non-persistent	16	2	0

p=0.15

Table 6.5 Two by three tables of aiielic association with lactase persistence
Fisher's exact test was performed on the derived two by two table.

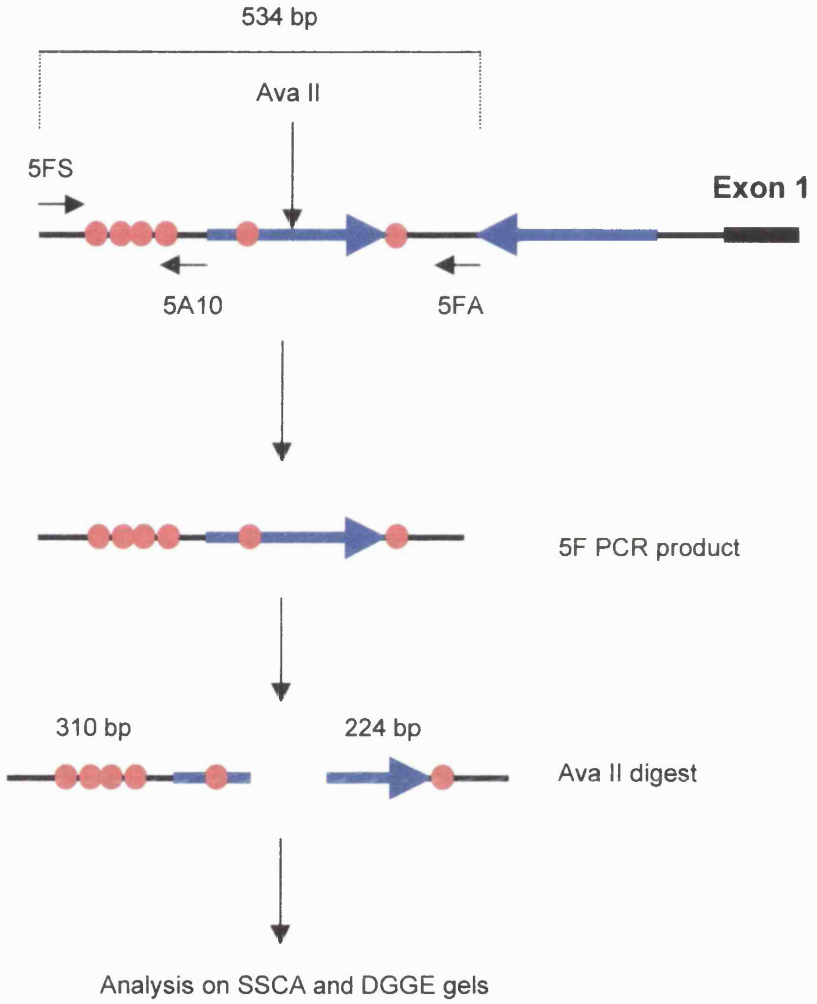


Figure 6.2 **The 5' flanking region of lactase and polymorphisms analysed within the region.**
Alu elements are shown in blue and polymorphic sites shown as red dots. The position of the primers used in amplification, the size of the PCR product and the restriction enzyme digest fragments are shown.

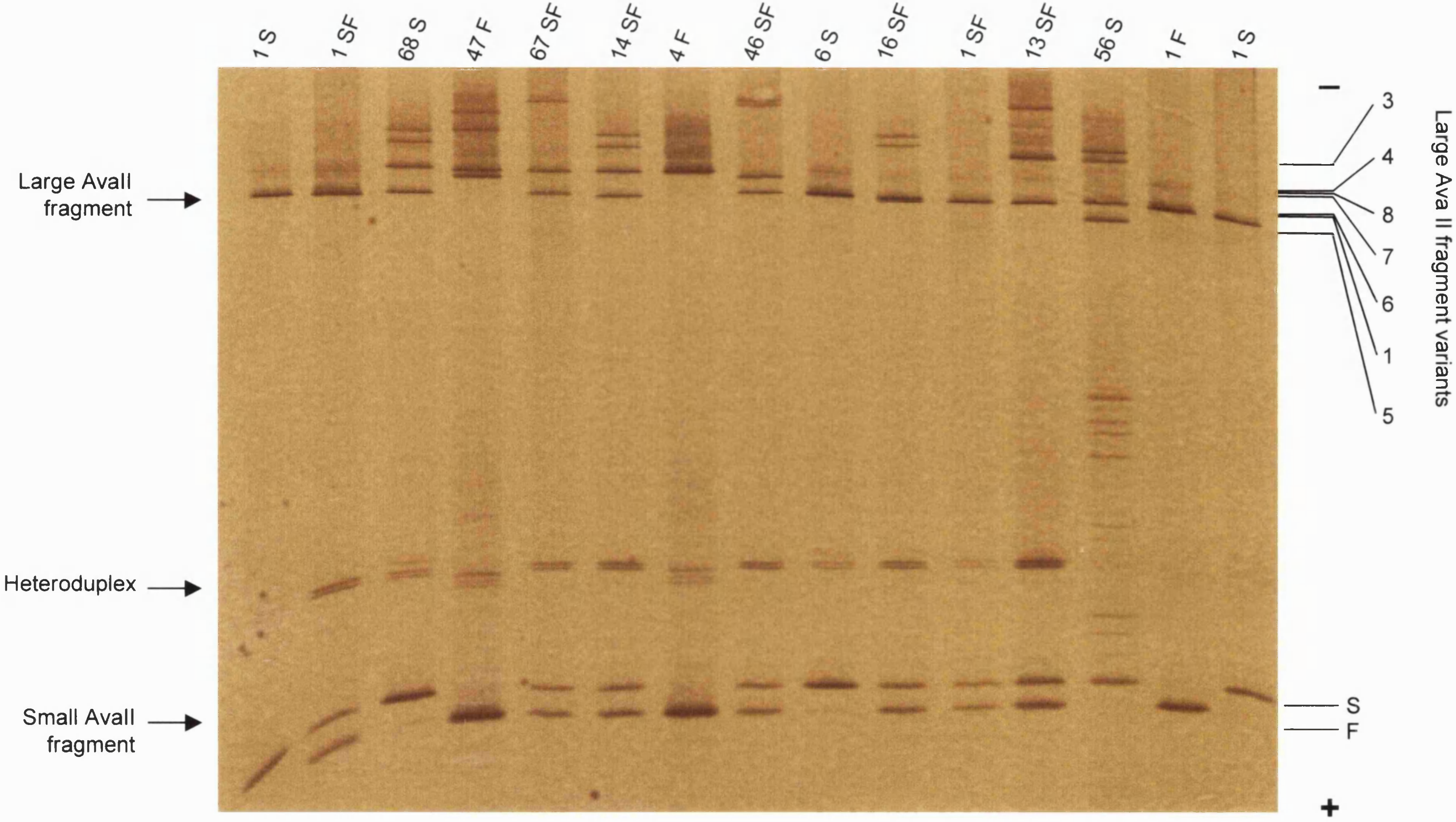


Figure 6.3 **A silver stained DGGE gel showing the variants detected.**
Large fragment variants are indicated, together with the common small fragment variants S and F. Heteroduplex bands are also shown.

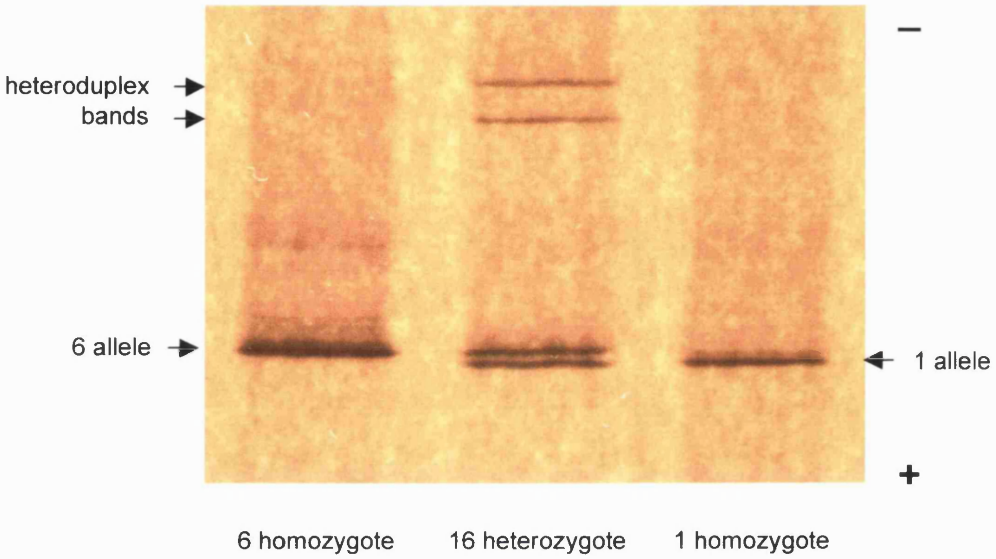


Figure 6.4 **Detail of a DGGE gel showing the resolution of the 1 and the 6 alleles**
Heteroduplex bands observed in heterozygous individuals are shown.

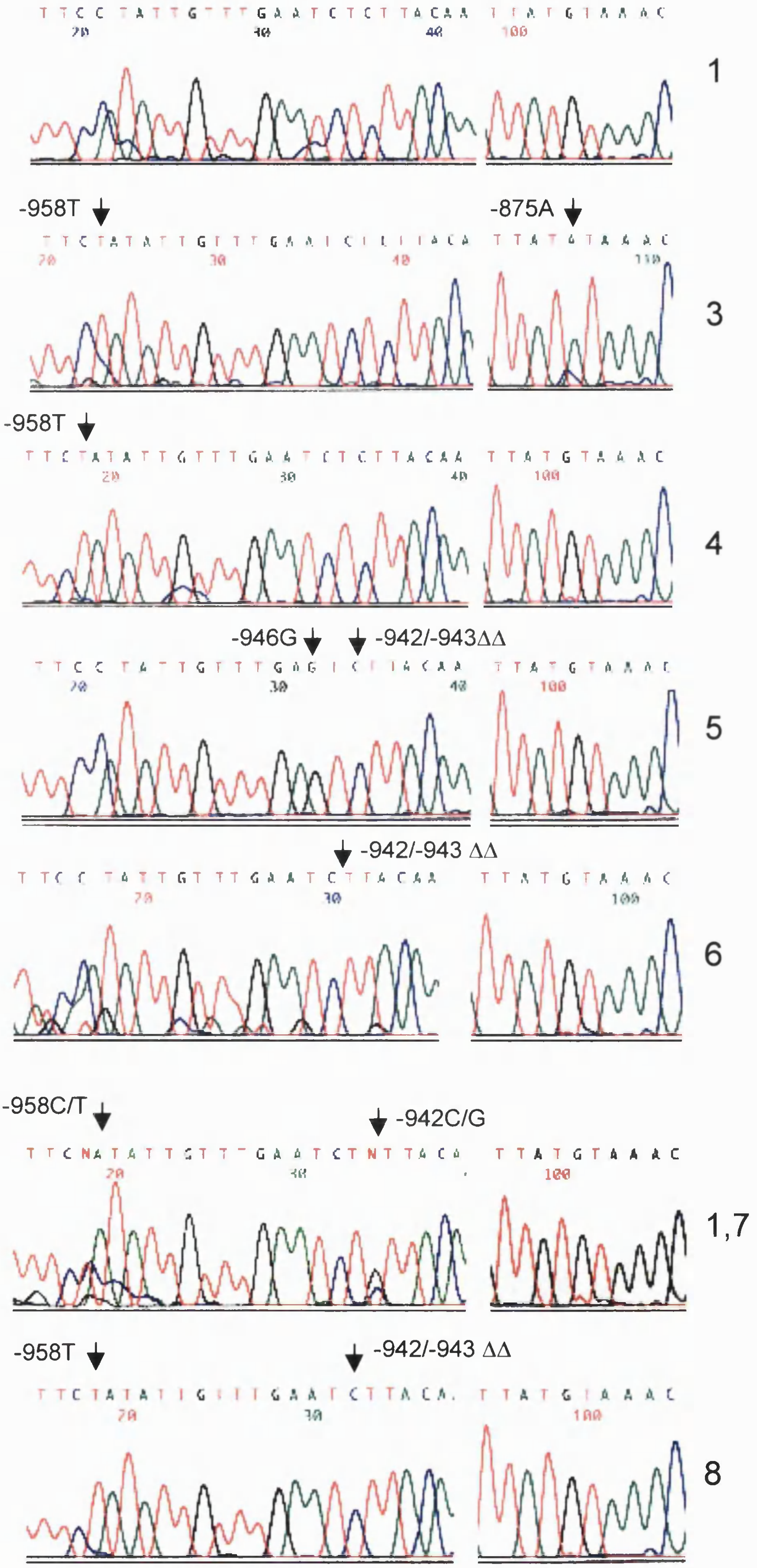


Figure 6.5
Sequence of DGGE variants
Sequences from the PCR product of a clone of each variant except 1,7, which is from genomic PCR product.

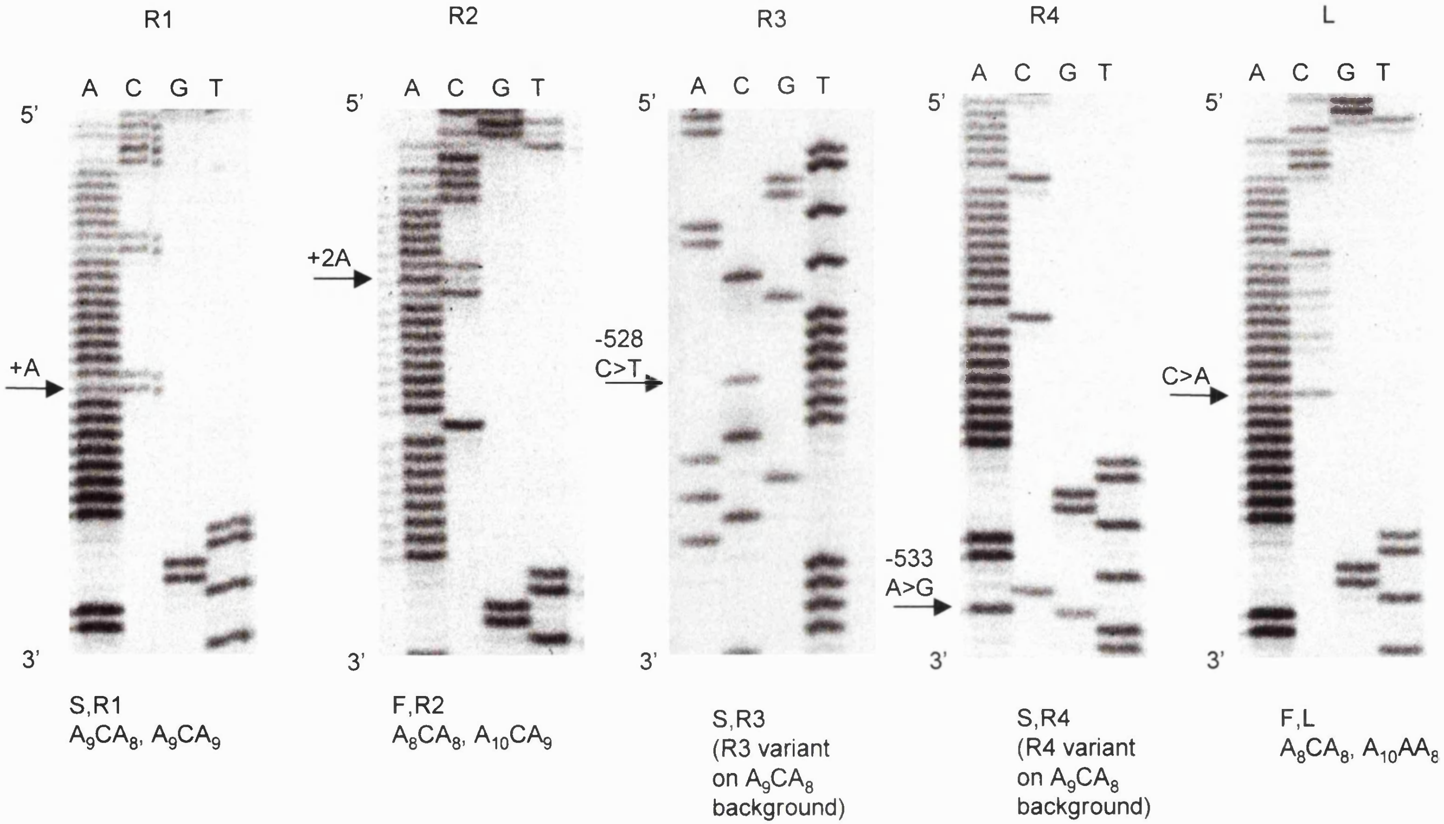


Figure 6.6 **Sequencing of 5F AvaII small fragment DGGE variants**
All samples shown are heterozygotes, with the genotype shown below the gel.

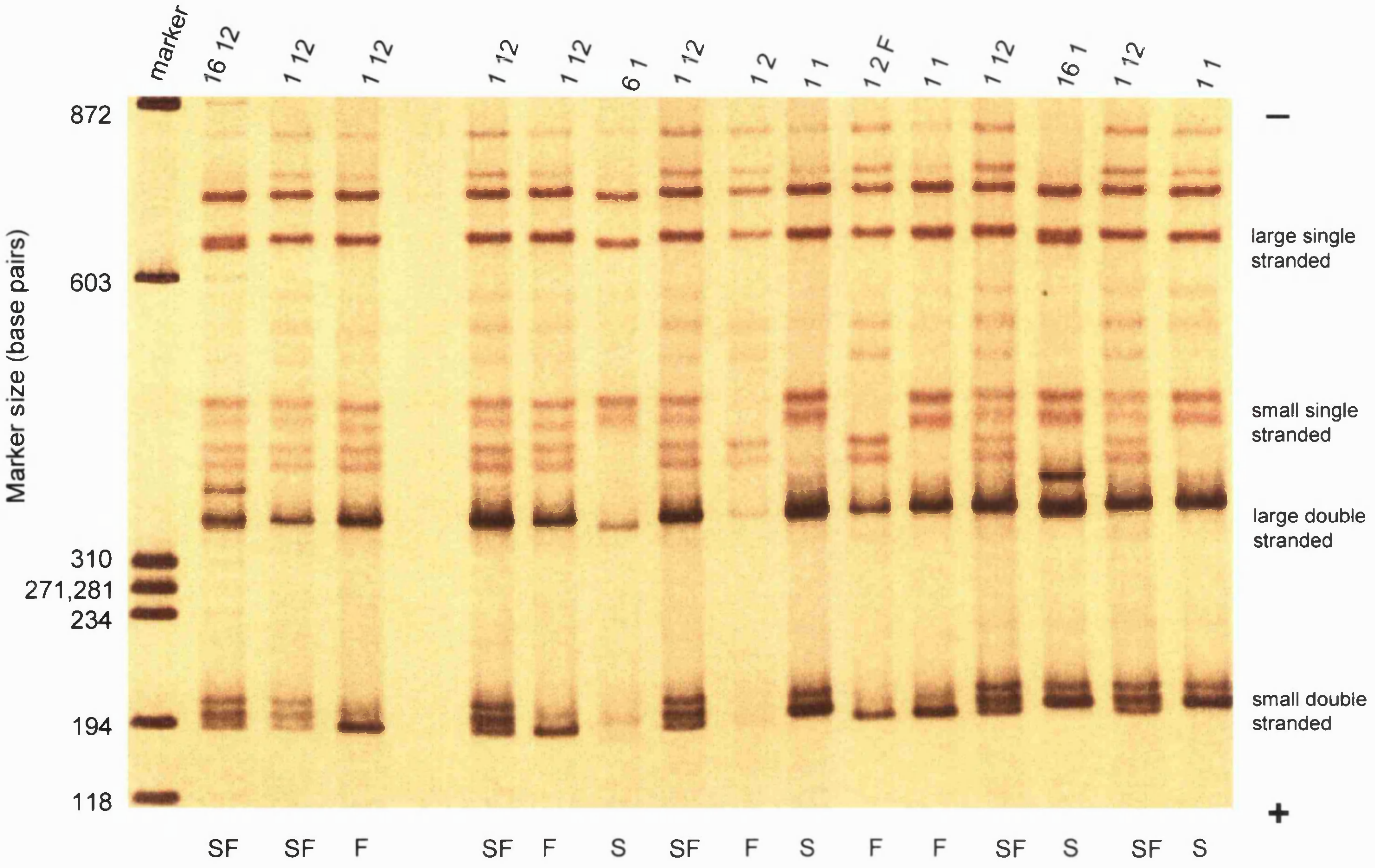


Figure 6.7
Silver stained SSCA gel of 5F AvaII digests.
Gel phenotypes are shown aboveand below the gel. The single stranded and double stranded fragments are shown, and can be distinguished by their colour. The 6 allele is detectable in the double starnded DNA by heteroduplex formation, and the SF polymorphisms is detectable in the small double stranded fragment.

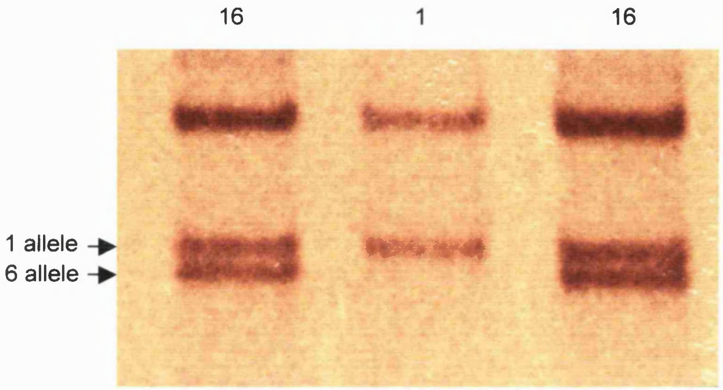


Figure 6.8 **Detail of silver stained SSCA gel showing resolution of the 1 and the 6 alleles**
Note that in SSCA gels the lower band is the 6 allele, whilst in DGGE gels the upper band is the 6 allele.

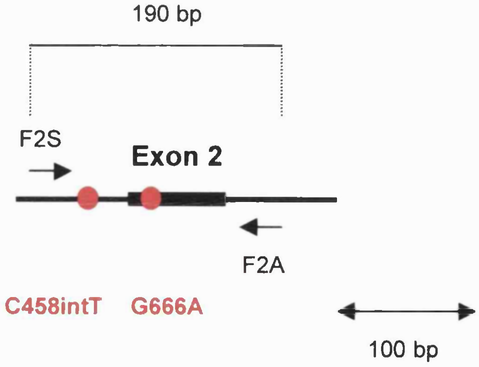


Figure 6.9 **Polymorphisms within and around exon 2.**
Polymorphic sites shown as red dots. The position of the primers used in amplification and the size of the PCR product are shown.

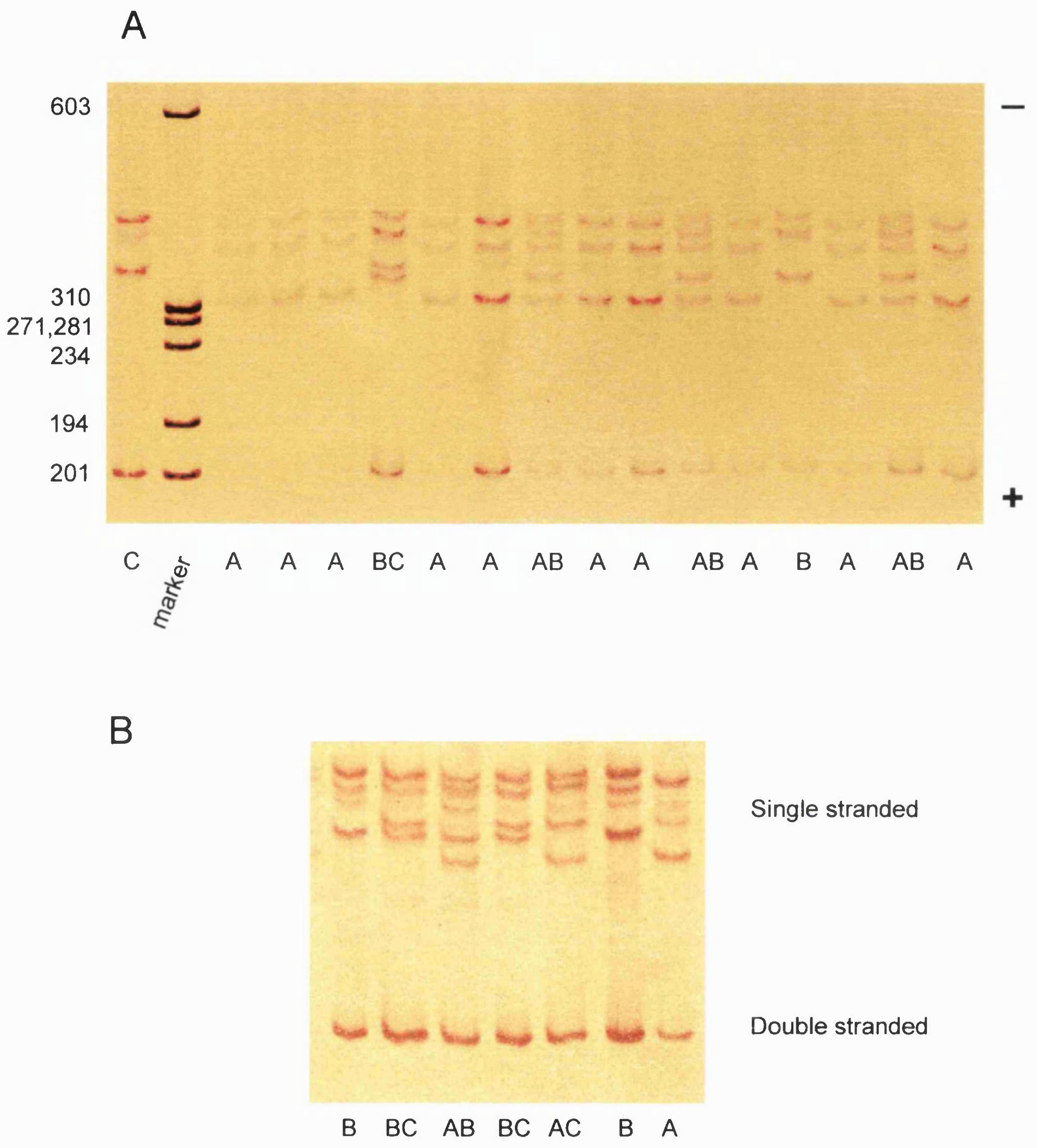


Figure 6.10 SSCA gels of the F2 PCR product
Gel phenotypes are shown under the gel.
A. A typical gel showing sensitive detection of weak PCR products.
B. A gel showing three types of heterozygotes in more detail.

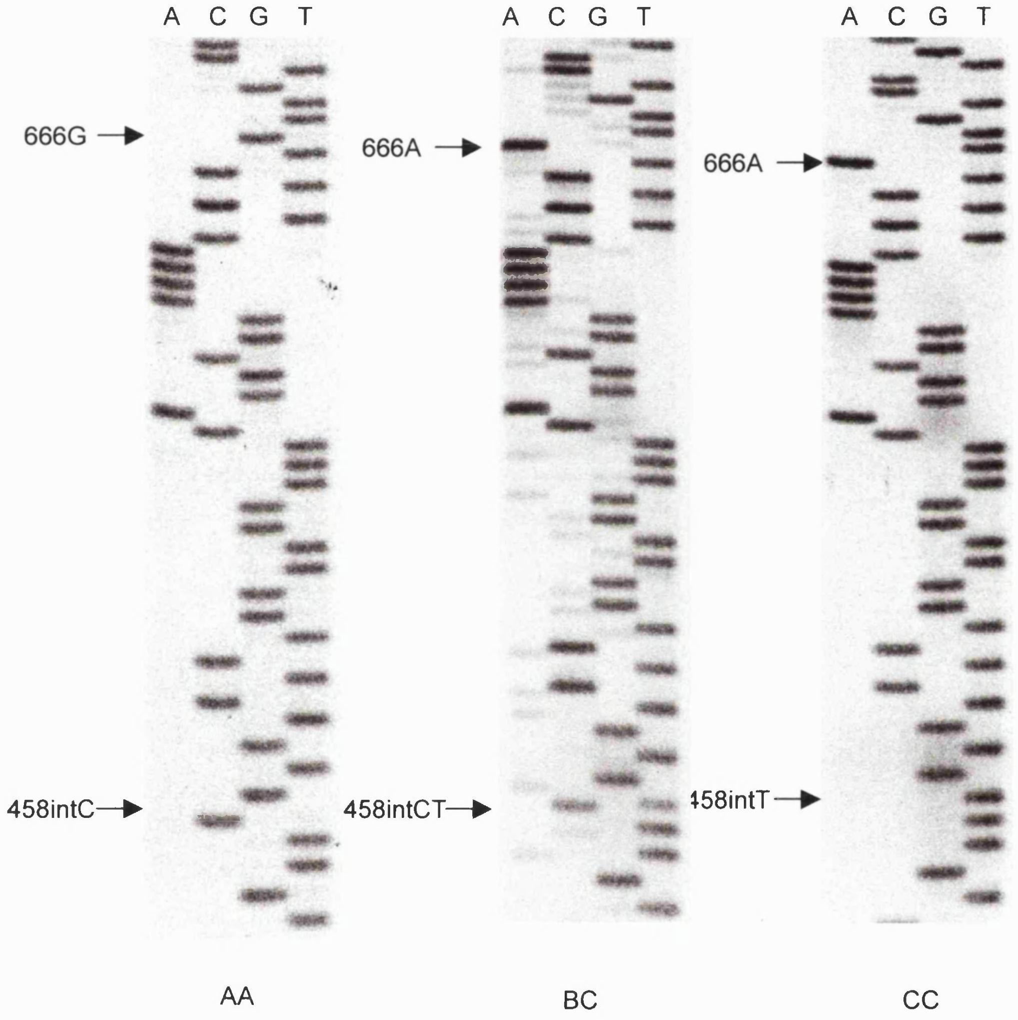


Figure 6.11 ^{33}**P Cycle sequencing of C SSCA variant in intron 1**
Sequencing using the F2S primer on F2S/F2A PCR product

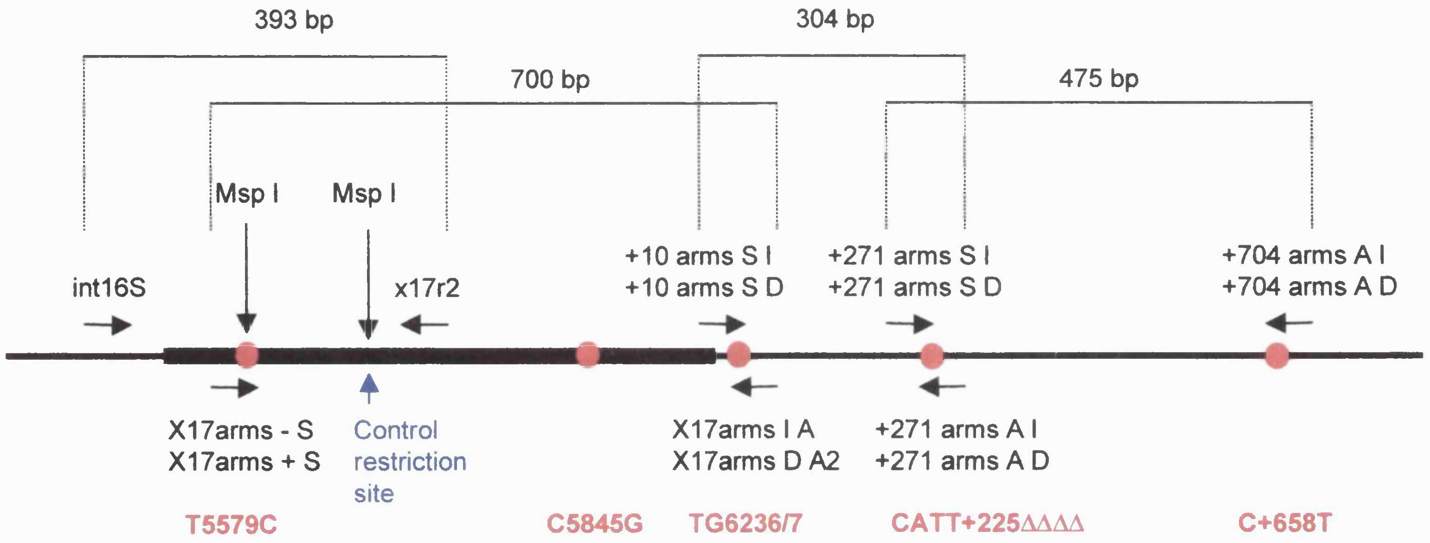


Figure 6.12 **Exon 17 and 3' flanking sequence of lactase**
Polymorphic sites are shown as red dots, together with the primers used for analysis. C5845Gwas not analysed on population or cohort samples. The sizes of the PCR products are also shown, together with the restriction enzyme sites used for analysis of T5579C.

