The Evolutionary Genetics of Lactase Persistence in Africa and the Middle East

Catherine Janet Ellen Ingram

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

May 2008

Galton Laboratory
Department of Biology
University College London

UMI Number: U591579

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI U591579
Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

I, Catherine Ingram, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Lactase, the enzyme responsible for milk digestion, is expressed in the small intestine of nearly all neonate mammals, and normally down-regulates following weaning. This is the ancestral state and in humans is described as lactase non-persistent. However, some people continue to have high expression of the enzyme for life due to a genetically inherited variation known as lactase persistence.

A single nucleotide polymorphism, *-13910*T* was identified as the causal variation in Europeans due to a very tight association with phenotype and evidence of a functional effect *in vitro*. Subsequently, an apparent disparity was observed between *-13910*T* frequency and reported lactase persistence frequency in some African populations, raising doubts about the causal nature of the allele. Two possible explanations were proposed; either *-13910*T* is not causal, but in Europeans is tightly linked to the true cause of lactase persistence, or, *-13910*T* is causal in Europeans, but the trait has evolved independently elsewhere.

The primary aim of this thesis was to investigate the causes of lactase persistence in sub-Saharan Africa. The occurrence of only one *-13910*T* carrier out of 45 lactase persistent people from a cohort of phenotyped Sudanese individuals provided confirmation that the allele is not causal worldwide. Haplotype analysis of a 70kb region spanning the lactase gene in the phenotyped cohort and in non-European pastoralist groups provided no evidence for a shared origin with the European mutation.

Resequencing of the –13910 locus led to the identification of a number of candidate SNPs; *–13915T>G*, *-13913T>C* and *–13907C>G*, all located within 5bp of the original variant. Despite being clustered within the same OCT1 protein binding site as *-13910*T*, gel shift experiments revealed that the new alleles did not have a common effect on protein binding. However, *–13915*G* showed a significant association with lactase persistence.

Resequencing of a second phenotyped cohort revealed the presence of many variant alleles at the locus, the occurrence of which is significantly higher in persistent individuals. Nearly every allele associates with an independent haplotype, providing strong evidence that multiple unrelated evolutionary events gave rise to lactase persistence.

The frequency and distribution of all newly identified alleles was surveyed in more than 700 individuals from a total of 18 African and Middle Eastern populations, and gives a preliminary indication of the geographic origin of some alleles.

The clustering of lactase persistence associated alleles within a single regulatory element implies that they are causal, and possible mechanisms and future approaches are discussed.

Acknowledgements

It is a pleasure to thank the many people whose kind assistance and support made this thesis possible.

First and foremost I would like to thank my PhD supervisor Professor Dallas Swallow, whose enthusiasm for this project has been both inspiring and contagious, and who has provided invaluable support and advice in academic and non-academic areas alike.

I also wish to thank my second supervisors: Dr Mark Thomas, for sharing his ideas and comments on many aspects of this project; and Dr Mike Weale for his expert help with statistical analysis and for reading earlier chapter drafts.

I am very grateful to Dr Neil Bradman for his continued interest in and discussion of this project, and for providing the opportunities to present at meetings and collect samples that I may otherwise not have had.

I would also like to thank all the individuals involved in collecting the samples used in this project: in particular Mohammed Elamin, who single-handedly collected the phenotyped Sudanese cohort, and Ayele Tarekegn and Tamiru Olijira Raga who helped collect the second cohort in Ethiopia, along with Sarah Browning, to whom I am especially indebted for her good humour and dubious cockroach catching skills.

I am very appreciative of the many individuals who made this project possible by kindly donating their DNA, and some of whom spared extra time (and risked their digestive calm) to be lactose tolerance tested.

There are a number of members of the TCGA who I would like to thank for their support and advice: Abigail Jones for my induction into the 'ways of the TCGA'; Krishna Veeramah for the statistical advice and for writing indispensable scripts for data analysis; Chris Plaster for all his patient help with Y-chromosome and mtDNA typing; Yuval Itan for assistance with databases, alignments and R, and Naser Pour for the impromptu lessons in basic statistics.

I would also like to thank Fiona Pring, Felicity Copp, Rhonda Sturley, Caroline Freeman and Kifaia Rashid, all of whose enthusiastic efforts in their undergraduate projects made an important contribution towards this thesis.

I am very grateful to all the past and present members of the Galton Laboratory who have contributed considerably to my enjoyment of the last 4 years, but I would particularly like to thank Lynne Vinall and Karine Rousseau for their assistance at the beginning of my PhD, and Laura Horsfall, Ana Teixeira and Andrew Loh who each put up with sharing an office with me at various points and who were all excellent company. Special thanks as well go to Ranji Arasaretnam who tirelessly helped and advised with many laboratory issues including DNA extraction and troublesome gels.

On a personal level there are many friends and family members who have supported me and to whom I am very grateful. I would particularly like to thank Michela Perani, who encouraged me to do a PhD in the first place, and Lorenzo Zanette, just for being him.

Special thanks go to Sarah Goodley and Becky Crowson for their continued support and friendship, which has been essential to me over the last few years.

Finally and most importantly I wish to thank my family; my brothers, John, David and James who have each provided support, encouragement and quick-witted counsel on various topics, and my parents, Janet and Jack, for the infinite number of ways in which they have shown their love and support for me. This thesis is for them.

Catherine Ingram
May 2008

Table of Contents

Contents of Tables

Contents of Figures

Abbreviations and Acronyms

3C	chromatin conformation capture
AMOVA	Analysis of Molecular Variance
ARMS-PCR	Amplification Refractory Mutation System-PCR
bp	base pair
CDX2	caudal type homeobox 2
ChIP	Chromatin ImmunoPrecipitation
CNV	Copy Number Variation
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribose Nucleic Acid
DTT	1,4-Dithiothreitol
EDTA	ethylene diamine tetraacetic acid
EHH	Extended Haplotype Homozygosity
EM	Expectation Maximisation
EMSA	Electrophoretic Mobility Shift Assay
ER	Endoplasmic reticulum
EtOH	Ethyl alcohol (ethanol)
GATA4/5/6	GATA binding protein 4/5/6
HNF1A	HNF1 homeobox A
HVR	Hyper variable region
HWE	Hardy-Weinberg Equilibrium
ID	identification
kb	kilobase
kDa	kilodalton
kya	thousand years ago
LCT	Lactase
LD	Linkage Disequilibrium
LPH	Lactase Phlorizin Hydrolase
LRH	Long Range Haplotype
Mb	megabase
MCM6	Minichromosome maintenance deficient 6
MET	methionine

mRNA	messenger RNA
mtDNA	mitochondrial DNA
OCT1	Octamer-binding transcription factor 1/POU domain, class 2, transcription factor 1 (encoded by the gene *POU2F1*)
PBS	Phosphate Buffered Saline
PCO	Principle co-ordinates
PCR	Polymerase Chain Reaction
PEG	Poly ethylene glycol
PMSF	phenylmethanesulphonylfluoride
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SDS	Sodium dodecyl sulfate
SNP	Single nucleotide polymorphism
SSCP	Single strand conformation polymorphism
STE	Sodium Chloride-Tris-EDTA
STR	Short tandem repeat
T.E.	Tris-EDTA
TBE	Tris-borate EDTA
TCGA	The Centre for Genetic Anthropology
UCLH	University College London Hospital
UEP	Unique Event polymorphism
URL	Uniform Resource Locator
UVR	Ultra-violet radiation

1 Introduction

This thesis is concerned with the molecular and population genetics of the variation in expression of the human enzyme lactase. Lactase variation is a classic example of how changes in the diet of ancestral populations have influenced the pattern of genetic diversity we observe today. A number of factors cause allele frequency differences between populations, including demography and drift, but it is becoming increasingly clear that selection has played an important role. This introduction reviews current knowledge on genetic variation in humans and describes some of the ways in which the distribution of this variation has been studied, with a particular focus on selection. Examples are given of other genetic variations that have been studied in parallel with the work on lactase. The background on lactase persistence is then described in detail, followed by the specific aims of this thesis.

1.1 Part I: Population genetics and human genetic adaptation

1.1.1 Extent of molecular variability in humans

Originally revealed by protein polymorphisms in the 1960s, genetic variation in humans is widespread and exists in a number of forms. The most common of these are substitutions of one nucleotide for another, commonly known as simple nucleotide polymorphisms (SNPs) for which there now exists an enormous amount of information (The International HapMap Consortium, 2005). Other types of genetic variation exist in the form of tandem repeat sequences known as micro-satellites or short tandem repeats (STRs). These consist of runs of a short (2-5bp) nucleotide motif repeated a number of times in succession. STRs are highly polymorphic and make a large contribution to inter-individual genome variation. Mini-satellites also consist of a repeated motif, but the repeat unit is larger (up to 70bp in length). Variation is also seen in LINES and SINES (**L**ong and **S**hort **IN**terspersed **E**lements), distinct mobile DNA sequences known as transposons which can re-locate to non-homologous parts of the genome. Nearly half of the human genome is composed of these elements (Smit, 1999; Lander *et al.*, 2001), and the repeat units can vary between individuals (and populations) in terms of sequence composition as well as presence/absence of the repetitive element itself (Novick *et al.*,

2008). More recently it has been shown that copy number variation (CNV) makes a large contribution to genetic diversity in humans. The term CNV describes variation which can range from kilobases to megabases in size and includes deletions and insertions, duplications and multi-allelic loci as well as complex loci whose precise nature is not easy to discern. It has been estimated that 12% of the genome is copy number variable (Redon *et al.*, 2006), and it has been shown that CNV is a major source of genetic variation in humans and has an important role to play in population genetics (Jakobsson *et al.*, 2008). CNV is also likely to be a common contributor to phenotypic variation (de Smith *et al.*, 2007).

A recent study suggests that when all types of variation are examined a minimum of 0.5% variation exists between two haploid genomes (Levy *et al.*, 2007). Past estimates of diversity in humans yielded nucleotide diversity values (π) of approximately 1/1000, indicating that, on average, a pair of humans will differ at about one in 1000 DNA base pairs (Jorde, 2005), however it now seems that the real figure is more like five in 1000. For most autosomal loci the greatest genetic diversity exists within populations. A smaller proportion of variation exists between continental populations and the least is found between populations within the major continents (Jorde *et al.*, 2000).

1.1.2 The origins and dispersal of modern humans

It is now generally accepted that modern humans originated in Africa some 100-200 thousand years ago, and subsequently migrated to Europe and Asia, eventually dispersing and settling throughout the world (Figure 1.1) (Cavalli-Sforza *et al.*, 1994; Templeton, 2002; Jobling *et al.*, 2004). The fossil record is compatible with this version of events, and consistent genetic data continues to accrue: global genetic diversity is a sub-set of that found within Africa (e.g. (Batzer *et al.*, 1996; Jorde *et al.*, 2000)), genetic diversity decreases as a function of distance from Africa (Handley *et al.*, 2007), and greater genetic diversity and lower linkage disequilibrium is observed within African populations (Gabriel *et al.*, 2002)).

In consideration alongside the fossil record, genetic diversity observed within and between contemporary human populations is used in an attempt to estimate when

particular geographic locations have been settled, and to summarise migrations and gene flow between populations inhabiting them. Genetic diversity data also informs upon genetic variation that persists due to selection, which may be population specific owing to local differences in environment. This has long been an area of interest for evolutionary geneticists, but it is now increasingly considered that past adaptations may have direct implications on present day health, with consequent increase in interest to a wider audience.

Figure 1.1 Possible history and routes of expansion of modern humans in the last 100,000 years (Figure from Cavalli-Sforza et. al., 1994).

1.1.3 Population genetic theory and selection

Population genetics uses variation in allele and haplotype frequencies to infer the evolutionary processes that have occurred or are occurring within a population. The main forces that create and maintain genetic variation between populations are mutation, gene flow, drift and selection. Population genetic models aim to account for and differentiate between the effects of these forces in order to understand the evolutionary and demographic histories of different populations and reconstruct the events that may have produced the patterns of genetic diversity observed. Of particular relevance to this

thesis are methods devised to detect selection. Numerous models exist which aim to identify the imprint of selection on molecular data. These 'signatures' or patterns of genetic diversity can be classified into five broad categories (Sabeti *et al.*, 2006):

1.1.3.1 Function-altering mutations (e.g. McDonald Kreitman; McDonald and Kreitman, 1991)

Variants that alter function are usually deleterious and are therefore unlikely to increase in frequency within a population. Thus when an excess of non-synonymous (potentially function-altering) mutations is detected, this can indicate an adaptive change. Determination of an excess of non-synonymous substitutions can be achieved by calculating the ratio of these against synonymous (assumed non-function altering) mutations in the same region, or by comparison with the number of non-synonymous mutations observed in other (closely related) species. This test however does represent an over-simplifed model of amino-acid substitution, as synonymous mutation may alter protein function and in some cases non-synonymous substitutions have little effect.

1.1.3.2 Diversity and frequency spectrum (e.g. Tajima's D; Tajima, 1989)

A consequence of an advantageous allele increasing in frequency within a population is that the surrounding linked alleles also increase or 'hitchhike' in frequency. This has the effect of reducing diversity in the region until new mutations or recombination events occur, and eventually restore diversity. As the mutation rate is low (for SNPs, 10^{-8} per nucleotide per generation; Jobling *et al.*, 2004) these new substitutions are rare, and so a region of low overall diversity with an excess of rare alleles can indicate a past selective event. Tests in this category evaluate whether there is an excess of alleles in a number of ways, e.g. by comparing the ratio of polymorphism to divergence (with a closely related out-group), or by calculating the number of singletons as a proportion of the total number of variant sites.

1.1.3.3 High frequency derived alleles (e.g. Fay and Wu's H; Fay and Wu, 2000)

Derived alleles occur by new mutations, and under neutral evolution these take a long time to rise to high frequency. However as described above (section 1.1.3.2), during a selective sweep derived alleles that are linked to a beneficial allele 'hitch-hike' to high frequency. Many of the derived alleles will not reach complete fixation due to an incomplete sweep or because of recombination of the selected allele during the sweep. Thus a region containing many high frequency derived alleles is created. This can be measured by determining the fraction of markers with a high derived allele frequency within a given region.

1.1.3.4 Population differentiation (e.g. F_{ST}; Wright, 1951)

Different environmental or cultural pressures may be imposed on geographically separated populations resulting in positive selection acting upon a given allele in one population but not in another. For this reason, large allele frequency differences between populations may indicate the presence of selection (although this pattern of genetic diversity can also be caused by demographic effects, in particular population bottle necks). The simplest measure of genetic distance would be to take two populations (X and Y) where the frequency of the ith allele is X_i or Y_i and sum the difference between allele frequencies; i.e. $\Sigma (X_i-Y_i)^2$, squaring to avoid differences in sign cancelling each other out. However this approach does not give sufficient weight to alleles with frequencies close to 0 or 1. There are a number of ways of calculating F_{ST} (see section 2.8.6 for the method used within this thesis) which is essentially a weighted form of the above measure.

1.1.3.5 Haplotype length and diversity (e.g. Long Range Haplotype test (LRH); Sabeti et al., 2002)

A third category of test, the long range haplotype test, also uses the pattern of genetic diversity caused by a selective sweep, i.e. a high frequency allele with an extended region of linkage disequilibrium surrounding it, known as extended haplotype homozygosity (EHH), to identify alleles under positive selection. In most cases, high frequency alleles are old and therefore recombination events have occurred which break

down the associations of nearby alleles on the chromosome. Under positive selection a young allele at very high frequency can exhibit LD across very extended distances. To evaluate whether a haplotype is unusually long, comparisons are made both with simulated and empirical data (if available). This method of detecting selection is used with increasing frequency in the wake of publicly available SNP databases.

1.1.4 Human Adaptation

Recently many of these tests have been applied on genome wide data, e.g. (Sabeti *et al.*, 2007; Barreiro *et al.*, 2008), but in parallel with lactase research there are also many examples of studies on individual loci in which the molecular mechanisms and selective forces (such as disease and diet) are both examined. In the following section some examples are given of how important selective factors that are thought to have influenced human diversity and evolution have been identified and investigated.

A growing body of research indicates that infectious diseases have played a considerable role in shaping the distribution of genetic variation in different geographic regions by providing a selective force that maintains protective alleles at relatively high frequencies. The most comprehensive evidence of this is the variety of alleles that have been shown to be advantageous to humans inhabiting areas of malarial endemicity. In 2002, malaria ranked 12th in the global rankings of the most frequent causes of death (Mathers and Loncar, 2006), although its ranking would be higher amongst populations inhabiting only tropical and sub-tropical areas. Malaria infection has impacted human health since historically ancient times, and was once far more widespread than its present day distribution (Carter and Mendis, 2002). A number of genetic variations, mainly affecting proteins expressed within and on the surface of erythrocytes, have been shown to associate with malaria resistance. Most of these diseases were originally identified as candidates of selection due to the parallel distribution of the disease and areas of endemic malaria.

1.1.4.1 Sickle-cell anaemia

The allele causing sickle-cell anaemia is the classic example of positive selection in humans caused by exposure to malaria. The amino acid substitution that characterises haemoglobin S (HbS) causes the protein to crystallise at low oxygen tension, giving blood cells a sickle-shaped appearance. Heterozygotes for the HbS allele are phenotypically normal carriers (except at very at low oxygen tension conditions; Martin *et al.*, 1989), however HbS homozygotes have sickle-cell anaemia, which in the past was frequently fatal, often before age 30. It was first suggested a very long time ago that sickle-cell disease offered some protection against malaria due to the observation that in the absence of malaria, the HbS allele is absent, or present only at very low frequency (Haldane, 1949).

This advantage was confirmed by comparison of the infection rates of individuals of different genotypes, which revealed that fatal infection is significantly reduced among HbS carriers (Allison, 1954; Ruwende *et al.*, 1995). It is now known that carriers of the sickle gene defend themselves better because infection by *P. falciparum* induces sickling of the sickle-trait cells which are then selectively destroyed by the spleen. If an affected sickle-trait cell escapes this barrier then polymerisation of haemoglobin (under extreme deoxygenation) offers a secondary protection, as parasite proteases cannot digest the polymerised protein and starve (reviewed in Nagel, 2005). More recently the EHH approach has revealed that two different HbS alleles have greater haplotype homozygosity (extending across 400Kb) than other haplotypes at the same locus (Hanchard *et al.*, 2007), consistent with selection for the allele.

1.1.4.2 Thalassaemia

A large body of indirect evidence (including allele frequency and distribution data) exists that suggests that thalassaemias also bestow a protective effect against malaria, although the precise mode of protection remains to be fully understood. In this group of globin diseases, there is usually no structural change to the haemoglobin subunits themselves, but red blood cells have either very low or zero concentrations of either the α or β chains which form the normal haemoglobin molecule, HbA ($\alpha_2\beta_2$). There are two copies of the α-globin gene on chromosome 16 (i.e. two copies of α-globin are inherited from each parent) and α-thalassaemias are caused by deletion of either one or both of

these. α-thalassemia is most common in Africa, the Middle East, India and S. East Asia. The β-thalassaemias are most common in Europe and Africa. They involve various mutations of the β-globin gene, located on chromosome 11. Some of these mutations result in the elimination of β -globin production, whereas others alter β -globin production levels. The result in both cases is an excess of α-chain globin molecules (for reviews, see Weatherall *et al.*, 2002; Weatherall, 2004).

Extensive population studies, case-control and cohort analyses have enabled elucidation of the relative roles of founder effects, genetic drift and selection in achieving the observed frequencies and distribution of thalassaemia alleles. The resulting data reveals that many populations have their own 'private' polymorphisms that cause thalassaemia, indicating multiple evolutionary origins of the disease (Weatherall, 2004).

(Flint *et al.*, 1986) found that α^+-thalassaemia in the S.W. Pacific correlates highly with malaria endemicity, and this correlation is not observed for genetic markers at other loci. The authors were able to reject the null hypothesis that α^+-thalassaemia had spread from mainland S.E. Asia because of the finding that a different mutation was prevalent in Melanesia and Papua New Guinea. In contrast, it was concluded that the intermediate frequencies of α^+-thalassaemia ($\leq$0.15) observed in Fiji and west Tahiti (where malaria has never been recorded) were best explained by population migration. In this case it was found that nearly all α^+-thalassaemia in the region was accounted for by a single mutation previously defined in Vanuatu on an identical haplotype background.