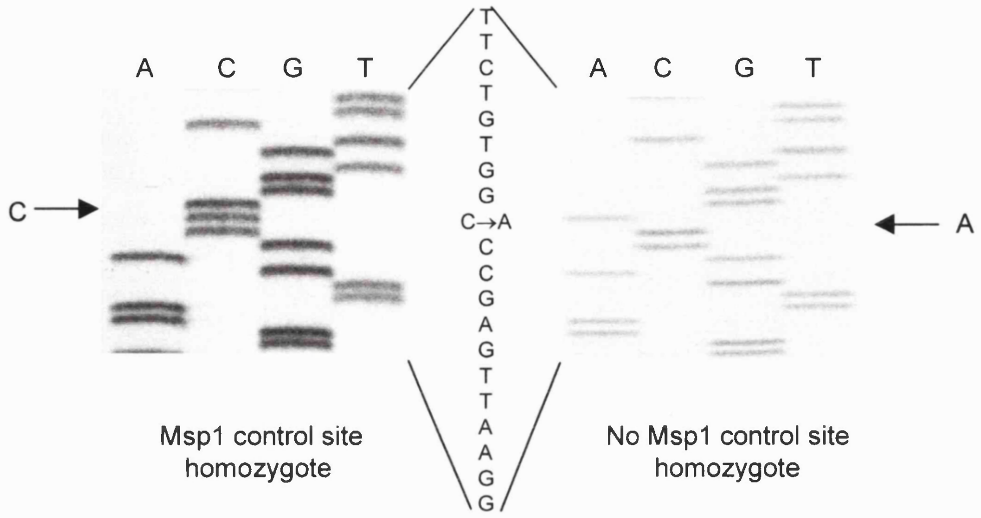


Figure 6.13 MspI restriction fragment length polymorphism
Sequencing gel showing base change abolishing MspI restriction enzyme site.

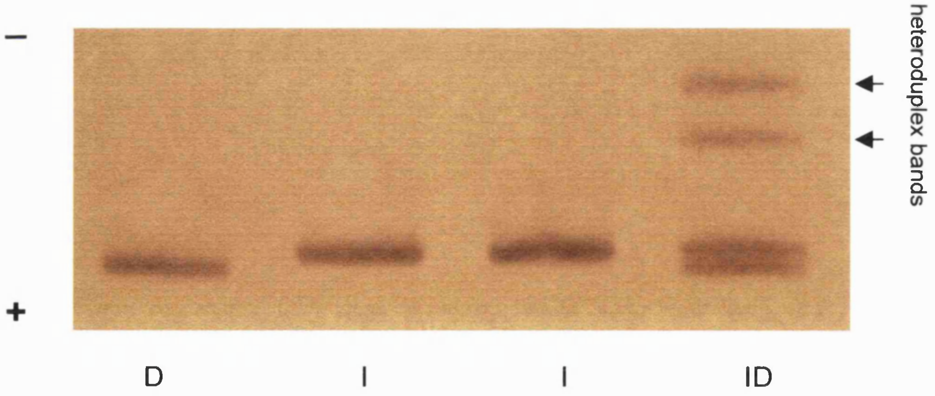


Figure 6.14 Silver stained gel analysis of TG6236/7ΔΔ polymorphism in UT PCR product
Gel phenotypes are shown below the gel, and heteroduplex bands produced in heterozygotes are shown.

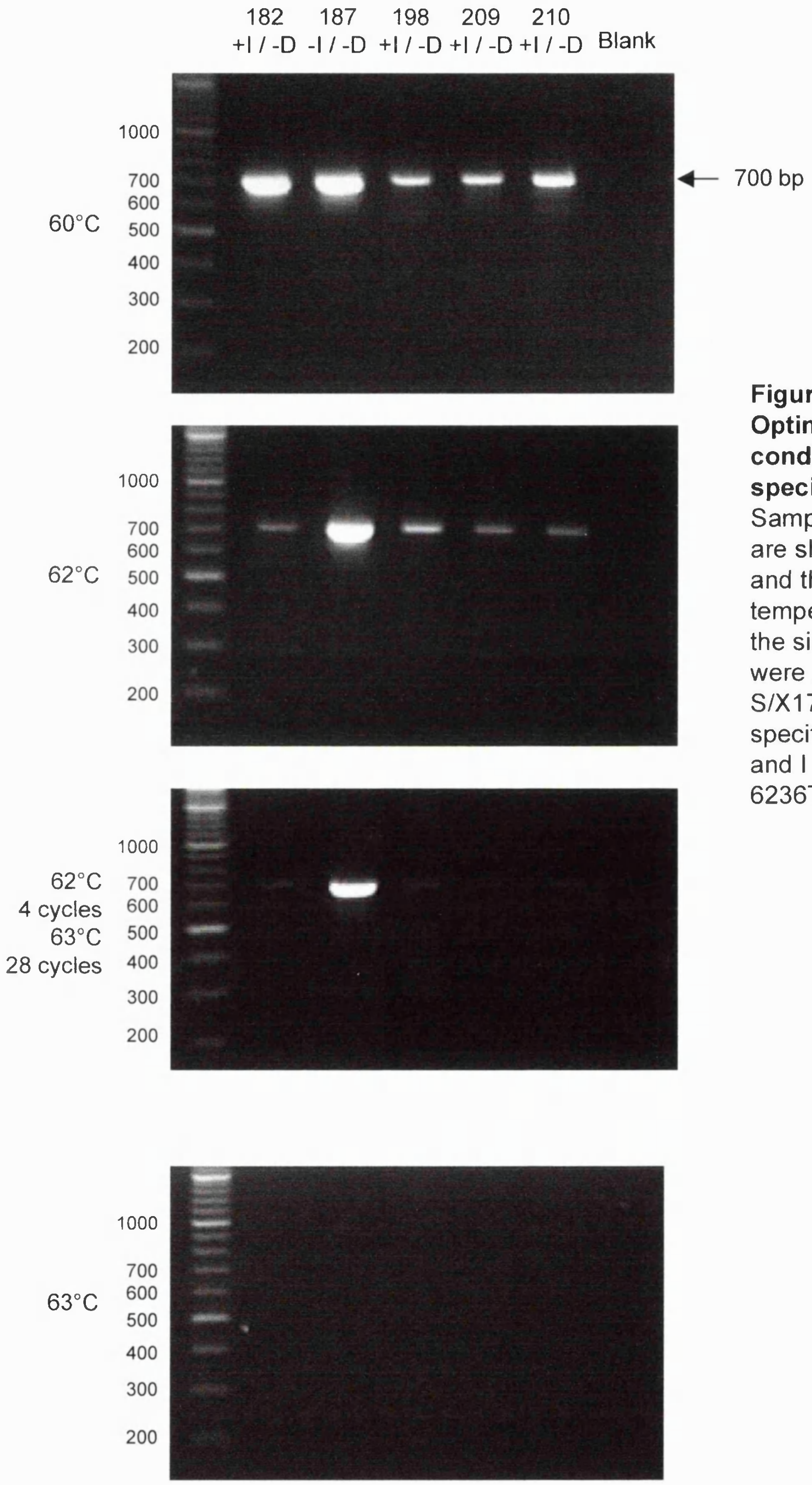
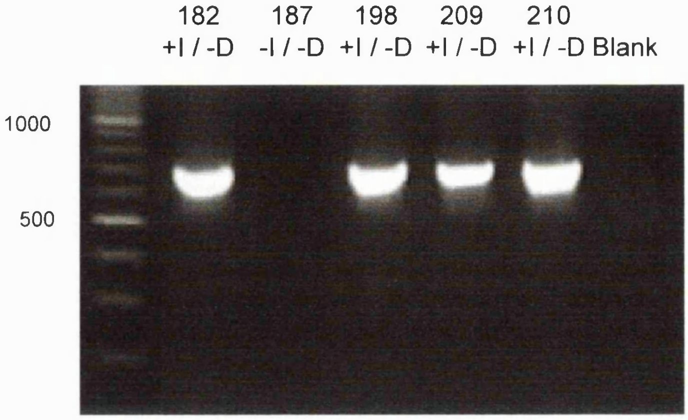
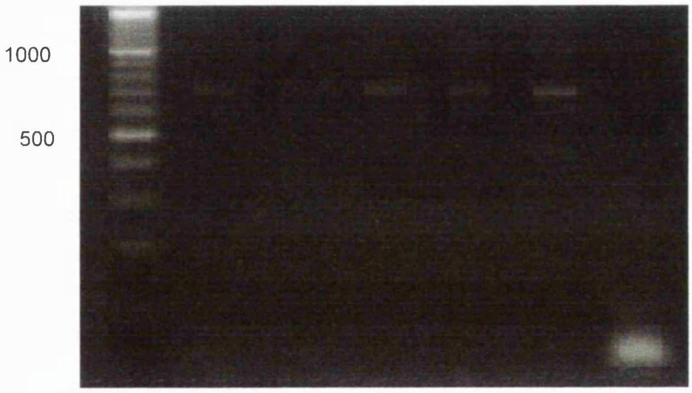


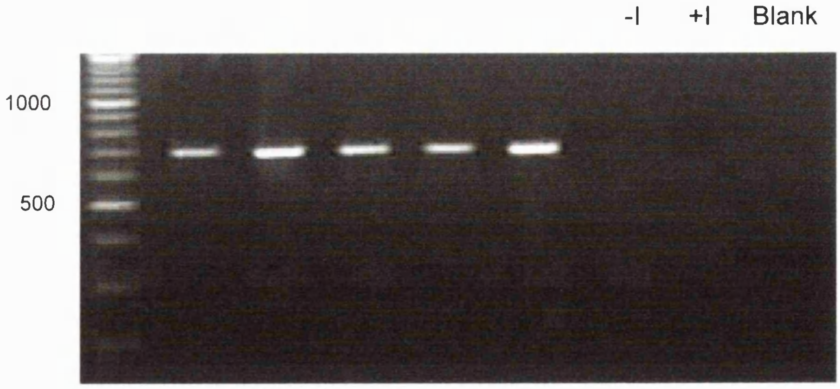
Figure 6.15 Optimisation of conditions for allele-specific PCR.
Samples and genotypes are shown above the gel, and the annealing temperature is shown at the side. Primers used were X17ARMS-S/X17ARMSIA to specifically amplify the - and I alleles (5579T and 6236TG) together.



X17ARMS+S/X17ARMSIA



X17ARMS+S/X17ARMSDA2



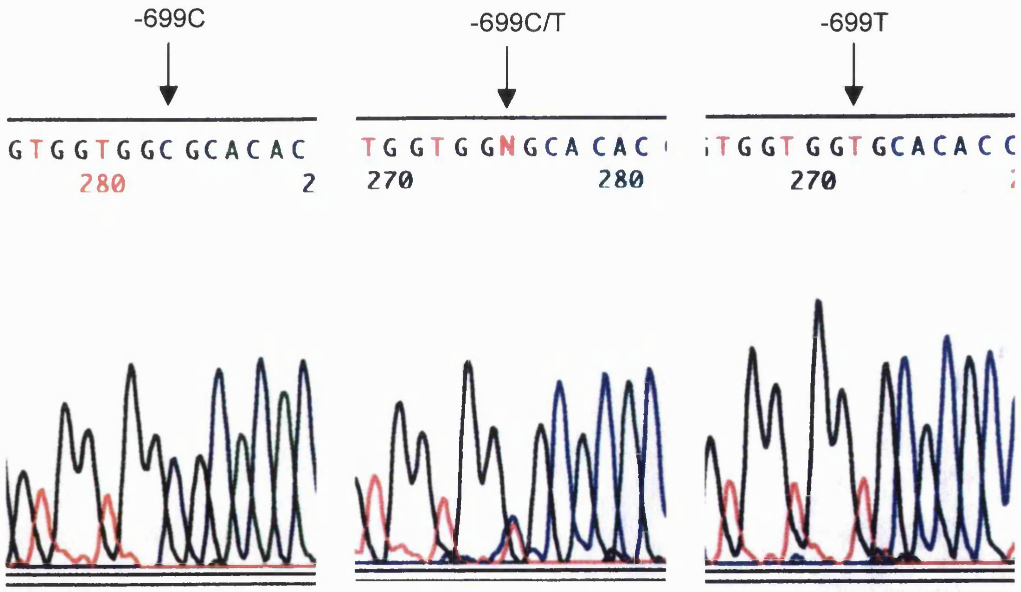
X17ARMS-S/X17ARMSDA2

Figure 6.16
Allele specific PCR products
Genotypes are shown above the gel, and primers used in each experiment are shown below the gel.

A

PCR product	Base change	Human	Tank	Harv	Colin	Carl	Kasey	Buttons	Amino acid change
5FS/5FA	-699	C	CT	C	CT	C	T	C	n/a
F2S/F2A	688	T	CT	T	T	T	T	-	Isoleucine>Threonine

B



chimpanzee Carl chimpanzee Tank chimpanzee Kasey

Figure 6.17 **A. Polymorphisms found in chimpanzee sequence**
The PCR product in which the polymorphism is detected is shown together with the base change number corresponds to the equivalent base in human sequence.
B Sequence analysis of one example of a chimpanzee polymorphism
The base change is shown above each sequence.

7 Linkage disequilibrium and haplotype analysis in populations

7.1 Patterns of linkage disequilibrium across the gene in different populations

7.1.1 Measuring linkage disequilibrium

Linkage disequilibrium (LD) is where the frequencies of haplotypes of two or more loci observed in a population are different from frequencies of haplotypes expected due to a random combination of the component alleles. LD can be due to recent admixture of two populations of different allele frequencies, or due to small physical distance between the loci preventing significant recombination between the loci.

Pairwise LD (LD between two polymorphic loci) can be measured in several ways. The simplest parameter is D, which is the difference between observed haplotype frequency and expected haplotype frequency assuming random assortment of alleles. D′ is the standardised LD coefficient and equals D/D_{max} where D_{max} is the maximum disequilibrium at the given allele frequencies (Ott, 1991). The absolute value of D′, which is a value between 1 (complete disequilibrium) and 0 (complete equilibrium) is used in this thesis (section 2.7.2). Other LD statistics include various correlation coefficients such as Δ^2 , Q and d, but, of these measurements, D′ is least sensitive to variation in allele frequency (Devlin and Risch, 1995), and so is ideal for a study on small populations, where there is a degree of error in allele frequency estimates.

7.1.2 Difference and similarities between populations

Table 7.1 shows the D′ values for combinations of all the polymorphic loci within and immediately upstream of LCT that have an allele frequency above 0.05 (except A-678G in Northern Europeans) together with the published information from the CEPH families (Harvey *et al.*, 1995). All of the populations were tested except the San, since a sample of such a small size (15 individuals) would not provide meaningful results. p values at 1 degree of freedom were derived from the χ-squared

CEPH Families

	-958				
-678	1	-678			
-552	1	1	-552		
666	1	1	1	666	
5579	0.99	0.79	0.98	0.91	5579
6236/7	0.97	1	0.94	0.95	0.91

S. European

	-958				
-678	1	-678			
-552	1	1	-552		
666	1	1	0.60	666	
5579	0.88	0.06	1	0.88	5579
6236/7	0.96	1	1	0.96	1

Bantu

	-942				
-678	1	-678			
-552	0.91	1	-552		
666	0.92	1	0.56	666	
5579	0.35	1	0.86	0.80	5579
6236/7	0.79	0.38	0.76	0.59	1

Russian

	-958				
-678	1	-678			
-552	1	1	-552		
666	1	1	0.93	666	
5579	0.88	0.23	0.73	0.90	5579
6236/7	0.86	1	0.85	0.92	1

Mordavian

	-958				
-678	0.99	-678			
-552	1	1	-552		
666	1	0.19	0.91	666	
5579	0.90	1	0.88	0.91	5579
6236/7	0.92	0.17	1	0.81	1

Roma

	-958				
-678	1	-678			
-552	0.97	1	-552		
666	0.97	1	1	666	
5579	0.96	0.71	0.97	1	5579
6236/7	0.94	1	1	0.97	1

Yakut

	-958				
-678	1	-678			
-552	1	1	-552		
666	0.87	1	0.11	666	
5579	1	0.89	0.86	1	5579
6236/7	0.81	0.59	1	0.83	1

Northern European

	-958				
-678	1	-678			
-552	1	1	-552		
666	1	1	1	666	
5579	1	0.51	1	1	5579
6236/7	1	1	1	1	1

Malay

	-958				
-678	1	-678			
-552	1	1	-552		
666	1	1	0.40	666	
5579	1	0.73	0.85	1	5579
6236/7	0.82	1	0.80	0.95	1

Papua New Guinean

	-958				
-678	1	-678			
-552	0.97	1	-552		
666	1	1	0.82	666	
5579	1	1	1	1	5579
6236/7	1	1	1	1	1

North Indian

	-958				
-678	1	-678			
-552	1	1	-552		
666	0.97	0.83	1	666	
5579	0.81	0.87	0.92	0.88	5579
6236/7	0.85	0.68	1	0.92	1

South Indian

	-958				
-678	1	-678			
-552	1	1	-552		
666	1	1	1	666	
5579	1	1	1	1	5579
6236/7	1	1	1	0.78	1

Chinese

	-958					
-942	1	-942				
-678	0.98	1	-678			
-552	1	1	1	-552		
666	1	1	1	0.40	666	
5579	1	1	0.80	0.86	1	5579
6236/7	0.73	0.39	0.46	0.87	1	1

Japanese

	-958					
-942	1	-942				
-678	1	1	-678			
-552	1	1	1	-552		
666	1	1	1	0.35	666	
5579	1	1	0.72	0.54	1	5579
6236/7	0.87	0.39	1	0.49	0.80	1

Table 7.1 **Pairwise linkage disequilibrium values for each population**
Pairwise linkage disequilibrium scores for each population tested, except San, are shown as the normalised disequilibrium statistic D'. All pairs of loci show significant departure from D'=0 (p≤0.05) except where shaded.

values calculated by ASSOCIATE to test for the significance of linkage disequilibrium. All values of D′ were tested and showed significant linkage disequilibrium ($D' \neq 0$, $p \leq 0.05$) except those values shaded.

These analyses revealed several interesting features in the different populations. All populations show a generally high level of linkage disequilibrium ($D' \approx 1$) across the gene, with disequilibrium highest in CEPH families, Northern Europeans, Southern Indians and Papua New Guineans. Bantu-speaking South Africans showed a lower level of disequilibrium between most of the loci, and several other interesting differences in D′ are revealed.

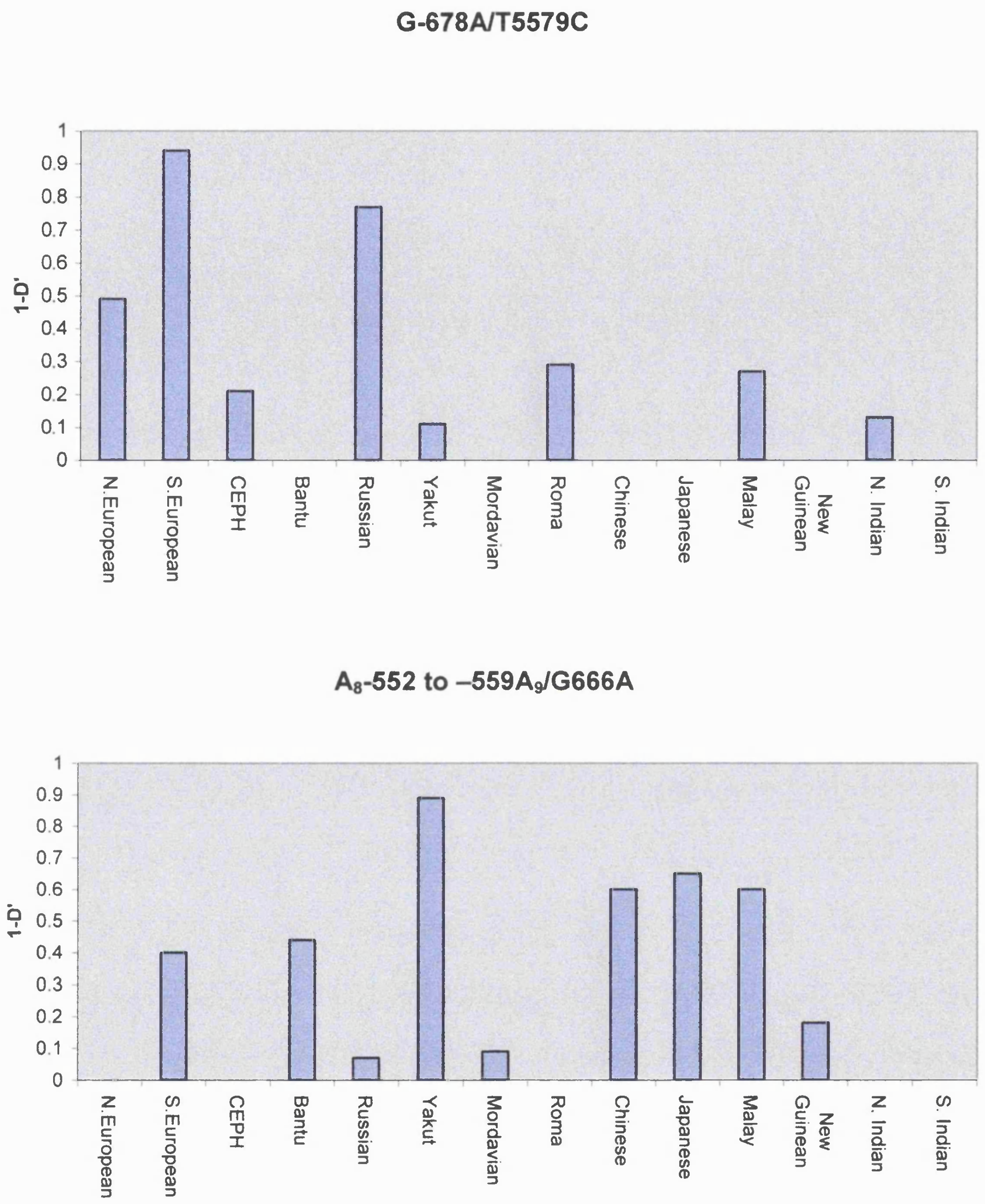
Figure 7.1 shows, in the form of a bar chart, the amount of linkage equilibrium (1-D′) between A-678G and T5579C, and between A_8-552/-559A_9 and G666A in the different populations. The values for $1\text{-}D'_{A\text{-}678G/T5579C}$ are highest in the Northern Europeans, Southern Europeans and Russians, and are lowest in South Indians, New Guineans, Mordavians and Bantu-speaking South African. Bantu-speaking South Africans would be expected to show high levels of linkage equilibrium because of their high haplotype diversity, although these two pairs of loci show that this is not always the case. There is a different pattern of $1\text{-}D'_{A8\text{-}552 \text{ to } \text{-}559A9/G666A}$ values across the populations. In this case, there is a high amount of linkage equilibrium in Yakut, followed by the Chinese, Japanese and Malay, each with similar values. The lowest values are Northern Europeans, CEPH population, Roma, and Indians, all of whom show complete disequilibrium between these two loci.

The different patterns of pairwise LD value, even between loci spanning the same genomic region indicate that each polymorphism has a different age and so has a different history. Analysis of differences in the frequencies of the haplotypes on which certain alleles occur helps to clarify some of the differences in LD values.

7.2 Patterns of haplotype frequencies

7.2.1 Review of Family data

The CEPH families allowed direct determination of haplotypes A to I, and estimation of the haplotype frequencies of unrelated individuals from those families. They are mostly from Utah Mormons with some from France, and although differences in haplotype frequencies were found between the two groups (Harvey, 1994), they were considered further as one group in that thesis.


G-678A/T5579C

A_8-552 to –559A_9/G666A

Figure 7.1 **Linkage equilibrium (1-D') between two pairs of loci in the populations tested.**

Pairwise linkage disequilibrium analysis of all CEPH families showed that LD was high (0.91 to 1) between all pairs of loci except between G-678A and T5579C where the D′ value fell to 0.79.

7.2.2 Methods of assigning haplotypes

When family data is not available, haplotypes of individuals who are not homozygous or heterozygous for more than one site must be deduced using population information. Haplotypes were deduced by two methods: haplotype counting by observation and maximum-likelihood analysis. Haplotype counting firstly involved noting the haplotypes of totally homozygous individuals and those heterozygous for only one site. In individuals heterozygous for more than one site, all possible haplotypes were considered and the two haplotypes which occur at the highest frequencies in that population were assumed to be present. Because of the high amount of LD within the populations, there are a relatively small number of haplotypes and so estimation of haplotype frequencies is reasonably simple.

The maximum-likelihood program EH uses essentially the same technique and assumes all loci are in Hardy-Weinberg equilibrium. Table 7.2 shows that the frequency estimations produced by haplotype counting (HC columns) and maximum-likelihood analysis (ML columns) agreed extremely well. The only exceptions are the Bantu-speaking South Africans and the San, where low LD and high haplotype diversity made haplotype counting difficult. Because of this, in further analyses of these two populations the maximum-likelihood haplotype frequency estimates are used. In two cases (Southern Europeans and Yakut) the maximum-likelihood method identified two suggested haplotypes that had not been found by haplotype counting in any population nor by the maximum-likelihood method in other populations. In this instance, where general LD is high, the haplotype counting method is probably more accurate because in ambiguous situations there is knowledge of the haplotype frequencies in neighbouring populations, in contrast to the maximum-likelihood method, which was used on a population by population basis.

In both methods, haplotype estimation of the eleven loci was facilitated by the polymorphism typing techniques that revealed information about allelic phase. This allowed several diallelic loci to be replaced in haplotype estimation procedures by a

	N.Europe		S.Europe		N.Indian		S. Indian		Malay		Chinese		Japanese		N.Guinean		Roma		Mordavian		Yakut		Russian		Bantu		San	
	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML
A	0.86	0.86	0.38	0.36	0.44	0.44	0.43	0.43	0.50	0.49	0.48	0.47	0.37	0.37	0.29	0.29	0.48	0.49	0.53	0.56	0.42	0.42	0.44	0.44	0.10	0.10	0.07	0.06
B	0.06	0.06	0.32	0.32	0.22	0.22	0.18	0.19	0.14	0.13	0.11	0.10	0.10	0.10	0.36	0.36	0.30	0.30	0.22	0.20	0.10	0.09	0.21	0.21	0	0	0.03	0.03
C	0.03	0.03	0.12	0.12	0.23	0.23	0.30	0.31	0.19	0.18	0.10	0.09	0.17	0.15	0.18	0.18	0.10	0.09	0.07	0.07	0.35	0.36	0.14	0.15	0.36	0.31	0.03	0.03
D	0.03	0.03	0.06	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0.02	0.02	0	0	0.01	0.01	0	0	0	0
E	0.02	0.02	0.07	0.06	0.01	0.01	0.01	0	0.01	0.02	0.02	0.02	0.02	0.03	0	0	0.03	0.02	0.04	0.01	0.01	0	0.05	0.05	0.01	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0.01	0.04	0.04	0.04	0.01	0.01	0.02	0.05	0.06	0	0	0.06	0.06	0.02	0.02	0.02	0.03	0	0	0.04	0.04	0	0	0	0
H	0	0	0	0	0	0.01	0	0	0	0	0	0	0.01	0.01	0	0	0	0	0	0	0	0	0.01	0.01	0	0	0.03	0.04
I	0	0	0.02	0.02	0.01	0.01	0	0.01	0	0	0	0	0.01	0.02	0	0	0	0	0.02	0.02	0	0	0.01	0.01	0	0	0	0
J	0	0	0.02	0.03	0	0	0	0	0.01	0.01	0.01	0.01	0	0.01	0	0	0.02	0.02	0	0	0	0	0	0	0.01	0	0	0
K	0	0	0	0	0	0	0.01	0.01	0.01	0	0	0	0	0	0	0	0.02	0.01	0	0	0	0	0.03	0.03	0.03	0.08	0	0
L	0	0	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0.01	0.01	0.02	0.01	0	0	0	0.01	0	0.01	0	0	0	0	0.01	0.01	0	0.02	0.01	0	0.01	0	0.04	0.05	0	0
N	0	0	0	0	0	0	0	0	0.01	0	0.02	0.02	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0.06	0.14	0.20
P	0	0	0	0	0.01	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.10	0.04	0.10	0.09
Q	0	0	0	0	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.07	0.07	0.01	0.01	0.02	0.02	0.01	0.01	0.03	0.03	0.06	0.04	0.03	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0.02	0.02	0.01	0	0.04	0.04	0	0	0	0	0	0	0	0	0	0	0.04	0.02	0.10	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0.03	0.03
U	0	0	0	0	0	0	0	0	0.08	0.08	0.18	0.17	0.24	0.24	0.04	0.04	0	0	0.04	0.02	0.06	0.05	0.03	0.01	0.08	0.07	0.17	0.17
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0.03	0.03
W	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.02	0	0	0	0	0	0	0.03	0.03	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0	0.06	0.07	0.10	0

	N.Europe		S.Europe		N.Indian		S. Indian		Malay		Chinese		Japanese		N.Guinean		Roma		Mordavian		Yakut		Russian		Bantu		San	
	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML	HC	ML
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.10	0.14
Z	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	0	0	0	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0.04
e	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.02	0	0	0	0	0	0	0	0
j	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0	0
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.10
others	-		-	0.03	-		-		-	0.02	-		-		-		-	0.02	-	0.03	-	0.03	-		-		-	
N	104		106		156		84		200		104		82		72		170		68		78		80		72		30	
H1	0.25		0.74		0.70		0.69		0.70		0.71		0.75		0.72		0.67		0.66		0.69		0.74		0.84		0.88	
H2	0.65		0.97		0.97		0.96		0.96		0.96		0.96		0.97		0.97		0.96		0.94		0.97		0.96		0.95	
H1/H2	0.38		0.76		0.72		0.72		0.73		0.74		0.78		0.74		0.69		0.69		0.73		0.76		0.88		0.92	

Table 7.2 **Haplotype frequencies in different populations.** HC indicates frequency estimated by haplotype counting, ML indicates frequency estimated by the EH program. H1 is the heterozygosity, assuming each haplotype is an allele at a multiallelic locus. H2 is the heterozygosity of all eleven sites in the haplotype assuming no association, and H1/H2 is a measure of association or LD across the haplotype, with 1 representing complete linkage equilibrium. *N* is number of chromosomes tested.

single multiallelic locus, which increased the power of both haplotype estimation methods. The DGGE method detects a variant that is a combination of five loci (see section 6.1.2), and similarly the SSCA technique detects a variant which is a combination of alleles at G666A and C458intT. The allele-specific PCR technique reveals allelic phase between C5579T and TG6236/7ΔΔ, so these two loci can be treated as one locus with four alleles. The allelic data was therefore treated as five segregating sites in the maximum likelihood analysis except for the Northern European, Southern European and Yakut populations. In these population samples, allele-specific PCR was not performed on C5579T and TG6236/7ΔΔ, and the data was therefore treated as six segregating sites. Figure 7.2 shows the allelic state of all the haplotypes deduced from the population data.

7.2.3 Haplotype frequencies in different populations

Table 7.2 shows estimates of haplotype frequencies in different populations. It also shows the haplotype diversity H1 compared to H2 (from table 6.2, probability of at least one heterozygous site in an individual, assuming random association of alleles). Since haplotype diversity H1 is in effect the probability of at least one heterozygous site (assuming the non-random association observed in the populations), H1/H2 may be regarded as an indicator of the general degree of linkage disequilibrium across all the loci. By this measure, Northern Europeans show the highest amount of LD and sub-Saharan Africans the lowest.

It is clear that four haplotypes account for most of the global diversity: A, B, C and U. The geographical distribution of these common haplotypes is shown by the map in figure 7.3, together with the lactase persistence allele frequency data for each population from other studies (Swallow and Hollox, 2000, Kozlov, A. and Markova, S., personal communication).

7.2.3.1 Distribution of haplotype A

Previous analysis of the haplotype frequencies in Northern and Southern Europeans had suggested that there was a cline, with a high frequency of haplotype A in Northern Europeans and a lower frequency in Southern Europeans (Harvey *et al.*, 1998). The data presented here extends the analysis and shows that the A haplotype is at a fairly constant frequency in all non-African populations except for

Northern Europeans. The haplotype is observed in both the San and Bantu-speaking South Africans, but at low frequencies (0.06 and 0.01 respectively).

7.2.3.2 Distribution of haplotype U

The U haplotype is absent from Northern Europeans, Southern Europeans, Roma, North Indians and South Indians. It is present in both Bantu-speaking South Africans and San, and in East Asian populations where it shows a cline in frequency from Malaysia and Papua New Guinea through China rising to its highest frequency in Japan. This distribution matches the distribution of its component allele, -942/-943ΔΔ.

7.2.3.3 Distribution of B and C haplotypes

Both B and C haplotypes are present in all populations, except Bantu-speaking South Africans, where no B haplotypes were observed. With the exception of Northern Europeans, where they are rare, both haplotypes vary in frequency between approximately 0.1 and 0.3. Papua New Guineans have the highest frequency of haplotype B (0.36), and Yakut have the highest frequency of haplotype C (0.35).

7.2.3.4 Distribution of haplotype D

The D haplotype distribution reflects the distribution of its component allele -875G. This haplotype (and component allele) is observed only in populations living in Europe (Northern and Southern Europeans, Roma, Russians and Mordavians), and indicates some kind of shared ancestry between these populations. Interestingly, Mordavians are not Indo-European speakers but Finno-Ugric speakers, so it may be that they have less common ancestry with other Europeans and the presence of the D haplotype may a result of recent admixture.

Found in Harvey *et al.* 1995/1998

	C -958 T	A -946 G	C -942 G	TC -942/3 ΔΔ	G -875 A	A -678 G	A$_8$-552/-559 A$_9$	C458intT	G 666 A	T 5579 C	TG 6236/7 ΔΔ
A	○	○	○	○	○	○	●	○	○	●	○
B	●	○	○	○	○	○	○	○	●	○	●
C	○	○	○	○	○	●	○	○	○	○	○
D	●	○	○	○	●	○	○	○	●	○	●
E	○	○	○	○	○	●	○	○	○	●	○
F	●	○	○	○	○	○	○	○	●	●	●
G	●	○	○	○	○	○	○	○	●	○	○
H	○	○	○	○	○	○	●	○	○	○	○
I	○	○	○	○	○	○	●	○	○	○	●
J	○	○	○	○	○	○	○	○	○	●	○
K	○	○	○	○	○	○	○	○	○	○	○
L	●	○	○	○	○	○	○	○	○	○	○
M	○	○	○	○	○	●	○	○	○	○	○
N	●	○	●	○	○	○	○	○	●	○	●
O	○	○	○	●	○	○	●	●	●	○	○
P	○	○	○	○	○	○	○	○	●	○	●
Q	○	○	○	○	○	○	●	○	○	○	○
R	●	○	○	●	○	○	●	○	●	○	●
S	○	○	○	●	○	○	●	○	●	○	●
T	○	○	○	●	○	○	●	○	○	○	○
U	○	○	○	●	○	○	●	○	●	○	○
V	○	●	○	●	○	○	●	○	●	●	○
W	○	○	○	○	○	○	●	○	●	○	○
X	○	○	○	●	○	○	●	○	●	●	○
Y	○	○	○	○	○	○	○	○	●	○	○
Z	○	○	○	○	○	○	○	○	○	○	●
a	●	○	○	○	○	○	○	○	○	●	○
b	○	○	○	●	○	○	●	○	○	○	●
c	○	○	○	○	○	●	○	○	●	○	●
d	○	○	○	○	○	○	○	○	●	●	○
e	○	○	●	○	○	○	○	○	●	○	○
f	○	○	○	●	○	○	●	●	●	○	●
g	○	○	○	●	○	○	○	○	●	○	○
h	○	○	○	○	○	●	○	○	●	●	○
i	●	○	○	○	○	○	○	○	○	○	●
j	○	○	○	○	○	●	●	○	○	○	○
k	○	○	○	○	○	●	●	○	○	○	●
l	○	○	○	○	○	●	●	○	●	○	●
m	○	○	○	●	○	○	○	○	●	●	○
n	○	○	○	●	○	○	●	●	●	●	○

Figure 7.2 Haplotypes observed in all individuals. Haplotypes are named in order of discovery. The empty circles represent the ancestral allele, and the filled circles represent the derived alleles.

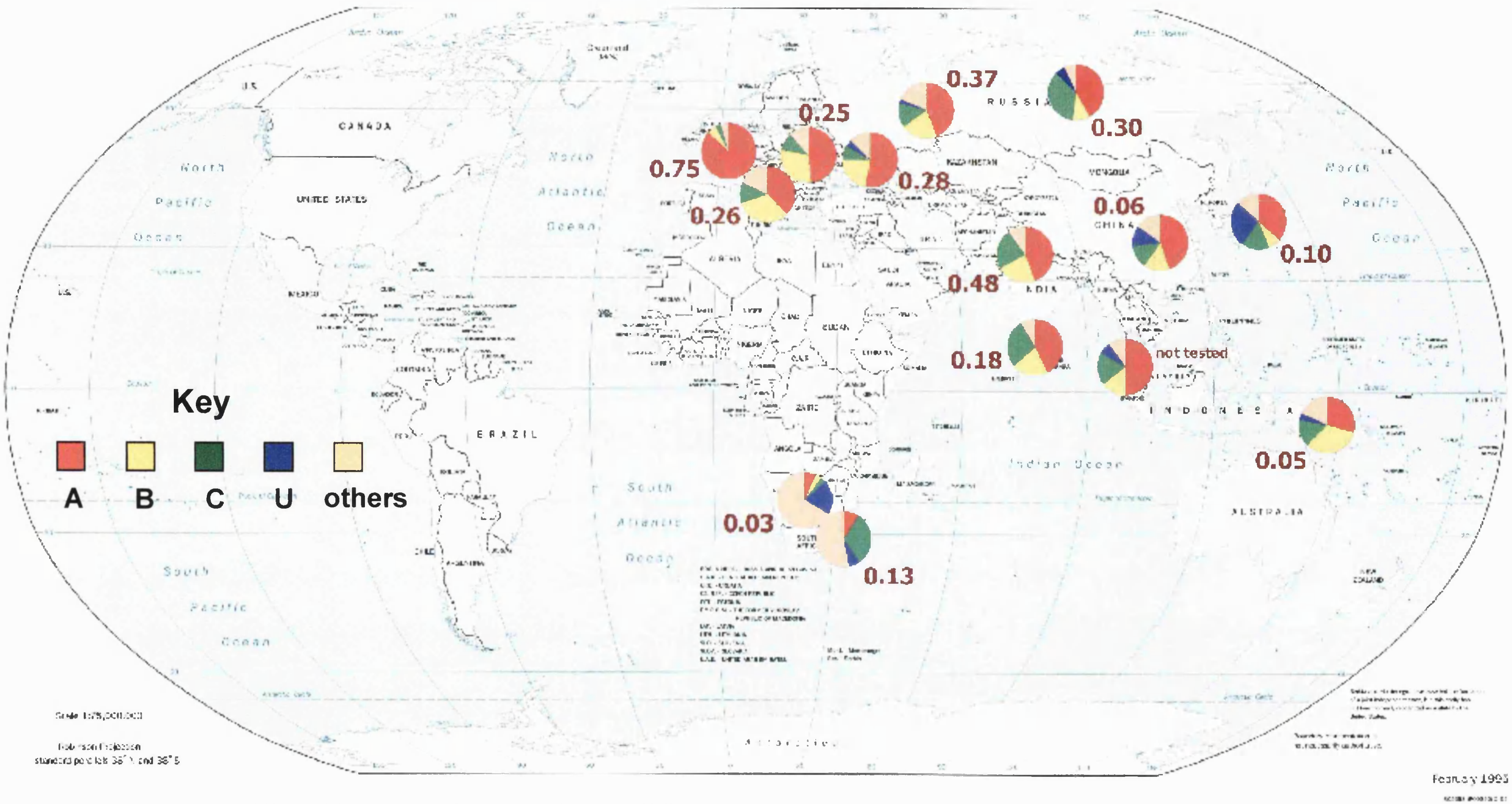


Figure 7.3 **Haplotype frequencies in different populations**
Frequencies shown are observed estimates, except those for the San and Bantu-speaking South Africans, where the maximum-likelihood estimates are shown. The frequency of the lactase persistence allele 1in each population, estimated by other studies, is shown next to each pie chart.

7.2.4 Nature of common haplotypes

The four common haplotypes are not closely related (figure 7.2) and differ by a minimum of three sites from each other (figure 7.4). Some of the rarer haplotypes, or 'others' in figure 7.3, are potential intermediate steps between the four common haplotypes.

7.3 Further analysis of haplotypes

7.3.1 Relating haplotypes to one another

Although linkage disequilibrium is observed across the whole haplotype, it is clear by examining the list of haplotypes shown in figure 7.2 that recombination has played a part in generating diversity. Because of this, a cladistic approach to reconstructing the evolution of the haplotypes would not work: recombination shuffles the allelic background of older mutations so that its ancestral chromosome (the chromosome of which the mutation occurred) cannot be determined with any certainty. However, there are some exceptions where a haplotype has been generated by a recent mutation on a common haplotype background. For example, the D haplotype is the B haplotype with an additional -875A change. The haplotypes that can be related in this way are shown in figure 7.5.

7.3.2 Analysis of simple recombinants

Examination of the rarer haplotypes suggested that some arose by recombination of two common haplotypes. Further information on the identity of the 3′ end of the gene was needed to attempt to distinguish simple recombinants of the four common haplotypes. Two polymorphic sites 3′ to exon 17 were analysed by allele specific PCR (section 2.4.6) in a small subset of samples containing both rare and common haplotypes from different populations (shown in table 7.3). This allows one to judge whether both the 3′ and the 5′ ends of a rare haplotype were derived from two different common haplotypes, which suggests that the rare haplotype was generated by a recombination between two common haplotypes. Of the two sites 3′ to exon 17 tested, CATT+225ΔΔΔΔ was in complete disequilibrium with the haplotype in 123 out of 124 chromosomes carrying the four common haplotypes. The C+658T site breaks up the haplotypes, with 12 of the 124 chromosomes carrying the common

New Guinean	1		1			1															1	
San																			1	1		1
Bantu	3					7		1						1	1	1		1	1	1		1
Japanese	2		1			3		1									1					
Chinese	4					1		1						1								
Malay	9		1	1	1	4		1		3	1			1			1				2	1
S. Indian	2	2	1			4					1				1						1	
N. Indian	3		3			4							1									
Indian British									2													
Black British												1						1				1
S. European	2		1	1		2	1						1									
N. European	6		5			1																
Total	32	2	13	2	1	27	1	4	2	3	2	1	2	3	2	1	2	2	2	2	4	4
C+658T CATT+225ΔΔΔΔ	● ○ A -	○ ○ A -	● ● B -	○ ● B -	○ ○ B -	○ ○ C -	● ● D -	● ○ E -	● ● E -	○ ○ G -	● ○ G -	● ● H -	● ○ I -	● ○ J -	○ ○ K -	● ● M -	● ● N -	○ ○ O -	● ○ P -	○ ○ P -	● ○ Q -	○ ○ Q -

	Total	N. European	S. European	Black British	Indian British	N. Indian	S. Indian	Malay	Chinese	Japanese	Bantu	San	New Guinean
R -●●	1			1									
S -●●	5							1	1		1	2	
T -○○	1											1	
U -○●	10							2	1	5	1		1
U -○○	4			1							1	2	
V -○●	2										1	1	
X -○●	2											2	
Y -○○	3											3	
Z -○●	1						1						
a -○●	1					1							
b -○●	1							1					
c -●●	1							1					
d -○○	1											1	
e -○●	1								1				
f - ●●	1										1		
g -○●	1										1		

Table 7.3 Analysis of two polymorphic sites 3' to exon 17

The identity of each chromosome in terms of the assumed haplotype and the two new sites is shown in ewach row, and the number tested and origin are shown.

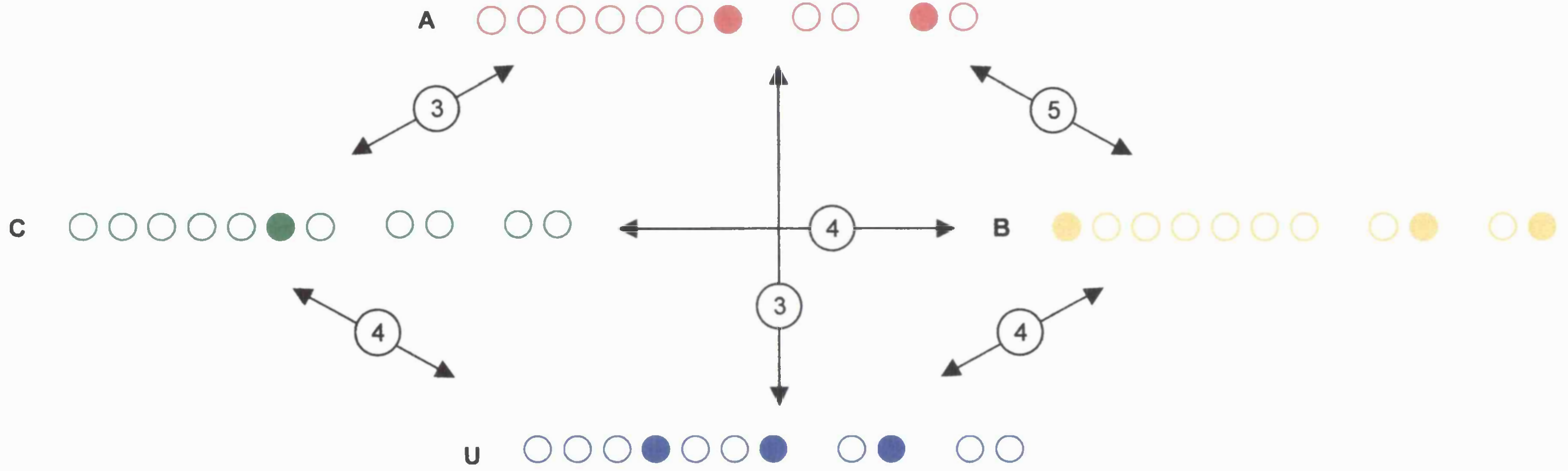


Figure 7.4 **Comparison between the four common haplotypes**
The number of differences between each haplotype are shown by each arrow. For example, four out of eleven sites differ between the B haplotype and the C haplotype.

haplotypes not carrying the expected allele. Nevertheless, from the 33 further chromosomes tested 11 haplotypes can be inferred to be likely simple recombinants of the four common haplotypes.

The list of haplotypes that are interpreted as simple recombinants, together with representations of their allelic composition, are shown in figure 7.6. Analysis of simple recombinants shows that recombination has occurred along the length of the lactase gene. Several haplotypes appear to have been generated by recombination between A_8-552/-559A_9 and G666A, a distance of 5kb. Section 3.1.4 analyses this sequence in more detail, and reveals that it is rich in repetitive elements, which are thought to promote recombination.

The recombination point of the recombinant haplotype could be localised more finely by analysis of two known polymorphisms within exon 6 and exon 11. Because phase of these polymorphisms could not be determined, and certain haplotypes shared the same allele, discrimination of the recombination point was only possible in a few individuals. Figure 7.7 shows some individuals where the recombination point of one haplotype has been localised. It demonstrates that the two I haplotypes tested (one Italian, one North Indian) were formed by independent recombination events and therefore do not share ancestry. This suggests that other types of other recombinant haplotype may not be equivalent.

7.3.3 Analysis of ancient haplotypes

The ancestral human haplotype was deduced from analysis of the polymorphic human sites in the chimpanzee sequence (section 6.2.1) and was found to be equivalent to the human K haplotype. Several haplotypes appear to be ancient since they are one mutational step away from K (figure 7.8). These include the C haplotype, as well as the rarer Y and Z haplotypes. Interestingly, the Q haplotype, a possible recombinant, also fits into this category, and it is possible that some Q chromosomes are recombinants and some are ancestral. In the population analysis they are assumed to be recombinant because analysis of CATT+225ΔΔΔΔ and C+658T show that in 4 out of 8 Q haplotypes, the 3′ end is the same as, and may be derived by recombination from, the A haplotype.

Fifteen unclassified haplotypes remain which are not simple recombinants, but could be intermediate steps in the formation of the four common haplotypes.

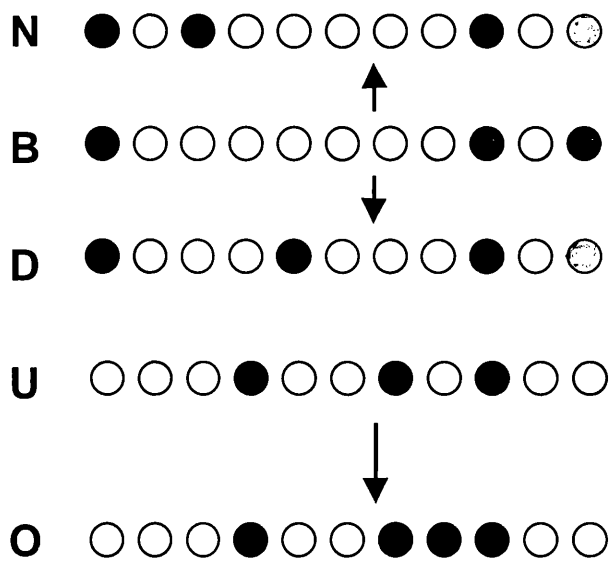


Figure 7.5 Haplotypes that are simple derivatives of common haplotypes

7.3.4 Generating a network of relatedness

It is possible to relate all the observed haplotypes to one another by linking together haplotypes that differ in only one position. The result, shown in figure 7.9a, is a complicated network of links.

The haplotypes that are probably recombinants of the common haplotypes presumably played no part in the evolution of the common haplotypes and can be removed from the network, together with all links to them and other haplotypes linked uniquely to them. The Q haplotype remains because some Q haplotypes may not be simple recombinants of A and C (see above, section 3.5). The resulting network is shown in figure 7.9b.

This network can be simplified further by removing haplotypes for rare pairwise combinations of close alleles using the 'four-gamete' test. Rare combinations that are distant from K, the ancestral haplotype, could be generated by recombinations, possible recurrent mutation of an Alu-tail polymorphism, or they may be maximum-likelihood estimation artefacts. Identifying haplotypes carrying these rare combinations allows several to be removed from the network. This removes haplotype d and haplotype b, which contain combinations of 666A and 5579C; haplotypes j,k and l, which conatin combinations of -678G and -552/-559A$_9$;

and haplotype g, which contains a combination of -942/-943ΔΔ and -552/-559A_8. The resulting network is shown in figure 7.9c.

The dismissal of haplotype g is the most controversial, but if it were the route to the evolution of the U haplotype then there would have to have been a recombination over 300bp to generate the -942/-943ΔΔ and -552/-559A_9 combination or a recurrent generation of the -552/-559A_9 allele, rather than recombination over 5kb, as discussed below.

A four gamete test shows that the four common haplotypes have all four possible combinations of alleles at A_8-552/-559A_9 and G666A (figure 7.11). This could be due to a recurrent mutation at A_8-552/-559A_9 locus regenerating an allele. However given the linkage disequilibrium between this locus and loci 5′ to it (table 1), this seems unlikely. It is more likely that a recombination event links the haplotypes A, B, C and U, especially since the 5kb between the two loci is Alu-rich sequence and probably recombinogenic. Theoretically, any haplotype could be the recombinant, but U is the best candidate because both -552/-559A_9 and 666A, which are constituents of the U haplotype, are not the ancestral alleles at these loci. In this scenario, the -552/-559A_9 mutation occurred on a proto-A haplotype, the 666A occurred on a proto-B haplotype, and recombination between them generated the proto-U haplotype. The C haplotype is one mutational step from the ancestral haplotype, as discussed above.

So if we accept that haplotype U was generated by recombination, then, since the -942/-943ΔΔ occurs on only the U haplotype or its obvious close relatives, then the 5′ end of a progenitor haplotype before any ancestral recombination had the -942/-943TC allele. The W haplotype differs from the U haplotype at this position only, and so the W haplotype may have been a proto-U haplotype. A simplified version of the network, considering this possibility, is shown in figure 7.10.

7.3.5 Relating haplotype frequency to linkage disequilibrium

In certain cases, population variation in linkage disequilibrium can be directly attributed to the frequencies of the different haplotypes. The bar charts in figure 7.12 demonstrate this for two pairs of loci that were discussed in section 1.2, G-678A/T5579C and A_8-552 to -559A_9 /G666A. The bars are as figure 7.1, with 1-D′ shown on the y axis and different populations shown on the *x* axis. Superimposed are the

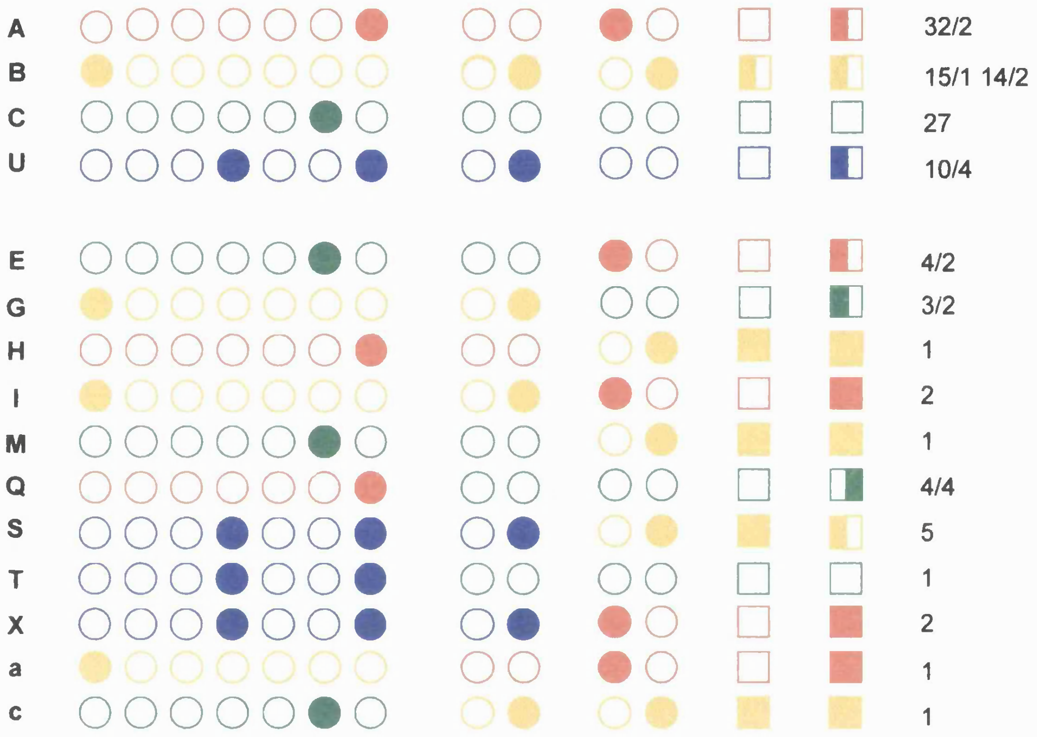


Figure 7.6 **Identification of possible recombinant haplotypes by analysis of two further polymorphic loci.**

Haplotypes are represented as before, with each common haplotype represented by a colour. The two additional loci are shown as squares and the number of chromosomes of each haplotypes on which these extra loci were tested is shown on the right. Where one or both of the loci split haplotype, the number of chromosomes showing each allele is shown: for example 5/2 shows that five haplotypes had the allele identified by the first half of the box and that two haplotypes had the allele identified by the second half of the box.

frequencies of haplotype E and haplotype U, with the second *y* axis showing the scale.

The $D'_{\text{G-678A/T5579C}}$ reflects the frequency of the E haplotype quite well. Haplotype E is a recombinant of haplotype A and C (figure 7.6) and, in populations with haplotypes A,B,C and E, all four combinations of the two alleles at G-678A and T5579C are present. If haplotype E is present at reasonably high frequencies, then the recombinant combination of -678G and 5579C is present and the linkage disequilibrium lowers. This is observed in the Russian and Southern European populations. An exception is the Mordavians who show linkage disequilibrium between the loci yet a haplotype E frequency of 0.04 was obtained by haplotype counting. Interestingly, the maximum likelihood method calculates the haplotype E frequency to be only 0.01 (table 7.2). This may be due to discrepancies in interpreting haplotypes of individuals who are heterozygous for both sites.

Similarly, the $1\text{-}D'_{\text{A8-552 to -559A9/G666A}}$ value reflects the frequency of the U haplotype (figure 7.6). Both the 1-D′ value and the frequency of the U haplotype are high in East Asian populations. The only exception is the Yakut, where the frequency of the U haplotype is higher than expected when compared with the linkage equilibrium statistic. This may be due to the low frequency of B haplotypes (0.10) and hence the rarity of association between -552/-559A_9 and 666A.

7.3.6 Further analysis of population distribution

Table 7.4 groups haplotypes as common haplotypes (together with simple derivatives, as shown in figure 7.5), simple recombinants (figure 7.6), ancient haplotypes (figure 7.8) and other haplotypes, and displays the percentage of each in the populations tested. There are several interesting observations.

Firstly, the frequency of the other haplotypes, which may be intermediate steps between the four common haplotypes, is high in sub-Saharan Africans (30% in San and 28% in Bantu-speaking South Africans), but very low (≤3%) in other populations.

Secondly, the frequency of simple recombinants is highest in the Bantu-speaking South Africans and second highest in the Russians. Assuming there are no differences in recombination rates between populations, this may be a reflection of differing population histories.

Finally, the frequency of possible ancient haplotypes is much higher in the San (13%) than in any other population, with the Papua New Guineans second with only 7%. This may be due to a reasonably steady population size combined with genetic isolation, as would be expected for a hunter-gatherer population.

7.3.7 Haplotype association with lactase persistence

Table 7.5 shows Two by three tables of lactase persistence against genotype. The one-sided p values are calculated, as before, by converting two-by-three to two-by-two tables and using Fisher's exact test. The Finns show slight association of persistence with the A haplotype, but the Yakut show no significant association with the A haplotype. There are several reasons for this which are essentially the same as the reasons for the lack of association with individual alleles, and are discussed in section 6.4.

The two other tables are from previous work (Harvey *et al.*, 1998), and show that when haplotype carrying persistence can be unambigously determined, both families and biopsy from random individuals show strong association of persistence with the A haplotype.

Table 7.4 Frequency of different haplotype classes in different populations

Population sample	n	% common haplotypes	% simple recombinants	% ancient haplotypes	% others
N. European	104	98	2	0	0
S. European	106	88	9	2	1
N. Indian	156	89	9	1	1
S. Indian	84	90	5	4	1
Malay	200	90	6	3	1
Chinese	114	88	8	3	1
Japanese	82	89	9	1	1
N. Guinean	72	88	5	7	0
Roma	170	89	7	2	2
Mordavian	68	88	9	1	2
Yakut	78	93	3	1	3
Russian	80	82	13	2	3
Bantu	72	52	16	4	28
San	30	50	7	13	30

Table 7.5 **Two-by three and Two-by-two contingency tables showing association of haplotype with iactase persistence or iactase non-persistence.**

Finns	A	A/non-A	NonA/nonA
Persistent	4	5	0
Non-persistent	2	6	3

p=0.08

Yakut	A	A/non-A	NonA/nonA
Persistent	1	11	7
Non-persistent	3	12	3

p=0.13 (non-A)

Families	A	Non A
Persistent	5	5
Non-persistent	1	14

p=0.02

Biopsies	A	non-A
Persistent	11	0
Non-persistent	2	9

p=0.0001

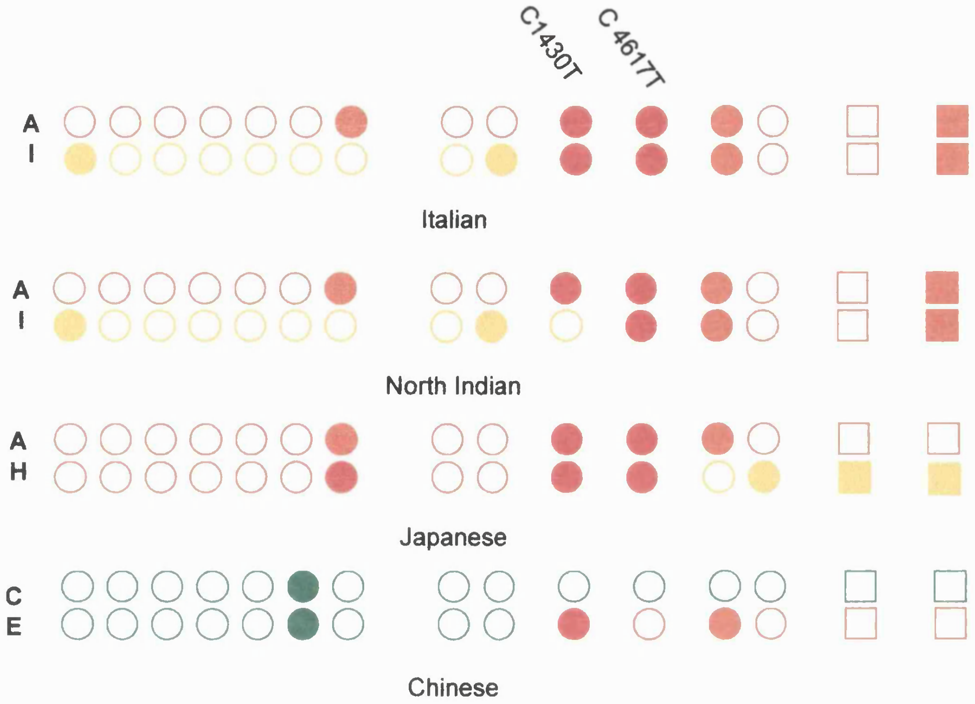


Figure 7.7 **Localisation of recombination breakpoints** Individuals carrying recombinant haplotypes where the point of recombinaton can be localised using polymorphisms in exon 6 (C1430T) and exon 11 (C4617T). The ancestral state of both polymorphisms was determined by sequencing five unrelated chimpanzees.

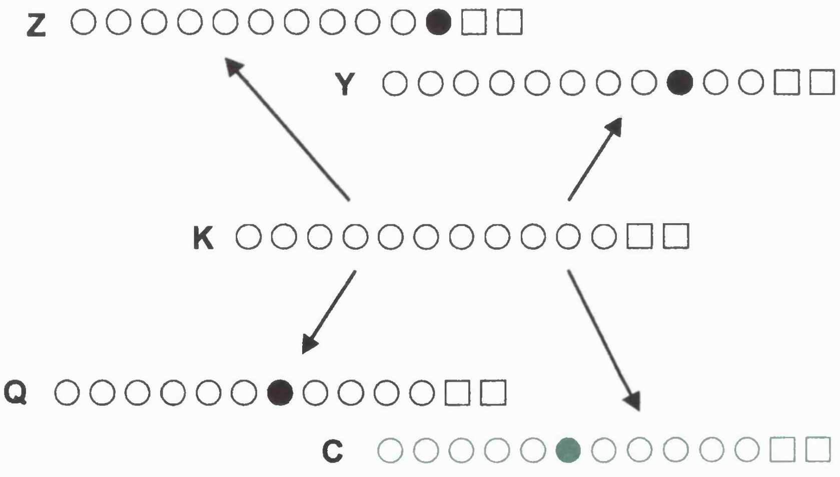


Figure 7.8 Possible ancient haplotypes that are related to the ancestral haplotype by one mutational step

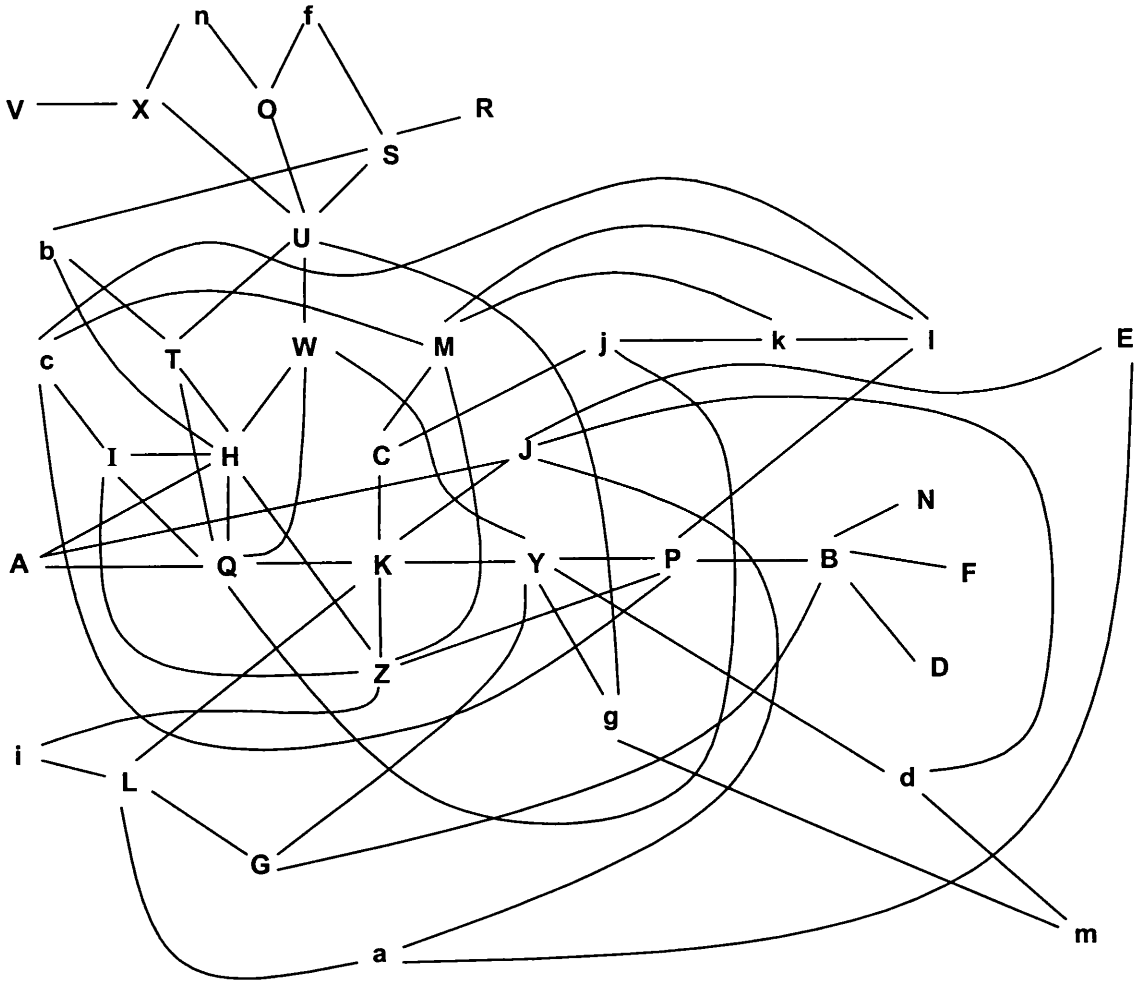


Figure 7.9a Haplotype network showing haplotypes related by one mutational step.