More recently, a cohort study of children in Papua New Guinea found that both heterozygotes and homozygotes for the α^+-thalassaemia allele were afforded a markedly reduced risk of contracting severe malaria (Allen *et al.*, 1997). This protective effect has recently been confirmed by (Wambua *et al.*, 2006), in a Kenyan cohort. The authors found that carriers of the α^+-thalassaemia allele were not more resistant to malarial infection but did have a significantly reduced risk of contracting fatal malaria.

1.1.4.3 Duffy Antigen Receptor for Chemokines (DARC)

The distribution of the *DARC* gene alleles clearly demonstrates how positive selection can cause large allele frequency differences between populations. The *DARC* gene

encodes a receptor molecule, known as the Duffy antigen, which is expressed on the surface of erythrocytes and other cells. There are three main alleles of the of the *DARC* gene which give rise to three isoforms of the Duffy antigen: A, B and O, and the global distribution and frequencies of each are highly variable. FY*O shows the highest population differentiation and has reached/ is close to fixation in Africa, but rare in other populations. The FY*A allele is at fixation in eastern Asia and the Pacific, and the ancestral FY*B allele is absent in many populations (Jobling *et al.*, 2004), but peaks in frequency between northern Brazil and the Guianas (Cavalli-Sforza *et al.*, 1994).

Molecular characterisation of the FY*O allele reveals a T/C substitution in the transcription initiation site, which prevents transcription factor binding and ultimately abolishes transcription of the *DARC* gene product in reticulocytes (Tournamille *et al.*, 1995). Erythrocytes lacking the Duffy antigen on their surface are resistant to *Plasmodium Vivax* infection (Barnwell *et al.*, 1989). Molecular evidence of selection at the FY*O locus include the observation that the DNA sequence shows reduced variation in African groups compared to Europeans (Hamblin and Di Rienzo, 2000), and that the large differences in F_{ST}s evident at the *DARC* locus are not observed at other (unlinked) loci (Hamblin *et al.*, 2002).

1.1.4.4 Glucose-6-Phosphate dehydrogenase (G6PD)

Variations of the G6PD gene are also thought to have been maintained within populations due to selection for resistance to malarial infection. G6PD is a housekeeping enzyme which catalyses the first step of the pentose phosphate pathway (glucose-6-phosphate + NADP $\rightarrow$ 6-phospho-gluconate + NADPH). In red blood cells the pentose phosphate pathway is the only mechanism by which NADP is reduced to NADPH and is essential for preventing oxidative damage (Stryer, 1995; Mason and Vulliamy, 2005). A number of different mutations in the G6PD gene have been identified which result in deficiency of the enzyme. The effect of this (haemolytic anaemia) is manifested when triggered by foods, drugs or infection (reviewed in Cappellini and Fiorelli, 2008).

The gene encoding G6PD is located on the X chromosome, and more than 400 distinct variants have been characterised at the DNA level. G6PD was also identified as a candidate for selection by virtue of the coincident distribution of G6PD deficiency and malaria (for review, see Beutler, 1994; Ruwende and Hill, 1998). Molecular and cohort studies are compatible with this hypothesis: Patients with G6PD mutations tend to have a lower parasite load than those with 100% functional G6PD alleles (Beutler, 1994; Ruwende and Hill, 1998); parasite growth is inhibited in G6PD-deficient cells compared with G6PD-sufficient cells (Luzzatto *et al.*, 1969; Roth *et al.*, 1983); and case-control data shows that the most common form of G6PD deficiency in Africa (G6PD 202A-) is associated with a 46-58% reduction in risk of severe infection for both female heterozygotes and male hemizygotes (Ruwende *et al.*, 1995).

The G6PD 202A- haplotype also exhibits a significantly exaggerated level of homozygosity (Sabeti *et al.*, 2002), and both SNP and microsatellite variation display reduced diversity in comparison to simulated data and other haplotypes at the same locus (Tishkoff *et al.*, 2001; Verrelli *et al.*, 2002). It has been hypothesised that severe malaria has only impacted heavily on human populations since the development of agriculture (~10,000 years ago), and estimates of the date of expansion of the G6PD 202 A- allele at ~4000-12,000 years old (Tishkoff *et al.*, 2001) are compatible with this time frame.

1.1.4.5 Other diseases

A number of infectious diseases other than malaria are thought to have influenced the pattern of genetic variation observed in present day populations. However, the evidence supporting selection is often less substantial, and the nature of the selective advantage not well characterised. For example, it has been proposed that the high frequency of Cystic Fibrosis in European populations is due to balancing selection acting on variation within the cystic fibrosis transmembrane conductance regulator gene (*CFTR*). Heterozygous individuals are thought to have an increased resistance to Cl- secreting diarrhoeas (Romeo *et al.*, 1989): It has been shown that infection of the gastro-intestinal epithelial cells by typhoid causing *S. typhi* is reduced in the presence of variant CFTR receptors (Pier *et al.*, 1998), and that in mouse models heterozygotes for variant *CFTR* secret less fluid in response to cholera infection (Gabriel *et al.*, 1994). However, a

selective advantage of the allele (using 3 STR markers) is only detected under certain demographic models and dates for expansion of the allele (Wiuf, 2001) and the putative benefits of *CFTR* mutations remain unresolved.

Similarly, debate also surrounds the selective advantage that led to the current distribution of the *CCR5-Δ32* allele which provides resistance against HIV infection in homozygotes and delayed onset of AIDS in heterozygotes (reviewed in de Silva and Stumpf, 2004). The 32bp deletion results in a null allele of the chemokine receptor CCR5. Molecular indications of selection at the *CCR5* locus include the high proportion of non-synonymous mutations, suggesting adaptive evolution (Carrington *et al.*, 1997) and large between-population differences in frequency of *CCR5-Δ32*. The allele is suggested to have been the subject of a selective sweep due to its high frequency and the EHH associated with it (Stephens *et al.*, 1998). However, AIDS is a relatively recent cause of wide-spread mortality and therefore resistance to HIV infection is not thought to be the selective advantage responsible for the current distribution of the allele. It has been suggested that resistance to infection with either plague (Stephens *et al.*, 1998) or smallpox (Galvani and Slatkin, 2003) may have increased the frequency of *CCR5-Δ32*, however the subsequent finding that the allele is widespread in Bronze age populations and at equal frequency in plague and non-plague related deaths have discredited the hypothesis that this was the selective agent (Hummel *et al.*, 2005), and the true nature of the selective force, (if any) is still under debate (Schliekelman *et al.*, 2001; Sabeti *et al.*, 2005; Hedrick and Verrelli, 2006).

Balancing selection has also been reported at the prion protein gene (*PRNP*). Heterozygosity at codon 129 confers resistance to the prion disease Kuru, prevalent in the Papua-New-Guinean populations who traditionally attended endocannibalistic feasts. Strong evidence for selection of this allele is provided by the excess of non-synonymous substitutions observed, reduced microsatellite diversity and extensive linkage disequilibrium surrounding the allele. Furthermore, the authors found an over-representation of heterozygotes in generations exposed to the disease (>50) compared to unexposed younger generations (Mead *et al.*, 2003).

1.1.4.6 Skin pigmentation

Human adaptation has not only occurred in response to disease, and one of the more obvious examples of phenotypic differences between modern human populations is variation in skin pigmentation, thought to be caused by localised adaptation to ultra-violet radiation (UVR) intensity. Dark skin pigmentation is thought to provide protection against the harmful effects of UVR, (i.e. skin cancer and folic acid degradation), and at higher latitudes light skin pigmentation may have been beneficial in facilitating vitamin D synthesis (reviewed in Barsh, 2003; Diamond, 2005).

Mutations which affect skin pigmentation have been identified in a number of genes (Parra, 2007), however, two recent independent studies use SNP genotype data to show that variation causing light skin pigmentation has evolved independently in European and Asian populations (Norton *et al.*, 2007; Lao *et al.*, 2007). Lao *et al.* identified eight candidate genes (*SLC45A2*, *OCA2*, *TYRP1*, *DCT*, *KITLG*, *EGFR*, *DRD2* and *PPARD*) that showed higher levels of inter-population differentiation than expected (by comparison with other loci) between African, European and Asian populations. Four of the genes (*SLC45A2, TYRP1, DCT* and *OCA2*) have function related to pigmentation, and show extended haplotype homozygosity in European and Asian (but not African) populations. The authors conclude from these data that dark skin is likely to have been ancestral in humans and, because different extended haplotypes are observed at the *KITLG* locus in Asia and in Europe, suggest that light skin has evolved independently in each location. Consistent with these results, a separate study using F_{ST} and admixture mapping identified three genes (*SLC24A5, MATP* and *TYR*) involved in regulating skin-pigmentation that also showed evidence of selection in Europeans but not east Asians (Norton *et al.*, 2007).

1.1.4.7 Diet and Digestion

The enormous variation in dietary habits of ancestral human populations has led to speculation that this provided an important selective pressure for human adaptation of genes involved in digestion and metabolism (Eaton and Konner, 1985; Stover, 2006). For example, starch consumption makes a variable contribution to the calorific intake of different human populations: it is significant in hunter-gatherer populations in arid

environments, whereas other groups, such as pastoralists, rely more heavily on different food sources. The observation that an enzyme involved in starch digestion, the human salivary amylase gene (*AMY1*) shows extensive copy number variation (Groot *et al.*, 1989) within the human lineage (Fortna *et al.*, 2004; Wilson *et al.*, 2006) has led to the hypothesis that this variation was directly affected by the starch composition of different populations' diets.

A recent study demonstrated a significant correlation between expression of salivary amylase and copy number of *AMY1*, and reported that in populations with traditionally starch-rich diets (European Americans, Japanese, Hadza) 70% of individuals had six or more copies, compared to 37% in groups where starch consumption was traditionally low (Biaka, Mbuti, Datog and Yakut). The authors found that diet was a better predictor of *AMY1* copy number than a geographical migration model incorporating drift and concluded that copy number in low-starch groups was probably subject to neutral selection, but that in high-starch consumption groups diet had provided a positive selection pressure for increased copy number (Perry *et al.*, 2007). Similar studies carried out independently agreed that populations were more alike at the locus when grouped by diet rather than geography. However, no evidence for selection was found either using microsatellites to type for allelic diversity, or using F_{ST}s calculated for the *AMY1* locus in comparison with the empirical genomic distribution of F_{ST}s (Caldwell, 2005).

Another variation thought to be affected by dietary habit occurs in the alanine: glyoxylate aminotransferase gene (*AGXT*). *AGXT* is responsible for converting the intermediate metabolite glyoxylate to glycine, and to do this efficiently, AGT must be present at the site of glyoxylate synthesis. In herbivores glycolate is the major precursor of glyoxylate synthesis and is converted to glyoxylate in peroxisomes (Noguchi, 1987), whereas in carnivores the major precursor is proline, and is converted to glyoxylate in the mitochondria (Takayama *et al.*, 2003).

A non-synonymous polymorphism has been identified that results in an alteration of the cellular location of a proportion of AGT from the peroxisomes to mitochondria, and it has been suggested that targeting of AGT to the mitochondria may confer an advantage

for individuals consuming a meat-rich diet (Takayama *et al.*, 2003). Investigation of this variation in animals (Danpure *et al.*, 1990; Danpure *et al.*, 1994; Holbrook *et al.*, 2000; Takayama *et al.*, 2003; Birdsey *et al.*, 2004) reveals that the location of AGT is mainly peroxisomal in herbivores and mitochondrial in carnivores. A study of allele frequency of the targeting polymorphism in human populations showed that the Saami (who traditionally have a meat-rich diet), had a higher frequency of the mitochondrial-targeted allele than the other populations tested, and that the difference in allele frequency between populations is greater at the *AGXT* locus than at comparator sites (Caldwell *et al.*, 2004). These observations would appear to indicate that the frequency of the mitochondrial-targeting allele has been increased by dietary selection in populations who rely heavily on meat as a food source.

There are a number of other examples of genetic variation in digestive enzymes (Swallow, 2003), for example, sucrase isomaltase and lactase. Polymorphisms associated with reduced expression or activity of these enzymes may be more prevalent in populations who did not traditionally consume the enzyme substrate in their diets, e.g. Inuit populations who gain most of their calorific requirements from animal products have a high frequency of sucrase isomaltase deficiency (0.1) in comparison with other populations where this is rather rare (McNair *et al.*, 1972). This suggests that where sucrose was not part of the diet, deficiency allele(s) were able to reach high frequency in the absence of selection and possibly in combination with genetic drift.

The situation for lactase contrasts starkly with this. Investigation of the genetic variation within and surrounding the lactase gene indicates the presence of positive selection in those populations who traditionally consumed the enzyme substrate, lactose, in their diets. Indeed, recent HapMap studies reveal that the signature of selection at the locus is one of the strongest observed in the genome (The International HapMap Consortium, 2005). The remainder of this introduction gives a detailed account of the research which has focussed on understanding variation within this gene.

1.2 Part II – Lactase Persistence

1.2.1 The small intestine and digestion of lactose

Lactose in milk is digested and absorbed in the small intestine. The surface of the small intestine has a specialised structure, composed of hundreds of 'villi', tiny finger-like structures that protrude from the wall of the intestine and have additional extensions called microvilli which make up the apical 'brush border' of the absorptive epithelial cells (enterocytes) lining the villi. The small intestinal enzymes such as lactase that facilitate digestion and absorption of carbohydrates are anchored to the surface of the brush border.

The enzyme lactase is responsible for cleaving lactose into its constituent monosaccharides, glucose and galactose, which are transported across the epithelial cell membranes into the enterocytes and then into the blood stream via active transport by a sodium-dependent galactose transporter (Wright *et al.*, 2007). Lactose itself cannot be transported across the cell membrane, and hence lactase is essential for the nourishment of neonatal mammals whose sole source of nutrition is milk, in which lactose is the major carbohydrate component. However, in most mammals the enzyme usually decreases significantly in quantity following weaning so that the level is low in adult life (Sebastio *et al.*, 1989; Buller *et al.*, 1990; Lacey *et al.*, 1994; Pie *et al.*, 2004).

1.2.2 Lactase and its structural gene, *LCT*

The gene *LCT*, which encodes the enzyme lactase, a β-glycosidase capable of hydrolysing a variety of substrates and often known as lactase-phlorizin hydrolase (LPH), maps physically to chromosome 2q21 (Kruse *et al.*, 1988; Spurr and White, 1991; NIH/CEPH Collaborative Mapping Group, 1992; Harvey *et al.*, 1993). The nucleotide sequence exhibits four-fold internal homology, suggesting that two partial gene duplication events occurred in the evolution of *LCT*. Two of the homologous domains (I and II) occur in the pro- region of the molecule and the others (III and IV) are found in the mature polypeptide (Figure 1.2, Mantei *et al.*, 1988).

1.2.3 Lactase-Phlorizin hydrolase enzyme (LPH)

The 5787bp pre-pro-lactase-phlorizin hydrolase mRNA transcript is encoded by 17 exons (Boll *et al.*, 1991). The pre-pro-protein, composed of 1927 amino acids (Mantei *et al.*, 1988) contains a putative signal sequence of 19 amino acids, and a large 'pro' portion of 847 amino acids, both of which are proteolytically removed before the protein assumes its mature form. The 19 amino acid pre- sequence is first removed in the ER by signal peptidase, yielding pro-LPH molecules, which become N-glycosylated and pair up to form homodimers (Grunberg and Sterchi, 1995). This dimerisation is essential for acquisition of transport competence and full enzymic activity of LPH (Naim and Naim, 1996). Further (O-linked) glycosylation occurs once the pro-LPH-homodimer has been translocated to the Golgi apparatus, which is also the predominant site of proteolytic cleavage of the pro-sequence (Naim *et al.*, 1987). The pro-sequence has been shown to play a vital role in the maturation of LPH, being involved in folding, targeting and dimerisation of the molecule (Panzer *et al.*, 1998).

Residues Ala-867 onwards comprise the 160 kDa glycoprotein found anchored to the brush border of the jejunum as mature LPH homodimers (Mantei *et al.*, 1988). LPH is an amphiphillic molecule, consisting of a short cytoplasmic domain followed by a membrane-spanning hydrophobic domain (residues 1883-1901) at its C-terminus, orientating the molecule such that the bulky, hydrophilic N-terminal projects into the lumen (Skovbjerg *et al.*, 1981; Wacker *et al.*, 1992). It is within this N-terminal portion that both catalytic activities reside, and it has been demonstrated that the active site for phlorizin hydrolysis is distinct from that of lactose (Columbo *et al.*, 1973; Leese and Semenza, 1973; Skovbjerg *et al.*, 1981). The active sites are situated within the homologous domains III (Glu1271) and IV (Glu1747), respectively (Arribas *et al.*, 2000).

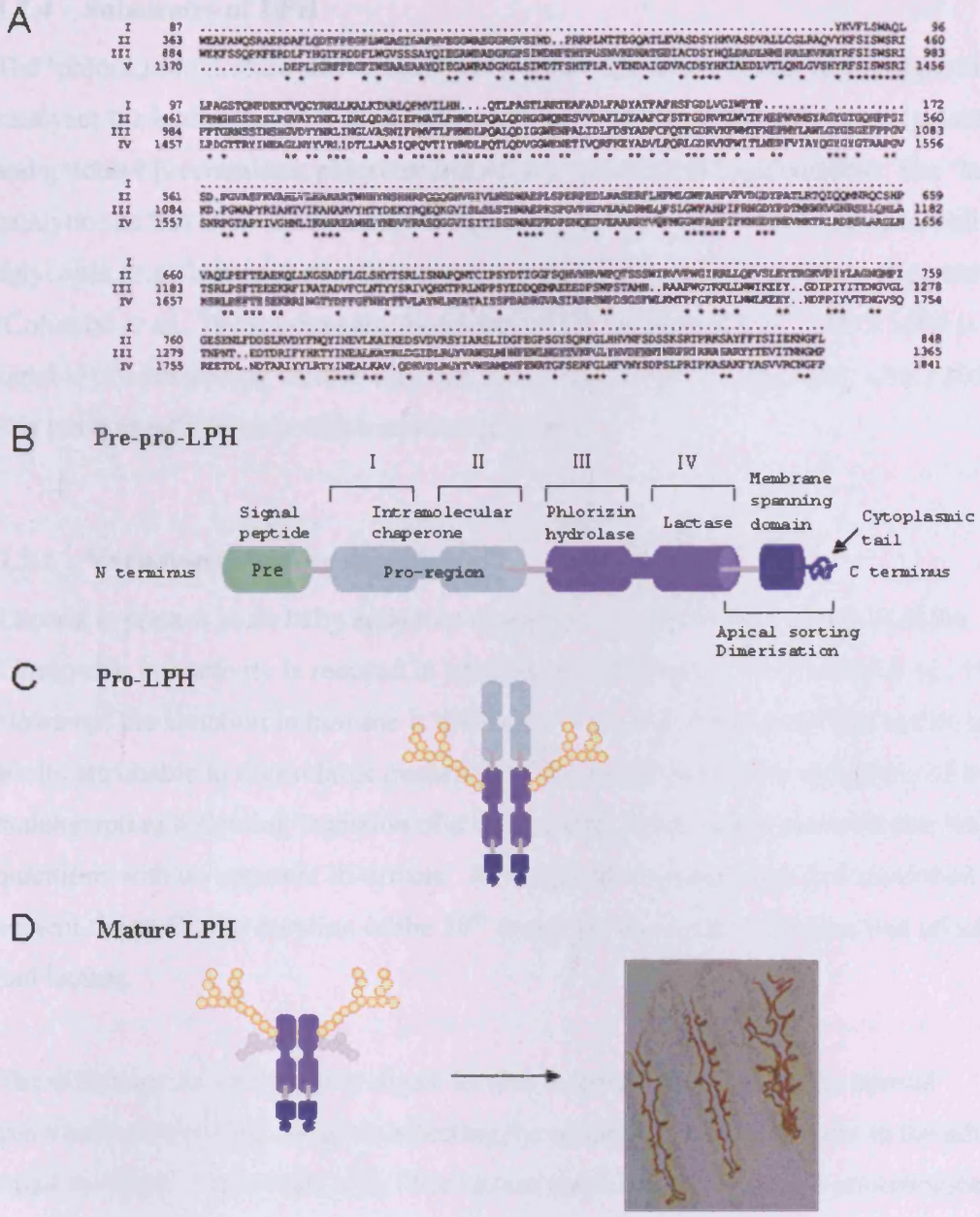


Figure 1.2.