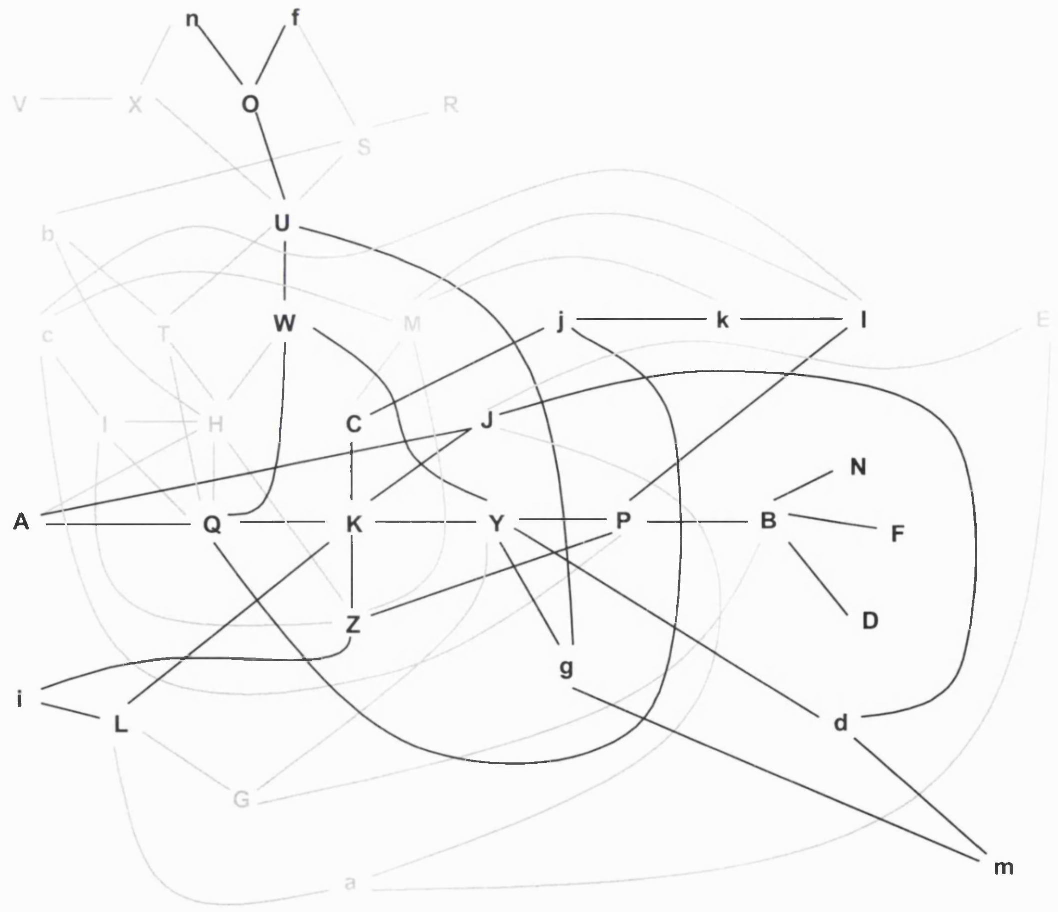


Figure 7.9b Haplotype network showing haplotypes related by one mutational step with simple recombinant haplotypes and links shown in grey.

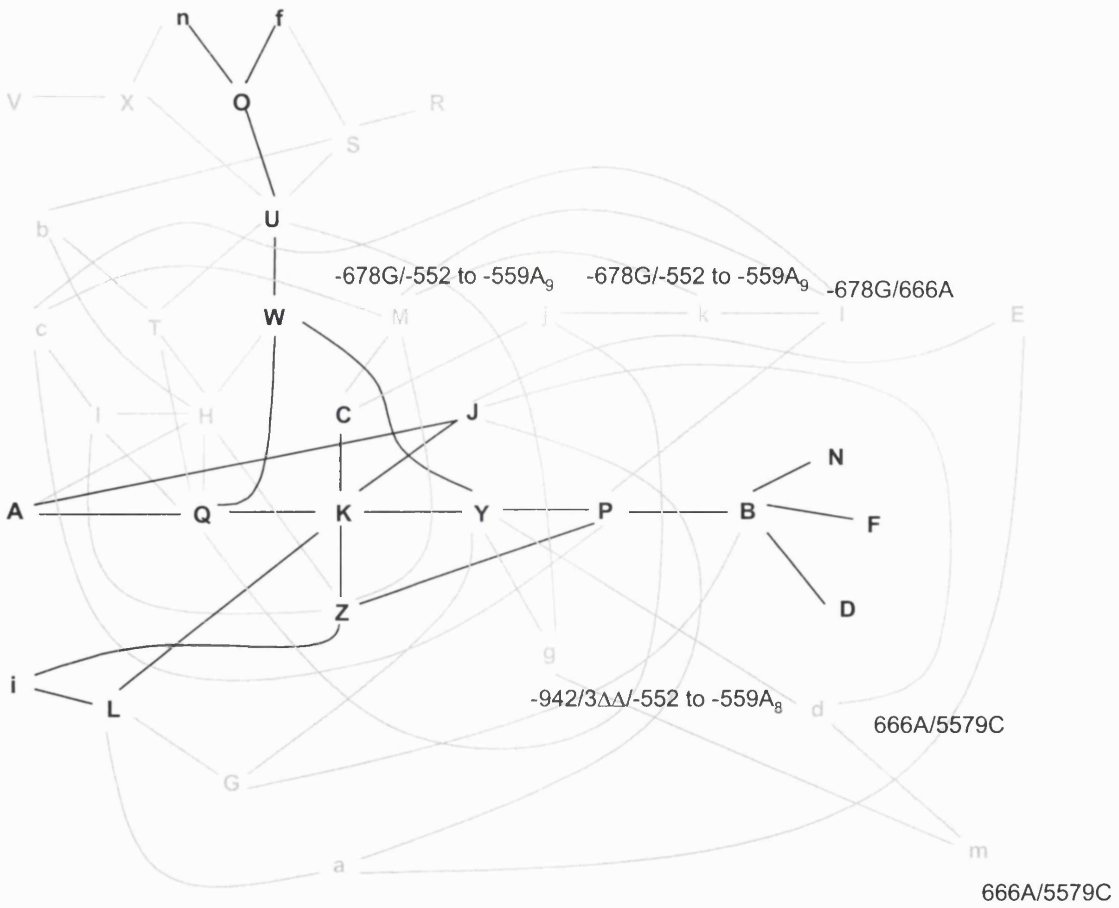


Figure 7.9c Haplotype network showing haplotypes related by one mutational step, with simple recombinant haplotypes and links shown in grey, together with haplotypes in grey due to an unusual combination of alleles. The respective allelic combinations are shown next to the haplotype.

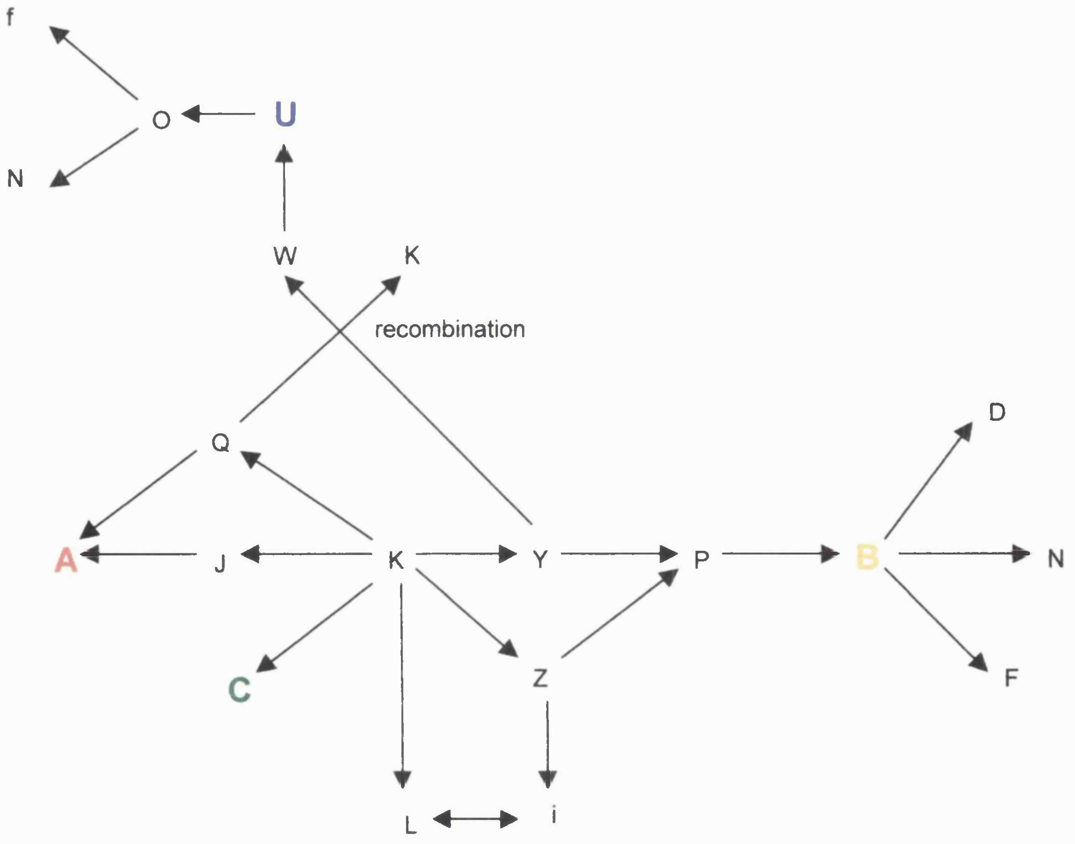
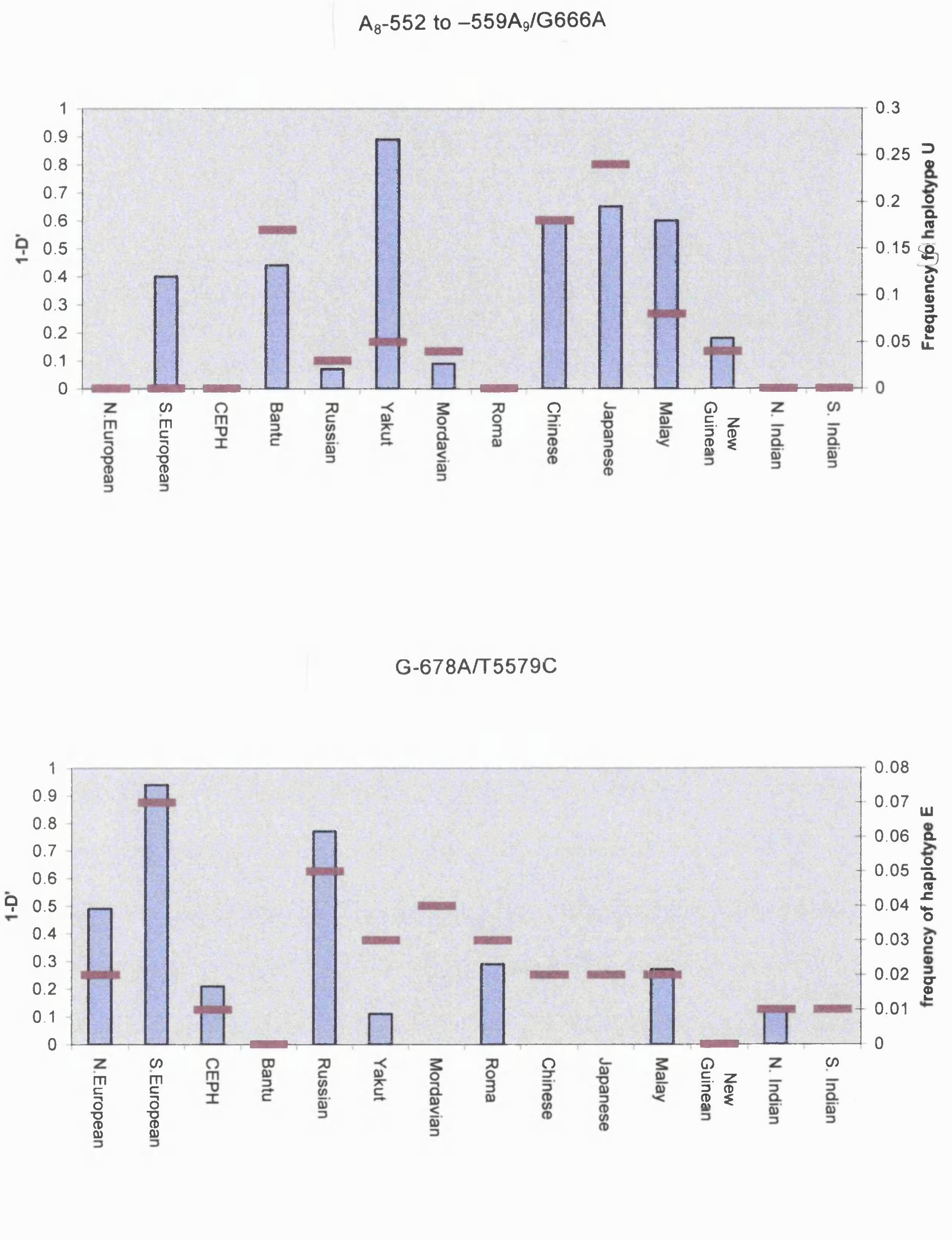


Figure 7.10 Simplified haplotype network showing the possible evolution of the four common haplotypes



A_8-552 to –559A_9/G666A

G-678A/T5579C

Figure 7.11 Linkage equilibrium (1-D') between two pairs of loci Linkage equilibrium is shown (blue bars, left y axis) compared with frequency of a specific haplotype (red lines, right y axis).

8 Discussion

At the start of this project several polymorphic loci were known across the lactase gene, and these had been studied in families to analyse linkage disequilibrium and determine haplotypes (Harvey *et al.*, 1995). Cohorts of Northern Europeans, Finns and Italians of known lactase persistence status were tested for association for any allele at these polymorphic sites. Alleles at two different loci, together defining the A haplotype, were found to show association with lactase persistence (Harvey *et al.*, 1998). Because both lactase persistence and non-persistence occurred on the A haplotype, a model of the lactase persistence mutation occurring on an A haplotype background was proposed. A further study on Yakut of known lactase persistence status was inconclusive. No association of lactase persistence with any allele was found, although whether this was due to experimental artefact or a reflection of the history of the population was unclear.

Further characterisation of the lactase gene was necessary in order to identify the polymorphism causing lactase persistence/non-persistence. 1.8kb upstream and 2kb downstream was sequenced from PCR products generated from the panel of known lactase persistence genotype cell lines. Several polymorphisms were found which associated with known haplotypes across the lactase gene but none associated better with lactase persistence or non-persistence. Part of a fosmid clone was used to generate the sequence for intron 1, a region which often contains *cis*-acting regulatory elements. However, only limited progress was made in identifying polymorphisms in this region due to difficulties in PCR and sequencing. Very recent work has shown that there is a very large polymorphic deletion in intron 1 with the sequence generated from the fosmid representing the deleted allele. Now this is understood, progress is being made and preliminary studies suggest association with lactase haplotypes but again no stronger association with lactase persistence or non-persistence (Poulter, M., personal communication). However characterisation of the deletion and the remainder of intron 1 by sequencing representative alleles is underway.

DGGE analysis described in this thesis and previously has shown that there are five variable sites -974 to -852 bases upstream of exon 1, two of which have common alleles. This region was found have a high identity with human sequence and sequence of the intervening Alu elements show that these have less identity with the human sequence. Analysis of this same 1kb upstream region with the published pig sequence show sequence conservation both in the upstream promoter and in the region -974 to -852. No variation in this region caused the phenotypic polymorphism, since both lactase persistence and lactase non-persistence could be found on one haplotype (A). Nevertheless, the variation could have some functional role in expression of the gene. For example, there have been several reports concerning the variation timing of the developmental program of lactase downregulation in humans: as early as 5 years in Thai children and as late as 13 years in Finnish children (Keusch *et al.*, 1969a; Sahi and Launiala, 1978). Indeed, several polymorphisms that have functional consequences have been identified in promoter regions of genes, such as a TG insertion which reduces expression of HLA-DQB1 (Beaty *et al.*, 1995), and a T to C transition which disrupts a GATA motif and abolishes expression of the Duffy antigen/chemokine receptor gene (Tournamille *et al.*, 1995).

By using EMSA and oligonucleotides corresponding to the variant alleles, possible allelic effects on protein binding were analysed. The newly described polymorphic site (TG-942/3ΔΔ) was within a small inverted repeat which was very similar to the functional CE-LPH elements further downstream that had been identified in the pig and bind to Cdx2 and HNF-1 transcription factors (Spodsberg *et al.*, 1999; Troelsen *et al.*, 1997). This polymorphism did not affect protein binding. In contrast, the other polymorphism at -958 was found to alter protein binding: the oligonucleotide with a C at the position equivalent to -958 bound a protein, whilst substitution of a T abolished protein binding. The possibility that this allele may be responsible for variation in timing of downregulation could be directly investigated by examining the data obtained by the analysis of lactase mRNA in children. There was no association of early or late onset of downregulation with the occurrence of a C or T allele at -958, although the sample set was small and ethnically mixed.

To further investigate any functional significance of this polymorphism, it would be interesting to determine, using protein extracts from human tissues and cell lines, whether each of the protein binding activities was intestine specific. However ubiquitous expression would not rule out function. The next experiment would be to separately transfect cells with two different forms of the construct corresponding to the -958C and the -958T alleles. If a protein binding activity affects gene expression, the *trans*-acting factor could be isolated either by standard protein chromatographic techniques using the EMSA as an assay for the protein, or by molecular biological methods as used by Mitchelmore *et al.* (1998) to isolate HOXC-11 binding to CE-LPH1a. The possible function of a *cis*-element containing the C-958T polymorphism may become clearer when the causative element is found and should be reviewed at that time.

While experiments to map the genetic region around the lactase gene were in progress in order to identify the causal element (as outlined below), my project concentrated on examining diversity of the lactase gene in different populations. A cline in haplotype A frequency across Europe had been observed previously (Harvey *et al.*, 1995; Harvey *et al.*, 1998), and it was clear that understanding the global diversity of the lactase haplotypes would illuminate the history of the lactase gene, both in terms of selection for the lactase persistence allele and drift events in human evolution.

The sub-Saharan African samples show much greater haplotype diversity than other populations with many extra haplotypes not seen in the rest of the population samples. This pattern of higher diversity has been observed at other loci and has been interpreted as evidence for an 'out of Africa' model for peopling of the Old World (section 1.6.4.3). Analysis of the unweighted means of haplotype composition shows that the four common haplotypes and their simple derivatives comprise 89% of total haplotype diversity in non-Africans, but only 51% in sub-Saharan Africans . Both San and Bantu have high frequency of haplotypes (30% and 28% respectively) which may be intermediate steps between the four common haplotypes, but these haplotypes are almost absent in non-African populations (≤3%). In addition, the San have a high frequency of haplotypes that are simply related to the ancestral haplotype (13%). This evidence suggests that the four common haplotypes originated in Africa before a migration into the rest of the Old World, and the distribution arose by genetic drift. Direct selection of alleles is unlikely to account for the loss of haplotype diversity, since there are no alleles shared between the four common haplotypes that are not present in other haplotypes. The migration occurred relatively recently since there has not been enough time to reduce linkage disequilibrium between loci.

The association of the A haplotype in Europeans with lactase persistence and the high frequency of both haplotype A and lactase persistence in Northern Europeans leads to the attractive hypothesis that haplotype A is at unusually high frequencies due to genetic hitch-hiking (selective sweep) with selection for lactase persistence. Alternatively, if haplotype A is at high frequencies in Europeans due to genetic drift, the association of lactase persistence with the A haplotype could be an artefact of the high frequency of that haplotype. However, the frequency of the A haplotype is reasonably constant throughout other non-African populations, and no similar clines exist. The north-west/ south east cline of lactase persistence (and presumably the frequency of A haplotype, although the data is not of sufficient resolution to define the gradient) does agree with Cavalli-Sforza's 1st principal component map of classical allele frequencies, which is thought to be due to the movement of farming from the near East. However, if the frequency of lactase persistence was due to this, then the Palaeolithic hunter-gatherer people of Europe would have been lactase persistent and the farmers non-persistent: the exact opposite of that expected. Similarly, this means that the Palaeolithic population would have been almost entirely haplotype A. This merely removes the timing to several hundred thousand years earlier, but does not explain the distribution. Nevertheless, it would be interesting to examine the haplotype frequencies of a sample whose gene pool has a large amount of Palaeolithic alleles, such as the Basques.

Another problem with the selection hypothesis is that the reported high gradient in lactase persistence frequency in India from North to South which is not reflected in a cline of haplotype A frequency. However, the cline is less drastic than the European cline, so there may not have been selection for lactase persistence and there are in fact 'sufficient' A haplotypes in the North Indian population for A to carry lactase persistence. Furthermore, the groups tested in this study are not the same as those used for the reported lactose tolerance tests, my samples having come from Singapore. It will be extremely interesting to see whether the A haplotype is at similar high frequencies in other populations with a high frequency of lactase persistence such as the Bedoiun of Arabia or the Beja of Sudan. The U haplotype is present in all populations except Europeans and Indians. The only exception is the Russian sample from Perm in the Ural Mountains. This may be due to admixture with Asiatic tribes, such as the Yakut, which have a high frequency of the U haplotype.

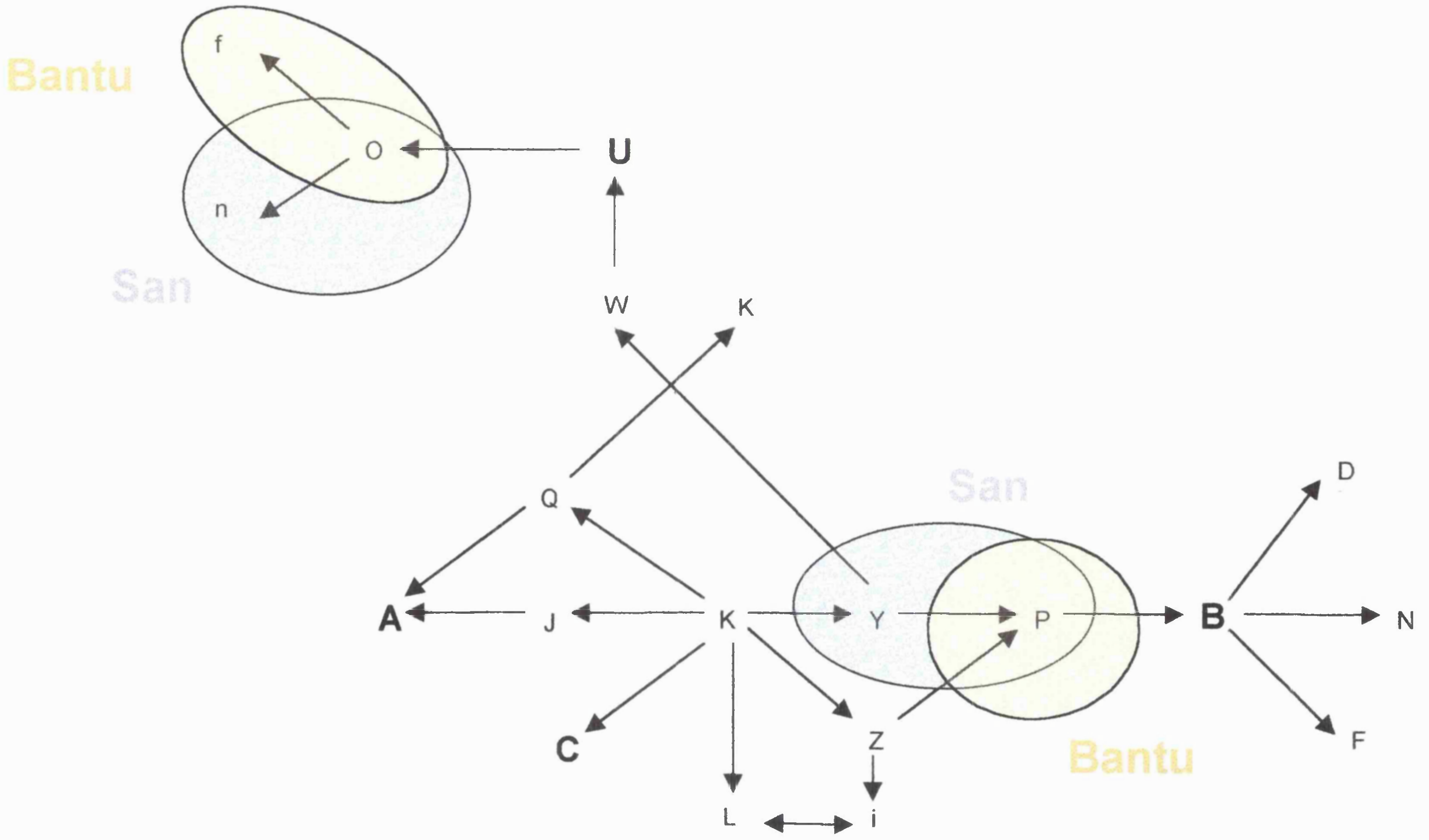


Figure 8.1 Haplotypes in the reduced web diagram that are unique to San and Bantu-speaking South Africans

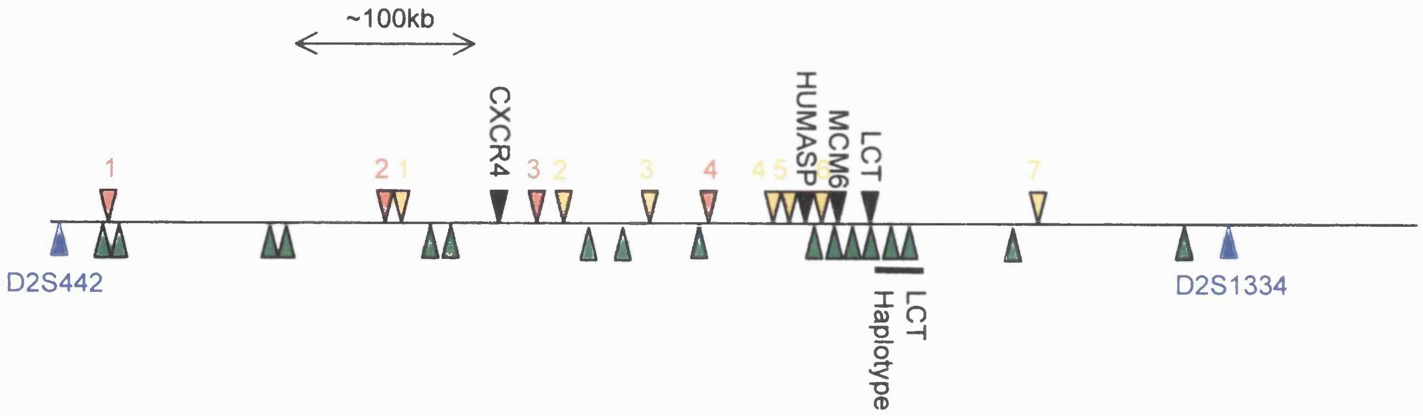
One can speculate on the cause of the interesting distribution of the U haplotype. A Y chromosome variant, YAP 3, has a rather similar distribution (Hammer *et al.*, 1998), and it has been suggested that YAP 3 arose in Asia and is observed in sub-Saharan populations due to a migration from Asia back into Africa (after an initial migration of modern humans from Africa). This was based on the fact that YAP 3 derivative chromosomes were observed in Asia but not in Africa. With the U haplotype, a different pattern is seen: U haplotype derivatives are observed and are common in San and Bantu-speaking South Africans but very rare in the East Asian population samples (figure 8.1), so the Asian migration hypothesis seems an unlikely explanation.

There are two other hypotheses. The first involves selection, either for the U haplotype directly via the -924/3ΔΔ component allele, or for an allele within linkage disequilibrium of the lactase gene raising U haplotype frequencies by genetic hitchhiking. Direct selection via the -942/3ΔΔ allele is unlikely since there is no evidence that it affects function: it does not affect binding of putative transcription factors (section 4.2). Selection for an associated allele at another gene is more likely. There are several genes within ~150kb of lactase, including the chemokine receptor CXCR4 (Poulter, M. unpublished data). A variant of one of these genes could be either selected for in East Asian populations or selected against in Indo-European-speaking populations. The second hypothesis is that genetic drift was responsible for the distribution, either in the peopling of East Asia, or in relatively more recent expansions or bottlenecks, such as the development of agriculture or the spread of Indo-European speaking populations.

There has recently been a study comparing frequencies of apolipoprotein A-IV-2 allele (on chromosome 11) with frequencies of the lactase persistence allele, and the distribution is remarkably similar, in Europe at least (Weinberg, 1999). The gene plays a part in lipid absorption by the intestine, and this allele encodes a histidine to glutamine substitution that alters biological function. Weinberg suggests that this allele was selected for in northern Europe because it gave a nutritional advantage in a diet rich in milkfat. Archaeological methods can begin to test and refine hypotheses put forward by analysis of the genetic data about the role of selection in diversity of lactase, apolipoprotein IV-2, and indeed other genes. For example, milk consumption by historic people has mainly been inferred from ceramic artefacts and art, but recently the first direct evidence that pottery fragments once contained products was published by Dudd and Evershed (1998). By analysis of the stable ^{13}C isotope in fatty acid deposits on the pottery and comparison with the expected isotope patterns given a biosynthetic origin of these fatty acids, contact of these pottery fragments with milk was deduced. In addition, it was shown that adipose fats and milk fats showed distinctly different ^{13}C isotope patterns even from animals raised on the same pasture, so the fatty acids on the pottery were from milk not adipose tissue.

While the work described in this thesis was in progress, the congenital alactasia locus was mapped in Finnish families to 2cM upstream of LCT by a multipoint linkage analysis and microsatellite haplotype sharing approach (Jarvela *et al.*, 1998). No mutation in lactase cDNA had been found that was responsible for congenital alactasia in a patient (Poggi and Sebastio, 1991). Jarvela's study showed no recombination in families and the haplotype sharing analysis gave large error bars. Since our study suggested that a *cis*-acting element far upstream of lactase was probably involved in lactase downregulation, it seemed likely that the congenital alactasia locus was in fact a mutation in the same *cis*-element which is polymorphic in lactase persistence/non-persistence. In order to provide a resource for further study of sequence variation, a PAC/BAC contig spanning about 600kb around the lactase locus between markers D2S442 and D2S1334 was constructed by Mark Poulter. Construction of this contig showed that the physical size of this region was considerably less than that predicted by recombination and radiation hybrid maps, as quoted by Jarvela. The congenital alactasia locus could therefore be much nearer, or even within, the lactase gene.

Poulter identified many single nucleotide polymorphisms (SNPs) across the region (figure 8.2). Regions of haplotype sharing in the panel of five individuals and in the Finnish cohort are currently being defined to determine where association with the lactase haplotypes A, B, and C breaks down. The breakdown point at the 3' end and at the 5' end has been determined, and preliminary data shows that lactase persistence/non-persistence does not show better association with the SNPs that break up the haplotypes. Therefore, the region that contains the causative polymorphism has apparently been localised to a 200 kb region, but with no extra directional information. Genomic sequence of this region has begun appearing on the public database, which will considerably aid progress by allowing simpler identification of SNPs and microsatelllite repeat polymorphisms.



Expressed Sequence Tags

1 EM:S74678 93% identity with hnRNP complex K
2 EM:AI184702
3 EM:HSXT02304
4 EM:1163295
5 EM:AI589561
6 EM:HSA69273homology to endoα-D-mannosidase
7 EM:HSD684

Genomic Survey (October 1999)

1 CIT-HSP-2166NS.TR
2 CIT-HSP-2197E9.MF
3 RPCI 11 43A12.TK
4 CIT-HSP-2377N1.TF

Figure 8.2 **Physical map between D2S442 and D2S1334**
Single nucleotide polymorphisms are shown in green

Complete sequencing of several chromosomes may be the most effective strategy for identifying the causal element. Once the causal element is found, microsatellite analysis across this region on lactase persistent and non-persistent backgrounds will date the origin of the mutation.. Its demographic history can then be inferred by examining patterns of variation at linked microsatellites in different populations (Stephens *et al.*, 1998), and then a more complete history of this functional polymorphism may be written.