A Internal homologies within the human pro-LPH amino acid sequence (Mantei *et al* 1988). Regions I (residues 87-172), II (363 -848), III (884-1365) and IV (1370-1841) of the human pre-pro-LPH sequence are shown. Residues which match in three or more sequences are indicated by asterisks.

B Transcription of *LCT* occurs in differentiated small intestinal enterocytes, yielding a 6.2kb mRNA. An overview of the translation product of the lactase mRNA; pre-pro-LPH is shown, with the domain structure indicated.

C The signal peptide mediates translocation of the pre-pro-LPH over the endoplasmic reticulum, and is cleaved off during the process. The pro-LPH is then *N*-glycosylated and homodimers are formed. .

D. The pro-LPH is then transported to the Golgi complex where it is complex and *O*-linked glycosylated and the large pro-region is cleaved off. The mature LPH is then transported to the microvillus membrane. An example of immunohistochemical staining of mature LPH at the brush border is shown.

1.2.4 Substrates of LPH

The 'phlorizin-hydrolase site' is situated closest to the brush border and preferentially catalyses the hydrolysis of β-glycosides with large, hydrophobic aglycones (galactosyl and glucosyl β-ceramides, phlorizin and other aryl- or alkyl-β-glycosides). The 'lactase' catalytic site has been shown to have a preference for β-glycosides with hydrophilic aglycones (e.g., lactose, cellobiose and some β-1, 4-linked small glucose oligomers) (Columbo *et al.*, 1973; Leese and Semenza, 1973; Skovbjerg *et al.*, 1981). LPH is also capable of hydrolysing various flavonol and isoflavone glucosides (Day *et al.*, 2000), but it is not currently known which active site is used.

1.2.5 Variation in lactose digestion

Lactase is present in all baby eutherian mammals tested with the exception of the Pinnipedia, but activity is reduced in adult animals (Plimmer, 1906; Crisp *et al.*, 1987). However, the situation in humans is different. It has long been noted that whilst some adults are unable to digest large quantities of fresh milk, and show symptoms of lactose malabsorption following ingestion of a lactose load, other people can consume large quantities with no apparent ill-effects. Although this variation was first described in ancient times, for the duration of the 20th century it was widely assumed that all adults had lactase.

The difference in our ability to digest lactose is most commonly due to normal genetically determined variation affecting the quantity of lactase present in the adult small intestine. Individuals with little lactase are described as lactase non-persistent or are said to have primary adult hypolactasia, or be lactose mal-digesters, while those who maintain a high level of lactase are said to be lactase persistent (terminology reviewed in Sahi, 1994).

This normal variation is quite distinct from the absence of lactase from birth, which is an extremely rare and potentially fatal inborn error of metabolism and is referred to as congenital alactasia. A number of mutations have been identified within the coding region of *LCT* in Finnish patients suffering from this condition. All the mutations affect

the primary structure of the protein, resulting in a non-functional lactase enzyme (Kuokkanen *et al.*, 2006).

Lactase level can also be reduced if damage occurs to the brush border due to gastrointestinal disease, and this condition (which is usually reversible) is referred to as secondary or acquired hypolactasia (Villako and Maaroos, 1994).

People who have either primary or secondary lactase deficiency are lactose mal-digesters, as judged by a lactose tolerance test, and may exhibit symptoms of lactose intolerance when they consume lactose. This thesis focuses on the former scenario, and considers lactose malabsorption in the context of genetically controlled variation of lactase expression in adults.

1.2.6 Symptoms of lactose intolerance

The symptoms of lactose intolerance due to lactose malabsorption, caused when milk is consumed by a lactase non-persistent person, vary greatly between individuals, but if they are evident, they usually manifest themselves within one to two hours of ingestion.

Undigested lactose passing through the small intestine into the colon has two physiological effects. Firstly, an osmotic gradient is set up across the gut wall, which results in a large influx of water (Christopher and Bayless, 1971), and can cause symptoms of diarrhoea. Secondly, the lactose is utilised as an energy source by colonic bacteria, which ferment it to produce fatty acids and gaseous by-products, potentially causing discomfort, bloating and flatulence (Hammer *et al.*, 1996).

1.2.6.1 Variation in symptoms of lactose malabsorption

Most lactase non-persistent individuals can tolerate small amounts of lactose (as in tea or coffee). Diarrhoea and discomfort are not seen in all individuals who are diagnosed as lactose mal-digesters in a lactose-tolerance test, even after consumption of 50 g of lactose, and it has been suggested that variation in the composition of the gut flora between individuals may be responsible for some of this variation (Hertzler and

Savaiano, 1996; Hertzler *et al.*, 1997), as well as a psychosomatic component (Briet *et al.*, 1997). Two separate studies suggest that around 40% of people self-diagnosed as lactose intolerant are actually lactose digesters (Saltzman *et al.*, 1999; Peuhkuri, 2000). This effect is perhaps due to prevalence in the public consciousness of 'dairy intolerance' caused by extensive media coverage of the condition, leading individuals to assign any symptoms of gastrointestinal discomfort to this cause.

The symptoms of lactose malabsorption described above were first attributed to a lack of lactase following the observation that a proportion of intestinal tissue samples from healthy Southern European adults had virtually absent lactase activity despite showing histologically normal mucosa (Auricchio *et al.*, 1963; Dahlqvist *et al.*, 1963). The discovery of this enzyme deficiency or 'abnormality' ignited the interest of researchers into the inter-individual differences in our capacity to tolerate milk and its derived products. In the following years, data on lactase persistence frequencies in other populations began to accumulate and a global picture began to develop that challenged the original perception of lactase non-persistence as the 'abnormal' phenotype.

1.2.7 Diagnosis of lactase non-persistence/persistence

In order to collect information on the worldwide frequencies, alternatives to direct quantification via biopsy of the small intestine were used. Biopsies are the most accurate method of establishing lactase activity. However, they are invasive and are not usually a preferred routine diagnostic for lactose intolerance, normally being obtained only when a patient is undergoing endoscopy to exclude another gastro-intestinal complaint.

Several indirect methods have been developed for the purpose of diagnosis, all of which utilise lactose digestion to inform on an individual's lactose tolerance status, and by implication lactase persistence status. The general practice is to give a lactose load after an overnight fast. The two most widely used methods are described below.

1.2.7.1 The blood glucose test (Dahlqvist, 1974)

A baseline measurement of blood glucose is taken before ingestion of a lactose load, and then at various time intervals (usually every 30 minutes) for the following 2 hours. An increase in blood glucose indicates lactose digestion (lactase cleaves the lactose molecules into glucose and galactose allowing absorption into the bloodstream and subsequent detection in a blood sample), and no increase, or a 'flat line', is indicative of a lactose non-digester/mal-digester or intolerant phenotype.

1.2.7.2 The breath hydrogen test (Metz et al., 1976; Howell et al., 1981)

This method measures hydrogen production by colonic bacteria. If the lactose dose is hydrolysed by the enzyme lactase in the small intestine, no changes in breath hydrogen will be observed, and the individual can be classified as a lactose digester/lactose tolerant. Conversely, in lactose mal-digesters, the lactose load passes through the small intestine and into the colon where it is digested by bacterial fermentation, a by-product of which is hydrogen. Some of this hydrogen is absorbed into the blood stream and released into the breath (where it can be detected) as the blood passes through the lungs. A baseline measurement of breath hydrogen is taken, prior to ingestion of the lactose load, and further readings are taken at 30-minute intervals from the time of ingestion for the following 3 hours.

1.2.8 Inaccuracies of indirect testing

In both cases somewhat arbitrary cut-off points have to be set for distinguishing the two phenotypes and both methods inform on the person's ability to digest lactose rather than that persons lactase expression. We infer the lactase persistence status of an individual from these tests, and must therefore keep in mind that there will be an error rate in both directions.

1.2.8.1 Lactose load

Some of these observed errors could be attributed to test design, particularly the quantity of lactose administered. Non-persistent individuals express a residual amount of lactase,

approximately 10% of the persistent adult levels (Semenza *et al.*, 1999) and so when only a low lactose dose is used the quantity passing through to the colon may not be large enough to increase breath hydrogen by the standard >20 ppm increment. Using the blood glucose method would, on the other hand, have the opposite effect; a low lactose dose in lactase persistent individuals may not yield a sufficient rise in blood glucose to cross the nominated 'digester' threshold (usually 1.1 mmol/L). High doses are used to avoid these issues, however some studies suggest that even lactase-persistent people fail to digest a proportion of consumed lactose, and therefore an increase in breath hydrogen could be observed in persistent subjects when a high dose challenge (such as 50 g) is used (Bond and Levitt, 1976).

1.2.8.2 Other factors

Aside from dose, many other factors can impact upon the test result; gastric emptying and intestinal transit times can exert an effect both on blood glucose and breath hydrogen measurements (Ladas *et al.*, 1982; Labayen *et al.*, 2001; He *et al.*, 2006). Diarrhoeal disease is known to reduce lactase expression temporarily as a result of villus flattening and loss of the cells which express lactase (Villako and Maaroos, 1994) and hence a genetically persistent individual may be classified as a mal-digester in this instance. Also, the use of antibiotics may disrupt the gut flora and result in erroneous results. Colonic adaptation to dairy products may affect breath hydrogen production by increasing bacterial populations that have increased metabolic activity for lactose (Hertzler and Savaiano, 1996). Also, some individuals will be 'hydrogen non-producers' (the reasons for this are not fully understood, but factors include having a hostile gut pH with acidity too severe for the existence of hydrogen-producing bacteria (Vogelsang *et al.*, 1988), and in this situation the breath hydrogen test would be uninformative.

In the clinical setting, there are ways of improving the quality of the test. These include retesting, and giving a dose of lactulose (an indigestible carbohydrate) to test for hydrogen production, and investigation of other causes of the lactose intolerance.

1.2.8.3 Error rates

A recent study by members of our group (Mulcare *et al.*, 2004) attempted to estimate the error rates of both the blood glucose and the breath hydrogen test from published data. Results were pooled from papers that compared either indirect method with each other or with a verified phenotype based on direct enzyme assays from jejunal biopsy. Exact protocols varied between the pooled data set, but all included a minimum 50 g lactose load and measured a change in parameter one or more times between 30 min and 4 h after ingestion. The blood glucose method error rates were 7% false positive (i.e., non-persistent individuals classified as lactose digesters) and 9% false negative (i.e., persistent individuals classified as lactose malabsorbers). The breath hydrogen method was found to give a slightly more accurate assessment of lactase persistence status, with approximately 5% false positive and 7% false negative error rates. Thus, the evidence suggests that to obtain the most accurate indirect assessment of lactose tolerance status, a breath hydrogen test should be undertaken.

According to our own experience, the most accurate method requires a fast of 12 h to be observed prior to consumption of the lactose dose, 50 g of which is the widely accepted standard (equivalent to approximately one litre of cow's milk). A baseline breath hydrogen measurement should be taken prior to the lactose dose, and at 30 min intervals afterwards for the following 3 hours. Test results for subjects with a H_2 baseline of zero (possible non-producers), or greater than 20 ppm (suggestive of failure to fast, or bacterial overgrowth of the colon), should be interpreted with caution and followed up if possible.

1.2.9 Worldwide distribution of lactase persistence

A number of surveys of lactase persistence phenotype frequencies have been carried out in many populations throughout the years, so that the global distribution of lactase persistence is now fairly well characterised (Figure 1.3) (Swallow and Hollox, 2000). These frequencies reveal clearly that lactose intolerance is the most commonly observed phenotype in humans, with lactase persistence being frequent only in those populations with a long history of pastoralism and where milking has been practised. Lactase persistence is at highest frequency in north-western Europe, with a decreasing cline to

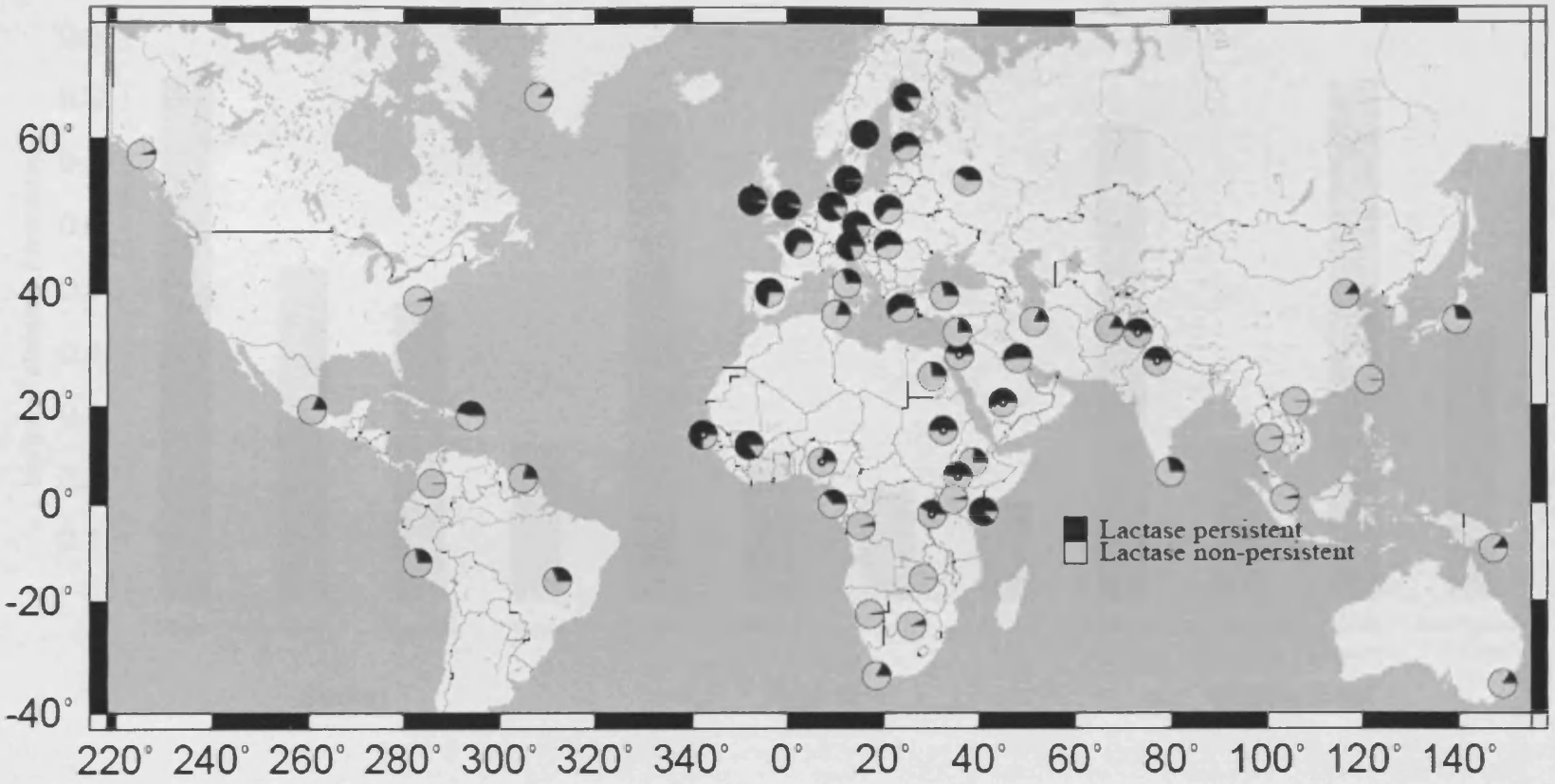


Figure 1.3 Worldwide distribution of lactase persistence frequency. Dark grey indicates the proportion of lactose digesters (presumed lactase persistent) in a given population and light grey represents mal-digesters (presumed non-persistent). A central circle indicates that the overall frequency for that country is comprised of different ethnic groups with very different phenotype frequencies. For examples of these cases, see Figure 1.4. (Data taken from Bloom and Sherman, 2005; Mulcare *et. al.*, 2006)

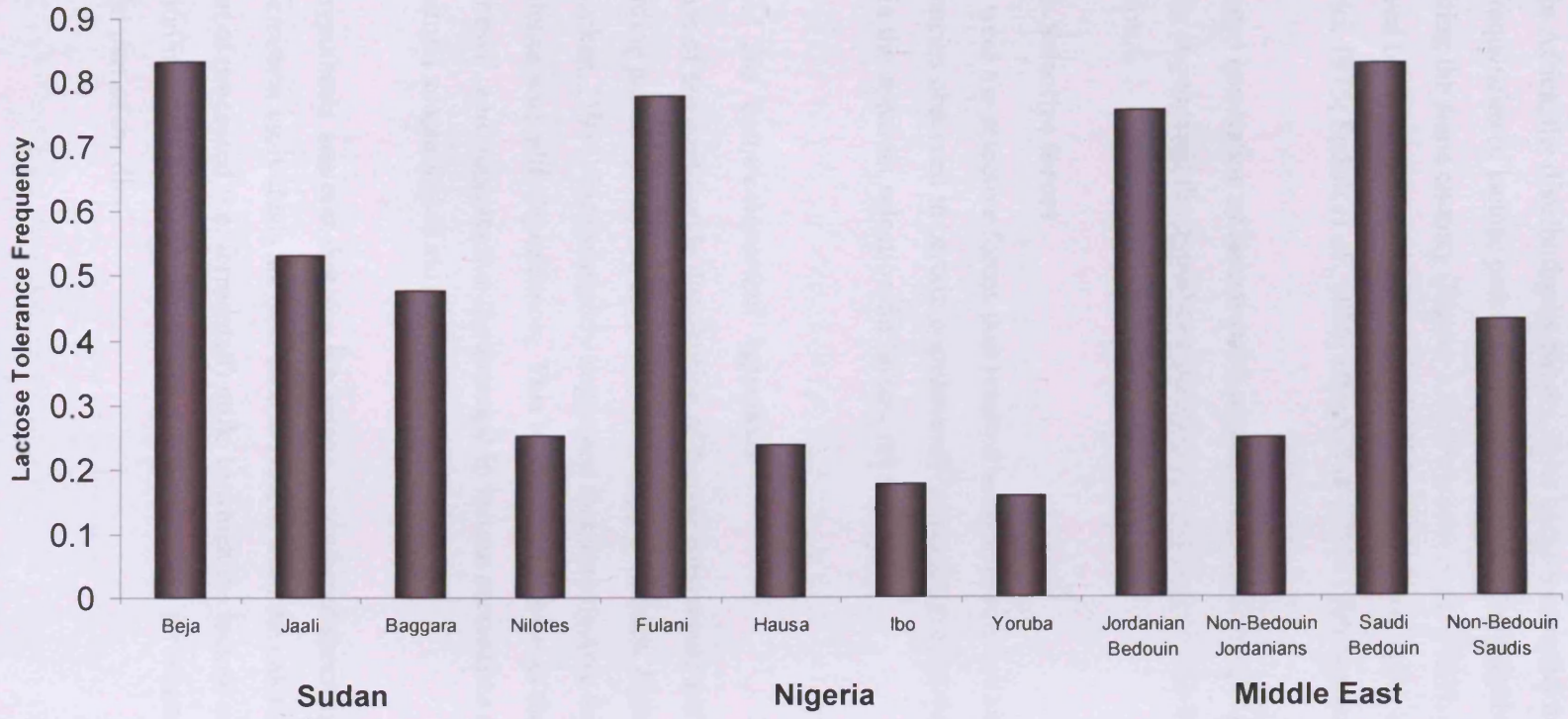


Figure 1.4 Examples of countries/geographic regions in which individual ethnic groups display large differences in inferred lactase persistence status. Data compiled by Mulcare (2006) who extracted adult data from source references.

the south and east. In India, the frequency of lactase persistence is higher in the north than the south, and in the rest of the world lactase persistence frequency is generally low. In Africa, the distribution is patchy, with some pastoralist nomadic tribes having high frequencies of lactase persistence compared with the neighbouring groups inhabiting the same country (Figure 1.4) (Bayoumi *et al.*, 1981), with a similar pattern observed between Bedouin and neighbouring populations in the Middle East (Cook and al-Torki, 1975; Snook *et al.*, 1976; Hijazi *et al.*, 1983; Dissanyake *et al.*, 1990).

The noted correlation of lactase persistence phenotype with the cultural practice of milking engendered the hypothesis that this trait has been subject to strong positive selection.

1.2.10 Selective forces

What were the selective forces that resulted in the elevated lactase persistence frequencies observed in certain populations? A number of theories have attempted to explain the apparent selection for lactase persistence.

1.2.10.1 The 'culture-historical' hypothesis

Because of the world-wide distribution of lactase persistence and the generally coinciding pattern of historically milk-drinking populations, both Simoons, (1970) and McCracken, (1971) independently suggested that the selective force for lactase persistence was milk dependence. This has become known as the 'culture historical hypothesis', and suggests that the increase in lactase persistence co-evolved alongside the cultural adaptation of milk drinking.

This hypothesis assumes that any advantage is conferred specifically by fresh milk, as non-persistent individuals are also able to benefit from the calorific, vitamin and mineral content of processed (i.e. fermented) milk, in which the lactose content is reduced. Thus selection is most likely to have occurred in populations for whom fresh milk formed an integral part of the diet.

A problem with this hypothesis is the 'non-fit' populations, who have either a high lactase persistence frequency without being milk dependent, or who rely heavily on milk products but who have a low reported frequency of lactase persistence. Statistical modelling has been used to consider this problem and suggests that an incomplete correlation does not necessarily provide evidence against the culture-historical hypothesis. Non-fits may be expected if some lactase persistent populations have recently stopped milking or if predominantly non-persistent groups have only recently adopted the habit, therefore allowing insufficient time for lactase persistence to be driven to high frequency (Aoki, 1986). Furthermore, this model does not account for migration of either genes or culture, which could both result in an imperfect correlation between lactose absorption capacity and milk-use.

Other statistical analysis took a phylogenetic approach, placing all populations included in the analysis on linguistic and genetic trees (constructed using data from classical markers; Holden and Mace, 1997). This provides a more realistic model compared with a non-phylogenetic approach which assumes all populations are equally related. The correlation of high lactose digestion frequency was tested with percentage dependence on pastoralism, levels of solar radiation and dry months/year or average rainfall and adjusted for relatedness in the analysis. These data revealed that percentage reliance on pastoralism best explained the variation observed between populations and concluded that lactose digestion capacity had most likely evolved as an adaptation to dairying, and that high frequency lactose digestion capacity had not evolved in the absence of milking.

More recent research has sought to address the question of why some populations and not others adopted the cultural habit of milk drinking. Bloom and Sherman, (2005) found that frequencies of lactose malabsorption were higher in populations where environmental conditions, such as extremes of climate or high incidence of endemic cattle disease made it impossible to raise livestock, thereby supporting the culture-historical hypothesis by providing further evidence that lactase persistence is selected for only in environments conducive to dairying. The exceptions to the general distribution were a number of African groups who had high lactase persistence and maintained herds despite environmental conditions being unfavourable. The authors suggest that the

nomadic lifestyle of these groups allowed them to circumvent harsh environmental conditions and maintain their herds.

Further evidence in support of the culture-historical hypothesis has been provided by the observation that high intra-allelic diversity of cattle milk protein genes is observed in native cattle from north central Europe (NCE), coincident with the locations of European Neolithic cattle farming sites and with the geographic incidence of high lactase persistence frequency (Beja-Pereira *et al.*, 2003).. The authors suggest the increased diversity observed at the cattle milk protein genes (which is not observed at other loci in the NCE cattle) is due to the larger herd sizes kept in dairying cultures and from selection for increased milk yield and altered milk protein compostion. This is presented as one of the few non-disease related examples of genetic co-evolution between domestic animals and humans

1.2.10.2 The arid climate hypothesis

The arid climate hypothesis, first suggested by (Cook and al-Torki, 1975) speculated that in desert climates (i.e., Middle and Near East) where water and food were scarce, nomadic groups could survive by utilizing milk as a food source, and in particular, as a source of clean, uncontaminated water. The benefits to persistent individuals may have become even more pronounced during outbreaks of diarrhoeal disease, when non-persistent individuals would be unable to utilize milk as a water source without exacerbating their condition. This scenario could be particularly pertinent to desert nomads who consume camel milk, as these animals continue to lactate for several days in the absence of water (ref?).

1.2.10.3 The calcium absorption hypothesis

Obviously, the benefits of drinking milk cannot be explained by the arid climate hypothesis in northern Europe. Here, the calcium absorption hypothesis has been suggested to explain the distribution of the trait (Flatz and Rotthauwe, 1973). The low-light levels experienced in the Northern Hemisphere are associated with an increased risk of developing rickets and osteomalacia due to a lack of vitamin D (which is

synthesised by the skin in the presence of sunlight). Calcium may help to prevent rickets by impairing the breakdown of vitamin D in the liver (Thacher *et al.*, 1999), and is itself an essential mineral required for bone health. Lactase non-persistent individuals could obtain calcium from yoghurt or cheese, dairy foods that contain reduced lactose. However, milk proteins and lactose are believed to facilitate the absorption of calcium (for review see Gueguen and Pointillart, 2000), and hence the ability to drink fresh milk which contains both calcium and components that stimulate its uptake (along with small amounts of vitamin D) may have provided an advantage to persistent individuals.

1.2.10.4 Selection for lactase non-persistence

In only one case has selection against lactase persistence been proposed. Anderson and Vullo (1994) suggested that selection had acted in favour of lactase non-persistence in malarial regions because of the observation that distribution of the two variables were correlated, along with the observation that individuals with flavin deficiency are at a slightly reduced risk of infection by malaria. The consumption of milk, which is rich in riboflavin was, therefore, said to be unfavourable as it would keep flavin level in the bloodstream high. This explanation is not widely supported, and a study in Sardinia found no association between lactase persistence frequencies and populations who had been exposed to malaria (Meloni *et al.*, 1998).

1.2.10.5 Cultural adaptation

As previously mentioned, the correlation between pastoralism, milk drinking and lactase persistence is not true for all populations, for example, the Dinka and Nuer in Sudan (Bayoumi *et al.*, 1982) have a low lactose persistence frequency despite cows or camels playing a very important role in their lifestyle. These populations are not completely dependent on milk despite its consumption being substantial, and therefore the selective pressure for lactase persistence may have been less strong. In these populations and in many other peoples, it seems that the lactose concentration is moderated by cultural adaptation. Milk is processed to sour milk, yoghurts and cheeses, which have reduced lactose content, and individuals also adapt their consumption habits by taking smaller quantities of milk at a time. These cultural adaptations enable non-persistent individuals

to benefit from the calorific, mineral and vitamin constituents of milk without inducing the associated symptoms of lactose malabsorption, and are complemented by adaptations of the large intestinal bacterial flora (see Diagnosis of lactase non-persistence/persistence, section 1.2.7).

1.2.11 Identifying the cause(s) of lactase persistence

As discussed above (section 1.2.5), early studies reported lactase activity in the intestine of young and suckling mammals, but not in the intestine of the corresponding adult of the species. This has been supported by recent studies, although in many cases substantial quantities of residual lactase are observed in adult tissue (Rossi *et al.*, 1997). These observations provide evidence that the ancestral state for mammals (and therefore humans) is for lactase expression to be down-regulated following the weaning period. A number of studies have examined *LCT* mRNA level (Sebastio *et al.*, 1989; Buller *et al.*, 1990; Lacey *et al.*, 1994; Pie *et al.*, 2004) and although there is much conflicting data, these also appear to decline in adult animals. However, the regulatory mechanisms controlling *LCT* expression are still not fully understood.

1.2.12 An inducible enzyme?

Many enzymes are regulated via an inducible system whereby their expression is activated or increased in the presence of high substrate concentration. For example, the lac operon in E. coli is an inducible system in which β-galactosidase (a bacterial enzyme capable of hydrolysing lactose to glucose and galactose) is expressed only in the presence of lactose (Jacob and Monod, 1961).

Initially, an inducible system was hypothesised to regulate lactase expression in humans (Gilat *et al.*, 1972; Cook, 1988), and would indeed provide a neat explanation of why lactase is down-regulated following weaning and why lactase persistence is seen more commonly in populations where milk forms an integral part of the diet. However, many animal studies have shown no increase in lactase activity in response to prolonged exposure to lactose (Leichter, 1973; Gutierrez *et al.*, 2002), and studies in human populations confirm this: amongst them a study of 50 adult Thai individuals who

voluntarily ingested daily lactose doses for one month with no reported improvement in lactose absorption or increased lactase activity (Keusch *et al.*, 1969).

1.2.13 Evidence of a genetic cause

By the early 1970s it had been established using family studies that the lactase persistence polymorphism in humans had a genetic cause, and was inherited in an autosomal dominant manner (Ferguson and Maxwell, 1967; Sahi, 1974). In another study, monozygotic twins showed 100% concordance of lactase persistence phenotype, and phenotype frequencies in dizygotic twins were found to agree with Hardy-Weinberg equilibrium (HWE) expectations of an autosomal dominant pattern of inheritance (Metneki *et al.*, 1984).

1.2.14 Evidence of a *cis*-acting effect

Further evidence that lactase persistence was a genetic trait, and more specifically that it was caused via a *cis*-acting element was produced in the early 1980's. Ho *et al.* (1982) reported a trimodal distribution of lactase: sucrase ratios in British natives. Both lactase and sucrase were extracted from autopsy material (from individuals without gastrointestinal disease) with the sucrase activity serving as an internal standard correcting for non-genetic variation. The trimodal distribution was interpreted to represent individuals homozygous for lactase persistence (highest lactase activity), heterozygtotes (mid-level activity) and non-persistent homozygotes (low lactase activity). The intermediate lactase activity observed in the heterozygotes indicated that only one copy of the lactase gene was being fully expressed, and concordant results were subsequently obtained in individuals of German ancestry (Flatz, 1984). Confirmatory evidence for the *cis*-acting nature was obtained from mRNA studies. Allelic variants of exonic SNPs were used to identify particular transcripts, and their expression levels. Europeans of the persistent phenotype who were heterozygous for exonic polymorphisms were used to demonstrate monoallelic expression at the mRNA level (Wang *et al.*, 1995).

1.2.15 *LCT* immediate promoter and upstream elements

Investigations first focused on finding these *cis*-acting elements included studies of the *LCT* immediate promoter. A conserved ~150 bp region exists immediately upstream of *LCT* transcription initiation site in human, rat, pig and mouse, suggesting that key regulatory elements important for lactase expression are encoded within this small region (reviewed in Troelsen, 2005). The region has been shown to drive low-level expression in an intestinal cell line (Troelsen *et al.*, 1992), while transgenic mouse experiments using rat and pig promoter constructs of different sizes show that elements outside this conserved 150 bp region are required for high and tissue-specific expression of lactase (Troelsen *et al.*, 1994; Krasinski *et al.*, 1997; Lee *et al.*, 2002; Wang *et al.*, 2006). A 1 kb pig promoter construct is sufficient to mimic endogenous gene expression in transgenic mice (Troelsen *et al.*, 1994). However, a 2 kb rat promoter is required to produce the same effect (Lee *et al.*, 2002). In humans this region is disrupted by two tail to tail Alu elements (Hollox *et al.*, 1999). Although DNA sequence is not conserved between species in the upstream regions, there is evidence that a similar pattern of upstream regulatory regions exist in pig, rat and human (Troelsen, 2005). The different enhancer sites encoded within the upstream regions are thought to make distinct contributions to the spatial and temporal expression of *LCT* (Wang *et al.*, 2006). However, these between species differences in promoter structure outside the proximal region complicate studies of lactase persistence using model organisms.

1.2.16 Transcription factors and *LCT* expression

A number of transcription factors have been identified that are involved in lactase expression (Figure 1.5) (Troelsen, 2005), some of which are discussed below.

Figure 1.5 Upstream regulatory regions affecting pig, rat and human lactase expression. The identity of transcription factors binding the *cis*-elements are indicated (? indicates unknown factors). Arrows represent functional regulatory regions identified by promoter analyses (+ indicates enhancer function, - indicates repressor activity). Figure originally published in Troelsen, 2005.

1.2.16.1 CDX2

The transcription factor CDX2 is implicated in the regulation of many intestinally expressed genes (Freund *et al.*, 1998; Beck, 2004) and has a number of binding sites upstream of the *LCT* initiation codon, including one (in humans) within the 150 bp conserved proximal promoter region (Troelsen *et al.*, 1997). CDX2 has been suggested to be involved in differentiation of the absorptive cells of the intestinal epithelium (Mutoh *et al.*, 2005) and has been shown to upregulate *LCT* expression *in vitro* (Fang *et al.*, 2000).

1.2.16.2 HNF1A

The transcription factor HNF1A is also known to modulate expression of lactase (Spodsberg *et al.*, 1999; Krasinski *et al.*, 2001; Bosse *et al.*, 2006). One HNF1A binding site occurs in the proximal promoter region, and both human and pig promoters contain distal HNF1A sites, although only the proximal site shows conservation between the two species (Spodsberg *et al.*, 1999). HNF1A is probably the main homologue binding the promoter HNF1 site (Spodsberg *et al.*, 1999; Bosse *et al.*, 2006), and it has been shown that HNF1A and CDX2 act synergistically to activate *LCT* promoter activity *in vitro* (Mitchelmore *et al.*, 2000).

1.2.16.3 GATA factors

The GATA4/5/6 transcription factors have been shown to play a critical role in the development of a number of endoderm-derived tissues, of which the small intestine is one (reviewed in Burch, 2005), and they are also implicated in the transcriptional regulation of a number of intestinally expressed genes including sucrase-isomaltase (Krasinski *et al.*, 2001; Boudreau *et al.*, 2002), intestinal fatty acid-binding protein (Gao *et al.*, 1998), trehelase (Oesterreicher and Henning, 2004) as well as lactase. GATA4, 5, and 6 all bind to a number of GATA recognition sites which occur upstream of *LCT*, including two within the proximal promoter (Fitzgerald *et al.*, 1998; Fang *et al.*, 2001). It is thought that GATA4 is the primary GATA factor responsible for modulating *LCT* expression, due to the highly correlated expression pattern of the two genes in small intestinal epithelia, in combination with the observation that GATA4 binding to the proximal promoter is more evident than binding of either GATA5 or 6 in EMSAs using nuclear extracts prepared from mouse intestinal epithelia (van Wering *et al.*, 2004). Further to this, it has been shown that transgenic mice producing an inducible mutant form of GATA4 have significantly reduced *LCT* expression levels (Bosse *et al.*, 2006). However, *in vitro*, all GATA factors bind the proximal sites with similar affinity, and it has been found that GATA5 and HNF1A are capable of co-operating to stimulate expression from the proximal promoter in this context (Krasinski *et al.*, 2001).

1.2.17 Variation in *LCT* and the immediate promoter with respect to lactase persistence

Whilst the identified recognition sites and the corresponding transcription factors of the proximal *LCT* promoter are undoubtedly important in the basal regulation of lactase expression they are, as discussed above, not sufficient for the correct temporal and spatial expression of the gene, and hence the immediate promoter region is not thought to be involved in causing lactase persistence.

Sequencing of *LCT* and the immediate promoter region in Europeans showed no nucleotide changes that were absolutely associated with persistence/non-persistence, although several useful variants were identified which were powerful in discriminating distinct haplotypes (Boll *et al.*, 1991; Lloyd *et al.*, 1992; Poulter *et al.*, 2003).

Subsequent research has, therefore, focused more intensely on the upstream regions of *LCT*, looking for regulatory elements that influence the lactase persistence phenotype. Interestingly, many enhancer motifs occur upstream of the proximal promoter, which are different depending on the species, and are recognised by various transcriptional regulators (including Cdx-2, HNF-1α and GATA factors; Lewinsky *et al.*, 2005; Troelsen, 2005). One such highly variable region was identified ~900 bp upstream of the *LCT* start site (Harvey *et al.*, 1995; Hollox *et al.*, 1999). A nucleotide change, *-958C>T*, was found to greatly affect interaction with an unidentified DNA-binding protein. However, this polymorphism was not considered to be causal of lactase persistence because the ancestral allele, C, was present in both lactase persistent and non-persistent people. If functional *in vivo*, this SNP perhaps affects the timing of down-regulation or spatial expression along the length of the intestine or modulates the effect of other nucleotide changes.

In fact, several polymorphisms exist across the 50 kb *LCT* gene and although no single allele was 100% associated with lactase persistence, association studies revealed that very few haplotypes occur in most of the human populations tested, although greater diversity was observed in African populations (Hollox *et al.*, 2001). One particular

combination of alleles, designated the '**A**' haplotype, is particularly common in northern Europe and is found to associate with lactase persistence (Harvey *et al.*, 1998).

1.2.18 Identification of a candidate causal allele

Sequencing upstream of the **A** haplotype revealed a putative causative single nucleotide polymorphism (*-13910C>T*) located 13.9kb away from the *LCT* transcription initiation site (Enattah *et al.*, 2002). The SNP is located within an intron of an adjacent gene, *MCM6* (Figure 1.6) (Poulter *et al.*, 2003). The *-13910*T* allele was found to associate completely with directly ascertained lactase persistence in 196 Finnish individuals, and subsequent studies have confirmed a tight association between *-13910*T* and lactase persistence in populations of northern European ancestry (Poulter *et al.*, 2003). The **A** haplotype extends far beyond the 50kb *LCT* gene region, and carriers of the *-13910*T* allele tend to have completely identical chromosomes extending for nearly 1 Mb (Poulter *et al.*, 2003; Bersaglieri *et al.*, 2004), an observation which is quite remarkable in comparison to other regions of the genome (The International HapMap Consortium, 2005).

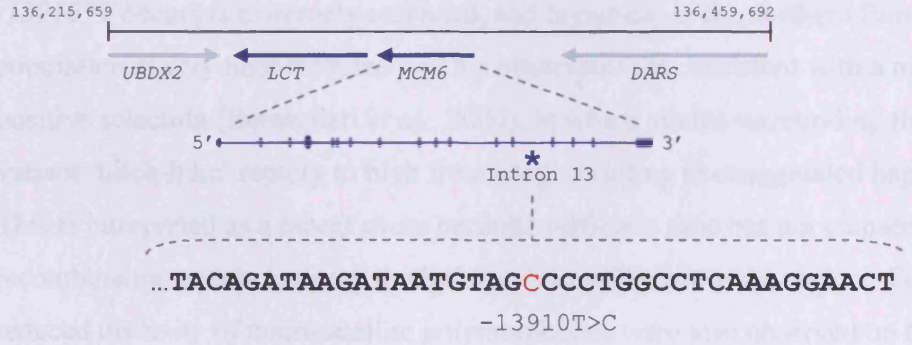


Figure 1.6 Map of *LCT* and surrounding genes. The position of the putative causative SNP (*-13910C>T*) is indicated.

1.2.19 Evidence of *-13910*T* function

In vitro studies demonstrated that the *-13910*T* allele increases transcription relative to *-13910*C* in promoter-reporter construct assays in a colon carcinoma cell line (CaCo2) (Olds and Sibley, 2003; Troelsen *et al.*, 2003), providing evidence that it may have

enhancer activity *in vivo*. A transcription factor, OCT1, was identified which bound more strongly to the *-13910*T* -containing motif than to the alternative C allele, providing a possible mechanism for up-regulation of *LCT* expression (Lewinsky *et al.*, 2005).