9 Appendices

Appendix 1 Sequences of all oligonucleotide primers referred to in this thesis

Name	Number	Sequence 5' to 3'
+10armsSD	236	TCCCTAGCTTCACATCTTGTT
+10armsSI	235	TCCCTAGCTTCACATCTTGTG
+271armsAD	238	GAGGAAGGAAACAGACTRAACC
+271armsAI	237	GAGGAAGGAAACAGACTRAATG
+271armsSD	315	CCCATCTGACCTTGGTTYAG
+271armsSI	314	CCATCTGACCTTGGTTCATT
+704armsAC	316	TCAGCACATTTGAGGGGATTG
+704armsAT	317	TCAGCACATTTGAGGGGATTA
DB2a2	77	CCCTTTTGTTGAGTATATTTGG
DB2a3	96	TCTGTTTGCTCTCTGCTATAC
DB2f	93	GTCCTTTCTCCTATGGTTGC
DB2r	94	GAGACAATGAGACATGGTCC
DB2s	74	CTAGGATGCCCTTTCTTCTG
DB2s1	76	ATTTGCTTCTTGTAAGCATTCC
DB2s2	78	AGGAGAATGGTGTGAACCTG
DB2s3	95	AAACTTGTAAGAAGAGCCAGG
F2A	84	CTCTCCTCAGATGTTACAGG
F2S	83	CAGTGGTTTCCACAGTCAGA
INT16S	111	ACCTCCACCTCGGCATCC
INT1A	81	AGTTGGCAAGGAGAGAACTC
INT1A12	334	CATTCATCCATCAGTGGATATC
INT1A5	157	ATGCTTTAAAATCTTAAAGAATGC
INT1A9	190	GTCTCGATCTTCTGACCTTGT
INT1S	80	CAGAATCTTAGAATTGAAACTAC
INT1S11	397	TATTTGATTATAGAGACGTTAAAC
LCT3S	-	TACAGTGACCCTTCTCTGCCAG
MCM6Wf	105	GTCCAACCTAAACATGTGAAG
MCM6Wr	106	CATTGATGCCACCAGCACC
MCM6ZS	66	AAGGTCAGCATCCTACTAGG
P2P2f	127	GTGTCAAAAAGATGGATTGTTC
P2P2r	128	AGAGCTATTTCTGTCTCATATG
P3f	26	TGGCAAAACACTCAGCATTTC
P3r2	41	GGTAGCAGGACATAGACGG
P4f	28	CCCTAAGTGTATGTTAGGTAC
P4r	29	ATGAGAATATAGTCATAAACTATG
PROA	-	GACTACATGCCAAGACAGCTCC
PROS2	-	TCTTCAGACATTTTCCGGGTTCC
W1A	64	AGGTGTGTGATGAAGGTTGC
W1S	63	GAAAACAGTGCAGTGCTACC
XINTPS	91	TGGAGGAAGTGATCAAGGAG
X1PS	-	CTGGTTCACCTTCAGTGACTTGGA
X10f	107	TTGAAGGTGCGTGGAGAGC
X10r	108	AGCAGTGTATCGATGAGCCT
X13f	163	AGGGAAGAGAGCTTAACCTG
X13r	164	GCTCAGTCATGGTAACTTGC
X17arms+S	286	GTCACACTCTCCTAGATGCC
X17armsDA2	287	GCTGTTTTTATTTTCTGGAAAACAAG
X17armsIA	274	GCTGTTTTTATTTTCTGGAAAACACA
X17arms-S	273	GTCACACTCTCCTAGATGCT
X17r2	44	TAGATGAAGAAACTAGGCCTG
X17r4	92	TTGTGTGACTTCATAACTTCTC
X6f	174	CAGCCTGACCAGCAGCAG
X6r	104	CCATCCACCATGATCCTGC
3UXA	65	CACAGTTCCTCAGCTCTGG
338f	214	TCTGAAACCTGAAGGACC
338r	215	TTTTGAAATCTGTCATAGATGC
5A10	70	CCAGCGCACCCAGCCACAC
5A4	11	ACTGGCAAAACAGGCGTGAT
5A5	16	TCTACAGGTGACAAAATAGAGG
5FA	53	GACCAACACAAAAACCTCAGAC
5FS	52	GGAGGGTGAAGGAATTTGCAAG
5seqA	-	GAACGTTTAAGGAAGTGGGAGG

Appendix 2 Addresses of all clones referred to in this thesis

Lab code	Library	Antibiotic	Address (384 well format)
PAC 1	LL02NP04	Kanamycin	AI4-F22
PAC 3	LL02NP04	Kanamycin	AI4-F3
PAC 4	LL02NP04	Kanamycin	AI4-C6
PAC 7	LL02NP04	Kanamycin	AI9-B1
PAC 8	LL02NP04	Kanamycin	AI2-I3
PAC 12	RPCI 1	Kanamycin	238-C18
PAC 13	RPCI 1	Kanamycin	55-P3
PAC 15	RPCI 1	Kanamycin	236-J19
PAC 19	LL02NP04	Kanamycin	AI5-H23
PAC 21	LL02NP04	Kanamycin	AI11-K24
PAC 23	RPCI 1	Kanamycin	110-E19
PAC 27	LL02NP04	Kanamycin	AI5-K11
PAC 31	RPCI 1	Kanamycin	106-O20
BAC 1	RPCI 11	Chloramphenicol	43-A12
FOS 6	LL02NC03	Kanamycin	AG52-H4
FOS 7	LL02NC03	Kanamycin	AG1-P22
FOS 8	LL02NC03	Kanamycin	AG11-K12
FOS 9	LL02NC03	Kanamycin	AG15-B20
FOS 10	LL02NC03	Kanamycin	AG15-B22
FOS 12	LL02NC03	Kanamycin	AG33-D19
FOS 13	LL02NC03	Kanamycin	AG33-O4
FOS 15	LL02NC03	Kanamycin	AG66-B5
COS 3	LL02NC02	Chloramphenicol	AE17-G6
COS 4	LL02NC02	Chloramphenicol	AE40-G7
COS 5	LL02NC02	Chloramphenicol	AE1-B16
COS 6	LL02NC02	Chloramphenicol	AE2-B16
COS 7	LL02NC02	Chloramphenicol	AE12-D7
COS 9	LL02NC02	Chloramphenicol	AE21-K24
COS 11	LL02NC02	Chloramphenicol	AE86-E16
COS 12	LL02NC02	Chloramphenicol	AE34-O4
COS 15	LL02NC02	Chloramphenicol	AE82-I10

Appendix 3 Lactose tolerance tests of Yakut samples

The individuals indicated by the shaded rows were not included in further analyses due to relatedness with other individuals.

Number	sex	Year of birth	Blood glucose levels (mmol/dm^3)		Persistence	Haplotype
			Before load	After load		
1	M	1967	4.9	6.5	P	CU
2	M	1967	5.9	6.1	N	AC
3	M	1967	5.5	5.4	N	AC
4	M	1969	6.1	8.5	P	C
5	M	1964	5.7	8.0	P	AG
6	M	1953	Low	Low	-	AC
7	M	1972	5.4	6.3	N	A
8	M	1960	5.9	7.2	P	AC
9	M	1964	5.8	6.9	P	AB
10	M	1968	Low	5.2	P	AC
11	M	1962	6.0	7.0	N	AC
12	M	1960	5.3	6.1	N	A
13	M	1973	5.8	7.0	P	AC
14	M	1957	5.7	5.8	N	AC
15	M	1973	5.6	6.2	N	CU
16	M	1966	5.5	6.9	P	CU
17	M	1968	4.5	5.6	P	BC
18	M	1966	6.1	5.9	N	CX
19	M	1956	7.1	9.0	P	A
20	M	1971	6.1	6.9	N	AC
21	M	1967	5.0	8.0	P	B
22	M	1957	6.6	7.2	N	AW
23	M	1964	5.5	5.9	N	AB
24	F	1965	5.9	7.2	P	AC
25	M	1939	7.1	7.2	N	AQ
25b	F	1955	6.1	6.9	N	A
27	F	1971	5.5	6.0	N	A
28	M	1966	5.4	5.0	N	AB
29	F	1974	8.8	7.5	N	AC
30	M	1942	4.4	5.9	P	C
31	F	1971	5.6	6.7	P	AC
32	F	1965	7.0	9.8	P	BC
33	F	1945	8.2	13.5	P	AC
34	F	1948	6.0	5.0	N	AC
35	F	1952	6.7	13.3	P	EU
36	F	1973	10.2	10.6	N	AU
37	F	1975	4.0	6.2	P	AB
38	F	1970	5.5	4.7	N	AC
39	F	1960	7.1	5.5	N	AC
40	F	1971	5.1	8.8	P	AW
41	F	1972	6.5	9.2	P	C
42	F	1954	5.2	15.9	P	AQ
43	F	1973	4.2	9.6	P	AC

10 References

OMIM: On-line Mendelian Inheritance in Man:
http://www.ncbi.nlm.nih.gov/Omim/searchomim.html

Ammermann, A. and L. Cavalli-Sforza (1971) Measuring the rate of spread of early farming in Europe, Man, 674-688.

Anderson, B. and C. Vullo (1994) Did malaria select for primary adult lactase deficiency? Gut, 35, 1487-1489.

Anderson, C.L. and C.J. Brown (1999) Polymorphic X-chromosome inactivation of the human TIMP1 gene, Am. J. Hum. Genet., 65, 699-708.

Aoki, K. (1986) A stochastic model of gene-culture coevolution suggested by the culture-historical hypothesis, Proc. Natl. Acad. Sci. USA, 83, 2929-2933.

Armour, J.A.L., T. Anttinen, C.A. May, E.E. Vega, A. Sajantila, J.R. Kidd, K.K. Kidd, J. Bertranpetit, S. Paabo and A.J. Jeffreys (1996) Minisatellite diversity supports a recent African origin for modern humans, Nat. Genet., 13, 154-160.

Auricchio, S., A. Rubino, M. Landholt, G. Semenza and A. Prader (1963) Isolated intestinal lactase deficiency in the adult, Lancet, 2, 324-326.

Axelsson, U. and C.B. Laurell (1965) Hereditary variants of serum alpha-1-antitrypsin, Am. J. Hum. Genet., 17, 466-472.

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz and E. Zuckerkandl (1996) Standardized nomenclature for Alu repeats, J. Mol. Evol., 42, 3-6.

Beaty, J.S., K.A. West and G.T. Nepom (1995) Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB21, Mol. Cell. Biol., 15, 4771-4782.

Bengtson, B., B. Steen, A. Dahlqvist and M. Jagerstad (1984) Does lactose intake induce cateract in man? Lancet, 1, 1293-1294.

Bodmer, W.F. and L.L. Cavalli-Sforza (1976) Genetics, Evolution, and Man, W. H. Freeman, San Francisco.

Boll, W., P. Wagner and N. Mantei (1991) 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., 48, 889-902.

Boukamel, R. and J.-N. Freund (1994) The cis-element CE-LPH1 of the rat intestinal lactase gene promoter interacts in vitro with several nuclear factors present in endodemal tissues, FEBS Lett., 353, 108-112.

Bowcock, A.M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd and L.L. Cavalli-Sforza (1994) High resolution of human evolutionary trees with polymorphic microsatellites, Nature, 368, 455-7.

Briet, F., P. Pochart, M. Marteau, B. Flourie, E. Arrigoni and J.C. Rambaud (1997) Improved clinical tolerance to chronic lactose ingestion in subjects with lactose intolerance: a placebo effect? Gut, 41, 632-635.

Brook, J.D., M.E. McCurrach, H.G. Harley, A.J. Buckler, D. Church, H. Aburatani, K. Hunter, V.P. Stanton, J.P. Thirion, and T. Hudson (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member [published erratum appears in Cell 1992 Apr 17;69(2):385], Cell, 68, 799-808.

Brunner, J., H. Hauser, H. Braun, K.J. Wilson, H. Wacker, B. O'Neill and G. Semenza (1979) The mode of association of the enzyme complex sucrase-isomaltase with the intestinal brush border membrane, J. Biol. Chem., 254, 1821-1828.

Buller, H.A., M.J.C. Kothe, D.A. Goldman, S.A. Grubman, W.V. Sasak, P.T. Matsudaira, R.K. Montgomery and R.J. Grand (1990) Coordinate expression of lactase-phlorizin hydrolase mRNA and enzyme levels in rat intestine during development, J. Biol. Chem., 265, 6978-6983.

Cambien, F., O. Poirier, V. Nicaud, S.M. Herrmann, C. Mallet, S. Ricard, I. Behague, V. Hallet, H. Blanc, V. Loukaci, J. Thillet, A. Evans, J.B. Ruidavets, D. Arveiler, G. Luc and L. Tiret (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders, Am. J. Hum. Genet., 65, 183-91.

Carlson, D.P. and J. Ross (1986) Point mutation associated with hereditary persistence of fetal hemoglobin decreases RNA polymerase III transcription upstream of the affected gamma-globin gene, Mol. Cell. Biol., 6, 3278-82

Castiglione, C.M., A.S. Deinard, W.C. Speed, G. Sirugo, H.C. Rosenbaum, Y. Zhang, D.K. Grandy, E.L. Grigorenko, B. Bonne-Tamir, A.J. Pakstis, J.R. Kidd and

K.K. Kidd (1995) Evolution of haplotypes at the DRD2 locus, Am. J. Hum. Genet., 57, 1445-56.

Cavalli-Sforza, L.L., P. Menozzi and A. Piazza (1994) The history and geography of human genes, Princeton University Press, Princeton, New Jersey.

Chandrasena, G., D.E. Osterholm, I. Sunitha and S.J. Henning (1994) Cloning and sequencing of a full-length rat sucrase-isomaltase-encoding cDNA, Gene, 150, 355-360.

Chantret, I., M. Lacasa, G. Chevalier, J. Ruf, I. Islam, N. Mantei, Y. Edwards, D. Swallow and M. Rousset (1992) Sequence of the complete cDNA and the 5' structure of the human sucrase-isomaltase gene. Possible homology with a yeast glucoamylase, Biochem. J., 285, 915-23.

Chomczynski, P. and N. Sacchi (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction, Anal. Biochem., 162, 156-159.

Clark, A.G., K.M. Weiss, D.A. Nickerson, S.L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartianen, M. Perola, E. Boerwinkle and C.F. Sing (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase, Am. J. Hum. Genet., 63, 595-612.

Clebert, J.-P. (1967) The Gypsies, Penguin, London.

Cook, G.C. (1978) Did persistence of intestinal lactase into adult life originate in the Arabian peninsula? Man, 13, 418-427.

Cooper, D.N. and M. Krawczak (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes, Hum. Genet., 83, 181-8.

Dahlqvist, A. (1974) Enzyme deficiency and malabsorption of carbohydrates, in: H.L. Sipple and K.W. McNutt (Eds.), Sugars in Nutrition, Academic Press, New York, pp. 187-214.

Dahlqvist, A., B. Hammond, R. Crane, J. Dunphy and A. Littman (1963) Intestinal lactase deficiency and lactose intolerance in adults: preliminary report., Gastroenterology, 45, 488-491.

Daniels, S.E., S. Bhattacharrya, A. James, N.I. Leaves, A. Young, M.R. Hill, J.A. Faux, G.F. Ryan, P.N. le Souef, L. G.M., A.W. Musk and W.O. Cookson (1996) A genome-wide search for quantitative trait loci underlying asthma, Nature, 383, 247-250.

Das, B.S., D.B. Das, R.N. Satpathy, J.K. Patnaik and T.K. Bose (1988) Riboflavin deficiency and severity of malaria, Eur. J. Clin. Nutr., 42, 277-283.

Deka, R., P.P. Majumder, M.D. Shriver, D.N. Stivers, Y. Zhong, L.M. Yu, R. Barrantes, S.J. Yin, T. Miki, J. Hundrieser, C.H. Bunker, S.T. McGarvey, S. Sakallah, R.E. Ferrell and R. Chakraborty (1996) Distribution and evolution of CTG repeats at the myotonin protein kinase gene in human populations, Genome Res., 6, 142-54.

Devlin, B. and N. Risch (1995) A comparison of linkage disequilibrium measures for fine-scale mapping, Genomics, 29, 311-22.

Diamond, J. (1998) Guns, Germs and Steel: A short history of everybody for the last 13000 years, Random House, London.

Drummond, F., J. Sowden, K. Morrison and Y.H. Edwards (1996) The *caudal*-type homeobox protein Cdx-2 binds to the colon promoter of the carbonic anhydrase 1 gene, Eur. J. Biochem., 236, 670-681.

Dudd, S.N. and R.P. Evershed (1998) Direct demonstration of milk as an element of archaelogical economies, Science, 282, 1478-1481.

Duluc, I., B. Jost and J.-N. Freund (1993) Multiple levels of control of the stage- and region-specific expression of rat intestinal lactase, J. Cell. Biol., 123, 1577-1586.

Ellis, J., K.C. Tan-Un, A. Harper, D. Michalovich, N. Yannoutsos, S. Philipsen and F. Grosveld (1996) A dominant chromatin-opening activity in 5' hypersensitive site 3 of the human beta-globin locus control region, EMBO J., 15, 562-8.

Fajardo, O., H.Y. Naim and S.W. Lacey (1994) The polymorphic expression of lactase in adults is regulated at the mRNA level, Gastroenterology, 106, 1233-1241.

Feldman, M.W. and L.L. Cavalli-Sforza (1989) On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem., in: M.W. Feldman (Ed.), Mathematical Evol. Theory, Princton U P, Princton, N J, pp. Chapter 8.

Ferak, V., D. Sivakova and Z. Sieglova (1987) The Slovak gypsies (Romany)--a population with the highest coefficient of inbreeding in Europe, Bratisl. Lek. Listy., 87, 168-175.

Fitzgerald, K., L. Bazar and M.I. Avigan (1998) GATA-6 stimulates a cell-line specific activation element in the human lactase promoter, Am. J. Physiol., 274, G314-G324.

Flatz, G. (1984) Gene dosage effect on intestinal lactase activity demonstrated *in vivo*, Am. J. Hum. Genet., 36, 306-310.

Flatz, G. (1987) Genetics of lactose digestion in humans, Adv. Hum. Genet., 16, 1-77.

Flatz, G. and H.W. Rotthauwe (1971) Evidence against nutritional adaption to tolerance to lactase., Humangenetik, 13, 118-125.

Flatz, G., C. Saengudom and T. Sanguanbhokai (1969) Lactose intolerance in Thailand, Nature, 221, 758-759.

Foley, K.P., S. Pruzina, J.D. Winick, J.D. Engel, F. Grosveld and P. Fraser (1994) The chicken beta/epsilon-globin enhancer directs autonomously regulated, high-level expression of the chicken epsilon-globin gene in transgenic mice, Proc. Natl. Acad. Sci. U S A, 91, 7252-6.

Forde, C.D. (1934) Habitat, economy and society, Methuen, London.

Forster, J.R. (1767) *Specimen Historiae Naturalis Volgensis*, Phil. Trans. Roy. Soc., 57, 346.

Freund, J.-N., B. Jost, O. Lorentz and I. Duluc (1997) Identification of homologues of the mammalian intestinal lactase gene in non-mammals (birds and molluscs), Biochem. J., 322, 491-498.

Ghersa, P., P. Huber, G. Semenza and H. Wacker (1986) Cell-free synthesis, membrane integration and glycosylation of pro-sucrase-isomaltase, J. Biol. Chem., 261, 7969-7974.

Gimbutas, M. (1970) Proto-Indo-European culture: The Kurgan culture during the fifth, fourth and third millenia BC, in: G. Cardona, H.G. Hoenigswald and A. Senn (Eds.), Indo-European and Indo-Europeans, University of Pennsylvania Press, Philadelphia.

Gingrich, J.C., D.M. Boehrer, J.A. Garnes, W. Johnson, B.S. Wong, A. Bergmann, G.G. Eveleth, R.G. Langlois and A.V. Carrano (1996) Construction and characterization of human chromosome 2-specific cosmid, fosmid, and PAC clone libraries, Genomics, 32, 65-74.

Gonzalez, G.A. and M.R. Montminy (1989) Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133, Cell, 59, 675-80.

Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnel and C.P. Groves (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence, Mol. Phylogenet. Evol., 9, 585-598.

Grabnitz, F., M. Seiss, K.P. Rucknagel and W. Stauden Bauer (1991) Structure of the b-glucosidase gene bgl A of *Clostridium thermocellase,* Eur. J. Biochem., 200, 301-309.

Green, F.R., P. Greenwell, L. Dickson, B. Griffiths, J. Noades and D.M. Swallow (1988) Expression of the ABH, Lewis and related antigens on the glycoproteins of the human jejunal brush border, in: J.R. Harris (Ed.), Subcellular Biochemistry, Plenum Press, New York, pp. 119-153.

Grosveld, F., M. Antoniou, M. Berry, E. De Boer, N. Dillon, J. Ellis, P. Fraser, O. Hanscombe, J. Hurst, A. Imam and et al. (1993) The regulation of human globin gene switching, Philos. Trans. R. Soc. Lond. B Biol. Sci., 339, 183-91.

Gudmand-Hoyer, E., H.J. Fenger, P. Kern-Hansen and P.R. Madsen (1987) Sucrase deficiency in Greenland: Incidence and genetic aspects, Scand. J. Gastroenterol., 22, 24-28.

Gyllensten, U.B. and H.A. Erlich (1989) Ancient roots for polymorphism at the HLA-DQ alpha locus in primates, Proc. Natl. Acad. Sci. USA, 86, 9986-9990.

Habte, D., G. Sterky and B. Hjaalmarsson (1973) Lactose malabsorption in Ethiopian children, Acta Paediatr. Scand., 62, 649-654.

Hammer, M. (1995) A recent common ancestry for human Y chromosomes, Nature, 378, 376-378.

Hammer, M.F., T. Karafet, A. Rasanayagam, E.T. Wood, T.K. Altheide, T. Jenkins, R.C. Griffiths, A.R. Templeton and S.L. Zegura (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation, Mol. Biol. Evol., 15, 427-41.

Harding, R.M., S.M. Fullerton, R.C. Griffiths, J. Bond, M.J. Cox, J.A. Schneider, D.S. Moulin and J.B. Clegg (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans, Am. J. Hum. Genet., 60, 772-789.

Harpending, H.C., M.A. Batzer, M. Gurven, L.B. Jorde, A.R. Rogers and S.T. Sherry (1998) Genetic traces of ancient demography, Proc. Natl. Acad. Sci. U S A, 95, 1961-7.

Hartl, D.L. and A.G. Clark (1989) Principles of population genetics. Sinuar Associates, Sunderland, Massachusetts.

Harvey, C.B. (1994) The biochemical and genetical analysis of lactase phlorizin hydrolase: with specific reference to the lactase persistence/ non-persistence polymorphism in man, Genetics and Biometry, University of London, London.

Harvey, C.B., M.F. Fox, P.A. Jeggo, N. Mantei, S. Povey and D.M. Swallow (1993) Regional localization of the lactase-phlorizin hydrolase gene, LCT, to chromosome 2q21, Ann. Hum. Genet., 57, 179-185.

Harvey, C.B., E.J. Hollox, M. Poulter, Y. Wang, M. Rossi, S. Auricchio, T.H. Iqbal, B.T. Cooper, R. Barton, M. Sarner, R. Korpela and D.M. Swallow (1998) Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism, Ann. Hum. Genet., 62, 215-223.

Harvey, C.B., W. Pratt, I. Islam, D.B. Whitehouse and D.M. Swallow (1995) DNA polymorphisms in the lactase gene: linkage disequilibrium across the 70kb region, Eur. J. Hum. Genet., 3, 27-41.

Harvey, C.B., Y. Wang, D. Darmoul, A. Phillips, N. Mantei and D.M. Swallow (1996) Characterisation of a human homologue of a yeast cell division cycle gene, MCM6, located adjacent to the 5' end of the lactase gene on chromosome 2q21, FEBS Letts., 398, 135-140.

Harvey, C.B., Y. Wang, L.A. Hughes, D.M. Swallow, W.P. Thurrell, V.R. Sams, R. Barton and M. Sarner (1994) Studies on the expression of intestinal lactase in different individuals, Gut, 36, 28-33.

Harvey, P.H., R.D. Martin and T.H. Clutton-Brock (1987) Life histories in comparative perspective, in: B.B. Smuts, D.L. Cheney, R.M. Seyfarth, R.W. Wrangham and T.T. Struhsaker (Eds.), Primate Societies, University of Chicago Press, Chicago, London, pp. 181-196.

Hastbacka, J., A. de la Chapelle, M.M. Mahtani, G. Clines, M.P. Reeve-Daly, M. Daly, B.A. Hamilton, K. Kusumi, B. Trivedi, and A. Weaver (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping, Cell, 78, 1073-87.

Hecht, A., C.F. Torbey, H.A. Korsmo and W.A. Olsen (1997) Regulation of sucrase and lactase in developing rats: role of nuclear factors that bind to two gene regulatory elements, Gastroenterology, 112, 803-812.

Hiernaux, J. (1968) Bantu expansion: the evidence from physical anthropology confronted with linguistic and archaeological evidence, J. Afr. Hist., 9, 505-515.

Ho, M.W., S. Povey and D.M. Swallow (1982) Lactase polymorphism in adult British natives: estimating allele frequencies by enzyme assays in autopsy samples, Am. J. Hum. Genet., 34, 650-657.

Holden, C. and R. Mace (1997) Phylogenetic analysis of the evolution of lactase digestion in adults, Hum. Biol., 69, 605-628.

Hollox, E.J. and D.M. Swallow (2000) Lactase deficiency -Biological and medical aspects of the human adult lactase polymorphism Chapter 17., in: K.e. al (Ed.), Genetic basis of common diseases, Oxford University Press, Oxford.

Honkanen, R., H. Kroger, E. Alhava, P. Turpeinen, M. Tuppurainen and S. Saarkoski (1997) Lactose intolerance associated with fractures of weight-bearing bones, Bone, 21, 473-477.

Hore, P. and M. Messer (1968) Studies on disaccharidase activities of the small intestine of the domestic cat and other carnivorous mammals, Comp. Biochem. Physiol., 24, 717-725.

Horowitz, M., J. Wishart, L. Mundy and C. Nordin (1987) Lactose and calcium absorption in postmenopausal osteoporosis, Arch. Intern. Med., 147, 534-536.

Howell, J.N., T. Schockenhoff and G. Flatz (1981) Population screening for the adult lactase phenotypes with a multiple breaths version of the breath hydrogen test, Hum. Genet., 57, 276-278.

Huang, M.-M., N. Arnheim and M.F. Goodman (1992) Extension of base mispairs by *Taq* DNA polymerase: implications for single nucleotide discrimination in PCR, Nucleic Acids Res., 20, 4567-4573.

Hunziker, W., M. Spiess, G. Semenza and H.F. Lodish (1986) The sucrase-isomaltase complex: primary structure, membrane-orientation, and evolution of a stalked, intrinsic brush border protein, Cell, 46, 227-234.

Isahara, R., S. Taketani, M. Sasai-Takedatsu, M. Kino, R. Tokunaga and Y. Kobayashi (1997) Molecular cloning, sequencing and expression of cDNA encoding human trehalase, Gene, 202, 69-74.

Ito, T., Y. Hayashi, S. Ohmori, S. Oda and H. Seo (1998) Molecular cloning of sucrase-isomaltase cDNA in the house musk shrew *Suncus murinus* and identification of a mutation responsible for isolated sucrase deficiency, J. Biol. Chem., 273, 16464-16469.

Jacob, R., C. Brewer, J.A. Fransen and H.Y. Naim (1994) Transport, function, and sorting of lactase-phlorizin hydrolase in Madin-Darby canine kidney cells, J. Biol. Chem., 269, 2712-2721.

Jarvela, I., N.S. Enattah, J. Kokkonen, T. Varilo, E. Savilahti and L. Peltonen (1998) Assignment of the locus for congenital lactase deficiency to 2q21, in the vicinity of but separate from the lactase-phlorizin hydrolase gene, Am. J. Hum. Genet., 63, 1078-1085.

Jeganathan, D., M.F. Fox, J.M. Young, J.R.W. Yates, J.P. Osborne and S. Povey (2000) Nonsense-mediated mRNA decay - useful to the geneticist as well as the cell?, in preparation.

Jenkins, T. (1982) Human evolution in southern Africa, in: B. Bonne-Tamir, Cohen, T., Goodman, R.M. (Eds.), Human Genetics - Part A - The Unfolding Genome, Alan R. Liss, New York.

Jersky, J. and R.H. Kinsley (1967) Lactose deficiency in the South African Bantu, S. Afr. Med. J., 41, 1194-6.

Jost, B., I. Duluc, M. Richardson, R. Lathe and J.-N. Freund (1997) Functional diversity and interactions between the repeat domains of rat intestinal lactase, Biochem. J., 327, 95-103.

Jussila, J. (1969) Diagnosis of lactose malabsoption by the lactose tolerance test with peroral ethanol administration, Scand. J. Gastroenterology, 4, 361-368.

Kaessmann, H., F. Heisig, A. von Haeseler and S. Paabo (1999a) DNA sequence variation in a non-coding region of low recombination on the human X chromosome, Nat. Genet., 22, 78-81.

Kaessmann, H., V. Wiebe and S. Paabo (1999b) Extensive nuclear DNA sequence diversity among chimpanzees, Science, 286, 1159-1162.

Kalaydjieva, L., A. Perez-Lezaun, D. Angelicheva, S. Onengut, D. Dye, N.U. Bosshard, A. Jordanova, A. Savov, P. Yanakiev, I. Kremensky, B. Radeva, J. Hallmayer, A. Markov, V. Nedkova, I. Tournev, L. Aneva and R. Gitzelmann (1999) A Founder Mutation in the GK1 Gene Is Responsible for Galactokinase Deficiency in Roma (Gypsies), Am. J. Hum. Genet., 65, 1299-1307.

Kanbe, M. (1992) Traditional fermented milks of the world, in: Y. Nakazawa and A. Hosono (Eds.), Functions of fermented milk, Elsevier, London.

Kapitonov, V. and J. Jurka (1996) The age of Alu subfamilies, J. Mol. Evol., 42, 59-65.

Kato, G.J. (1999) Human genetic diseases of proteolysis. Hum. Mutat. 13, 87-98

Kerry, K.R. (1969) Intestinal disaccharidase activity in a monotreme and eight species of marsupials (with an added note on the disaccharidases of five species of sea birds), Comp. Biochem. Physiol., 29, 1015-1022.

Keusch, G.T., F.J. Troncale, L.H. Miller, V. Promadhat and P.R. Anderson (1969a) Acquired lactose malabsorption in Thai children, Pediatrics, 43, 540-545.

Kidd, K.K., B. Morar, C.M. Castiglione, H. Zhao, A.J. Pakstis, W.C. Speed, B. Bonne-Tamir, R.B. Lu, D. Goldman, C. Lee, Y.S. Nam, D.K. Grandy, T. Jenkins and J.R. Kidd (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus, Hum. Genet., 103, 211-27.

Kittles, R.A., M. Perola, L. Peltonen, A.W. Bergen, R.A. Aragon, M. Virkkunen, M. Linnoila, D. Goldman and J.C. Long (1998) Dual origins of Finns revealed by Y chromosome haplotype variation, Am. J. Hum. Genet., 62, 1171-9.

Kleinjan, D.J. and V. van Heyningen (1998) Position effect in human genetic disease, Hum. Mol. Genet., 7, 1611-8.

Kocian, J., I. Skala and K. Bakos (1973) Calcium absorption from milk and lactose free milk in healthy subjects and patients with lactase deficiency, Digestion, 9, 317-324.

Kolars, J.C., M.D. Levitt, M. Aouji and D.A. Savaiano (1984) Yogurt - an autodigestive source of lactase, N. Engl. J. Med., 310, 1-3.

Krasinski, S.D., G. Estrada, K.Y. Yeh, M. Yeh, P.G. Traber, E.H. Rings, H.A. Buller, M. Verhave, R.K. Montgomery and R.J. Grand (1994) Transcriptional regulation of intestinal hydrolase biosynthesis during postnatal development in rats, Am. J. Physiol., 267, G584-G594.

Krasinski, S.D., B.H. Upchurch, S.J. Irons, R.M. June, K. Mishra, R.J. Grand and M. Verhave (1997) Rat lactase-plorizin hydrolase/human growth hormone transgene is expressed on small intestinal villi in transgenic mice, Gastroenterology, 113, 844-855.

Krichevsky, A.M., E. Metzer and H. Rosen (1999) Translational control of specific genes during differentiation of HL-60 cells, J. Biol. Chem., 274, 14295-305.

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes, Nat. Genet., 22, 139-44.

Kruse, T.A., L. Bolund, K.-H. Grzeschik, H.H. Ropers, H. Sjostrom, O. Noren, N. Mantei and G. Semenza (1988) The human lactase-phlorizin hydrolase gene is located on chromosome 2, FEBS Letts., 240, 123-126.

Kuro-o, M., Y. Matsumura, H. Aizawa, H. Kawaguchi, T. Suga, T. Utsugi, Y. Ohyama, M. Kurabayashi, T. Kamane, E. Kume, H. Iwasaki, A. Iida, T. Shika-Iida, S. Nishikawa, R. Nagai and Y. Nabeshima (1997) Mutation of the mouse *klotho* gene leads to a syndrome resembling aging, Nature, 390, 45-51.

Laroia, G., R. Cuesta, G. Brewer and R.J. Schneider (1999) Control of mRNA decay by heat shock-ubiquitin-proteasome pathway, Science, 284, 499-502.

Lerman, L.S. and K. Silverstein (1987) Computational simulation of DNA melting and its application to denaturing gradient gel electrophoresis, Methods in Enzymol., 155, 482-501.

Li, W.-H. (1997) Molecular Evolution, Sinauer Associates, Sunderland, Massachusetts.

Lloyd, M., G. Mevissen, M. Fischer, W. Olsen, D. Goodspeed, M. Genini, W. Boll, G. Semenza and N. Mantei (1992) Regulation of intestinal lactase in adult hypolactasia, J. Clin. Invest., 89, 524-529.

MacDougall, C., D. Harbison and M. Bownes (1995) The developmental consequences of alternate splicing in sex determination and differentiation in *Drosophila*, Dev. Biol., 172, 353-76.

Madzarovova-Nohejlova, J. (1982) Small bowel disaccharidase activity in Czech population and in Gipsy [sic] population living in West Bohemia (abstract 1106), 7th Congress, Organisation Mondiale de Gastroenterologie, Stockholm.

Maiuri, L., M. Rossi, V. Raia, S. D'Auria, D. Swallow, A. Quaroni and A. Auricchio (1992) Patchy expression of lactase protein in adult rabbit and rat intestine, Gastroenterology, 103, 1739-1746.

Maiuri, L., M. Rossi, V. Raia, V. Garipoli, D. Swallow, L. Hughes, O. Noren and S. Auricchio (1994) Mosaic regulation of lactase in human adult-type hypolactasia, Gasteroenterology, 107, 54-60.

Mantei, N., M. Villa, T. Enzler, H. Wacker, W. Boll, P. James, W. Hunziker and G. Semenza (1988) Complete primary structure of human and rabbit lactase-phlorizin hydrolase: implications for biosynthesis, membrane anchoring and evolution of the enzyme, EMBO J., 7, 2705-2713.

March, R.E., W. Putt, M. Hollyoake, J.H. Ives, J.U. Lovegrove, D.A. Hopkinson, Y.H. Edwards and D.B. Whitehouse (1993) The classical human phosphoglucomutase (PGM1) isozyme polymorphism is generated by intragenic recombination, Proc. Natl. Acad. Sci. USA, 90, 10730-10733.

Markowitz, A.J., G.D. Wu, A. Bader, Z. Cui, L. Chen and P.G. Traber (1995) Regulation of lineage-specific transcription of the sucrase-isomaltase gene in transgenic mice and cell lines, Am. J. Physiol., 269, G925-39.

Matunis, M.J., W.M. Michael and G. Dreyfuss (1992) Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein, Mol. Cell. Biol., 12, 164-71.

McNair, A., E. Gudman Hoyer, S. Jarnum and L. Orrild (1972) Sucrose malabsorption in Greenland, Brit. Med. J., 2, 19-21.

Meloni, T., C. Colombo, G. Ruggiu, M. Dessena and G.F. Meloni (1998) Primary lactase deficiency and past malarial endemicity in Sardinia, Ital. J. Gastroenterol. Hepatol., 30, 490-493.

Merriman, T., R. Twells, M. Merriman, I. Eaves, R. Cox, F. Cucca, P. McKinney, J. Shield, D. Baum, E. Bosi, P. Pozzilli, L. Nistico, R. Buzzetti, G. Joner, K. Ronningen, E. Thorsby, D. Undlien, F. Pociot, J. Nerup, S. Bain, A. Barnett and J. Todd (1997) Evidence by allelic association-dependent methods for a type 1 diabetes polygene (IDDM6) on chromosome 18q21, Hum. Mol. Genet., 6, 1003-10.

Metneki, J., A. Cziezel, S.D. Flatz and G. Flatz (1984) A study of lactose absorption capacity in twins, Hum. Genet., 67, 296-300.

Metz, G., M.A. Gassul, A.R. Leeds, L.M. Blendis and D.J.A. Jenkins (1976) A simple method for measuring breath hydrogen in carbohydrate malabsorption by end expiratory sampling, Clin. Sci., 50, 237-240.

Mian, I.S. (1998) Sequence, structural, functional, and phylogenetic analyses of three glycosidase families, Blood Cells, Mol., and Dis., 24, 83-100.

Michalet, X., R. Ekong, F. Fougerousse, S. Rousseaux, C. Schurra, N. Hornigold, M. Van Slegtenhorst, J. Wolfe, S. Povey, J.S. Beckmann and A. Bensimon (1997) Dynamic molecular combing: stretching the whole human genome for high-resolution studies, Science, 277, 1518-1523.

Mitchelmore, C., J.T. Troelsen, H. Sjostrom and O. Noren (1998) The HOXC11 homeodomain protein interacts with the lactase-phlorizin hydrolase promoter and stimulates HNF1a-dependent transcription, J. Biol. Chem., 273, 13297-13306.

Montgomery, R.K., H.A. Buller, E.H.H.M. Rings and R.J. Grand (1991) Lactose intolerance and the genetic regulation of intestinal lactase-phlorizin hydrolase, FASEB J., 5, 2824-2832.

Naim, H., E.E. Sterchi and M.J. Lentze (1988a) Structure, biosynthesis and glycosylation of human small intestinal maltase-glucoamylase, J. Biol. Chem., 263, 19709-19717.

Naim, H.Y., R. Jacob, H. Naim, J.F. Sambrook and M.-J.H. Gething (1994) The pro-region of human intestinal lactase-phlorizin hydrolase, J. Biol. Chem., 269, 26933-26943.

Naim, H.Y. and M.J. Lentze (1992) Impact of O-glycosylation on the function of human intestinal lactase-phlorizin hydrolase: characterization of glycoforms varying in enzyme activity and localization of O-glycoside addition, J. Biol. Chem., 267, 25494-25504.

Naim, H.Y., J. Roth, E.E. Sterchi, M. Lentze, P. Milla, J. Schmitz and H.P. Hauri (1988b) Sucrase-isomaltase deficiency in humans. Different mutations disrupt intracellular transport, processing, and function of an intestinal brush border enzyme, J. Clin. Invest., 82, 667-679.

Nan, X., P. Tate, E. Li and A. Bird (1996) DNA methylation specifies chromosomal localization of MeCP2, Mol. Cell. Biol., 16, 414-21.

Neel, H., P. Gondran, D. Weil and F. Dautry (1995) Regulation of pre-mRNA processing by *src*, Curr. Biol., 5, 413-22.

Neele, A.M., A.W. Einerhand, J. Dekker, H.A. Buller, J.N. Freund, M. Verhave, R.J. Grand and R.K. Montgomery (1995) Verification of the lactase site of rat lactase-phlorizin hydrolase by site-directed mutagenesis, Gastroenterology, 109, 1234-1240.

Nei, M. and N. Saitou (1986) Genetic relationship of human populations and ethnic differences in relation to drugs and food., in: W. Kalow, H.W. Goedde and D.P. Agarwal (Eds.), Ethnic Differences in Reactions to Drugs and other xenobiotics, Alan R Liss, New York, pp. 21-37.

Newcomer, A.D., S.F. Hodgson, D.B. McGill and P.J. Thomas (1978) Lactase deficiency: prevalence in osteoporosis, Ann. Intern. Med., 89, 218-220.

Newcomer, A.D. and D.B. McGill (1966) Distribution of disaccharidase activity in the small bowel of normal and lactase deficient subjects, Gastroenterology, 51, 481-488.

Nichols, B.L., M.A. Dudley, V.N. Nichols, M. Putnam, S.E. Avery, J.K. Fraley, A. Quaroni, M. Shiner and F.R. Carrazza (1997) Effects of malnutrition on expression and activity of lactase in children, Gastroenterology, 112, 742-751.

Nichols, B.L., J. Eldering, S. Avery, D. HAhn, A. Quaroni and E. Sterchi (1998) Human small intestinal maltase gluocoamylase cDNA cloning homology to sucrase isomaltase, J. Biol. Chem., 273, 3079-3081.

Nickerson, D.A., S.L. Taylor, K.M. Weiss, A.G. Clark, R.G. Hutchinson, J. Stengard, V. Salomaa, E. Vartiainen, E. Boerwinkle and C.F. Sing (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene, Nat. Genet., 19, 233-40.

Nurminsky, D.I., M.V. Nurminskaya, D. De Aguiar and D.L. Hartl (1998) Selective sweep of a newly evolved sperm-specific gene in Drosophila, Nature, 396, 572-5.

Ober, C., N.J. Cox, M. Abney, A. Di Rienzo, E.S. Lander, B. Changyaleket, H. Gidley, B. Kurtz, J. Lee, M. Nance, A. Pettersson, J. Prescott, A. Richardson, E. Schlenker, E. Summerhill, S. Willadsen and R. Parry (1998) Genome-wide search for asthma susceptibility loci in a founder population, Hum. Mol. Genet., 7, 1393.

Oberholzer, T., N. Mantei and G. Semenza (1993) The pro sequence of lactase-phlorizin hydrolase is required for the enzyme to reach the plasma membrane: an intramolecular chaperone?, FEBS Lett., 333, 127-131.

O'Connor, T.P. and J. Diamond (1999) Ontogeny of intestinal safety factors: lactase capacities and lactose loads, Am. J. Physiol., 276, R753-765.

Oesterreicher, T.J., N.N. Nanthakumar, J.H. Winston and S.J. Henning (1998) Rat trehalase: cDNA cloning and mRNA expression in adult rat tissues and during intestinal ontogeny, Am. J. Physiol., 274, R1220-R1227.

Ott, J. (1991) Analysis of human genetic linkage, The Johns Hopkins University Press, Baltimore and London.

Ouwendijk, J., W.J. Peters, C.P. Hollenberg, L.A. Ginsel, J.A. Fransen and H.Y. Naim (1996) Congenital sucrase-isomaltase deficiency. Identification of a glutamine to proline substitution that leads to a transport block of sucrase-isomaltase in a pre-Golgi compartment, J. Clin. Invest., 93, 633-641.

Penny, D., M. Steel, P.J. Waddell and M.D. Hendy (1995) Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*, Mol. Biol. Evol., 12, 863-82.

Plimmer, R.H.A. (1906) On the presence of lactase in the intestines of animals and on the adaptation of the intestine to lactose, J. Physiol., London, 35, 20-31.

Poggi, V. and G. Sebastio (1991) Molecular analysis of the lactase gene in the congenital lactase deficiency, Am. J. Hum. Genet. Suppl., 49, 105.

Potter, J. and H. Bolton (1982) Adult human lactase : purification and properties, Manuscript.

Potter, J., M.-W. Ho, H. Bolton, A.J. Furth, D.M. Swallow and B. Griffiths (1985) Human lactase and the molecular basis of lactase persistence, Biochem. Genet., 23, 423-439.

Ransome-Kuti, O. (1977) Lactose intolerance - a review, Post. Grad. Med. J., 53, 73-87.

Razin, A. (1998) CpG methylation, chromatin structure and gene silencing-a three-way connection, EMBO J, 17, 4905-8.

Renfrew, C. (1987) Archaeology and language: the puzzle of Indo-European origins, Jonathan Cape, London.

Rieder, M.J., S.L. Taylor, A.G. Clark and D.A. Nickerson (1999) Sequence variation in the human angiotensin converting enzyme, Nat. Genet., 22, 59-62.

Rings, E.H.H.M., P.A.J. d Boer, A.F.M. Moorman, E.H. van Beers, J. Dekker, R.K. Montgomery, R.J. Grand and H.A. Buller (1992) Lactase gene expresssion during early development of rat small intestine, Gastroenterology, 103, 1154-1161.

Rogers, A.R. and L.B. Jorde (1995) Genetic evidence on modern human origins, Hum. Biol., 67, 1-36.

Rossi, M., L. Mauiri, M.I. Fusco, V.M. Salvati, A. Fuccio, S. Auricchio, N. Mantei, L. Zecca, S.M. Gloor and G. Semenza (1997) Lactase persistence versus decline in human adults: multifactorial events are involved in downregulation after weaning, Gastroenterology, 112, 1506-1514.

Ruf, J., H. Wacker, P. James, M. Maffia, P. Seiler, G. Galand, A. v Kieckebusch, G. Semenza and N. Mantei (1990) Rabbit small intestinal trehalase: purification, cDNA cloning, expression and verification of glycosylphosphatidylinositol anchoring, J. Biol. Chem., 265, 15034-15039.

Ruhlen, M. (1987) A guide to the world's languages, Stanford University Press, Stanford, California.

Sabol, S.Z., S. Hu and D. Hamer (1998) A functional polymorphism in the monoamine oxidase A gene promoter, Hum. Genet., 103, 273-9.

Sahi, T. (1974) The inheritance of selective adult-type lactose malabsorption, Scand. J. Gastroenterol., 9, 1-73.

Sahi, T., M. Isokoski, J. Jussila, K. Launiala and K. Pyorala (1973) Recessive inheritance of adult-type lactose malabsorption, Lancet, 823-826.

Sahi, T. and K. Launiala (1978) Manifestation and occurence of selective adult-type lactose malabsorption in Finnish teenagers. A follow-up study., Dig. Dis. Sci., 23, 699-704.

Sahi, T., K. Launiala and H. Laitinen (1983) Hypolactasia in a fixed cohort of young Finnish adults, a follow-up study, Scand. J. Gastroenterol., 18, 865-870.

Sajantila, A., A.H. Salem, P. Savolainen, K. Bauer, C. Gierig and S. Paabo (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population, Proc. Natl. Acad. Sci. U S A, 93, 12035-9.

Sambrook, J., E.F. Fritsch and T. Maniatis (1989) Molecular cloning: a laboratory manual, Cold Spring Harbor Laboratory Press.

Savaiano, D.A., A. Abou El Anouar, D.E. Smith and M.D. Levitt (1984) Lactose malabsorption from yogurt, pastuerized yogurt, sweet acidophilus milk, and cultured milk in lactase-deficient individuals, Am. J. Clin. Nutr., 40, 1219-1223.

Sawcer, S., H.B. Jones, R. Feakes, J. Gray, N. Smaldon, J. Chataway, N. Robertson, D. Clayton, P.N. Goodfellow and A. Compston (1996) A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22 [see comments], Nat. Genet., 13, 464-8.

Scozzari, R., F. Cruciani, P. Santolamazza, P. Malaspina, A. Torroni, D. Sellitto, B. Arredi, G. Destro-Bisol, G. De Stefano, O. Rickards, C. Martinez-Labarga, D. Modiano, G. Biondi, P. Moral, A. Olckers, D.C. Wallace and A. Novelletto (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations, Am. J. Hum. Genet., 65, 829-46.

Sebastio, G., W. Hunziker, A. Ballabio, S. Auricchio and G. Semenza (1986) On the primary site of control in the spontaneous development of small-intestinal sucrase-isomaltase after birth [published erratum appears in FEBS Lett. 1987 Nov 2;223(2):423], FEBS Lett., 208, 460-4.

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

Seielstad, M., E. Bekele, M. Ibrahim, A. Toure and M. Traore (1999) A view of modern human origins from Y chromosome microsatellite variation, Genome Res., 9, 558-67.

Sheratt, A. (1981) Plough and pastoralism: Aspects of the secondary products revolution, in: I.e.a. Hodder (Ed.), Pattern of the Past, Cambridge University Press, Cambridge, pp. 261-305.

Simoons, F. (1982) A geographic approach to senile cataracts. Possible links with milk consumption, lactase activity and galactose metabolism., Dig. Dis. Sci., 27, 257-264.

Simoons, F.J. (1971) The antiquity of dairying in Asia and Africa, The Geographical Review, 61, 431-439.

Sjostrom, H., O. Noren, L. Christiansen, H. Wacker and G. Semenza (1980) A fully active, two-active-site, single chain sucrase-isomaltase from pig small intestine: implications for the biosynthesis of a mammalian integral stalked membrane protein, J. Biol. Chem., 255, 11332-11338.

Skovbjerg, H. (1981) Immunoelectrophoretic studies on human small intestinal brush border proteins - the longitudinal distribution of peptidases and disaccharidases, Clin. Chem. Acta, 112, 205-212.

Skovbjerg, H., E. Gudmand-Hoyer and H. Fenger (1980) Immunoelectrophoretic studies on human small intestinal brush border proteins - amount of lactase protein in adult-type hypolactasia, Gut, 21, 360-364.

Skovbjerg, H., H. Sjostrom and O. Noren (1981) Purification and characterisation of amphiphilic lactase-phlorizin hydrolase from human small intestine, Eur. J. Biochem., 114, 653-661.

Smit, A.F.A. and P. Green http://ftp.genome.washington.edu/RM/RepeatMasker.html.

Spencer, N., D.A. Hopkinson and H. Harris (1964) Phosphoglucomutase polymorphism in man, Nature, 204, 742-745.

Spiro, M.J., V.D. Bhoyroo and R.G. Spiro (1997) Molecular cloning and expression of rat liver endo-alpha-mannosidase, an N-linked oligosaccharide processing enzyme, J. Biol. Chem., 272, 29356-63.

Spodsberg, N., J.T. Troelsen, P. Carlsson, S. Enerback, H. Sjostrom and O. Noren (1999) Transcriptional regulation of pig lactase-phlorizin hydrolase. Involvement of HNF-1 and FREACs, Gastroenterology, 116, 842-854.

Stallings-Mann, M.L., R.L. Ludwiczak, K.W. Klinger and F. Rottman (1996) Alternative splicing of exon 3 of the human growth hormone receptor is the result of an unusual genetic polymorphism, Proc. Natl. Acad. Sci. USA, 93, 12394-9.

Stecher, P.G. (1968) The Merck Index, Merck & Co., Rahway, New Jersey.

Stephens, J.C., D.E. Reich, D.B. Goldstein, H.D. Shin, M.W. Smith, M. Carrington, C. Winkler, G.A. Huttley, R. Allikmets, L. Schriml, B. Gerrard, M. Malasky, M.D. Ramos, S. Morlot, M. Tzetis, C. Oddoux, F.S. di Giovine, G. Nasioulas, D. Chandler, M. Aseev, M. Hanson, L. Kalaydjieva, D. Glavac, P. Gasparini and M. Dean (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes, Am. J. Hum. Genet., 62,1507-1515.

Suh, E., L. Chen, J. Taylor and P.G. Traber (1994) A homeodomain protein related to caudal regulates intestine-specific gene transcription, Mol. Cell. Biol., 14, 7340-51.

Swallow, D.M. and C.B. Harvey (1993) Genetics of adult-type hypolactasia, Dyn. Nutr. Res., 3, 1-7.

Swallow, D.M. and E.J. Hollox (2000) The genetic polymorphism of intestinal lactase activity in adult humans, in: C.R. Scriver, A.L. Beaudet, W.S. Sly and D. Valle (Eds.), The Metabolic and Molecular Bases of Inherited Disease, McGraw-Hill, New York.

Takahata, N. and Y. Satta (1997) Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences, Proc. Natl. Acad. Sci. U S A, 94, 4811-5.

Tanaka, T., S. Takase and T. Goda (1997) A possible role of a nuclear factor NF-LPH1 in the regional expression of lactase-phlorizin hydrolase along the small intestine, J. Nutr. Sci. Vitaminol. (Tokyo), 43, 565-573.

Tanimoto, K., Q. Liu, J. Bungert and J.D. Engel (1999) Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice, Nature, 398, 344-8.

Tavare, S., D.J. Balding, R.C. Griffiths and P. Donnelly (1997) Inferring coalescence times from DNA sequence data, Genetics, 145, 505-18.

Templeton, A.R. (1997) Out of Africa? What do genes tell us?, Curr. Opin. Genet. Dev., 7, 841-7.

Terwilliger, J.D. and J. Ott (1994) Handbook of human genetic linkage, Johns Hopkins University Press, Baltimore.

Terwilliger, J.D. and K.M. Weiss (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality?, Curr. Opin. Biotechnol., 9, 578-94.

Tishkoff, S.A., E. Dietzsch, W. Speed, A.J. Pakstis, J.R. Kidd, K. Cheung, B. Bonne-Tamir, A.S. Santachiara-Benerecetti, P. Moral and M. Krings (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins, Science, 271, 1380-7.

Tishkoff, S.A., A. Goldman, F. Calafell, W.C. Speed, A.S. Deinard, B. Bonne-Tamir, J.R. Kidd, A.J. Pakstis, T. Jenkins and K.K. Kidd (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations, Am. J. Hum. Genet., 62, 1389-402.

Tomoda, K., Y. Kubota and J. Kato (1999) Degradation of the cyclin-dependent-kinase inhibitor p27Kip1 is instigated by Jab1 [see comments], Nature, 398, 160-5.

Torp, N., M. Rossi, J.T. Troelsen, J. Olsen and E.M. Danielsen (1993) Lactase-phlorizin hydrolase and aminopeptidase N are differentially regulated in the small intestine of the pig, Biochem. J., 295, 177-182.

Tournamille, C., Y. Colin, J.P. Cartron and C. Le Van Kim (1995) Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals, Nat. Genet., 10, 224-8.

Tournev, I., L. Kalaydjieva, B. Youl, B. Ishpekova, V. Guergueltcheva, O. Kamenov, M. Katzarova, Z. Kamenov, M. Raicheva-Terzieva, R.H. King, K. Romanski, R. Petkov, A. Schmarov, G. Dimitrova, N. Popova, M.

Uzunova, S. Milanov, J. Petrova, Y. Petkov, G. Kolarov, L. Aneva, O. Radeva and P.K. Thomas (1999) Congenital cataracts facial dysmorphism neuropathy syndrome, a novel complex genetic disease in Balkan Gypsies: clinical and electrophysiological observations, Ann. Neurol., 45, 742-50.

Troelsen, J.T., A. Mehlum, J. Olsen, N. Spodsberg, G.H. Hansen, H. Prydz, O. Noren and H. Sjostrom (1994a) 1 kb of the lactase-phlorizin hydrolase promoter directs post-weaning decline and small intestinal-specific expression in transgenic mice, FEBS Lett., 342, 291-196.

Troelsen, J.T., C. Mitchelmore, N. Spodsberg, A.M. Jensen, O. Noren and H. Sjostrom (1997) Regulation of lactase-phlorizin hydrolase gene expression by the caudal-related homoeodomain protein Cdx-2, Biochem. J., 322, 833-838.

Troelsen, J.T., J. Olsen, C. Mitchelmore, G.H. Hansen, H. Sjostrom and O. Noren (1994b) Two intestinal specific nuclear factors binding to the lactase-phlorizin hydrolase and sucrase-isomaltase promoters are functionally related oligomeric molecules, FEBS Lett., 342, 297-301.

Tuan, D., M. J. Murnane, J. K. de Riel and B.G. Forget (1980) Heterogeneity in the molecular basis of hereditary persistence of fetal haemoglobin, Nature, 285, 335-357.

Tung, J., A.J. Markowitz, D.G. Silberg and P.G. Traber (1997) Developmental expression of SI is regulated in transgenic mice by an evolutionarily conserved promoter, Am. J. Physiol., 273, G83-92.

Uni, Z. (1998) Identification and isolation of chicken sucrase-isomaltase cDNA sequence, Poult. Sci., 77, 140-144.

Veitch, A.M., P. Kelly, I. Segal, S.K. Soies and M.J. Farthing (1998) Does sucrase deficiency in black South Africans protect against colonic disease? Lancet, 351, 183.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A.C. Wilson (1991) African populations and the evolution of human mitochondrial DNA, Science, 253, 1503-7.

Wacker, H., P. Keller, R. Falchetto, G. Legler and G. Semenza (1992) Location of the two catalytic sites in intestinal lactase phlorizin hydrolase: comparison with sucrase-isomaltase and other glycosidases, the membrane anchor of lactase phlorizin hydrolase, J. Biol. Chem., 267, 18744-18752.

Wade, P.A., A. Gegonne, P.L. Jones, E. Ballestar, F. Aubry and A.P. Wolffe (1999) Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation, Nat. Genet., 23, 62-6.

Wallrath, L.L. and S.C. Elgin (1995) Position effect variegation in Drosophila is associated with an altered chromatin structure, Genes Dev., 9, 1263-1277.

Wang, Y., C. Harvey, M. Rousset and D. Swallow (1994) Expression of intestinal mRNA transcripts during development: analysis by a semi-quantitative RNA PCR method, Pediatr. Res., 36, 514-521.

Wang, Y., C.B. Harvey, E.J. Hollox, A.D. Phillips, M. Poulter, P. Clay, J.A. Walker-Smith and D.M. Swallow (1998) The genetically programmed down-regulation of lactase in children, Gastroenterology, 114, 1230-1236.

Wang, Y., C.B. Harvey, W.S. Pratt, V.R. Sams, M. Sarner, M. Rossi, S. Auricchio and D.M. Swallow (1995) The lactase persistence/non-persistence polymorphism is controlled by a *cis*-acting element, Hum. Mol. Genet., 4, 657-662.

Weinberg, R.B. (1999) Apolipoprotein A-IV-2 allele: association of its worldwide distribution with adult persistence of lactase and speculation on its function and origin, Genet. Epidemiol., 17, 285-97.

Welsh, J.D., L.C. Russell and A.W. Walker (1974) Changes in intestinal lactase and alkaline phosphatase activity levels with age in the baboon (*Papio papio*), Gastroenterology, 66, 993-997.

Welsh, J.D. and A. Walker (1965) Intestinal disaccharidase and alkaline phosphatase activity in the dog, Proc. Soc. Exptl. Biol. Med., 120, 525-527.

Wen, C.-P., I. Antonowicz, E. Tovar, R.B. McGandy and S.N. Gershoff (1973) Lactose feeding in lactose intolerant monkeys, Am. J. Clin. Nutr., 26, 1224-1228.

West, L.F., M.B. Davis, F.R. Green, R.H. Lindenbaum and D.M. Swallow (1988) Regional assignment of the gene coding for human sucrase-isomaltase (SI) to chromosome 3q25-26, Ann. Hum. Genet., 52, 57-61.

Wilmsen, E. (1991) Pastoro-Foragers to "Bushmen": Transformations in Kalahari relations of property, production and labor, in: J.G. Galaty and P. Bonte (Eds.), Herders, warriors, and traders: Pastoralism in Africa, Westview Press, Oxford.

Wixman, R. (1984) The peoples of the USSR: an ethnographic handbook, Macmillan, London.

Wu, G.D., W. Wang and P.G. Traber (1992) Isolation and characterization of the human sucrase-isomaltase gene and demonstration of intestine specific transcription elements, J. Biol. Chem., 267, 7863-7870.

Yip, S.P., D.A. Hopkinson and D.B. Whitehouse (1999) Improvement of SSCP analysis by use of denaturants, Biotechniques, 27, 20-2, 24.

Yuh, C.-H., H. Bolouri and E.H. Davidson (1998) Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene, Science, 279, 1896-1902.

Zerjal, T., B. Dashnyam, A. Pandya, M. Kayser, L. Roewer, F.R. Santos, W. Schiefenhovel, N. Fretwell, M.A. Jobling, S. Harihara, K. Shimizu, D. Semjidmaa, A. Sajantila, P. Salo, M.H. Crawford, E.K. Ginter, O.V. Evgrafov and C. Tyler-Smith (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y- chromosomal DNA analysis, Am. J. Hum. Genet., 60, 1174-83.

Zietkiewicz, E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K.K. Kidd, D. Modiano, R. Scozzari, M. Stoneking, S. Tishkoff, M. Batzer and D. Labuda (1997) Nuclear DNA diversity in worldwide distributed human populations, Gene, 205, 161-71.

Zietkiewicz, E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K.K. Kidd, D. Modiano, R. Scozzari, M. Stoneking, S. Tishkoff, M. Batzer and D. Labuda (1998) Genetic structure of the ancestral population of modern humans, J. Mol. Evol., 47, 146-55.

Errata:

Fischer, S. G., and L. S. Lerman (1983). DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory, Proc. Nat. Acad. Sci. USA, 80, 1579-1583

Myers, R.M., S.G. Fischer, L.S. Lerman and T. Maniatis (1985). Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis, Nucleic Acids Res., 13, 3131-3145

The Genetically Programmed Down-regulation of Lactase in Children

YANGXI WANG,* CLARE B. HARVEY,* EDWARD J. HOLLOX,* ALAN D. PHILLIPS,‡ MARK POULTER,* PETER CLAY,§ JOHN A. WALKER-SMITH,‡ and DALLAS M. SWALLOW*

*Medical Research Council Human Biochemical Genetics Unit, University College London; ‡University Department of Paediatric Gastroenterology, Royal Free Hospital; and §Department of Biochemistry, Hospital for Sick Children, London, England

Background & Aims: Intestinal lactase activity is high in all healthy human babies, but in adults a genetic polymorphism, which acts in cis to the lactase gene, determines high or low messenger RNA (mRNA) expression and activity (lactase persistence and nonpersistence, respectively). Our aim was to investigate the onset of expression of this polymorphism in children. **Methods:** Activities were analyzed in relation to age in normal biopsy specimens from a 20-year collection of diagnostic specimens. In a smaller set of 32 samples, aged 2–132 months, RNA was extracted for semiquantitative reverse-transcription polymerase chain reaction. Marker polymorphisms were used to determine the allelic origin of lactase mRNA transcripts. **Results:** Analysis of 866 children showed evidence that the lactase persistence/nonpersistence polymorphism began before 5 years of age. The 32 children tested had high lactase mRNA and activity. Six children aged 2–16 months showed equal expression of two alleles, 2 children aged 7 and 14 months showed slightly asymmetric expression, and 7 children aged 22–132 months showed very asymmetric expression, the second allele being undetectable in the 11-year-old, as previously seen in lactase-persistent heterozygote adults. **Conclusions:** Genetically programmed down-regulation of the lactase gene is detectable in children from the second year of life, although the onset and extent are somewhat variable.

The intestinal enzyme lactase is responsible for the digestion of lactose, the main carbohydrate in milk. In most mammals, lactase activity declines after weaning when lactose is no longer part of the diet. In contrast, in many humans, particularly in northern Europe, lactase activity persists into adult life. The expression of lactase cannot be modulated by maintaining lactose in the diet.[1,2] Rather, there is substantial evidence from lactose tolerance tests and some from direct measurement of enzymatic activity that persistence or nonpersistence of lactase activity is genetically determined: persistence of lactase is dominant to nonpersistence.[1-3]

We have recently shown that the genetic difference responsible for the lactase persistence/nonpersistence polymorphism, which determines high or low lactase messenger RNA (mRNA) expression, is cis-acting to the lactase gene.[4] We used DNA marker polymorphisms within the exons of the lactase gene to distinguish between transcripts from different chromosome homologues. Individuals who were heterozygous for the marker polymorphisms were informative for this purpose and allowed us to identify allele-specific down-regulation of lactase expression in those who also were heterozygous for lactase persistence.[4]

In this study, we have examined age-related changes in lactase expression in normal intestinal biopsy specimens from children, at the level of both mRNA and enzyme activity. Two approaches were taken; one was a retrospective analysis of enzyme activity data taken from diagnostic records covering a 20-year period. The other was a prospective study in which pieces of the same biopsy sample were used for RNA analysis. RNA expression was compared in fetuses, children, and adults. We again made use of DNA marker polymorphisms to look for allele-specific down-regulation of lactase expression. The two original polymorphisms were tested, and two others which are located in exons 1 and 17.[5] Our observations provide clear evidence of genetically programmed down-regulation of the lactase gene in children.

Materials and Methods

The patients were referred to Queen Elizabeth Hospital for Children in London for chronic intestinal symptoms and/or failure to thrive. Their ages ranged from 1 to 203 months. Biopsies of the small intestine were performed as part of their routine investigations with fully informed consent after an overnight fast, using a double-port pediatric capsule positioned at the duodenojejunal flexure. A sample was immediately frozen in liquid nitrogen and stored at −70°C for enzyme assay.

Abbreviations used in this paper: L/S, lactase/sucrose ratio; RT-PCR, reverse-transcription polymerase chain reaction; S/L, sucrose/lactase ratio.

0016-5085/98/$3.00

Histopathologic examination was carried out on the remaining tissue. In a separate study of 32 cases selected for histological normality (age range, 2–132 months), half of one of the two double-port capsule samples was stored at −70°C for RNA extraction.

The adult specimens and the results relating to these have been described previously,[6] although the reverse-transcription polymerase chain reaction (RT-PCR) analysis shown in Figure 2 was redone in parallel with the samples tested here. RNA and DNA of 14 of the fetuses had been tested previously; 5 were informative for the DNA marker polymorphisms and showed a low level of expression from both alleles.[4,7] We tested another 18 fetuses (gestational age, 8–15 weeks), and from the full set of 32 we identified 12 informative individuals for allele expression analysis. The samples were from suction terminations, collected from the MRC Tissue Bank within a few hours of termination. The mid–small intestine was identified by orientation with respect to the appendix, and the morphology was checked under a microscope before being flash frozen and stored at −70°C.