1.2.20 Molecular evidence of selection

Aside from the original observation of a positive correlation between lactase persistence frequencies and milk drinking, much molecular evidence has accumulated over the years which would appear to corroborate the hypothesis that lactase persistence has been the subject of strong positive selection. However research focussing on the haplotype diversity observed around the lactase gene in various populations suggested that the extremely low haplotype diversity observed in northern Europeans compared to other populations is most probably explained by a combination of genetic drift and strong positive selection for lactase persistence (Hollox *et al.*, 2001).

The *-13910*T* carrying chromosome is a real outlier in the context of molecular signatures of selection compared with the rest of the genome. The haplotype on which *-13910*T* occurs is extremely extended, and is present in the northern European population at very high frequency. This observation is consistent with a model of recent positive selection (Bersaglieri *et al.*, 2004), in which alleles surrounding the causal variant 'hitch-hike' rapidly to high frequency, resulting in exaggerated haplotype length. This is interpreted as a recent event because sufficient time has not elapsed for recombination events to decay the linkage disequilibrium in the region. Further to this, reduced diversity of microsatellite polymorphisms were also observed on the *-13910*T* carrying chromosomes, indicating recent selection for the allele (Coelho *et al.*, 2005; Mulcare, 2006).

Both these observations are consistent with selection for lactase persistence occurring alongside the advent of dairying, approximately 9000 years ago in Europe, with date estimates of the age of *-13910*T* placing its spread within this time period (2,188-20,650 years old; Bersaglieri *et al.*, 2004, and 7,450-12,300 years old; Coelho *et al.*, 2005). Consistent with these datings, a recent study reported the absence of *-13910*T*

in ancient DNA samples of the early Neolithic period. These findings are in agreement with a model in which dairying was adopted prior to lactase persistence becoming frequent, i.e. there is a low frequency of *-13910*T* in older populations (Burger *et al.*, 2007).

1.2.21 *-13910*T* in non-Europeans

It has been noted that *-13910*T* is extremely rare in sub-Saharan African populations, even in those populations where lactase persistence frequency had previously been reported to be high. A statistical procedure designed to enable a comparison to be carried out between *-13910*T* allele frequencies and the expected allele frequencies given the previously published lactase persistence frequency for an ethnically matched group. This procedure corrected for genotyping and sampling errors, and despite being conservative found a highly significant difference between the observed and expected frequency of *-13910*T* if it was assumed to be causal of lactase persistence. The study concluded that *-13910*T* could not be causal of lactase persistence throughout sub-Saharan Africa (Mulcare *et al.*, 2004).

The exceptions to this observation, i.e., African populations in which *-13910*T* was able to account for the lactase persistence frequency were in the Fulani and Hausa populations of Cameroon (Mulcare *et al.*, 2004), and Berber populations from Algeria and Morocco (Myles *et al.*, 2005). In both cases, this is thought likely to be a reflection of the demography of those populations, and there is evidence to suggest contacts between these and Eurasian populations (Cruciani *et al.*, 2002; Myles *et al.*, 2005).

The distribution of *-13910*T* could be interpreted in one of two ways; either *-13910*T* is not truly causal of lactase persistence, but is very strongly associated with the causal element and in Europeans therefore acts as a marker for the trait. In this case the apparent rarity of *-13910*T* in sub-Saharan Africa could be explained by the variant appearing on the lactase persistence carrying chromosome after humans had begun to spread out of Africa. The other possibility is that the identification of *-13910*T* as a cause of lactase persistence is correct, but that there is heterogeneity of the trait, and that different causal variations exist in other human populations.

1.2.22 Aims

The general objective of this thesis is to examine the evolutionary origins and possible molecular mechanisms of lactase persistence, including the population genetic and anthropological aspects, with particular emphasis on sub-Saharan Africa.

The specific aims addressed in each chapter were as follows:

- To investigate the apparent discrepancy between the frequency of lactase persistence and the putative causative *-13910*T* allele in Africa using a phenotyped African cohort (chapter 3).
- To search for genetic variation associated with lactase persistence in a phenotyped African cohort (chapter 4).
- To collect an independent cohort of phenotyped African individuals in which to replicate association of previously identified SNPs and further characterise variation at the locus of interest (chapter 5)
- To understand the evolutionary relationships between *LCT* haplotypes and document the geographic distribution of alleles identified during the course of this research (chapter 6).

2 Materials and Methods

2.1 DNA samples and population histories

All DNA samples tested within the course of this project were obtained from either the Galton or TCGA laboratories or newly collected. Samples were anonymised at source and informed consent (verbal in Africa) was obtained from each sample donor. Ethics approval was obtained from UCLH (99/0196 and 01/0236), and also obtained locally for each individual study.

Table 2.1 gives an overview of all populations included in this thesis, and gives information on language families and agricultural practises, including milking status. The table also includes a description of settlement patterns of the populations. According to (Blench, 1999) the term pastoralism is fluid, and can 'be so inclusive as to cover most of the peoples of semi-arid Africa'. Therefore, within this thesis the accepted definition of pastoralism will be that proposed by Blench: that movement must be the defining characteristic of pastoralist societies. More detailed information regarding each of the populations (listed alphabetically) is given in the sections below. Tree diagrams (Figures 2.1a & b) are provided to illustrate linguistic relationships between populations. Figure 2.2 indicates the sampling locations of populations included in this thesis on a geographic map.

2.1.1 Afar

The Afar are a pastoralist population (Blench, 1999) located in the north east of Ethiopia. The Afar population number about 200,000 (Cavalli-Sforza *et al.*, 1994), and their primary livestock are camels. Afar is a Cushitic language closely related to the Somali and Galla languages, and some researchers suggest all these populations share a southern Ethiopian origin and subsequently migrated north (Lewis, 1966).The samples used herein were collected by A Tarekegn and T. Olijiria-Raga in Assayita, near Djibouti, and milk drinking information was collected from each volunteer.

Population	Language	Population Code & definition (Murdock)	Type & Intensity of Agriculture	Crop type	Settlement Pattern	Predominant domestic animal	Milking status
Afar	**Afar** Afro-Asiatic, Cushitic	Afar: Ca6	Casual Agriculture	cereal grains	Semi-Nomadic	Camels	Regular milking of domestic animals
Amhara	**Amharic** Afro-Asiatic, Semitic	Amhara: Ca7	Intensive agriculture on permanent fields	cereal grains	Separated hamlets forming a permanent single community	Bovine	Regular milking of domestic animals
Bedouin	**Arabic** Afro-Asiatic, Semitic	Rwala, Mutatir: Cj2, Cj5	Complete absence of agriculture	-	Fully migratory or nomadic bands	Camels	Regular milking of domestic animals
Beni Amer	**Bedawi** Afro-Asiatic, Cushitic	Beni Amer: Ca36	Casual Agriculture	cereal grains	Fully migratory or nomadic bands	Camels	Regular milking of domestic animals
Druze	**Arabic** Afro-Asiatic, Semitic	Druze: Cj8	Intensive agriculture on permanent fields	cereal grains	Compact and relatively permanent settlements	Bovine	Regular milking of domestic animals
Fulani	**Fulfulde** Niger-Congo, Atlantic	Bororo Fulani: Cb8	Extensive or shifting cultivation	cereal grains	Semi-sedentary (transhumant)	Bovine	Regular milking of domestic animals
IsraeliUrban Arab & Palestinian	**Arabic** Afro-Asiatic, Semitic	Sedentary Arab type, Jordanians:Cj6	Intensive agriculture on permanent fields	cereal grains	Compact and relatively permanent settlements	Bovine	Regular milking of domestic animals
Jaali, Shaigi, Donglawi	**Arabic** Afro-Asiatic, Semitic	-	-	-	-	-	-
Mambila	**Mambila** Niger-Congo, Atlantic	Mambila: Ah4	Intensive agriculture on permanent fields	cereal grains	Separated hamlets forming a permanent single community	Sheep and/or goats	Absence of milking
Shuwa Arabs	**Shuwa Arabic** Afro-Asiatic, Semitic	Shuwa: Cb16	Extensive or shifting cultivation	cereal grains	Fully migratory or nomadic bands (some semi-nomadic)	Bovine	Regular milking of domestic animals
Somali	**Somali** Afro-Asiatic, Cushitic	Somali: Ca2	Intensive agriculture on permanent fields	cereal grains	Fully migratory or nomadic bands	Camels	Regular milking of domestic animals
Wolof	**Wolof** Niger-Congo, Atlantic	Wolof: Cb2	Extensive or shifting cultivation	cereal grains	Compact and relatively permanent settlements	Bovine	Regular milking of domestic animals

Table 2.1 Language classification and agricultural information for samples used within this thesis. Language classifications obtained from Ethnologue (**www.ethnologue.com**) and all other information obtained from (Murdock G, 1967).

Figure 2.1 shows abridged language trees for (a) the Atlantic-Congo language family and (b) the Afro-Asiatic language family. Yellow highlight indicates languages spoken by populations included herein. Branches are extended only in cases where a language group is included in this thesis, and … denotes that further information exists. Numbers within brackets denote the number of languages on a given branch. Branch length does not represent distance. All data extracted from www.ethnologue.com.

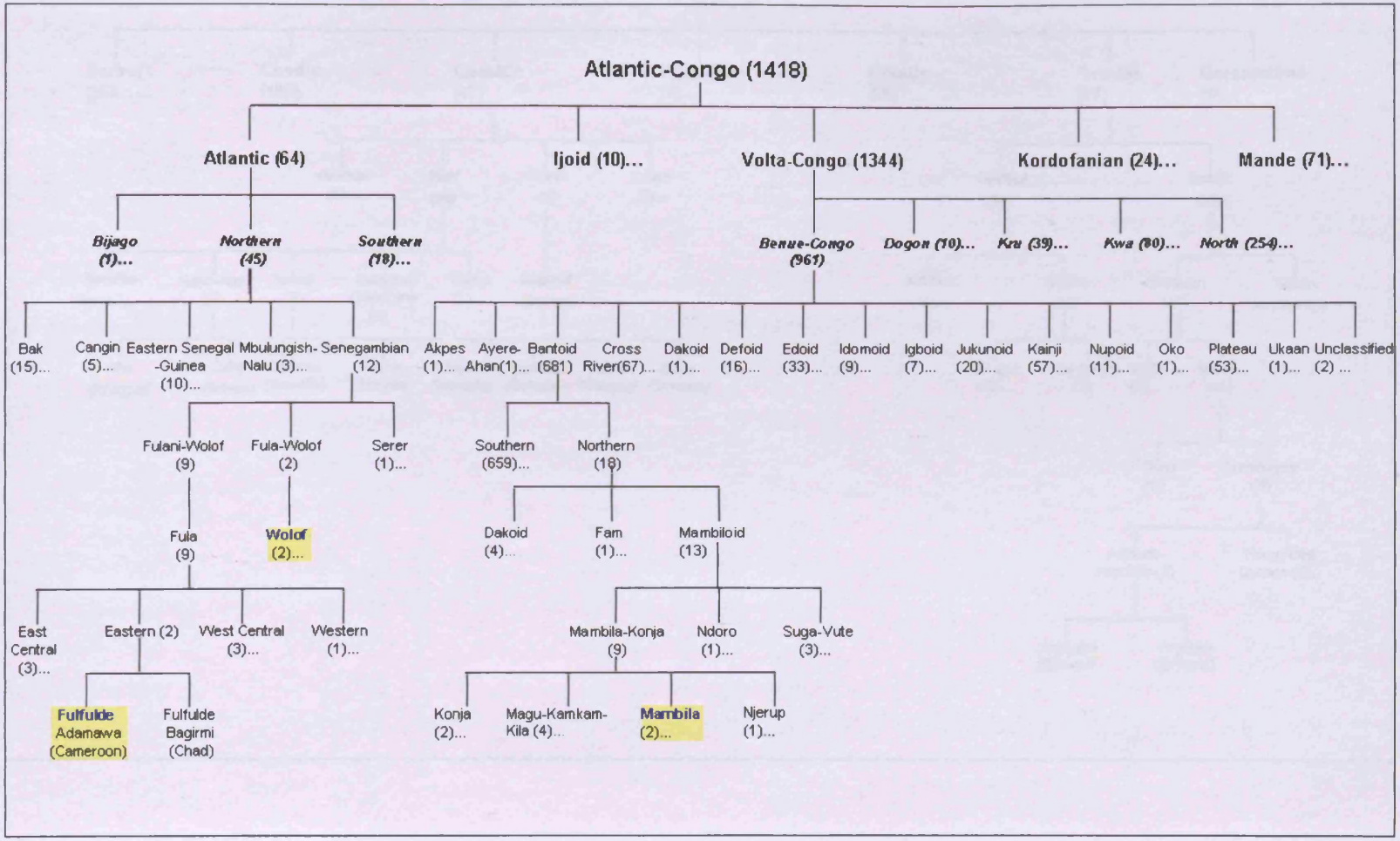
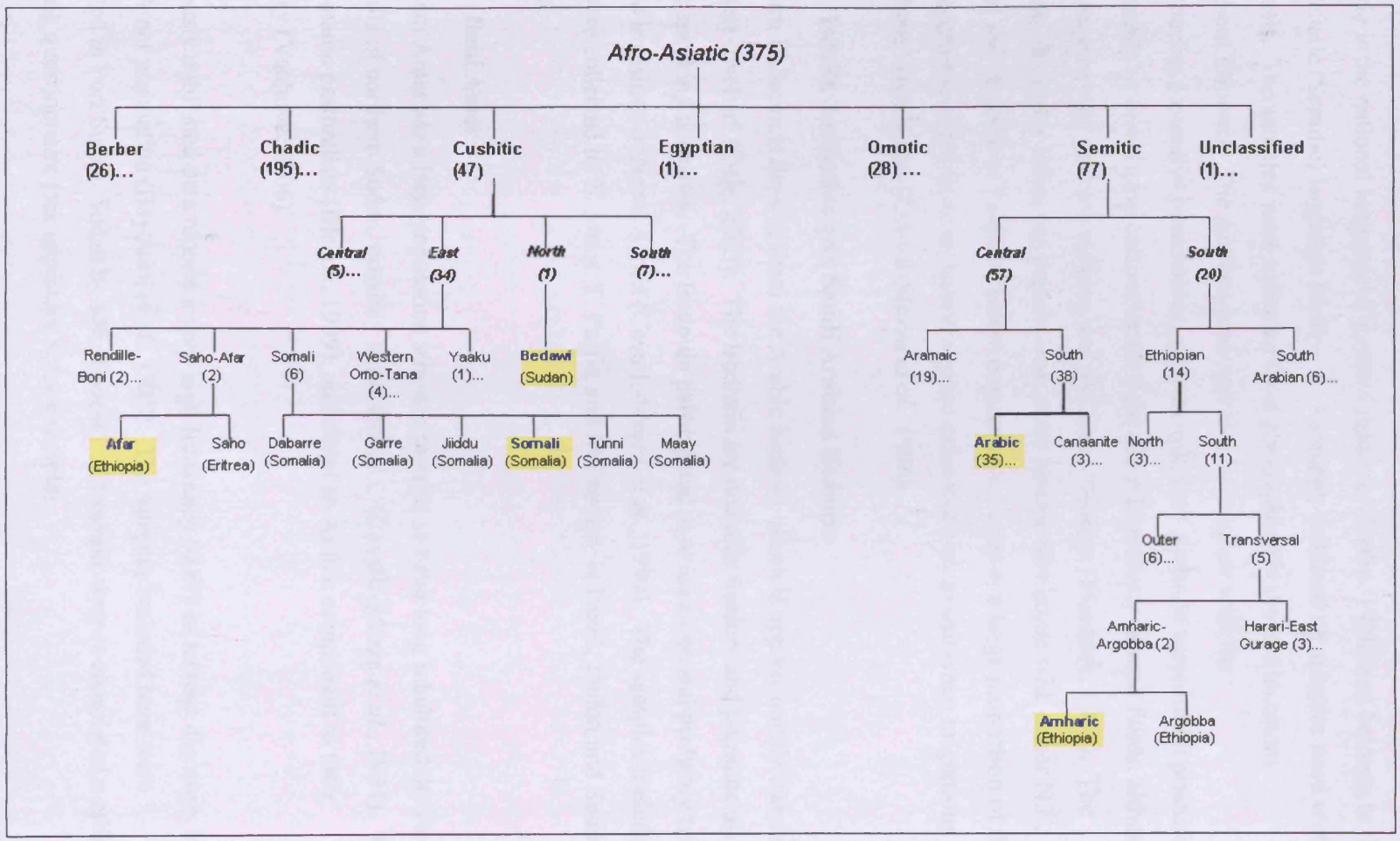


Figure 2.1b


Afro-Asiatic (375)
Berber (26)...
Chadic (195)...
Cushitic (47)
Egyptian (1)...
Omotic (28) ...
Semitic (77)
Unclassified (1)...
Central (5)...
East (34)
North (1)
South (7)...
Central (57)
South (20)
Rendille-Boni (2)...
Saho-Afar (2)
Somali (6)
Western Omo-Tana (4)...
Yaaku (1)...
Bedawi (Sudan)
Aramaic (19)...
South (38)
Ethiopian (14)
South Arabian (6)...
Afar (Ethiopia)
Saho (Eritrea)
Dabarre (Somalia)
Garre (Somalia)
Jiiddu (Somalia)
Somali (Somalia)
Tunni (Somalia)
Maay (Somalia)
Arabic (35)...
Canaanite (3)...
North (3)...
South (11)
Outer (6)...
Transversal (5)
Amharic-Argobba (2)
Harari-East Gurage (3)...
Amharic (Ethiopia)
Argobba (Ethiopia)

2.1.2 Amhara

Amharic is the national language of modern Ethiopia (Dalby, 1998) and belongs to the Afro Asiatic (Semitic) language family. A. Tarekegn collected all samples used within this thesis. The samples were collected from geographically diverse locations throughout Ethiopia. The Amhara were selected to compare with the neighbouring pastoralist populations, as descriptions of Amharic agricultural practises refer mainly to cereal crop cultivation on permanent intensively farmed fields, although cattle are also kept for both milking and plough cultivation (Murdock, 1967). The Amhara, like many Ethiopian populations, show genetic admixture with either N.E. African and Arabian or Middle Eastern populations, however a large proportion of their genes appear to come from an ancestral population that was *in situ* when migrations from these areas began (Cavalli-Sforza *et al.*, 1994).

2.1.3 Israeli, Jordanian and Saudi Arabian Bedouin

The term Bedouin is derived from the Arabic *badawi*, which is approximately translated to 'desert-dweller' (Cole, 2003). The Bedouin are nomadic herders and populate all Arabic speaking countries. The Bedouin participated in at least one and probably two historic invasions of North Africa (Cavalli-Sforza *et al.*, 1994). The samples included here were collected by S. Jones, T. Parfitt and M. Hawary in Israel, Jordan and Saudi Arabia.

2.1.4 Beni Amer

The Beni Amer are a Beja population who are thought to have long inhabited the Red Sea Hills of northern Sudan, possibly since 4000B.C (Cavalli-Sforza *et al.*, 1994). They are nomadic pastoralists (Blench, 1999), and claim an Arabic component to their ancestry (Vanhove, 2006).

Previously published data reports a very high frequency (0.87) of lactose digesters in the Beni Amer population (Bayoumi *et al.*, 1982). The samples included here were collected in Port Sudan, Sudan by ME. Elamin. All sample donors completed a milk drinking questionnaire (see appendix A for example).

2.1.5 Donglawi (Danagala, Danagla) and Shaigi (Shawaga, Shaigiya)

Both the Donglawi and the Shaigi populations were included in this thesis as comparators for pastoralist groups. They are not groups traditionally associated with milk-drinking and pastoralism (personal communication, Mohammed Elamin). Anthropological information on both these populations is sparse, so that which is available is summarised below.