Enzyme activities were determined as described previously.[8] Extraction of RNA and DNA from the intestinal samples, PCR, and semiquantitative RT-PCR were performed as described previously.[4,6,7]

The four polymorphic sites located at nucleotides 593 (exon 1), 666 (exon 2), 5579 (exon 17), and 5845 (exon 17) in the complementary DNA (cDNA) sequence[5] were typed on genomic DNA using PCR products. For exon 2, PCR products were obtained using primers described previously.[9] For exon 1, part of the exon was amplified using primers designed from Genbank/EMBL HSLCT01 (accession no. M61834). These were an exonic sense primer (nucleotides 1357–1376) and an intronic antisense primer (1729–1710). Exon 17 primers were designed from Genbank/EMBL entry HSLCT17 (accession no. M61850). The intronic sense primer (nucleotides 4–23) and the exonic antisense primer (445–425) flank both polymorphisms within this exon. All four sites were sequenced on the cDNA of the informative individuals using two RT-PCR products prepared using primers located in the cDNA (accession number X07994) at nucleotides 515–538 and 805–783, and 5454–5476 and 5913–5890. PCR products were analyzed by direct cycle sequencing (Thermo Sequenase radiolabeled terminator cycle sequencing kit; Amersham Int., Buckinghamshire, England) and/or by our previous electrophoretic methods.[4,9] The bands were quantified by phosphorimager analysis using a FUJIX BAS-1000 Phosphorimager (FUJIX, Tokyo, Japan). Reexposure of the same gel established the linear range, and measurements (minus background) were standardized for variable lane intensity.

Results

Retrospective Review of Lactase Activities in Histologically Normal Intestinal Biopsy Specimens

Over a 20-year period, a total of 1264 small intestinal biopsy specimens were assayed for lactase, sucrase, and maltase activities as one of the routine diagnostic tests on children with chronic intestinal symptoms and/or failure to thrive. In a retrospective review of the histology and enzyme activities of all these cases, it was possible to identify 866 samples from different individuals that were normal by two criteria: both normal histology and maltase activity of more than 10 U/g. (Maltase activities of 8 U/g wet weight have previously been taken as the lower end of the normal range of activities.[8]) These biopsy specimens came from patients in whom a temporary gastrointestinal disorder (e.g., cow's milk protein intolerance) had resolved, a possible diagnosis of celiac disease was excluded, or no gastrointestinal cause was found for their symptoms (mainly failure to thrive and/or diarrhea lasting 14 days or longer).

The enzyme activities in this series of normal samples were analyzed in several ways. Because the assays were performed on extracts made with a fixed weight per volume extraction of tissue, the activities were expressed as ratios of lactase to sucrase (L/S) or sucrase to lactase (S/L). There was an increasing scatter of results with age, with more samples showing low lactase in relation to sucrase; this finding is more clearly seen when the results are presented as S/L (data not shown). The population was therefore divided into four age groups, and the distribution of S/L ratios was plotted in histogram form, the method of data expression used previously in our studies on adults.[6,10] The data are shown in Figure 1. In the 189 children under 1 year old (1–12 months), the distribution of activities is unimodal. In the second age group of 218 children aged 13–22 months, more individuals have higher ratios, indicating lower lactase activity. This trend continues in the third group of 264 children aged 23–60 months (5 years). Not only is there a general shift in distribution, but two children had ratios greater than 10, previously considered diagnostic of lactase nonpersistence,[6] and 3 other children had S/L greater than 6.5. In the 195 children over the age of 5 years, the distribution approaches the trimodal distribution seen in adults,[6,10,11] with the suggestion of a split in the main peak followed by a shoulder and a distinct group of individuals with S/L greater than 10. Unfortunately, detailed ethnic origin information is not available, but the catchment area for the hospital is approximately 40% Northern European, 5% Southern European, 30% from the Indian subcontinent, 20% Afro-Caribbean and African, and 5% other.

Prospective Study of Intestinal Tissue From Fetuses and Children

Lactase mRNA expression. We assessed the level of lactase mRNA in 32 samples from children and 32

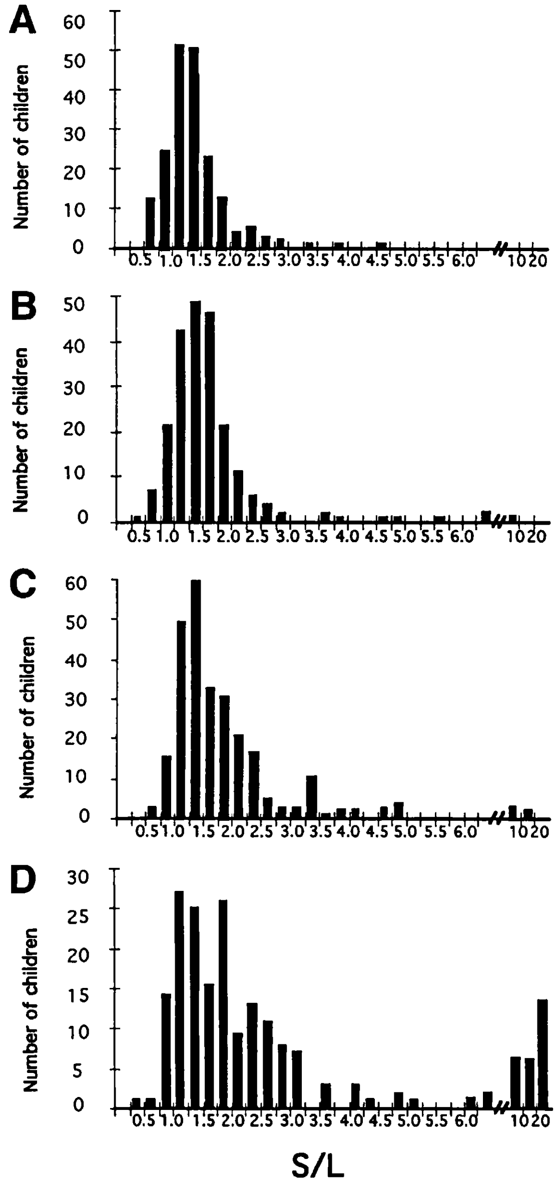


Figure 1. Distribution of S/L in four different age groups shown in histogram form. (*A*) 1–12 months (n = 189); (*B*) 13–22 months (n = 218); (*C*) 23–60 months (n = 264); and (*D*) 61–203 months (n = 195). The ratios are shown along the *x* axis and the numbers of individuals in each class are on the *y*-axis. All four histograms are to the same scale along the *x* axes, although the *y* axes are on different scales. The scale of the end of the *x* axis is condensed from the position marked (6.5 and above); the highest ratio group contains all with ratios above 20. In our previous studies, ratios of >10 have been considered diagnostic of lactase nonpersistence.

samples from fetuses by use of a semiquantitative RT-PCR assay in comparison with sucrase-isomaltase and dipeptidyl peptidase 4. We measured lactase and sucrase activities in all children and a sample of 10 fetuses. All of the samples from children had high lactase mRNA levels and high lactase activity, comparable with lactase-persistent adults (mean S/L, 1.8 [range, 0.78–5.0]; mean L/S, 0.62 [range, 0.2–1.28]). In the fetal small intestinal samples, the lactase mRNA levels were low, and lactase activities were correspondingly low (mean S/L, 10.0 [range, 5.3–33.3]; mean L/S, 0.1 [range, 0.03–0.19]). Figure 2 shows a representative experiment that includes a direct comparison of lactase mRNA levels in children with those from fetuses and from persistent and nonpersistent adults, together with the corresponding enzyme activities, expressed as L/S for easy assessment.

Allelic expression of lactase mRNA. Typing of four polymorphic sites identified 15 children who were heterozygous at one or more of the sites. The two cDNA PCR products were sequenced to cover the informative sites in these individuals and in the 12 fetuses (Table 1) and at least two sites in the children whose results were uninformative (not shown). Relative band intensity was determined by phosphorimage analysis. The results show reasonably good agreement between the different sites, with slightly more variability in the fetal samples, which require more PCR cycles because of very low expression. Nevertheless, the fetal samples showed roughly equal expression of both alleles, confirming our previous results[4] (Table 1). Six of the 15 informative children (aged 2–16 months) showed equal high expression of both alleles (Table 1 and Figure 3). Three children (aged 7 and 14 months) showed a suggestion of asymmetry of expression of the two alleles. Seven children aged 22–132 months showed very marked asymmetry, the second allele being below the threshold of detection in the oldest child (Table 1 and Figure 3). Although there is interindividual variation, there is a clear age-related down-regulation of one allele. The 7 older children are interpreted as lactase-persistent heterozygotes in whom there is relative down-regulation of the nonpersistent allele.

Discussion

A large number of lactose tolerance studies on children of various variety of ages indicate that although all healthy babies have lactase activity, genetically determined lactose intolerance caused by lactase nonpersistence can develop at variable ages after the first year of life.[2,12] In a few studies, enzymatic activity has been measured directly.[13–17] There has been some suggestion that the age of onset of lactose intolerance is different in different ethnic groups,[1] with Finnish nonpersistent individuals remaining lactose tolerant until they are adolescents, whereas Thais, for example, are all lactose intolerant by 5 years of age.[18,19] However, the evidence for this is not clear-cut and a certain amount of confusion has been caused by variations in the method of lactose tolerance testing.[20]

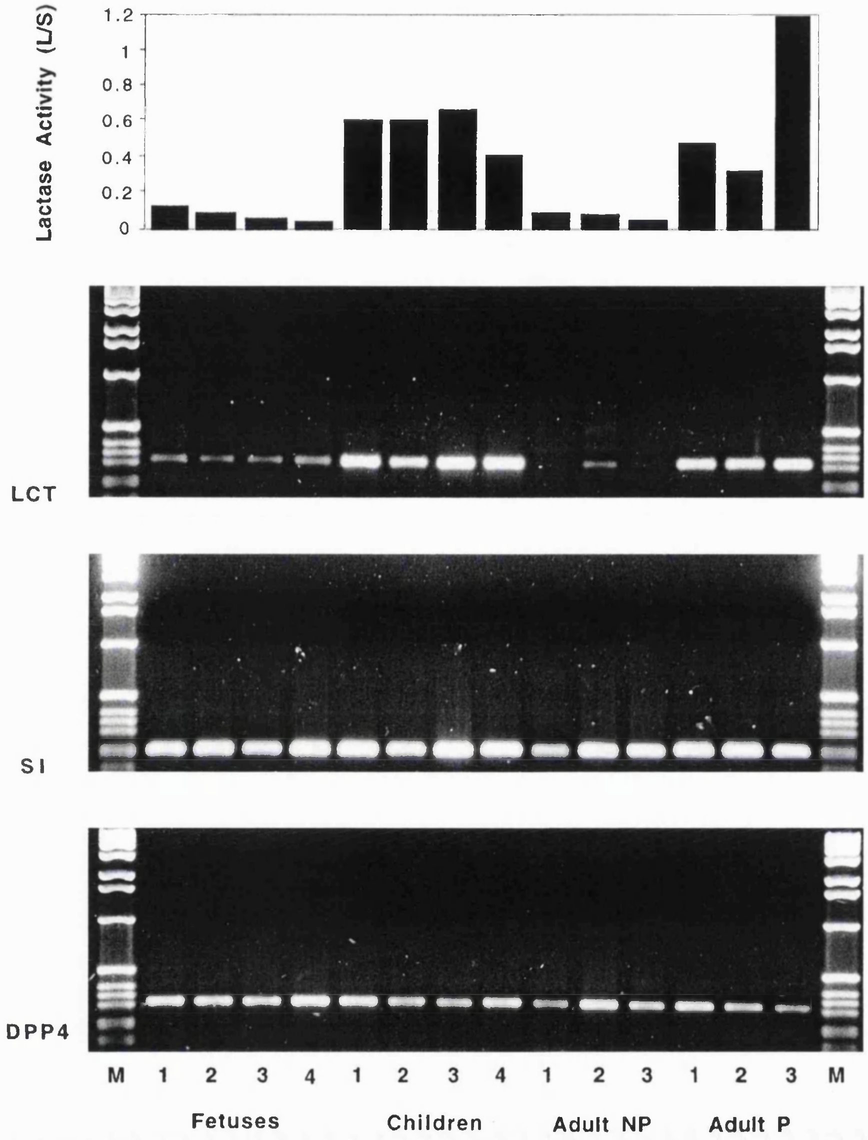


Figure 2. Developmental differences in lactase expression. Semiquantitative RT-PCR showing the developmental difference in lactase (LCT) mRNA levels in 4 typical samples from fetuses, 4 from children (aged 96, 8 , 8, and 3 months), 3 from adults who are lactase nonpersistent (NP), and 3 from adults who are lactase persistent (P), in comparison with sucrase isomaltase (SI) and dipeptidyl peptidase 4 (DPP4). M, size marker (1-kilobase ladder; GIBCO BRL, Paisley, Scotland). Lactase activities of the individual samples expressed as L/S are shown above the gels.

In this study, two different approaches were taken to evaluate developmental changes in lactase expression in children. Analysis of enzyme assay results from 866 normal intestinal biopsy samples showed a progressive increase in the scatter of activities with age, with an increasing number of individuals showing high S/L. In those aged 60–203 months (5–17 years), there is evidence of the lactase activity polymorphism observed in adults. The data obtained by previous investigators seem to show the same trend,[13-17] although they were not plotted in the same way and the data sets were smaller. This is particularly noticeable in the study of Lebenthal et al.,[17] who plotted S/L against age and showed appearance of individuals with high ratios by the age of 5 years.

Lactase mRNA levels were very low in all the fetuses and high in this series of children, in direct agreement with the levels of lactase activity. When measured quantitatively, there was reasonable correlation of lactase mRNA and enzyme activity in the children, although there was greater variance than in adults (Wang, unpublished findings), but none had low lactase mRNA and high activity as found by Olsen et al.[21] None of the

Table 1. Phosphorimaging Analysis Showing the Relative Expression of the Two Lactase Alleles in Informative (Heterozygous) Individuals

	Exon 1 nt 593		Exon 2 nt 666		Exon 17 nt 5579		Exon 17 nt 5845	
	Genomic	cDNA (%*C*)	Genomic	cDNA (%*A*)	Genomic	cDNA (%*T*)	Genomic	cDNA (%*G*)
Fetuses								
Gestational age (*wk*)								
10.1	CT	40	AG	46	CT	45	CG	—
11.2	CT	45	AG	47	CT	43	—	—
12.0	C	*h*	AG	46	T	*h*	C	*h*
12.6	CT	—	G	*h*	CT	47	C	—
12.8	CT	50	AG	42	CT	52	—	—
13.1	C	*h*	A	*h*	CT	47	G	*h*
13.1	C	*h*	AG	52	T	*h*	CG	57
13.2	CT	53	G	*h*	CT	52	C	—
13.3	CT	53	AG	51	CT	51	—	—
13.7	C	*h*	AG	57	T	*h*	G	*h*
13.8	CT	46	AG	53	CT	54	—	—
14.8	CT	52	AG	47	CT	54	CG	56
Children								
Age (*mo*)								
2	CT	49	AG	45	CT	48	CG	45
3	CT	49	AG	49	CT	52	CG	51
5	C	*h*	AG	50	T	—	CG	55
7	CT	59	G	*h*	CT	57	C	*h*
8	C	*h*	AG	51	T	*h*	CG	54
14	CT	40	AG	42	CT	40	CG	43
16	CT	47	AG	40	CT	46	CG	49
16	C	*h*	G	*h*	CT	51	C	—
22	CT	13	AG	15	CT	13	CG	17
42	CT	27	G	*h*	CT	27	C	—
49	CT	37	G	*h*	C	—	C	—
50	CT	20	G	*h*	CT	20	C	—
54	CT	40	G	*h*	CT	33	C	—
98	CT	14	AG	16	CT	20	CG	22
132	CT	6	AG	0	CT	2	CG	0

NOTE. The nucleotides (nt) present at each site in the genome are shown followed by the relative band intensity in the cDNA, expressed as a percentage of the total band intensity of the two alleles. The values obtained for the homozygous sites that were tested on the cDNA and calculated the same way range from 96% to 103% or −4% to 4%, according to which allele is homozygous, but are shown as *h* for clarity. —, Not tested.

children in this series showed the very low lactase activity (S/L > 10.0) and low lactase mRNA indicative of lactase nonpersistence.

However, by taking advantage of polymorphisms within the exons, we have found evidence of progressive, although variable, allele-specific down-regulation. All 7 children aged 22 months to 11 years who were heterozygous for two or more exonic polymorphisms showed very asymmetric expression of the lactase alleles. Five of these children were classified as white, 1 was from the Indian subcontinent (aged 11 years), and 1 was part white, part Afro-Caribbean (aged 54 months). In each case, the homologue expressed at high levels carries the same nucleotide at the four polymorphic sites (T, G, C, C). This is consistent with our previous results in adults: of the 12 lactase-persistent adults who were heterozygous at the two original sites tested, 10 showed preferential expression of the same alleles.[4] This combination of alleles is characteristic of the haplotype (A) that is most common in Northern and Southern Europeans[9] and in persons from the Indian subcontinent, although not in all human populations (Harvey and Hollox, unpublished data). These data together provide strong evidence of allelic association of lactase persistence and the A haplotype. The only young infant with slight asymmetric expression (who is of Indian extraction) showed higher expression of a non-A haplotype. Further studies on allelic association are in progress and will be published separately.

Our working hypothesis is that a developmentally regulated trans-acting DNA-binding protein binds to one kind of lactase allele and not the other and influences transcription and/or mRNA stability.

Using the approach described in this paper, it should be possible to determine whether the differences in timing of down-regulation relate to published population differences in age of onset of lactose intolerance.[1] This study also highlights the importance of considering the implications of the lactase persistence polymorphism in

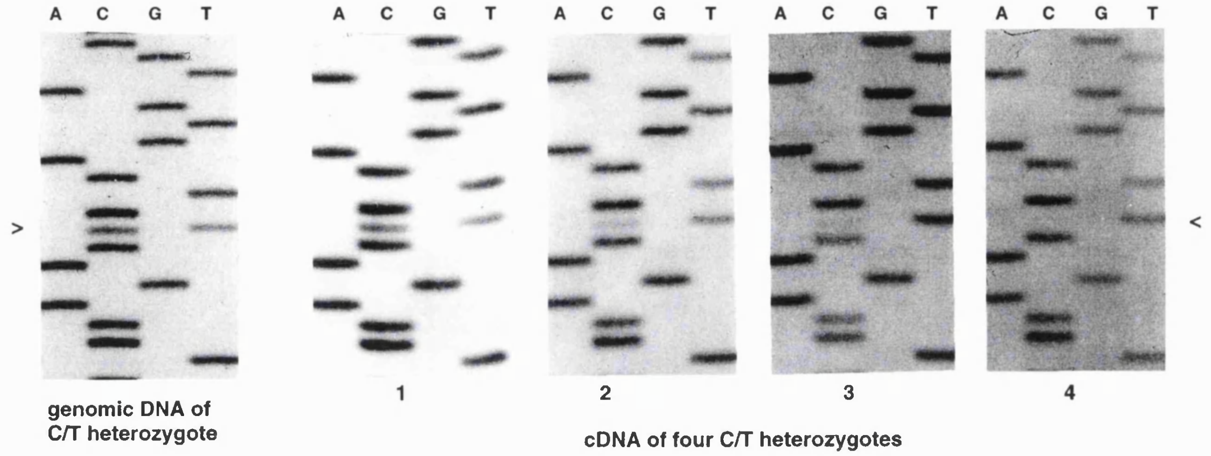


Figure 3. Asymmetrical expression of lactase mRNA transcripts. Sequencing gels showing genomic sequencing of a typical exon 1 CT heterozygote and four representative examples showing variable expression of the two mRNA transcripts detected by sequencing across the same site using a cDNA PCR product. All 4 individuals are heterozygous at this site (position marked with *arrows*) but show varying levels of expression of the C-containing allele. Individual 1 is 3 months old and shows equal C and T (%C = 49; Table 1). Individual 2 is 42 months old (%C = 27), and individual 3 is 50 months old (%C = 20). In individual 4, who is 132 months old (11 years), the C band is not detectable on the autoradiograph, although trace amounts are possibly detectable by phosphorimaging (%C = 6). Both bands show equal intensity in genomic DNA from heterozygotes (%C = 51.25 ± 1.29, n=4, as determined by phosphorimage analysis).

interpreting disaccharidase assay results even in very young children, especially if they are of non–Northern European origin. In this context it should be noted that the low expression engendered by secondary phenomena such as gut damage and malnutrition is likely to be superimposed on the normal genetically determined down-regulation. It is to be hoped that before long the nucleotide changes responsible for the lactase persistence polymorphism will be found and that this will enable a more certain interpretation of the observations relating to secondary lactase deficiency[22,23] and to posttranslational changes that have been reported in some studies.[24]

References

1. Flatz G. Genetics of lactose digestion in humans. Adv Hum Genet 1987;16:1–77.
2. Sahi T. Progress report. Dietary lactose and the aetiology of human small-intestinal hypolactasia. Gut 1978;19:1074–1086.
3. Swallow DM, Harvey CB. Genetics of adult-type hypolactasia. In: Auricchio S, Semenza G, eds. Common food intolerances 2: milk in human nutrition and adult-type hypolactasia. Basel, Switzerland: Karger, 1993:85–92.
4. Wang Y, Harvey CB, Pratt WS, Sams VR, Sarner M, Rossi M, Auricchio S, Swallow DM. The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. Hum Mol Genet 1995;4:657–662.
5. 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.
6. 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.
7. Wang Y, Harvey C, Rousset M, Swallow D. Expression of intestinal mRNA transcripts during development: analysis by a semi-quantitative RNA PCR method. Pediatr Res 1994;36:514–521.
8. Phillips AD, Avigad S, Sacks J, Rice SJ, France NE, Walker-Smith JA. Microvillous surface area in secondary disaccharidase deficiency. Gut 1980;21:44–48.
9. Harvey CB, Pratt W, Islam I, Whitehouse DB, Swallow DM. DNA polymorphisms in the lactase gene: linkage disequilibrium across the 70kb region. Eur J Hum Genet 1995;3:27–41.
10. Ho MW, Povey S, Swallow DM. Lactase polymorphism in adult British natives: estimating allele frequencies by enzyme assays in autopsy samples. Am J Hum Genet 1982;34:650–657.
11. Flatz G. Gene dosage effect on intestinal lactase activity demonstrated in vivo. Am J Hum Genet 1984;36:306–310.
12. Scrimshaw NS, Murray EB. Prevalence of lactose maldigestion. Am J Clin Nutr 1988;48:1086–1098.
13. Keane R, O'Grady JG, Sheil J, Stevens FM, Egan-Mitchell B, McNicholl B, McCarthy CF, Fottrell PF. Intestinal lactase, sucrase and alkaline phosphatase in relation to age, sex and site of intestinal biopsy in 477 Irish subjects. J Clin Pathol 1983;36:74–77.
14. Thomas S, Walker-Smith JA, Senewiratne B, Hjelm M. Age dependency of the lactase persistence and lactase restriction phenotypes among children in Sri Lanka and Britain. J Trop Pediatr 1990;36:80–85.
15. Zheng BY, KhinMaung U, Lu RB, Hill ID, Lebenthal E. Disaccharidase and glucoamylase enzyme levels in infants and children. Int Pediatr 1994;9:33–36.
16. Welsh JD, Poley JR, Bhatia M, Stevenson DE. Intestinal disaccharidase activities in relation to age, race, and mucosal damage. Gastroenterology 1978;75:847–855.
17. Lebenthal E, Antonowicz I, Schwachman H. Correlation of lactase activity, lactose tolerance and milk consumption in different age groups. Am J Clin Nutr 1975;28:595–600.
18. Sahi T, Launiala K. Manifestation and occurence of selective adult-type lactose malabsorption in Finnish teenagers. A follow-up study. Dig Dis Sci 1978;23:699–704.
19. Keusch GT, Troncale FJ, Miller LH, Promadhat V, Anderson PR. Acquired lactose malabsorption in Thai children. Pediatrics 1969; 43:540–545.

20. Tadesse K, Yuen RCF, Leung DTY. The status of lactose absorption in Hong Kong Chinese children. Acta Paediatr Int J Paediatr 1992;81:598–600.
21. Olsen WA, Li BUK, Lloyd M, Korsmo H. Heterogeneity of intestinal lactase activity in children: relationship to lactase-phlorizin hydrolase messenger RNA abundance. Pediatr Res 1996;39:877–881.
22. Nichols BL, Dudley MA, Nichols VN, Putnam M, Avery SE, Fraley JK, Quaroni A Shiner M, Carrazza FR. Effects of malnutrition on expression and activity of lactase in children. Gastroenterology 1997;112:742–751.
23. Northrop-Clewes CA, Lunn PG, Downes RM. Lactose maldigestion in breast feeding Gambian infants. J Pediatr Gastroenterol Nutr 1997;24:257–263.
24. Rossi M, Mauiri L, Fusco MI, Salvati VM, Fuccio A, Auricchio S, Mantei N, Zecca L, Gloor SM, Semenza G. Lactase persistence versus decline in human adults: multifactorial events are involved in down-regulation after weaning. Gastroenterology 1997;112: 1506–1514.

Received June 26, 1997. Accepted February 26, 1998.

Address requests for reprints to: Dallas M. Swallow, Ph.D., Medical Research Council Human Biochemical Genetics Unit, University College London, Wolfson House, 4 Stephenson Way, London NW12HE, England. e-mail: dswallow@hgmp.mrc.ac.uk; fax: (44) 171 387 3496.

Supported by a grant from the British Digestive Foundation (to C.B.H.).

Dr. Wang's current affiliation is: Harvard Institute of Medicine, Boston, Massachusetts.

Dr. Harvey's current affiliation is: Unitat de Biologia Cellular i Molecular, Barcelona, Spain.

Drs. Wang, Harvey, and Hollox contributed equally to this study.

The authors thank Dr. A. Johnson, Institute of Child Health, London, for help with the disaccharidase data; Wendy Pratt for help with PCRs; Jan Henry for help with data analysis; Jeff Pinxteren for the use of his Phosphorimager; and Dr. L. Wong for samples.

Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism

C. B. HARVEY[1]†, E. J. HOLLOX[1], M. POULTER[1], Y. WANG[1]‡, M. ROSSI[2], S. AURICCHIO[3], T. H. IQBAL[4], B. T. COOPER[3], R. BARTON[5]§, M. SARNER[5], R. KORPELA[6,7], AND D. M. SWALLOW[1]*

[1] *MRC Human Biochemical Genetics Unit, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK*
[2] *Instituto di Scienze dell'Alimentazione-CNR, via Romea 52, 83100 Avellino, Italy*
[3] *Dipartimento di Pediatria, Via S. Pansini, 5, 80131 Naples, Italy*
[4] *Gastroenterology Unit, City Hospital, Dudley Rd, Birmingham, B18 7QH, UK*
[5] *Gastroenterology Unit, The University College London Hospitals, Gower St London WC1, UK*
[6] *Valio Research Centre, PO Box 30, FIN-00039, Helsinki, Finland*
[7] *University of Helsinki, Department of Pharmacology and Toxicology, Helsinki, Finland*

(Received 19.3.98. Accepted 8.5.98)

SUMMARY

A genetic polymorphism is responsible for determining that some humans express lactase at high levels throughout their lives and are thus lactose tolerant, while others lose lactase expression during childhood and are lactose intolerant. We have previously shown that this polymorphism is controlled by an element or elements which act in cis to the lactase gene. We have also reported that 7 polymorphisms in the lactase gene are highly associated and lead to only 3 common haplotypes (**A**, **B** and **C**) in individuals of European extraction. Here we report the frequencies of these polymorphisms in Caucasians from north and south Europe and also from the Indian sub-continent, and show that the alleles differ in frequency, the **B** and **C** haplotypes being much more common in southern Europe and India. Allelic association studies with lactase persistence and non-persistence phenotypes show suggestive evidence of association of lactase persistence with certain alleles. This association was rather more clear in the analysis of small families, where haplotypes could be determined. Furthermore haplotype and RNA transcript analysis of 11 unrelated lactase persistent individuals shows that the persistence (highly expressed) allele is almost always on the **A** haplotype background. Non-persistence is found on a variety of haplotypes including **A**. Thus it appears that lactase persistence arose more recently than the DNA marker polymorphisms used here to define the main Caucasian haplotypes, possibly as a single mutation on the **A** haplotype background. The high frequency of the **A** haplotype in northern Europeans is consistent with the high frequency of lactase persistence.

INTRODUCTION

The intestinal enzyme lactase is responsible for the digestion of lactose which is the main carbohydrate in milk. In most mammals lactase activity declines after weaning when lactose is no longer part of the diet. In contrast, in many humans, particularly in northern Europe, lactase activity persists into adult life. In most non-European populations lactase non-persistence is

† Current address Unitat de Biologia Cel.lular i Molecular, IMIM, c/. Dr Aiguader, 80, E 08003 Barcelona, Spain
‡ Current address Harvard Institute of Medicine, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
§ Current address: Department of Medicine, North Tyneside General Hospital, Tyne and Wear, NE29, UK
* Correspondence: Tel: +171 504 5040; Fax +171 387 3496.
E-mail: dswallow@hgmp.mrc.ac.uk

the most common phenotype, and even in Europe the frequencies of lactase persistence decline progressively from north to south (for review, see Flatz 1987). 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 (reviewed in Flatz 1987; Swallow and Harvey 1993). We have recently shown that the genetic difference responsible for this polymorphism is cis-acting to the lactase gene, and determines high or low lactase mRNA expression in adults (Wang *et al.* 1995) and that progressive down-regulation of one allele can be detected during childhood (Wang *et al.* 1998). Healthy adult heterozygous individuals have intermediate levels of enzymic activity which are sufficient to hydrolyse dietary lactose and the lactose load of a lactose tolerance test.

Despite substantial sequence analysis of the lactase gene, including 1 kb upstream from exon 1, the sequence variation responsible for the persistence polymorphism has not yet been identified (Boll *et al.* 1991; Lloyd *et al.* 1992) However many polymorphisms have been identified across the 70 kb region spanned by the gene (Boll *et al.* 1991; Harvey *et al.* 1995) and our initial observations on seven of these in the series of 50 families from the Centre d'Etude du Polymorphisme Humain (CEPH) revealed only three common haplotypes (Harvey *et al.* 1995). The frequency of these haplotypes was different in the families of French origin from those from Utah (a region peopled by individuals largely from the UK and Scandinavia (McLellan *et al.* 1984)). The aim of this study was to determine whether these differences in haplotype frequency were a reflection of the frequency gradient of lactase persistence seen across Europe. We have determined the allele frequencies at seven polymorphic sites within the lactase gene in unrelated individuals of unknown lactase persistence status from different parts of Europe and from India. We have also looked for evidence of association of these alleles with lactase persistence.

Materials and Methods

DNA samples were prepared using standard procedures, in most cases from blood, or from lymphoblastoid cell lines or in a few cases from biopsy material (Wang *et al.* 1995). Specimens were available from a variety of sources and the study populations included unrelated hospital patients and volunteers. All the people tested were Caucasian but were divided into groups according to their ethnic or geographic origin (Northern European, Southern European, Indian and Finnish). The Finnish population were all female aged 22–67, mean age 36, and were volunteers taking part in a study relating to the symptoms of lactose intolerance. This population was deliberately selected to contain approximately half non-persistent ($n = 11$, age 22–67, mean = 37) and half persistent individuals ($n = 9$, age 22–61, mean = 35). Eight small families, five from Southern Europe, and three of Indians of Sikh faith and Punjabi origin resident in Birmingham, UK, usually of four individuals, were also tested. The unrelated members of these families were included in the population surveys.

A simple classification of lactase persistence status was made either by direct lactase activity measurement or by lactose tolerance testing. Lactase activity was measured on duodenal biopsy material from 49 hospital patients from London (Harvey *et al.* 1995) and 29 jejunal surgical specimens from Naples (Maiuri *et al.* 1991). Lactose tolerance testing was performed on volunteers and families by measuring breath hydrogen (Fernades *et al.* 1978; Robb and Davidson 1981) and in most cases also blood glucose (Weijers *et al.* 1961) and/or urinary galactose (Grant *et al.* 1989), after a lactose load of 50 g.