(Khogali M.M., 1991) reports that Danagla is the general name for the inhabitants of the Dongola district of Sudan, located between Abu Fatima and Al-Ghadar, the population of which was approximately 186,000 people in 1983. The Danagla are described as riverine peoples who cultivate the narrow strip of alluvial soils along the river where they have settled (Stanton, 1903). Khogali states that the Danagla are a mixture of ethnic groups, and that the Danagla 'proper' speak the Donglawi language. All individuals included here name Arabic as their first language (and the Donglawi language is not listed in the ethnologue; see section 2.10 for URL). (Bayoumi *et al.*, 1981) report the frequency of lactose digestion in the Donglawi to be 0.18.

The Shaigi groups inhabit both banks of the Nile from Korti to the Third Cataract, and portions of the Bayuda Desert (north of Khartoum). The Shaigi are described as sedentary Arabs (Stanton, 1903), and all the individuals included here spoke Arabic as their first language. The frequency of lactose digesters within the Shaigi population is reported to be 0.38 (Bayoumi *et al.*, 1981). Both the Shaigi and Donglawi samples were collected by H. Babiker.

2.1.6 Druze

The Druze are described as sedentary Arabs of the Near East. They belong to an Islamic sect that originated shortly after AD 1000 in Fatimid Cairo (Cavalli-Sforza *et al.*, 1994; Swayd, 1998). The Druze are now resident in a number of Arabic-speaking countries including Syria, Israel, Lebanon and Jordan. The samples included in this research were collected in Israel by N. Bradman.

2.1.7 Fulani (Fulbe, Peuhl)

M. Abakachi collected buccal samples and milk drinking information from nomadic Fulani individuals in the Provinces of the North and Extreme North of Cameroon. Samples were collected from 141 individuals belonging to 35 families. After DNA extraction and genotyping of three *LCT* SNPs it was noted that there appeared to be a large number of misfits between parents and children. Subsequently, all DNA samples were gender determined using a PCR of intron 1 of the amelogenin gene (Faerman *et al.*, 1995) which amplifies an Alu insertion present on both X and Y chromosomes, but with a 6bp deletion in the X chromosome copy. This confirmed that in all cases gender was in agreement with information recorded on the data sheet, and the possibility that the samples had been either mixed up during extraction or fraudulently collected were considered highly unlikely.

To further understand the relationships of individuals within families, analysis was carried out on a 13 allele STR marker located between MUC2 & MUC5AC in the mucin gene complex on chr11p15.5 (as described in Rousseau *et al.*, 2007), as well as 6 Y chromosome STRs (see section 2.5.4) and mtDNA (HVR1) (section 2.6.2). Only 13 families had children showing inheritance of alleles consistent with reported parentage, and we believe this is likely due to a linguistic misunderstanding of the relationships between family members, as polygynic marriages are common in Fulani culture. (Riesman, 1977) provides a detailed account of familial relationships within Fulani society. In view of these findings, 94 unrelated individuals who had been designated 'parents' on the collection sheets were selected from the original sample set and were used in this thesis.

The Fulfulde language belongs to the Atlantic branch of the Niger-Congo languages, and is closely related to the Wolof language of Senegal (Dalby, 1998). The Fulani are one of the dominant ethnic groups of west Africa. The traditional Fulani lifestyle includes periods of transhumance with the cattle, and heavy dependence on milk as a food source (Murdock, 1967; Blench, 1999), although the consumption of milk as a staple food is now mainly restricted to the nomadic Fulani (Keen and Zeitlyn, 2007). Lactose digestion frequency in nomadic Fulani is reported to be 0.78 (Kretchmer *et al.*, 1971).

It is commonly suggested that the Fulani have a non-African component to their ancestry, and are often referred to as 'fair-skinned' or 'Caucasoid-like' in the literature (Murdock, 1959; Adebayo, 1991). One interpretation of the Fulani legend of origin is that the group arose from a union between infiltrating Berber herdsmen from North Africa and an indigenous Senegambian population (Adebayo, 1991). This is consistent with observations that the *–13910*T* lactase persistence associated allele occurs in both Fulani and Berber populations (Mulcare *et al.*, 2004; Myles *et al.*, 2005), and described later in this thesis (chapter 3). It has also been suggested (from Y-chromosome data) that a migration from Eurasia to Cameroon may have occurred at some point in the ancient past (Cruciani *et al.*, 2002). In contrast mtDNA studies reveal that Fulani in Cameroon and Burkina Faso have mitochondrial haplogroups consistent with a west African origin (Cerny *et al.*, 2006).

2.1.8 Israeli urban Arabs and Palestinians

Israeli urban Arabs and Palestinians from Israel/region of the Palestinian Area Authority (PAA) were collected by P. Smith and N. Bradman. Both populations are permanent settled populations of the Near East which are included here to provide a comparison to the neighbouring nomadic pastoralist populations.

2.1.9 Jaali (Jaalin, Jaaliyin, Ja'lien, Ga'ali, Gaaliyin)

The Jaali are a Sudanese group with part Arab ancestry (Robinson, 1927; Warburg, 1978), traditionally resident in Shendi, a town approximately 150 km north of Khartoum. The Jaali form permanently settled communities (Stanton, 1903), and whilst camels are of cultural importance to the population, the Jaali are not identified as pastoralists in any of the available literature. The Jaali population have been previously reported to have an intermediate lactose digester frequency of 0.53 (Bayoumi *et al.*, 1981).

ME. Elamin collected buccal cell samples from 99 Jaali individuals resident in Shendi, Sudan and unrelated at the grandparental level. Lactase persistence status was ascertained using the breath hydrogen method (see section 2.4), and each individual completed a milk drinking questionnaire.

2.1.10 Mambila

Forty-two Mambila samples were collected in the Province of Adamawa, Cameroon by D. Zeitlyn. Sixteen of the samples were collected in the village Somié, and another 18 in the nearby village Atta. A further 8 samples were collected in Mayo Darlé.

It is generally accepted that the Nigeria - Cameroon borderland region is the hub of the expansion of the Bantu-speaking peoples (~2000B.C) and has been continuously occupied by related peoples since, although subject to population migrations (Ehret, 2001). Mambila is a bantoid language, and it is been suggested that the languages and peoples who inhabit the Nigeria - Cameroon borderland region may represent the 'Bantu who stayed home' (Zeitlyn and Connell, 2003). The Mambila form settled communities that cultivate cereal crops on permanent fields. The practise of milking is near absent (Murdock, 1967) and consumption of dairy products rare.

2.1.11 Shuwa Arabs (Baggara)

The Shuwa Arab samples used within this thesis were collected at a number of locations near Lake Chad, in the Province of the Extreme North of Cameroon, by EF. Caldwell and MF. Lepey. Shuwa Arabic is the name given in various other countries for the variety of Arabic spoken near Lake Chad (Ethnologue).

The Shuwa are directly related to the Bedouin groups of the Egyptian desert, but have intermarried heavily with local populations (Blench, 1999). The first records of the Shuwa/Baggara in the Chad basin date back to the 14th Century, and ethno historical records report that they migrated there from the east of the Chad Republic, near Sudan (Levy and Holl, 2002). The centrality of intensive cattle husbandry and pastoralism (Blench, 1999) as a lifestyle are said to have maintained the Shuwa as a distinct ethnic group, despite their adoption of local languages of the region (Levy and Holl, 2002).

2.1.12 Somali

There are 2-3 million Somalis living in Ethiopia (Cavalli-Sforza *et al.*, 1994). The Somali are a pastoralist population (Blench, 1999), although some papers suggest pastoralism has been adopted by the northen Somali in response to extreme poverty, and that mixed agriculture may have been the traditional form of agriculture within the

population (Lewis, 1966). Much of the literature suggests that the Somali population originated in the Gulf of Aden and migrated south and westwards (Lewis, 1960), however alternative interpretations of linguistic evidence and historic documents suggest that the Somali originated in S. Ethiopia/N. Kenya and migrated north and east to inhabit their current geographic locations (Lewis, 1966). Somali origins are often discussed alongside those of the Galla (Oromo) populations with whom the Somali have a long and intimate history (Lewis, 1960; Lewis, 1966; Besteman, 1993). Y-chromosome analysis of the Somali population shows that the Somali cluster with the Oromo and Borana (N. Kenyan) populations, with an E. African origin dated to approx 4-5kya (Sanchez *et al.*, 2005). The same study reported indications of some admixture with Eurasian and sub-Saharan African populations.

Two separate collections of Somali samples are included here. A. Tarekegn and T. Olijiria-Raga collected 39 samples with milk drinking information in Jijiga, Ethiopia. A further collection of 111 individuals was then carried out in Shenile, near Dire Dawa, Ethiopia. In the second collection lactose tolerance status was obtained using the breath hydrogen test, and milk drinking information was recorded on both collections. The second sample collection was carried out by SL. Browning, CJE. Ingram, A. Tarekegn and T. Olijiria-Raga.

2.1.13 Wolof

Wolof is the predominant language of Senegal and belongs to the Atlantic group of Niger-Congo languages (Dalby, 1998). In geographic terms, within this thesis the nearest populations to the Wolof are the Mambila and Fulani populations from Cameroon. mtDNA studies show that the haplotype frequency distribution in the Wolof is similar to the distribution generally observed in populations of Bantu-speaking peoples (Scozzari *et al.*, 1988).

The Wolof agricultural practises include extensive cultivation of cereal crops, and they also keep cattle which are regularly milked (Murdock, 1967). The Wolof population have been reported to have a lactose digestion frequency of 0.51 (Arnold *et al.*, 1980). The samples referred to in this thesis were collected in Senegal by D. Gomis.

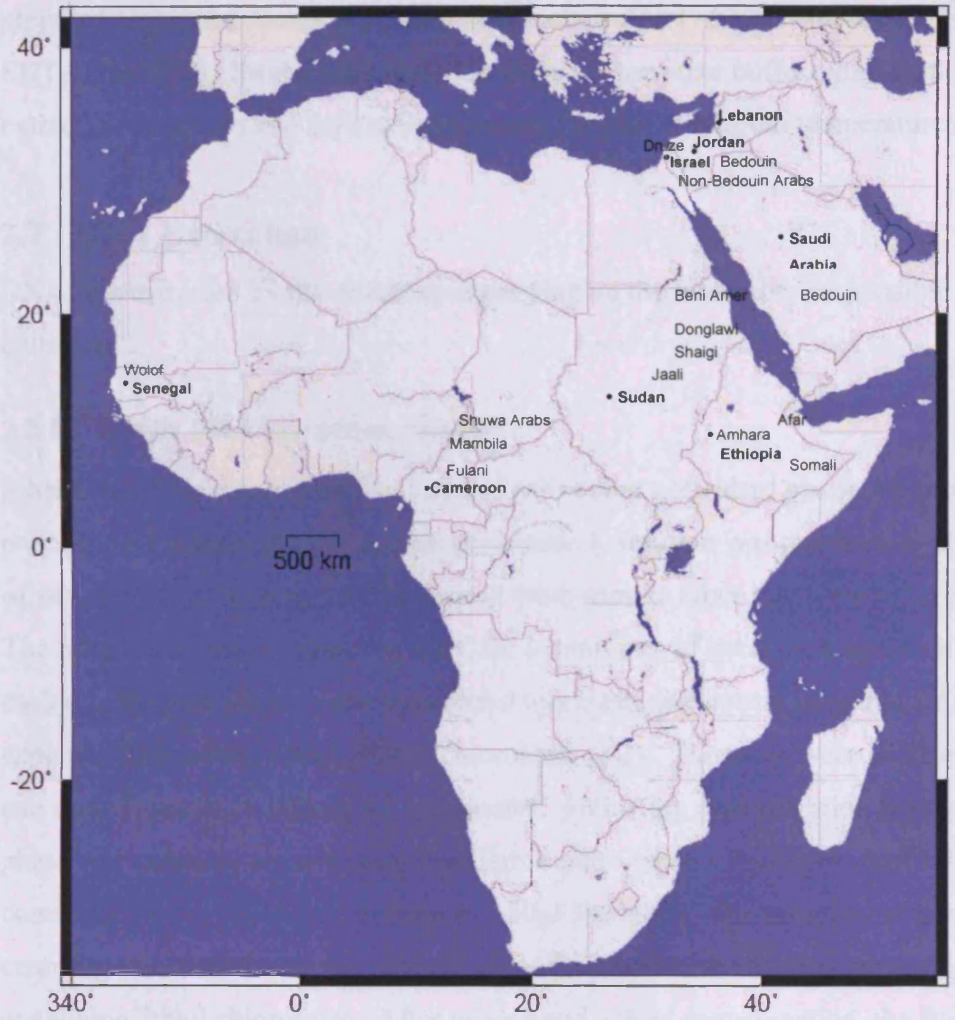


Figure 2.2 Geographic map detailing the sampling locations of populations included within this thesis. Country names are only indicated (in **bold**) in the cases where DNA samples were collected. The names of the ethnic groups sampled in each country indicate an approximate sampling location only.

2. Collection of DNA

All DNA samples used in this thesis were extracted from buccal cells. Cotton buds on sticks were rubbed over the surface of the inner cheek for approximately 30 seconds to collect buccal cells. A single swab was collected per individual for TCGA organised collections and ten swabs were collected per individual for Galton samples.

TCGA sample swabs were placed into a sterile tube to which 1.5ml extraction/preservative solution (50mM EDTA and 0.5% SDS) was immediately added. In the case of Galton lab sample swabs, all ten were placed into a single, previously

prepared 15ml tube containing 2.5ml Slagboom buffer (which also contains SDS and EDTA, see 2.10). Swabs were stored in these preservative buffers until further extraction in as cool and dark an environment as possible (room temperature or below).

2.2 DNA Extraction

DNA was extracted by two methods depending on the laboratory responsible for sample collection.

2.2.1 TCGA DNA extraction method

Single swabs were extracted from buccal cells using a standard phenol chloroform method. A stock solution of 2µg/ml proteinase K solution was prepared in dH_20, 800µl of which was added to each of the mouth swab sample tubes that were to be extracted. The tubes were then incubated at 60°C for a minimum of two hours, and then 800µl of each mouth swab solution was transferred to a clean, previously prepared 1.5ml eppendorf containing 600µl phenol/chloroform (1:1). The tubes were inverted to mix and centrifuged at 16,000 xg for 10 minutes. Following centrifugation the aqueous phase was carefully removed and transferred into a clean 1.5ml eppendorf tube containing 600µl phenol/chloroform and 30µl 5M NaCl. The samples were mixed and centrifuged as before and the aqueous phase transferred into another eppendorf containing 700µl chloroform. After mixing and a third centrifugation, the final aqueous phase was transferred to a clean screw-cap tube containing 700µl isopropanol. The tubes were inverted to mix and then cooled to -20°C for a minimum of 2 hours to facilitate DNA precipitation. The samples were then centrifuged at 16,000 xg for 12 minutes, after which the supernatant was carefully poured away and the tubes left to drain (inverted and at an angle) for one minute. 800µl of 70% EtOH wash was added to each tube, which was then centrifuged for another 10 minutes. The supernatant was discarded and the tubes were left to drain at an angle for 20 minutes. To resuspend the DNA pellet, 300µl of TE was added to each sample and the tubes were incubated at 56°C for 10 minutes with occasional mixing. Samples were pulse centrifuged before storage at -20°C. Remaining extraction solution and cotton buds of each sample were transferred into labelled 1.5ml eppendorf tubes and stored at -20°C.

2.2.2 Galton Laboratory DNA extraction method

Samples for which ten swabs had been collected were extracted using an adaptation of the phenol chloroform method as reported by (Freeman *et al.*, 2003). The 15ml collection tubes containing the cotton swabs in Slagboom buffer were incubated in a water bath at 65°C. After 2 hours incubation, the caps were removed and the tubes inverted into clean 50ml conical tubes that were centrifuged at 650 xg for 10 minutes. The original 15ml tube and swabs were drained and discarded and the liquid transferred into a clean 15ml tube to which 300µl of an organic deproteinisation solution (Yeast Reagent 3 (YR3) solution (Autogen, Bioclear) mixed 1:1 with EtOH was added. The contents were vigorously shaken for 1 minute, and then centrifuged at 8000 xg for 25 minutes. Following centrifugation the supernatant was transferred into a clean 15ml tube and another 300µl YR3 was added. The shaking and centrifugation steps were repeated as before, and the supernatant was again transferred to a clean 15ml tube. 1.8ml isopropanol was added and the tube inverted gently to mix, then centrifuged at 8000 xg for 25 minutes. The supernatant was discarded and the pellet washed by adding 1ml 70% ethanol, and centrifuging for 10 minutes at 8000 xg. The ethanol wash was discarded and the pellet left to air-dry for 15 minutes. DNA pellets were resuspended in 400µl TE and the sample was incubated on a rocking platform at 4°C overnight. Two aliquots were made of each sample, one for dilution and use at 4°C and another aliquot stored at -70°C.

2.3 Whole Genome Amplification

In some cases where DNA samples were in short supply, it was necessary to carry out amplification of the genomic DNA in order to provide enough material for use in this project. This was necessary only with samples obtained from TCGA stocks; specifically the Middle Eastern samples (Israeli and Palestinian non-Bedouin Arabs and Israeli, Jordanian and Saudi Arabian Bedouins).

The Genomiphi kit (Amersham Biosciences) was used to amplify genomic DNA. The method uses random hexamer primers to anneal to the template DNA at multiple random sites and bacteriophage Phi29 DNA polymerase to replicate the template strand in an

isothermal reaction. Random primers also bind to the newly synthesised single stranded DNA, thus allowing an exponential replication of the original material.

0.5µl sample DNA (approximately 1ng) was used for each amplification reaction, and added to a clean eppendorf containing 4.5µl sample buffer (Amersham Biosciences) containing random hexamer primers. The solution was incubated at 95°C for 3 minutes then cooled on ice. Meanwhile a master mix of 4.5µl of reaction buffer (Amersham Biosciences) containing hexamer primers, buffers, salts and dNTPs, and 0.5µl enzyme mix (Amersham Biosciences) were combined on ice (both multiplied up by the number of samples to be amplified). 5µl of this master mix was added to each sample, which was then incubated at 30°C for 16-18 hours. After amplification, the Phi29 DNA polymerase was inactivated by heating to 65°C for 10 minutes, and then cooled to 4°C. Samples were finally diluted 1:50 for use in PCR reactions.

2.4 Breath hydrogen lactose tolerance testing

Lactose tolerance tests were used to assess lactase persistence status. Fully informed consenting volunteers who fasted overnight and refrained from smoking were given 50g lactose as a solution in 250ml water brought to room temperature. Lactose tolerance was tested by the breath hydrogen method using a MicroH_2 meter (MicroMedical) (Peuhkuri *et al.*, 1998) to measure breath hydrogen before lactose ingestion and at 30-minute intervals thereafter. Reports of any recent intestinal complaints and antibiotic treatment were recorded and these data were consulted during analysis of results, but not used as a criterion for excluding individuals from the study. To be included individuals were usually required to have clearly detectable starting breath hydrogen (preferably above 2ppm) to exclude potential hydrogen non-producers, and not above 20ppm (to exclude individuals with potential bacterial overgrowth, and those who had not fasted, or had smoked, or recently eaten an excess of fermentable carbohydrate such as beans). A replicated increase of breath hydrogen of greater than 20ppm above baseline was considered to indicate lactose digestion.

2.5 Genotyping

SNPs were typed either by RFLP or tetra-primer ARMS-PCR. Both methods employ an initial PCR step and the resulting PCR products were electrophoresed on agarose gels. Genotypes were inferred from the gel phenotype of each sample. Microsatellelites were typed by PCR and electrophoresed on a DNA analyser.

2.5.1 Polymerase Chain Reaction

Standard PCR reactions were carried out in 10µl total volume of 1x Buffer IV (Abgene) containing 0.2mM dNTPs, 0.25u Taq DNA polymerase (Abgene) and approximately 10-20ng genomic DNA. Primer concentrations were optimised for each reaction (Table 2.2). All primers were obtained from Sigma-Genosys.