The seven polymorphic sites (see Fig. 1) were analysed as described previously by PCR amplification of sequence containing the sites concerned, followed by denaturing gradient gel electrophoresis, single stranded conformation analysis, simple acrylamide gel electrophoresis or restriction enzyme digestion followed by electrophoresis (Harvey *et al.* 1995). The exonic poly-

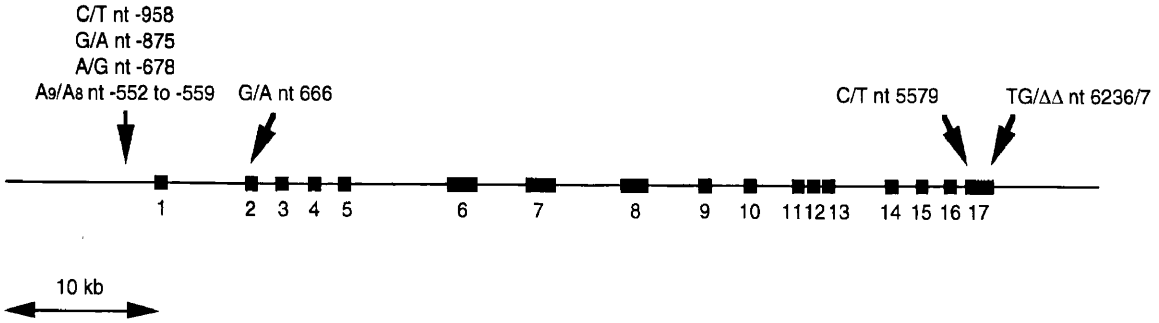


Fig. 1. Diagramatic representation of the lactase gene to show the position of the polymorphisms tested (size of the small exons not to scale).

morphisms are located at nucleotides 666 (G/A), 5579 (C/T) and 6236/7 (TG/ΔΔ) of the published cDNA sequence (Mantei *et al.* 1988, Genbank/EMBL accession no X07994) and the 5′flanking polymorphisms are located at nucleotide positions −958 (TC(C/T)AT), −875 (AT(G/A)TA), −678 (CT(A/G)CC), −552–559 (C(A_9/A_8)C) of the published 5′flanking sequence, Genbank/EMBL accession no M61834.

Ethical approval was obtained for each of the studies included in this paper (from the Ethical Committees of UCL Hospitals London, the City Hospital, Birmingham, University Hospital of Helsinki and 'Federico II' University of Naples Medical School.

Because of the small data sets all statistical analysis was done using 2 by 2 tables for analysis by Fisher's exact test. For analysis of the data on Table 3, the two DNA polymorphism alleles at each site were counted (assuming the genotypes indicated) to construct 2 by 2 Tables.

Results

Population frequencies

The people tested were divided into three groups according to their ethnic or geographic origin. These were northern European, the majority of whom were UK residents (London), southern European which includes 41 from Italy (Naples) as well as London residents who came from the Mediterranean countries and Indian who were people from the Indian sub-continent who are resident in the UK and include people who originate from all over India as well as Bangladesh and Pakistan. The Finnish group could not be included in this part of the study since they were selected to include a large number of non-persistent individuals. In each population the phenotype frequencies were not significantly different from those expected under Hardy–Weinberg equilibrium (not shown) but there were major differences in allele frequencies between the three groups (Table 1). The frequencies observed in the French CEPH samples is shown for comparison and it can be seen that they are intermediate between the Northern and Southern Europeans.

By analysis of the homozygous sites present in these individuals and the numbers observed for each population, as well as comparison with the haplotypes observed in the CEPH families, it was possible to infer the probable haplotypes and the numbers for each population are listed (Table 2). Four haplotypes were found in all three populations (**A**, **B**, **C** and **E**), with two other relatively common haplotypes, **D** found in Europeans and **G** in the Indians. In both European groups the most frequent haplotype was **A** but the **B** and **C** haplotypes were much more frequent in the Southern Europeans. In the Indians **A** and **C** were approximately equal in frequency with **B** the third most frequent.

Association of alleles at the polymorphic sites within the lactase gene with lactase persistence phenotype

Table 3 shows a comparison of the numbers of individuals with each phenotype (presumed genotype) at each of the polymorphic sites, in the groups of individuals classified as lactase per-

Table 1. *Comparison of the frequency of the alleles at seven polymorphic sites within the lactase gene in different populations*

The nucleotide positions of the 5′ sequence are taken from Genbank/EMBL M61834 and the cDNA sequence is from Genbank/EMBL X07994. The frequency of the allele at each site (nucleotide or sequence motif) which is the rarer in the Northern European population, is shown.

		5′ flanking region				Exon 2	Exon 17	
Population	Number of chromosomes	nt −958 T	nt −875 A	nt −678 G	nt −552 to −559 A8	nt 666 A	nt 5579 T	nt 6236/7 ΔΔ
Northern European	108–120	0.08	0.03	0.04	0.11	0.08	0.10	0.08
Southern European	108	0.40	0.06	0.22	0.63	0.39	0.53	0.38
Indian	52–54	0.25	0	0.40	0.65	0.28	0.63	0.19
French (CEPH)	48	0.33	0	0.08	0.42	0.33	0.38	0.33

sistent and non-persistent. It should be noted that since lactose tolerance (lactase persistence) is a dominant trait the group of lactase persistent/lactose tolerant individuals includes both homozygotes (LL) and heterozygotes (Ll) while all lactase non-persistent individuals are homozygous (ll). The frequency of each of the DNA polymorphisms in the Northern and Southern European and Indian groups characterized for lactase persistence was not significantly different from the complete data sets for each group (Table 1). However, the phenotype distributions and allele frequencies in the persistent and non-persistent groups are slightly different. These differences are in the same direction in each population, the alleles not present in the **A** haplotype tending to be more frequent in the non-persistent group, but statistical significance ($p = < 0.05$) was reached only for the A_9/A_8 polymorphism in the 5′flanking region of the gene in the Southern European populations ($p = 0.04$) and Finnish ($p = 0.014$) and the C/T transition in exon 17 in the Finnish population ($p = 0.009$) (Table 3).

Association of lactase gene haplotypes with lactase persistence alleles (L and l)

In some cases it was possible to infer haplotypes: two methods were used for this. The first made use of the limited family material available to us. Assuming Mendelian dominant inheritance of lactose tolerance, it was possible to deduce the lactase persistence genotype and thus the full haplotypes for some or all of the chromosomes in each family. All seven polymorphisms within the lactase gene were analysed on each family member and the haplotypes deduced from the joint segregation of alleles at each individual site. In most individuals it was also possible to deduce the genotype with respect to the lactase persistence polymorphism (LL, Ll or ll). All non-persistent individuals are assumed to be homozygous ll. Heterozygosity of lactase persistent individuals (Ll) was diagnosed by the identification of non-persistent individuals among the progeny. Homozygosity (LL) could not be determined. Table 4 shows all the chromosomes where both L/l genotype and haplotype could be deduced.

In all but one case the persistence allele was shown to be on the **A** haplotype (as defined in Table 2). The non-persistence allele was on a variety of haplotypes. Comparison of the numbers of lactase persistent alleles, on, and not on the **A** haplotype with the non-persistence alleles, on, and not on the **A** haplotype shows a statistically significant difference ($p = 0.01$ Fisher's exact test on pooled data). It should be noted that the lactase persistent and non-persistent alleles were inherited from the mother or from the father in approximately equal proportions.

The second approach involves the analysis of

Table 2. *Numbers of each haplotype identified or inferred in unrelated individuals in each population.*

(The numbers in brackets are those identified unambiguously since they were tested for all seven sites and shown to be homozygous for at least at 6, or they were deduced from the genotypes in other family members. Details of the polymorphisms are given in Table 1.)

Numbers of each haplotype per population				5′ flanking region				Exon 2	Exon 17	
Northern European	Southern European	Indian	Haplotype	nt −958	nt −875	nt −678	nt −552 to 559	nt 666	nt 5579	nt 6236/7
90 (76)	40 (19)	17 (9)	A	C	G	A	A_9	G	C	TG
6	34 (19)	9 (3)	B	T	G	A	A_8	A	T	ΔΔ
3	13 (2)	18 (8)	C	C	G	G	A_8	G	T	TG
3	6		D	T	A	A	A_8	A	T	ΔΔ
2	7 (3)	2	E	C	G	G	A_8	G	C	TG
			F	T	G	A	A_8	A	C	ΔΔ
		4	G	T	G	A	A_8	A	T	TG
			H	C	G	A	A_9	G	T	ΔΔ
	2		I	T	G	A	A_8	A	C	TG
	2		J	C	G	A	A_8	G	C	TG
			K	C	G	A	A_8	G	T	TG
	1		L	T	G	A	A_8	G	T	TG
	1 (1)		M	C	G	G	A_8	G	T	ΔΔ
			N	C	G	G	A_8	G	T	TG

Table 3. *Two by three tables showing DNA polymorphism phenotype (assumed genotype) distribution in lactase persistent and non-persistent individuals in four populations*

	5′ flanking nt −958			5′ flanking nt −875		
	CC	CT	TT	GG	GA	AA
Northern European UK						
Persistent	23	5		26	2	
Non-persistent	1					
Finnish						
Persistent	7	2		9		
Non-persistent	5	6		11		
Southern European						
Persistent	6	9		14	1	
Non-persistent	9	12	7	25	3	
Indian						
Persistent	3	2		5		
Non-persistent	4	2		6		

	5′ flanking nt −678			5′ flanking nt −552 to −559		
	AA	AG	GG	A9/A9	A9/A8	A8/A8
Northern European UK						
Persistent	30	2		25	7	
Non-persistent		1		1		
Finnish						
Persistent	6	3		4	5	
Non-persistent	4	6	1	1	6	4
Southern European						
Persistent	10	3	1	1	13	1
Non-persistent	17	8	3	2	12	14
Indian						
Persistent	3	2		1	4	
Non-persistent	1	3	2		3	3

	exon 2 nt 666			exon 17 nt 5579		
	GG	GA	AA	CC	CT	TT
Northern European UK						
Persistent	28	5		26	6	
Non-persistent	1				1	
Finnish						
Persistent	7	2		5	4	
Non-persistent	5	6		1	6	4
Southern European						
Persistent	6	9		3	11	1
Non-persistent	11	11	6	4	16	8
Indian						
Persistent	3	2		1	4	
Non-persistent	3	2			3	3

	exon 17 nt 6236/7		
	TG/TG	TG/ΔΔ	ΔΔ/ΔΔ
Northern European UK			
Persistent	29	5	
Non-persistent	1		
Finnish			
Persistent	7	2	
Non-persistent	5	6	
Southern European			
Persistent	6	9	
Non-persistent	10	12	6
Indian			
Persistent	3	1	
Non-persistent	3	2	

Table 4. *Haplotypes deduced from small families segregating lactase presistence/non-persistence.*

(Haplotypes are shown of all the chromosomes where both L/l genotype and haplotype could be deduced.)

	Italian		Indian		Total	
Haplotype	L	ℓ	L	ℓ	L	ℓ
A	3	5	2		5	5
B	1	7		1	1	8
C		1		4		5
M		1				1
Summary						
A	3	5	2		5	5
Not A	1	9		5	1	14

$p = 0.01$

Table 5. *Probable haplotypes of unrelated individuals defined as lactase persistent heterozygotes (Ll) and lactase persistent homozygotes (LL) by RNA expression analysis of intestinal biopsies.*

(In each case the expressed haplotype is infered from the sites in 2 exons (Harvey *et al.* 1995; Wang *et al.* 1995) and the identity of the second haplotype deduced by full typing of the 7 polymorphisms. L (persistent) is the allele expressed at high level and ℓ (non-persistent) is the allele expressed at lower level.)

(*a*) Lactase persistent heterozygotes (Lℓ)

	Chromosome		
Population	L	ℓ	Number of people
N. Europe	A	B	1
	A	C	1
	A	D	2
S. Europe	A	B	3
	A	D	1
Indian	A	B	1

(*b*) Lactase persistent homozygotes (LL)

	Chromosome		
Population	L	L	No. of observations
N. Europe	A	B	1
S. Europe	A	I	1

(*c*) Total of chromosomes carrying L and ℓ alleles

	L	ℓ
A haplotype	11	0
Non-A haplotype	2	9

$p = 0.0001$.

the inferred DNA marker polymorphism haplotypes from lactase persistent individuals whose lactase persistence genotypes was deduced by expression studies. These data are shown in Table 5. The genotype was deduced from the analysis of expression of RNA transcripts: persistent homozygotes (LL) express both alleles at high levels whereas the heterozygotes (Ll) show essentially mono-allelic expression. In all nine heterozygotes (Ll) the combination of alleles at the sites carried on the highly expressed transcript are characteristic of the **A** haplotype. The haplotype of the second chromosome was deduced in each case by analysis of all seven polymorphic sites on genomic DNA. In just two cases which are defined as persistent homozygotes (LL) both the **A** and the non **A** carrying chromosomes are expressed at high level. Comparison of the numbers of lactase persistent alleles on or not on the **A** haplotype with the non-persistent alleles on and not on the **A** haplotype shows a statistically significant difference ($p =$ 0.0001, Fisher's exact test on pooled data).

Discussion

The analyses described here show that there are substantial differences in allele frequencies of polymorphisms in the lactase gene in the populations tested. These differences appear to reflect the distribution of 3 major haplotypes, of which **A** is much the most common in the UK while haplotypes **B** and **C** are relatively more common in southern Europe than in the UK. Putting this together with our previous data of the allele frequency in a French population (Table 1), there is clearly a gradient in the distribution of these haplotypes from north to south. It is not possible to determine the frequency of the haplotypes in Finland from the data presented here because this was a selected sample, but it is clear that the same three haplotypes are also found in Finland and the **C** haplotype may be relatively more frequent than in Southern Europe. In the Indians tested, the **C** haplotype is also very common, being of similar frequency to the **A** haplotype.

Population studies are suggestive of allelic association of the DNA marker alleles and lactase persistence/non-persistence. This association was rather more clear in the analysis of small families where haplotypes could be determined. Using another approach, namely the analysis of transcript expression in lactase persistent individuals heterozygous for the marker polymorphisms, there was very clear association between the lactase persistence allele (L) and the **A** haplotype. It should be noted however that the very high level of statistical significance could partly be due to a population admixture artefact, which could not be infered from the limited family history data obtained.

Our results demonstrate the high incidence of an **A** haplotype chromosome carrying the lactase persistence allele in the population native to the UK, where some 95% of the people are lactase persistent (allele frequency 0.78) (Ho *et al.* 1982; Flatz 1987) and some 87% have the **A** haplotype (Table 2). This combination is also frequent in Italy and Finland and in Indian Sikhs. However it is clear that lactase persistence can occasionally occur on other haplotypes. There are two examples of persistence on a **B** haplotype chromosome (one from the UK and one from Italy) and one example (from Italy) of lactase persistence carried on the rare **I** haplotype. In addition, and rather surprisingly, homozygosity of the **C** haplotype was seen in one of the lactase persistent Italians. Lactase non-persistence can be found on all haplotypes including **A**. Even in the Finnish data set non-persistence is carried on at least three haplotypes despite the fact the modern population is believed to have arisen from relatively few founders due to a population bottleneck some four thousand years ago (Sajantila *et al.* 1996). Efforts to subdivide the European **A** haplotype by analysis of additional polymorphisms within and flanking the gene have so far been unsuccessful.

Thus, as for the marker polymorphisms on their own, which associate to form only three common haplotypes, there is clear evidence of associations with lactase persistence, but the pattern of associations unfortunately gives no clear clues as to the direction we should search for the nucleotide change(s) which cause the lactase persistence/non-persistence polymorphism. It is possible however that the very few non-**A** persistent alleles may provide useful information and careful definition of the regions which are in common in the non-**A** and **A** persistent alleles is in progress. Furthermore study of the Caucasian chromosomes alone does not give any clear indications of the phylogeny of the major European haplotypes, but suggests a more complex history than point mutational and reciprocal recombination events alone. The results do however indicate that lactase persistence arose on the **A** haplotype, possibly as a unique event, and is thus probably more recent than the DNA marker polymorphisms used in this study. The high frequency of the **A** haplotype in northern Europeans is consistent with the high frequency of lactase persistence. It is of interest in this context that other (non-Caucasian) populations show a greater haplotype diversity, which together with the study of primates will give clearer insight into the origins of the haplotypes. It will ultimately be possible to map the persistence polymorphism onto this phylogeny which may help to resolve old arguments about the role of selection in the origin of this fascinating functional polymorphism (Holden and Mace 1997).

The British Digestive Foundation is gratefully acknowledged for support. We thank the summer students Monica Shaw and Delphine Loirat for their help.

REFERENCES

Boll, W., Wagner, P. & Mantei, N. (1991). 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.* **48**, 889–902.

Fernades, J., Vos, C. E., Douwes, A. C., Slotema, E. & Degenhart, H. J. (1978). Respiratory hydrogen excretion as a parameter for lactose malabsorption in children. *Am. J. Clin. Nutrit.* **31**, 597.

Flatz, G. (1987). Genetics of lactose digestion in humans. *Adv. Hum. Genet.* **16**, 1–77.

Grant, J. D., Bezerra, J. A., Thompson, S. H., Lemen, R. J., Kodolvsky, O. & Udall, J. N. (1989). Assessment of lactose absorption by measurement of urinary galactose. *Gastroenterology* **97**, 895.

Harvey, C. B., Pratt, W., Islam, I., Whitehouse, D. B. & Swallow, D. M. (1995). DNA polymorphisms in the lactase gene: linkage disequilibrium across the 70 kb region. *Eur. J. Hum. Genet.* **3**, 27–41.

Harvey, C. B., Wang, Y., Hughes, L. A., Swallow, D. M., Thurrell, W. P., Sams, V. R., Barton, R., Lanzon-Miller, S. & Sarner, M. (1995). Studies on the expression of intestinal lactase in different individuals. *Gut* **36**, 28–33.

Ho, M. W., Povey, S. & Swallow, D. M. (1982). Lactase polymorphism in adult British natives: estimating allele frequencies by enzyme assays in autopsy samples. *Am. J. Hum. Genet.* **34**, 650–657.

Holden, C. & Mace, R. (1997). Phylogenetic analysis of the evolution of lactase digestion in adults. *Human Biology* **69**, 605–628.

Lloyd, M., Mevissen, G., Fischer, M., Olsen, W., Goodspeed, D., Genini, M., Boll, W., Semenza, G. & Mantei, N. (1992). Regulation of intestinal lactase in adult hypolactasia. *J. Clin. Invest.* **89**, 524-529.

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

Mantei, N., Villa, M., Enzler, T., Wacker, H., Boll, W., James, P., Hunziker, W. & Semenza, G. (1988). Complete primary structure of human and rabbit lactase-phlorizin hydrolase: implications for biosynthesis, membrane anchoring and evolution of the enzyme. *EMBO J.* **7**, 2705–2713.

McLellan, T., Jorde, L. B. & Skolnick, M. H. (1984). Genetic distance between the Utah Mormons and related populations. *Am J Hum Genet* **36**, 836–857.

Robb, T. A. & Davidson, G. P. (1981). Advances in breath hydrogen quantitation in paediatrics: sample collection and normalisation to constant oxygen and nitrogen levels. *Clin. Chim. Acta.* **111**, 281.

Sajantila, A., Salem, A. H., Savolainen, P., Bauer, K., Gierig, C. & Paabo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc. Nat. Acad. Sci.* **93**, 12035–12039.

Swallow, D. M. & Harvey, C. B. (1993). Genetics of adult-type hypolactasia. *Common Food Intolerances 2: Milk in Human Nutrition and Adult-type hypolactasia* Ed. S. Auricchio and G. Semenza. Karger. 85–92.

Wang, Y., Harvey, C. B., Hollox, E. J., Phillips, A. D., Poulter, M., Clay, P., Walker-Smith, J. A. & Swallow, D. M. (1998). The genetically programmed down-regulation of lactase in children. *Gastroenterology* **114**, 1230–1236.

Wang, Y., Harvey, C. B., Pratt, W. S., Sams, V. R., Sarner, M., Rossi, M., Auricchio, S. & Swallow, D. M. (1995). The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Hum. Mol. Genet.* **4**, 657–662.

Weijers, H. A., van der Kamer, J. H., Dicke, W. K. & Ijsseling, J. (1961). Diarrhoea caused by deficiency of sugar splitting enzymes. *Acta. Paediatr.* **50**, 55.

European Journal of Human Genetics (1999) 7, 791–800
© 1999 Stockton Press All rights reserved 1018–4813/99 $15.00
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

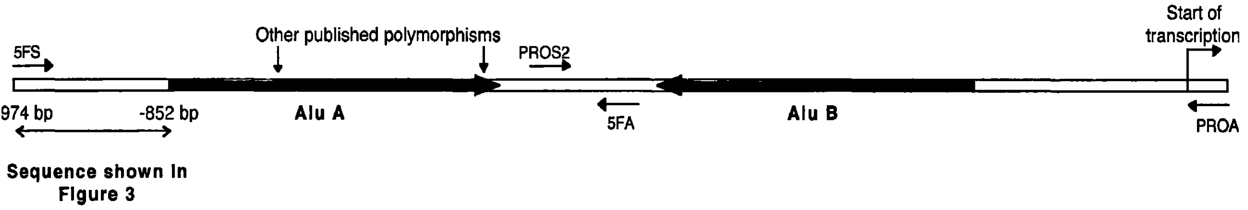


Figure 1 *Diagram showing the first 1 kb upstream from exon 1 of the human lactase gene. Positions of the PCR primers 5FA, 5FS, PROS2 and PROA are shown. The sequence shown in Figure 3 is underlined, and the other published polymorphisms are described previously.*[7]

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%

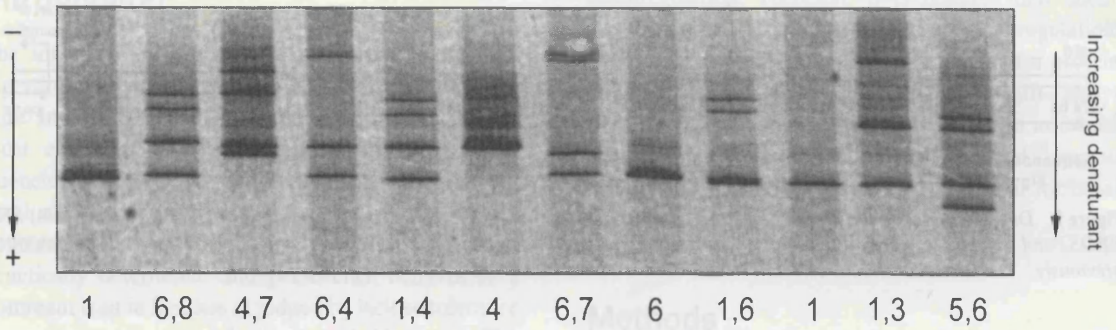


Figure 2 *The top half of a silver stained DGGE gel showing the seven variant alleles detectable in the Ava II digest fragment. The numbers assigned to the variants are below the appropriate lanes. Upper bands represent heteroduplexes. The differences between 1 and 6 could also be detected by SSCP analysis 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.

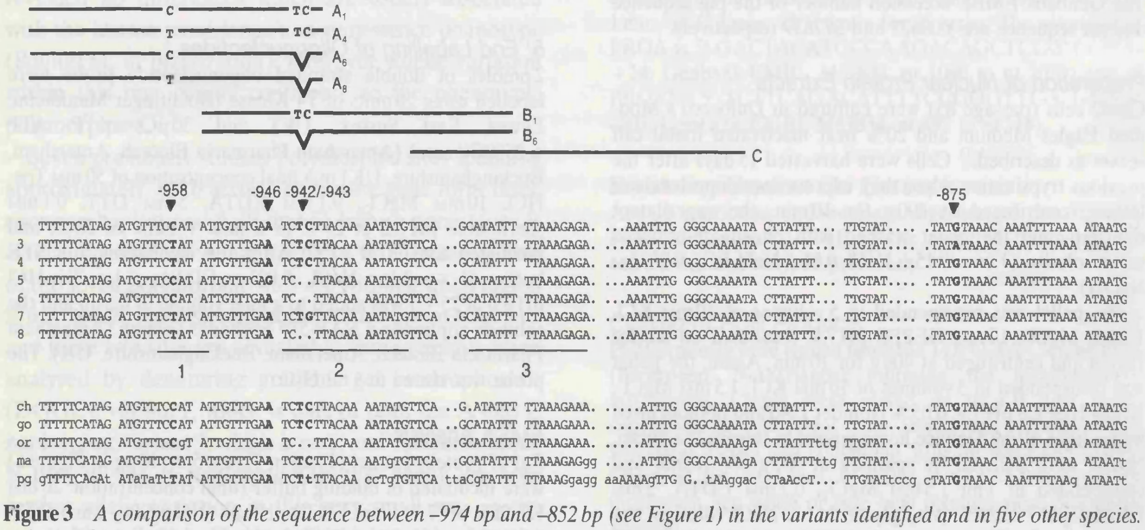
Electrophoresis was performed using Hoefer SL600 equipment (Amersham Pharmacia Biotech, Amersham, Buckinghamshire, UK) and kept at a constant temperature of 10°C using an LKB Bromma water bath.

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.



```
1  TTTTTCATAG ATGTTTCCAT ATTGTTTGAA TCTCTTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
3  TTTTTCATAG ATGTTTCTAT ATTGTTTGAA TCTCTTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATATAAAC AAATTTTAAA ATAATG
4  TTTTTCATAG ATGTTTCTAT ATTGTTTGAA TCTCTTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
5  TTTTTCATAG ATGTTTCCAT ATTGTTTGAG TC..TTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
6  TTTTTCATAG ATGTTTCCAT ATTGTTTGAA TC..TTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
7  TTTTTCATAG ATGTTTCTAT ATTGTTTGAA TCTGTTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
8  TTTTTCATAG ATGTTTCTAT ATTGTTTGAA TC..TTACAA AATATGTTCA ..GCATATTT TTAAAGAGA. ...AAATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG

ch TTTTTCATAG ATGTTTCCAT ATTGTTTGAA TCTCTTACAA AATATGTTCA ..G.ATATTT TTAAAGAAA. .....ATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
go TTTTTCATAG ATGTTTCCAT ATTGTTTGAA TCTCTTACAA AATATGTTCA ..GCATATTT TTAAAGAAA. .....ATTTG GGGCAAAATA CTTATTT... TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATG
or TTTTTCATAG ATGTTTCCgT ATTGTTTGAA TC..TTACAA AATATGTTCA ..GCATATTT TTAAAGAAA. .....ATTTG GGGCAAAAgA CTTATTTttg TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATt
ma TTTTTCATAG ATGTTTCCAT ATTGTTTGAA TCTCTTACAA AATgTGTTCA ..GCATATTT TTAAAGAGgg .....ATTTG GGGCAAAAgA CTTATTTttg TTGTAT.... .TATGTAAAC AAATTTTAAA ATAATt
pg gTTTTCAcAt ATaTaTtTAT ATTGTTTGAA TCTtTTACAA ccTgTGTTCA ttaCgTATTT TTAAAGgagg aaAAAAgTTG G..tAAggac CTaAccT... TTGTATtccg cTATGTAAAC AAATTTTAAg ATAATt
```

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

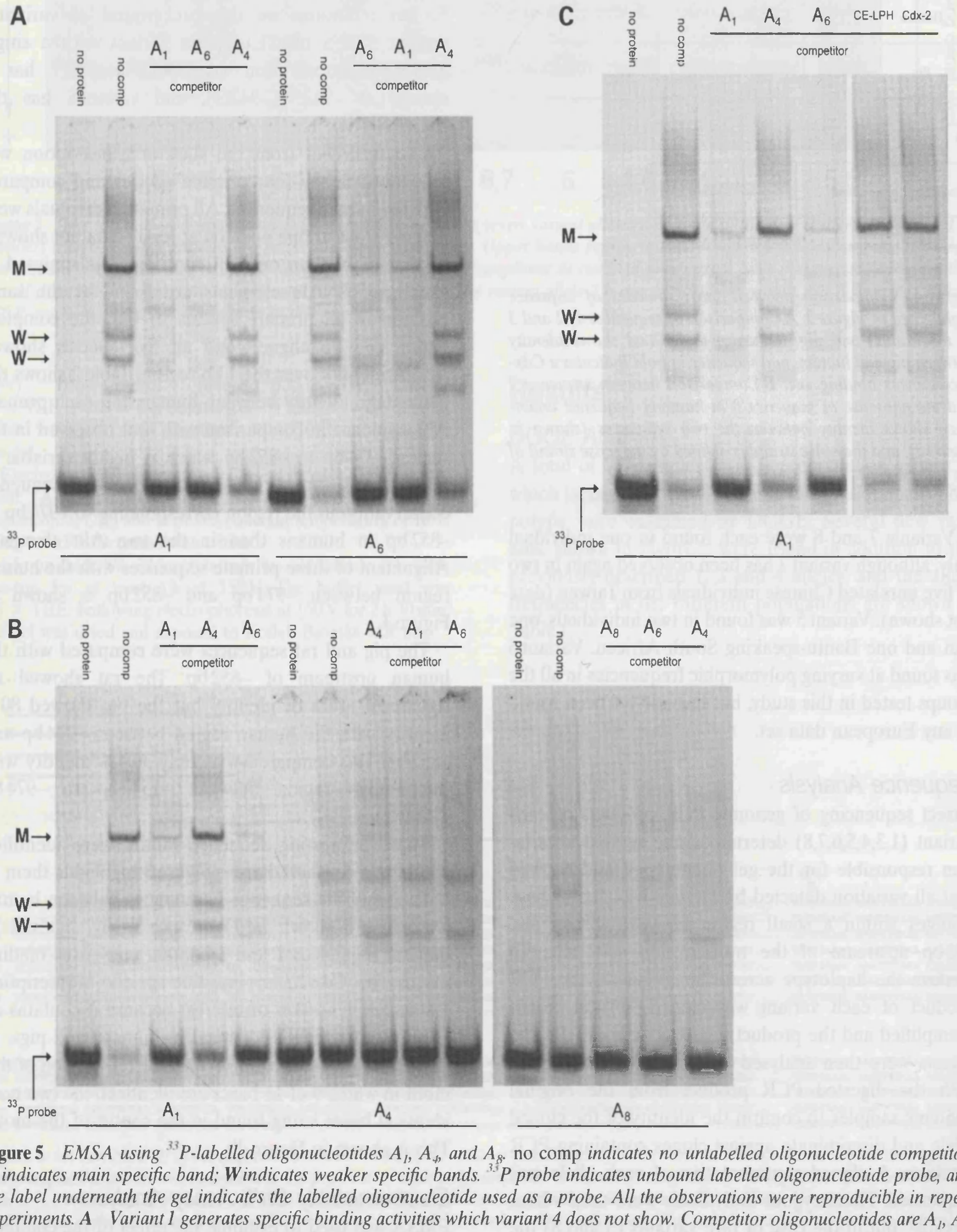


Figure 5 *EMSA using ^{33}P-labelled oligonucleotides A_1, A_4, and A_8.* no comp *indicates no unlabelled oligonucleotide competitor;* **M** *indicates main specific band;* **W** *indicates weaker specific bands.* ^{33}*P probe indicates unbound labelled oligonucleotide probe, and the label underneath 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

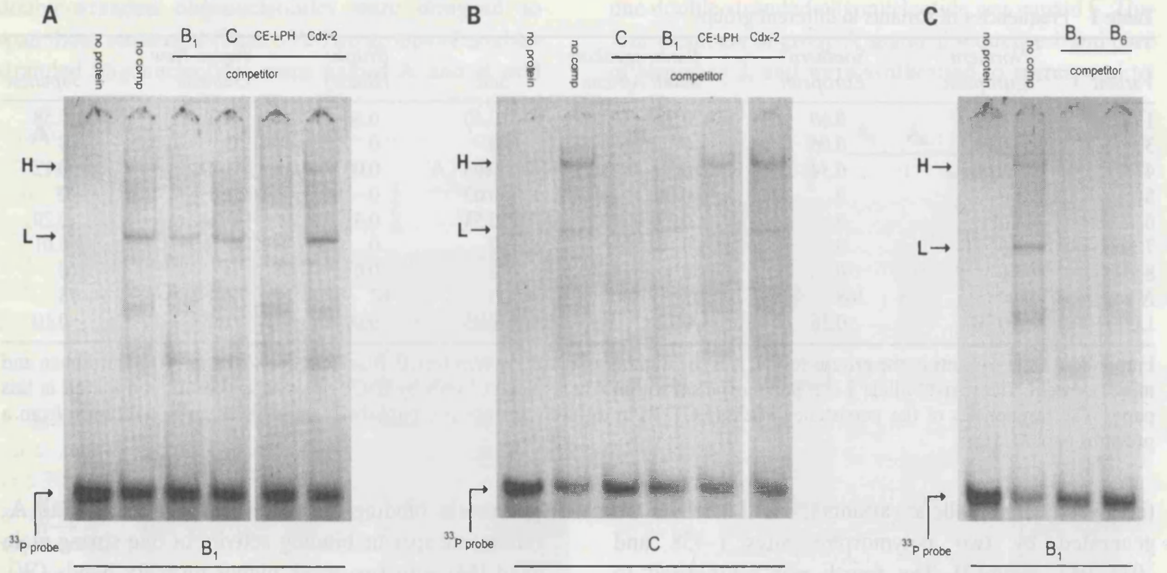


Figure 6 *EMSA using ^{33}P-labelled oligonucleotide B_1 or C.* no comp *indicates no unlabelled oligonucleotide competitor;* H *indicates the higher specific band,* L *indicates the lower specific band.* ^{33}*P probe indicates unbound labelled oligonucleotide probe, and the label underneath the gel 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.