In some cases, faint non-specific bands were produced using standard PCR conditions, and in these cases a 'hot start' PCR was used to increase specificity. Thermostart PCR reactions were carried out in 10µl total volume of 1x Thermostart reaction buffer (Abgene) with 0.2mM dNTPs, 2.5mM $MgCl_2$, 0.25u Thermostart Taq DNA polymerase (Abgene) and approximately 10-20ng genomic DNA. Primer concentrations were optimised for each reaction.

Cycling conditions were individually optimised for each PCR reaction, however, all PCR programs followed the same structure: 1 cycle of 95°C for 5 minutes followed by a variable number of cycles of 95°C for 30 seconds, X°C for 30 seconds (where X = optimised primer annealing temperature), and 72°C for 1minute/kb (for a minimum of 30 seconds), followed by a single cycle of 72°C for 5 minutes. Primer sequence and PCR-specific cycling conditions are listed in Table 2.2.

Special Note

Page 70 missing from the original

Polymorphism	Location (bp 5' of LCT start)	Primer name	Primer sequence	Primer concentration	Annealing temperature	Number of cycles
rs309180	-19516	MCM6Taq1left	5′-TCA-TGT-CCC-GAT-TCG-ATC-TCT-T-3′	0.50	58	35
		MCM6Taq1right	5′-CAG-TCT-CCA-GTG-AGA-GGC-T-3′	0.50		
rs4954493	-15226	MCM6i11FO	5′-TAA-ATT-TCT-TTG-GGA-CAG-TGA-GAG-C-3′	0.75	56	35
		MCM6i11RO	5′-GGA-AAT-TAA-CTG-AAC-CTG-TCA-TCT-ACC-3′	0.50		
		MCM6i11FI	5′-TCA-ATC-ACT-GTA-AAA-CAG-TAA-TTT-ATA-TTG-AC-3′	1.25		
		MCM6i11RI	5′-ACA-TTG-GAT-TCA-TCT-AGG-CCA-TTA-GGA-3′	0.50		
rs4988235	-13897	LAC-C-M-U	5′-GCT-GGC-AAT-ACA-GAT-AAG-ATA-ATG-*GA-3′	0.25	55	
(-13910C>T)		LAC-C-L-2	5′-CTG-CTT-TGG-TTG-AAG-CGA-AGA-T-3′	0.25		
rs4954490	-13482	4954490For	5′-AAT-GGA-AGG-CAG-GGG-CTC-TC-3′	0.50	60	35
		4954490Rev	5′-GCC-CTC-TTC-CCC-TGA-AGT-A*GC-3′	0.50		
rs3769005	-8617	ZA2SE	5′-AAT-TTG-TGG-AGT-ATC-AGC-ATA-CCT-GC-3′	0.50	58	35
		MCM6i15	5′-CTG-TGA-TGA-CAA-AAT-ATT-CCA-G-3′	0.50		
942/3TC>ΔΔ	-942/3	942Ff	5′-GAC-AAA-ATA-GAG-GCA-CAA-AGT-TAA-G-3′	0.50	53	30
		A942-TC	5′-GTT-TCC-ATA-TTG-TTT-GAA-TCA-TAC-3′	0.50		
		S942Fr	5′-CAG-CCA-CAC-ATT-ATT-TTA-AAA-TTT-G-3′	0.50		
		942+TC	5′-ATA-TTG-CTG-AAC-ATA-TTT-TGT-AAG-AGA-3′	0.50		
-678A>G	-678	C678f	5′-CGC-ACA-CCT-ATG-GTC-CCA-*GCT-3′	0.50	64	30
		C678r	5′-AAG-CAG-AGG-AAC-CCG-GAA-AAT-GTC-3′	0.50		
rs3754689	4003	Exon2f	5′-AGT-GGT-TTC-CAC-AGT-CAG-AGC-3′	0.25	52	34
(666G>A)		2Exon2r	5′-TCT-ACA-GCT-CGC-AGG-*TCA-A-3′	0.25		
rs2278544	48639	x17F	5′-CTG-AGA-ACT-CAA-ATC-AGC-GC-3′	0.50	58	34
(5579T>C)		Exon17r	5′-AAA-GCT-GTC-TGT-GCT-TCT-GTG- 3′	0.50		

Table 2.2 Primer sequences and PCR conditions for SNP genotyping. A * in the primer sequence indicates that the subsequent base has been substituted in order to engineer a restriction endonuclease site.

2.5.2 Tetra-primer ARMS PCR

Tetra–primer ARMS-PCR was designed for genotyping both the *-942/3TC>ΔΔ* indel, which had previously been typed by DGGE (Hollox *et al.*, 1999) and the rs4954493 SNP. Undergraduate students Fiona Pring and Caroline Freeman assisted in the design and optimisation of these assays respectively. The method employs two primer pairs to amplify the two different alleles of a SNP in a single PCR reaction (Ye *et al.*, 2001). Two allele specific primers recognise the different alleles of the SNP, on opposite strands of the DNA, and each elongates towards an outer primer located (at different distances) further away from the SNP. Three possible products can be produced during PCR, a control band, which is the product of the two outer primers, and the two products of the allele specific primers with their outer primer partners. PCR products were resolved on 2.5% agarose gels stained with 0.05mg/ml ethidium bromide. Primer sequences and PCR conditions are included in Table 2.2.

2.5.3 PCR-RFLP

All other SNPs were typed by restriction fragment length polymorphism (RFLP). The PCR products were digested overnight with an appropriate restriction endonuclease and genotyped by visualisation of the digestion products on agarose gels (for specific reaction conditions refer to Table 2.3). All genotyping assays were conducted with positive and negative controls and in most cases typing was confirmed by two investigators.

Polymorphism	Restriction endonuclease	Enzyme Units	Incubation temperature (°C)	Reaction Volume (& Buffer	Agarose Gel %
rs309180	TaqI (T/CGA)	1.0	65	15µl, NEB 3 + 100 µg/ml BSA	2.0
rs4988235 *(-13910C>T)*	HinfI (G/ANTC).	0.2	37	10µl, NEB 2	3.5
rs4954490	AluI (AG/CT)	1.0	37	10µl, NEB 2	3.5
rs3769005	NlaIII (CATG/)	2.5	37	15µl, NEB 4 + 100 µg/ml BSA	2.5
-678A>G	PvuII (CAG/CTG)	0.1	37	10µl, NEB 2	3.5
rs3754689 *(666G>A)*	HincII (GTY/RAC)	0.2	37	15µl, NEB 3 + 100 µg/ml BSA	3.5
rs2278544 *(5579T>C)*	MspI (C/CGG)	5.0	37	15µl, NEB 2	3.0

Table 2.3 Restriction endonuclease digestion conditions and agarose gel resolving concentrations for PCR-RFLP genotypings.

2.5.4 Y-chromosome high throughput STR and SNP genotyping

Typing of six microsatellite and six diallelic polymorphisms was carried out using a high-throughput technique as described in Thomas *et al.*, (1999). Briefly, two multiplex PCRs were performed (microsatellite kit 1 (MS1), and unique event polymorphism kit 1 (UEP1), using oligonucleotide primers labelled with three different ABI dyes (full details of primer sequences and dyes can be found in (Thomas *et al.*, 1999)). Following PCR amplification, 2.1µl of the UEP1 amplification products were transferred into a clean 384-well plate, which had been pre-prepared with 5.5µl per well of UEP1 digestion mix (Thomas *et al.*, 1999) and were digested overnight at 37°C.

Both the MS1 amplification products and the UEP1 digestion products were analysed in the presence of GS-500 ROX size standard (Applied Biosystems). 1.1µl of digestion/amplification product was added to a 10µl aliquot of 1:90 GS-500 ROX size standards: HiDi Formamide (Applied Biosystems). Immediately prior to electrophoresis, the samples were denatured at 96°C for 4 minutes then cooled on ice for at least 5 minutes. The samples were loaded onto an ABI 3730xl DNA analyser, and fragment analysis was performed using GeneMapper v4.0 software (Applied Biosystems).

2.6 Sequencing

2.6.1 *MCM6* intron 13 PCR

The *MCM6* intron 13 enhancer region was amplified using specific primers and the standard PCR reaction as described above (section 2.5.1), with all components scaled up by a factor of 1.5 to give a total final volume of 15µl. Sequencing primer pairs and cycling conditions are indicated in Table 2.4. All stages of sequencing were carried out on 96-well plates.

Primer name	Primer sequence	Primer concentration	Annealing temperature	Number of cycles
MCM6i13	5′-GGA-CAT-ACT-AGA-ATT-CAC-TGC-AAA-TAC-3′	0.5	54	34
MCM778	5′-CCT-GTG-GGA-TAA-AAG-TAG-TGA-TTG-3′	0.5		
MCM6i13	5′-GGA-CAT-ACT-AGA-ATT-CAC-TGC-AAA-TAC-3′	0.5	54	34
LAC-C-L2	5′-CTG-CTT-TGG-TTG-AAG-CGA-AGA-T-3′	0.5		
MCM778	5'-CCT-GTG-GGA-TAA-AAG-TAG-TGA-TTG-3'	0.5	58	35
MCM6ex13	5'-ATT-TCC-AAA-GAG-TCA-GAG-GAC-TTC-3'	0.5		

Table 2.4 Primer combinations and PCR conditions for sequencing of *MCM6* intron 13 enhancer region.

2.6.2 mtDNA HVR1 PCR

Hypervariable region 1 (HVR-1) was amplified in 10-µl reaction volumes containing 5-10ng DNA, 0.35 mM of primers ConL849 (5'-CTA-TCT-CCC-TAA-TTG-AAA-ACA-AAA-TA-3') and ConH1-mod (5'-CCC-TGA-AGT-AGG-AAC-CAG-ATG-3'), 0.13 units Taq polymerase (HT Biotech), 9.3 nM TaqStart monoclonal antibody (Clontech), 200 mM dNTPs, 0.5mM $MgCl_2$ and reaction buffer supplied with the polymerase (HT Biotech). Cycling parameters were as follows: preincubation for 5 min at 95°C, followed by 37 cycles of 1 min at 93°C, 1 min at 55°C, and 1 min at 72°C.

2.6.3 PCR Clean up

Following PCR, amplified product was purified using a polyethylene glycol (PEG) precipitation protocol. Three volumes of PCR precipitation solution (section 2.10) were added to the PCR reaction and mixed thoroughly, then centrifuged at 900/2000 xg (dependent on centrifuge availability) for 45 minutes. Following centrifugation the supernatant (which contains excess nucleotides, primers and salts) was discarded by inverting the plate onto a piece of tissue and centrifuging at 20g for 30 seconds. A 70% (150µl) EtOH wash was then added to the plate to remove any remaining PEG and contaminants. Following centrifugation at 900/2000 xg for 25 minutes the wash was discarded by inversion and centrifugation (as before). The DNA pellets were dried at 65°C for 5 minutes. The purified DNA pellets were resuspended in 10µl dH_20 for autosomal loci or 80µl for mtDNA.

2.6.4 *MCM6* intron 13 cycle sequencing

PCR products were subjected to electrophoresis to estimate the correct amount of template to use in the sequencing reaction, which was typically 3µl (corresponding to approximately 10-20ng template). Sequencing reactions were carried out in 15µl total volume consisting of 0.78µl BigDye v1.1 Terminator mix (Applied Biosystems, ABI), 2.4µM primer (sequences given in Table 2.4), and 5µl sequencing buffer (section 2.11). Cycling conditions were as follows (according to the ABI BigDye protocol): 96°C for 1 minute, followed by 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes.

2.6.5 mtDNA HVR-1 cycle sequencing

Forward-strand sequencing was performed in 15µl reaction volumes using 6µl of the resuspended mtDNA PCR product, 1µl BigDye termination mix, 5µl sequencing buffer (section 2.11), and 0.16 µM primer ConL884 (5'-GCC-TGT-CCT-TGT-AGT-ATA-A-3'). Reverse strand sequencing was performed using the reverse primer ConHmt3 (5'-CCA-GAT-GTC-GGA-TAC-AGT-TC-3'). Cycling conditions were as above (section 2.6.4).

2.6.6 Sequencing clean up

Following the sequencing reaction DNA was pelleted by isopropanol precipitation. 80µl of 80% isopropanol was added to the sequencing reaction and the mix was centrifuged at 900/2000 xg for 45 minutes. The supernatant was removed by inverting the plate onto tissue paper and centrifuging at 20 xg for 30 seconds. 150µl of 70% isopropanol was then added to remove any remaining contaminants and centrifuged at 900/2000 xg for 25 minutes. The supernatant was removed by inverted centrifugation as described in the previous step. Pellets were dried by incubation at 65°C for 5 minutes.

2.6.7 Analysis

Immediately prior to loading onto an ABI 3100, 3700 or 3730xl DNA analyzer (Applied Biosystems), the pellets were resuspended in 10µl HiDi formamide (Applied Biosystems), then denatured at 95°C for 5 minutes and immediately placed on ice for a minimum of 5 minutes. Chromatograms were analysed using Sequencher (Gene Codes Corporation), or ChromasPro (Technelysium Pty Ltd) software.

2.7 Electrophoretic Mobility Shift Assay (EMSA)

EMSAs were used to characterise the affinity of oligonucleotides containing single nucleotide substitutions for the transcription factor OCT1.

2.7.1 Probe preparation.

Double stranded oligonucleotides (designed with 5' overhangs) were used as probes and competitors (sequences are shown in Table 2.5). Two single stranded oligonucleotides were obtained for each probe and resuspended to a concentration of 100μM. Double stranded molecules were prepared by mixing equimolar amounts (5μmol) of the complementary oligos in a total volume of 125μl 1x STE buffer (section 2.11) such that the final concentration of each primer was equal to 40μM. This solution was heated to 95°C for 5 minutes. The heat source was then removed and the solution was left to return gradually to room temperature. The probes were electrophoresed on a 1xTBE, 20% (37.5:1) acrylamide:bis-acryalmide gel which was silver stained to verify that double stranded probes had been formed.

2.7.2 Silver Staining

The acrylamide gel was carefully removed from between the glass plates of the electrophoresis apparatus and immediately submerged in fixing solution (10% EtOH, 0.5% acetic acid) for 6 minutes with constant agitation. The fixing solution was carefully discarded and the gel incubated in silver nitrate solution (0.1% (w/v) $AgNO_3$ in dH_20) on a shaking platform for 10 minutes. The silver nitrate solution was discarded and the gel was very briefly washed in dH_20, before the final 20 minute incubation in developing solution (375mM NaOH, 2.6mM $NaBH_4$, 0.148% v/v formaldehyde).

The gel was then placed between two acetate sheets and scanned using an hp scanjet 7400c scanner to record the image. A back-up record of the gel was obtained by vacuum drying on filter paper at 80°C for 2 hours in a Biorad model 583 gel dryer.

Probe name	Sequence
13910C (ANCESTRAL)	5′-AAGATAATGTAGCCCCTGGCCTCAA-3′ 3′-TCTATTACATCGGGGACCGGAGTTT-5′
13910T	5′-AAGATAATGTAG**T**CCCTGGCCTCAA-3′ 3′-TCTATTACATC**A**GGGACCGGAGTTT-5′
13915G	5′-AAGATAA**G**GTAGCCCCTGGCCTCAA-3′ 3′-TCTATT**C**CATCGGGGACCGGAGTTT-5′
13913C	5′-AAGATAATG**C**AGCCCCTGGCCTCAA-3′ 3′-TCTATTAC**G**TCGGGGACCGGAGTTT-5′
13907G	5′-AAGATAATGTAGCC**C**GTGGCCTCAA-3′ 3′-TCTATTACATCGG**G**CACCGGAGTTT-5′
OCT1	5′-ATGTCGAATGCAAATCACTAGAACT-3′ 3′-ACAGCTTACGTTTAGTGATCTTGAA-5′
Nonspecific	5′-AACTCCGGTCCCCGATGTAATAGAA-3′ 3′-TGAGGCCAGGGGCTACATTATCTTA-5′
TAATGARAT	5′-TCGTCGTATCTCATTACCGCCGTCG-3′ 3′-GCAGCATAGAGTAATGGCGGCAGCT-5′

Table 2.5 Double stranded oligonucleotide probes used for EMSAs. Positions of nucleotide substitutions are indicated in bold.

2.7.3 Radioactive labelling

Double stranded oligonucleotide probes were end labelled using T4 polynucleotide kinase (Roche). 2pmols double stranded oligonucleotide probe were diluted into 6µl 1x phosphorylation buffer, pH8.2 (Roche) to which 3µl (10µCi/µl) γ-32P ATP (Amersham Biosciences) was added. Finally, 1µl T4 polynucleotide kinase (Roche) was added and the reaction was mixed, pulse centrifuged and incubated at 37°C for 1 hour.

2.7.4 Probe Purification

Unincorporated nucleotides were removed using a NAP-5 column (Amersham Biosciences) and eluted in 1ml 1x STE buffer (section 2.11). Efficiency of transfer of the radioactive label to the 5' termini of each probe was calculated by dividing the amount of radioactivity recorded in the purified probe by that recorded prior to purification. Measures of radioactivity were taken in counts per second using a Geiger counter placed 50cm away from the eppendorf tubes, and care was taken to ensure that the same type of tubes were used for radioactivity measurement of both the original solution and the purified probe.

2.7.5 Nuclear protein extract preparation

Nuclear protein enriched extracts were prepared from differentiated Caco-2 cells by Yangxi Wang by the method described in (Hollox *et al.*, 1999). Briefly, cells were cultured in Dulbecco's Modified Eagles Medium and 20% foetal calf serum for 15 days, at which time they had achieved maximal lactase expression. The cells were harvested, and washed in phosphate-buffered saline solution (PBS; section 2.11). The pellets were homogenised in 3 volumes of cell lysis buffer (section 2.11) to release the nuclei, which were pelleted by centrifugation. Nuclei were resuspended in nuclear extraction buffer (section 2.11), mixed well and incubated on ice for 30 min. Following centrifugation at 25, 000 xg 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 and 260 nm. DTT and PMSF were added to a final concentration of 0.5mM in all solutions prior to use.

2.7.6 Protein binding

8µg nuclear protein extract was pre-incubated in 10µl total volume of protein binding buffer (section 2.11) containing 2µg poly[d(I-C)] (Roche) and in some cases 100pmols of unlabelled competitor probe. Following 15 minutes pre-incubation on ice, 10fmol of ^{32}P labelled oligonucleotide was added and incubation on ice was continued for a further 35 minutes. Following incubation, 3µl EMSA loading buffer was added to each sample and the total volume was electrophoresed on a polyacrylamide gel.

2.7.7 Polyacrylamide gel electrophoresis

Samples were separated on a 0.5x TBE (section 2.11), 5% (v/v) 29:1 acrylamide: bisacrylamide gel. The gel was pre-run for 1 hour and the samples then loaded and run at 200v, 20mA for 1 hour 45 minutes at room temperature. After electrophoresis the gels were vacuum dried for 2 hours at 80°C and exposed to Fuji Super HR-E 30 x-ray film for a minimum of 24 hours.

2.8 Statistical Methods

2.8.1 Deviations from Hardy-Weinberg Equilibrium (HWE)

HWE defines the expected proportions of inferred genotypes with respect to observed allele frequencies in a randomly mating population, and is given by the formula:

$$p^2 + 2pq + q^2 = 1$$

where p is the major allele frequency and $q = 1-p$. Deviations from Hardy-Weinberg can be observed due to non-random mating, selection, population stratification, or a technical issue such as allele drop out, and provides a simple procedure that checks for technical errors or sampling problems that may occur in small samples. All markers genotyped for this thesis were tested for deviation from HWE using the Arlequin software (Excoffier *et al.*, (2005); freely available on the internet; section 2.10).

Arlequin uses an adaptation of the Fishers Exact test to evaluate departures from equilibrium, as suggested by (Guo and Thompson, 1992). In this method a contingency table for an autosomal locus with m alleles is built (A_1, $A_2 \ldots A_m$), where the m x m entries in the table are the observed genotype frequencies. A markov-chain random walk algorithm is then used to explore possible tables with identical margins to the observed. The p value of the test is the proportion of visited tables that have a probability less than or equal to the observed. Standard error is calculated using a system of batches from which mean and variance are calculated.

2.8.2 Fishers Exact Test

Fishers Exact Test was used to test for the significance of associations between phenotype and marker loci. The test involves creating a 2x2 contingency table of observed binary outcome data (in this case lactase persistent/non persistent and either allele at the marker locus). The null hypothesis is that there is no association between row and column classifications, and a table of expected cell frequencies with the same row and column totals as the observed is generated. Fishers test calculates the exact probability, under the null hypothesis, of obtaining the observed data plus all other tables that deviate more from the expected by chance. Two-tailed p values for the Fishers Exact Test were calculated using the statistical analysis software on the Graphpad website (section 2.10). If $p \leq 0.05$ the null hypothesis of independence was rejected.

2.8.3 Haplotype inference

When using unrelated individuals for population studies, pedigree information is not available; therefore alternative methods of haplotype reconstruction must be employed. Linkage phase can be established directly through molecular haplotyping, or inferred using statistical methods employed by computer simulations. Two methods were used to deduce haplotypes in these studies.

2.8.3.1 A Bayesian method of haplotype reconstruction

Phase (Stephens *et al.*, 2001) is a Bayesian method of haplotype reconstruction, which aims to evaluate the conditional distribution of the unknown haplotypes given the observed genotype data. The method also takes into account the idea that the next haplotype is likely to look the same, or similar to a haplotype previously observed in the sample set.

The program uses Gibbs sampling, a type of Markov chain-Monte Carlo (McMC) algorithm to obtain an approximate sample of the posterior distribution of haplotypes given the genotype data. The algorithm resolves all unambiguous haplotypes and initialises by taking a guess at the haplotype frequencies of the sample population, H0. It then selects (at random) an individual from the pool of ambiguous individuals and estimates the probability of that persons haplotypes under the assumption that all other haplotypes are correctly reconstructed, thus giving a new haplotype reconstruction, H1. This process is repeated until the Markov chain constructed (H0, H1, H2 , ...,) gives stable estimated haplotype frequencies. A point estimate of H (haplotype distribution) is obtained from the mean of the empirical haplotype frequencies, together with an estimate of the probability of the phase calls for each individual.

The case/control feature of PHASE was used to compare haplotype distribution between groups. The program performs a permutation test for significant differences in haplotype frequencies between 'case' and 'control' groups. It tests the null hypothesis that both case and control haplotypes are a random sample from a single set of haplotype

frequencies, versus the alternative that cases are more similar to other cases than to controls

2.8.3.2 The Expectation-Maximisation (EM) algorithm

PL-EM (Qin *et al.*, 2002) uses the expectation maximisation algorithm to estimate haplotype frequencies. Only doubly heterozygous individuals have an ambiguous haplotype assignment (see Table 2.6), and the EM algorithm is used to assign haplotypes for these individuals. This is an iterative process that begins by calculating haplotype proportions of doubly heterozygous individuals from the gene frequencies assuming complete linkage equilibrium. The haplotype frequencies in the ambiguous group are compared with the frequencies obtained in the unambiguous groups through a likelihood function which examines 'goodness of fit'. Haplotype frequencies for the ambiguous group are re-calculated until the change in likelihood used to assess the goodness of fit is negligible. PL-EM gives for each subject a list of the haplotype pairs predicted and confidence probabilities associated with the phase call.

Genotype	AA	Aa	aa
BB	ABAB	ABaB	aBaB
Bb	ABAb	ABab/aBAb	aBab
bb	AbAb	Abab	abab

Table 2.6 Table depicting possible haplotypes of individuals genotyped at two loci. All genotype combinations apart from AaBb can be resolved into unambiguous haplotypes. The two possible haplotype assignments for doubly heterozygous individuals are shown in the centre square.

2.8.4 Linkage Disequilibrium

DnaSP was used to estimate the degree of linkage disequilibrium (or nonrandom association between variants of different polymorphic sites) using the D' parameter (Lewontin, 1964). D is calculated by subtracting the expected frequency of a two-locus haplotype (based on allele frequencies) from the observed frequency of the haplotype (i.e. for two loci **A** (alleles A and a) and **B** (alleles B and b) the observed frequency of the haplotype AB is given by P_{AB}. Under linkage equilibrium, the expected frequency is

the product of the observed allele frequencies A and B, P_A x P_B. Therefore D is given by $D = P_{AB} - P_A \times P_B$. Linkage disequilibrium is said to exist if D is significantly different to zero, howver, D is dependent on allele frequencies and is therefore not always comparable between loci. To improve comparability the value is normalised to give the parameter D'. The absolute value of D' is given by dividing D by its maximum possible value given the allele frequencies of the two loci. DnaSP also computes a chi-square test to determine whether the associations between polymorphic sites are, or are not, significant.

2.8.5 Exact test of population differentiation

Exact tests of population differentiation as suggested by (Raymond and Rousset, 1995) were also performed using Arlequin software. Arlequin tests the random distribution of individuals between pairs of populations by constructing contingency tables of number of populations by number of haplotypes. A markov chain random walk is used to calculate the proportion of tables with less than or equal probability than that constructed from the observed data. Default settings of 10,000 steps in the Markov chain and 1000 dememorisation steps (iterations which are not counted toward the final probability, to adjust for the chain starting with the observed data) were used. Populations are considered to be significantly different if the p value obtained is smaller than the significance level (set at 0.05).

2.8.6 F_{ST}

F_{ST} measures the apportionment of genetic diversity between sub-populations within the total population. It describes the mean amount of genetic diversity within the sub-populations to the amount of genetic diversity in the meta-population.

$$F_{ST} = (H_T - H_S)/H_T$$

where $\mathbf{H_T}$ is an estimate of the total heterozygosity of the meta-population (generated by random sampling of alleles) and $\mathbf{H_S}$ is the estimated heterozygosity within a sub-population. F_{ST} varies from 0, (when gene flow is high and there is a complete absence of variation between sub-populations) and 1 (when populations are completely differentiated).

Arelequin was used to compute pairwise F_{ST}s. To test significance of the data, Arlequin generates a null distribution of pairwise F_{ST}s by permuting haplotypes between populations. The *p* value is the proportion of permutations giving an F_{ST} value greater than or equal to the value obtained for the observed data (Excoffier *et al.*, 2005).

2.8.7 Tests of neutrality

All tests of neutrality were calculated using DnaSP software (section 2.10). Tajima's *D* (Tajima, 1989) compares two estimates of θ (the expected level of diversity for a population), the number of segregating sites, S, which does not incorporate allele frequencies, and π, which does. Under neutral evolution different estimates of θ should be equal and Tajima's *D* should be equal to zero. DnaSP calculates the confidence limits of *D* (two-tailed test) assuming that this statistic follows a beta distribution. Significantly positive values can indicate population subdivision or balancing selection and significantly negative values indicate positive selection or population growth (Jobling *et al.*, 2004).

The *D**, *F** (Fu and Li, 1993) and *FS* (Fu, 1997) statistics were calculated to test whether an excess of rare alleles was apparent within a sample. The *D** statistic measures the number of variants observed only once in a sample versus the total number of variant sites and the *F** statistic measures the number of variants observed only once in the sample versus the average number of pairwise differences between sequences (Sabeti *et al.*, 2006). DnaSP uses the critical values obtained by (Fu and Li, 1993) to determine the statistical significance of the *D** and *F** statistics.

The *FS* statistic compares the number of pairwise differences, π, with the number of observed alleles (haplotypes) (Sabeti *et al.*, 2006). DnaSP and Arlequin were both used to calculate *FS* (which each gave identical values for the statistic). The significance of the *FS* statistic is tested in Arelquin by generating random samples under the hypothesis of selective neutrality and population equilibrium, using a coalescent simulation algorithm. The *p* value of the *FS* statistic is obtained as the proportion of random *FS* statistics less than or equal to the observation. Due to properties of the distribution of

the *FS* statistic (which are not fully understood), it should only be considered as significant at the 5% level, if its *p* value is below 0.02, and not below 0.05 (Fu, 1997).

2.8.8 GenoPheno

The program GenoPheno, written by Mike Weale (Mulcare *et al.*, 2004; software available from the TCGA website, see section 2.9) was used to compare the predicted lactose digester frequency based on the occurrence of candidate causative DNA variants, with the empirical phenotypic frequencies for published matched groups.

The GenoPheno program was designed to take into account sampling and phenotyping error rates. The authors performed a survey of the literature and averaged the rates of false negative and false positive phenotypes for both the blood glucose and breath hydrogen lactose tolerance tests. A statistical procedure was then devised that tests whether the frequency of lactose digesters predicted by the *-13910C>T* genotype data was sufficient to explain the observed frequency found in the phenotyped group, taking into account the phenotyping error described above.

Briefly, 'true' lactase persistence frequency is calculated based on the genotype frequency of the candidate causal allele ($p^2+2p(1-p)$), and this value is modified accounting for phenotyping error to give the apparent frequency of digesters. A simulated value for the number of lactose digesters observed in the phenotyped group is drawn from a binomial distribution (***n***, apparent digesters), where ***n*** is the number in the phenotyped group. These steps are repeated to build up a Monte-Carlo sampling distribution for the number of apparent digesters under the null hypothesis that genotype corrected for phenotyping error alone accounts for the observed number of lactose digesters. A two-tailed *p* value is obtained from the simulated distribution.

When using this program populations were accepted as ethnic 'matches' if they were the same ethnic group living in the same country or in an immediately neighbouring country.

2.8.9 TEST_h_DIFF

Is a set of functions written by M. Weale to test for a significant difference in genetic diversity, *h* (Nei, 1987), between two populations, based on samples of haplotypes at a single locus. The software is available from the TCGA website (section 2.10).
The test looks for differences in haplotype diversity between the two populations and significance of the difference is obtained using two methods:

I) Frequentist solution: A *p* value is obtained using both a bootstrapping method and a Z-test. A conservative value for P is obtained as the solution is equal to the larger of the two values.

II) Bayesian solutions: A posterior distribution in allele frequencies is obtained for both populations, allowing for unseen alleles. In method A simulations of both populations are made from which samples can be drawn. The measure of genetic diversity *h* is calculated for the simulated data, and a two-tailed *p* value generated. In method B, one population (A) is used to simulate allele frequencies and samples of the size of population B are drawn from it. If the observed data shows a significant departure from this distribution, it indicates that population B does not have the same frequency distribution as population A.

2.9 Web resources

Arlequin	http://lgb.unige.ch/arlequin/
DnaSP	http://www.ub.es/dnasp/
Ensembl	http://www.ensembl.org/index.html
Ethnologue:	http://www.ethnologue.com.
Graphpad	http://www.graphpad.com/quickcalcs/index.cfm
TCGA:	http://www.ucl.ac.uk/tcga/software/index.html

2.10 Buffers

Solutions prepared in lab

Agarose gel loading buffer: 0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll in H_20.

Cell lysis buffer: 10mM KCl, 1.5mM $MgCl_2$,10mM HEPES (pH7.9) and 0.05% Nonidet P-40.

EMSA loading buffer: 60% (w/v) glycerol, 0.2% (w/v) bromophenol blue, 0.25x TBE (pH 8.2-8.4).

Nuclear extraction buffer: 1.5mM $MgCl_2$, 0.2mM EDTA, 5mM HEPES (pH7.9), 25% (v/v) glycerol and 300mM NaCl.

PBS (10x): 1.5M NaCl, 0.1M NaH_2PO_4, 0.075M NaOH, pH7.4.

PCR precipitation solution (1x): 27% (w/v) PEG 8000, 0.7M NaCl, 1.3mM Tris-HCL, 0.13mM EDTA, 2.3mM $MgCl_2$, pH8.0

Protein binding buffer: 20mM HEPES (pH7.6), 1mM EDTA, 10mM $(NH_4)_2SO_4$, 1mM DTT, 0.2% (v/v) Tween-20, 30mM KCl.

Sequencing buffer: 200mM Tris-HCl (pH9.0), 5mM $MgCl_2$.

Slagboom buffer: 100mM NaCl, 10mM Tris-HCl (pH8.0), 10mM EDTA (pH8.0), 0.5% SDS and 0.2mg/ml Proteinase K.

STE (10x): 100mM Tris-HCl (pH8.0), 1M NaCl, 10mM EDTA, pH8.0.

TBE (5x): 0.44M Tris, 0.44M Boric Acid, 12.5mM EDTA, pH8.2-8.4.

TE: 10mM Tris-HCl (pH7.0-8.0), 1mM EDTA.

Commercial Solutions

ABgene Buffer IV (1x): 75mM Tris-HCl (pH8.8 at 25°C), 20mM $(NH_4)_2SO_4$, 0.01% (v/v) Tween 20, 1.5mM $MgCl_2$.

Abgene Buffer I (1x): 10mM Tris-HCl (pH8.3 at 25°C), 50mM KCl, 1.5mM $MgCl_2$

NEBuffer 1 (1x): 10mM Bis Tris Propane-HCl, 10mM $MgCl_2$, 1mM dithiothreitol (pH7.0 at 25°C)

NEBuffer 2 (1x): 10mM Tris-HCl, 10mM $MgCl_2$, 50mM NaCl, 1mM dithiothreitol (pH7.9 at 25°C)

NEBuffer 3 (1x:): 50mM Tris-HCl, 10mM $MgCl_2$, 100mM NaCl, 1mM dithiothreitol (pH7.9 at 25°C)

NEBuffer 4 (1x): 20mM Tris-acetate, 10mM magnesium acetate, 50mM potassium acetate, 1mM dithiothreitol (pH7.9 at 25°C)

Roche Phsophorylation buffer (10x): 500mM Tris-HCl, 100mM $MgCl_2$, 0.1mM EDTA, 50mM dithioerythritol, 1mM spermidine, (pH8.2 at 25°C)

AutoGen Yeast Reagent 3: (catalogue number AG00312) Potassium acetate, Phenol, Chloroform

2.11 Suppliers

Abgene Ltd: Abgene House, Blenheim Road, Epsom, KT19 9AP, United Kingdom.

Amersham Biosciences: Amersham Place, Little Chalfont, Buckinghamshire, HP7 9NA, United Kingdom.

Applied Biosystems: Lingley House, 120 Birchwood Boulevard, Warrington, WA3 7QH, United Kingdom.

Autogen Bioclear (UK) Ltd: Holly Ditch Farm, Calne, Wiltshire, SN11 0PY.

Clontech: Takara Bio Europe/Clontech, 2 Avenue du President Kennedy, 78100 Saint-Germain-en-Laye, France.

Gene Codes Corporation: 775 Technology Drive, Suite 100A, Ann Arbor, MI 48108, USA.

HT Biotech: Unit 4, 61 Ditton Wk, Cambridge, CB5 8QD, Cambridgeshire

Micro Medical Ltd: Quayside, Chatham Maritime, Chatham, Kent, ME4 4QY, United Kingdom.

NEB: New England Biolabs, 75-77 Knowl Piece, Wilbury Way, Hitchin, Herts, SG4 0TY, United Kingdom.

Sigma-Genosys Ltd: Sigma-Aldrich House, Homefield Business Park, Homefield Road, Haverhill, Suffolk, CB9 8QP, United Kingdom.

Roche: Roche Diagnostics Ltd, Charles Avenue, Burgess Hill, West Sussex, RH15 9RY United Kingdom.

Technelysium Pty Ltd: Technelysium Pty Ltd, PO Box 743, Tewantin, QLD 4565, Australia.

2.12 Equipment

Centrifuges

Heraeus Biofuge Pico

MSE-Europa 24M

MSE MISTRAL 2000

ALC PK 120

MSE MISTRAL 3000E

Other equipment

MJ Research PTC-225 Tetrad thermal cycler

MicroH_2 breath hydrogen monitor

Biorad gel dryer (model 583)

ABI 3100 Genetic Analyzer

ABI 3730xl DNA analyser

3 Lactase Persistence: A single cause or multiple origins?

3.1 Introduction

As discussed in chapter 1, in Europeans, lactase persistence is known to be controlled by a genetic trait, which is *cis*-acting to *LCT*, and is located within a region of strong linkage disequilibrium that is characterised by only three common *LCT* haplotypes, designated **A**, **B** and **C** (Harvey *et al.*, 1998; Hollox *et al.*, 2001). One of these haplotypes, **A**, forms the core of an extended (500kb+) haplotype, which is associated with lactase persistence and is at high frequency, probably due to the effect of selection for this trait (Poulter *et al.*, 2003; Bersaglieri *et al.*, 2004; Coelho *et al.*, 2005).

The single SNP (*-13910*T*) identified as potentially causal (Enattah *et al.*, 2002) occurs on the **A** haplotype 13.9kb upstream from *LCT*, and is located within intron 13 of the adjacent gene, *MCM6* (Figure 3.1).

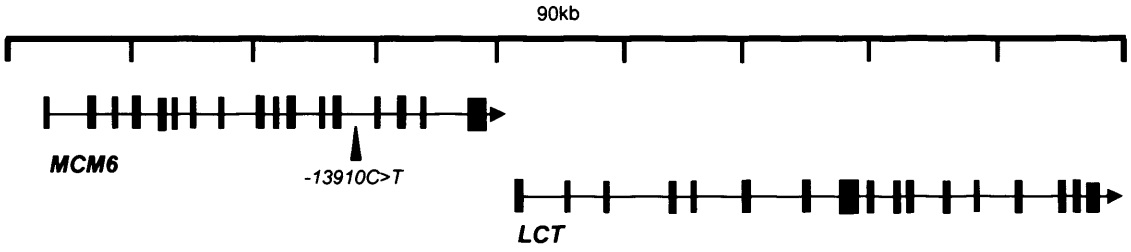


Figure 3.1 Diagrammatic representation of *MCM6* and *LCT*. An arrow indicates the location of *-13910*T*, the polymorphism thought to cause lactase persistence in Europeans.

The absence of *-13910*T*, in several groups in sub-Saharan Africa in which the lactase persistence trait is common (Mulcare *et al.*, 2004) suggested that either there is more than one cause of lactase persistence, with a different allele occurring in some parts of Africa, or that *-13910*T* is not in fact functional in relation to lactase persistence *in vivo*, and that the true causal mutation is located elsewhere on the extended **A** haplotype. Other research groups refuted the existence of a discrepancy between *-13910*T* and lactase persistence frequencies in African populations (Rasinpera *et al.*, 2004) and the suggestion of alternative causes of lactase persistence caused some controversy (Kolho and Jarvela, 2006; Swallow, 2006; Weale, 2006).

Resolving the debate regarding the nature of *–13910*T* with respect to lactase persistence is a crucial first step toward progressing our understanding of the evolution of the trait. In this chapter, the occurrence or not of *–13910*T* and the **A** haplotype, and their pattern of association with lactase persistence is examined in a cohort of phenotyped Sudanese individuals and in pastoralist groups from east Africa and also the Middle East.

Genotyping of *–13910*T* and other SNPs which discriminate between the common *LCT* haplotypes was first carried out in selected populations for whom lactase persistence frequency was well documented, and the program GenoPheno (see chapter 2.9.5) was used to analyse whether observed frequencies of either *–13910*T* or the **A** haplotype were able to account for lactase persistence in the non-phenotyped groups. The same SNPs were then genotyped in the phenotyped Sudanese cohort in which association between SNPs, haplotypes and lactase persistence was tested directly.

3.2 Population selection

Five populations from the Middle East and one from east Africa were selected because of the close proximity and long history of migrations between these regions. Three different Bedouin groups were included because of the known high incidence of lactase persistence in the Bedouin (Cook and al-Torki, 1975; Hijazi *et al.*, 1983; Dissanyake *et al.*, 1990) and two neighbouring urban Arab populations in which lactase persistence is reported to be low (Snook *et al.*, 1976; Hijazi *et al.*, 1983) were also included for comparison. A population from northern Sudan, the Beni Amer, previously defined in the literature as having very high frequency of lactase persistence (0.87) (Bayoumi *et al.*, 1982) were also included. One population from Cameroon, the Fulani, were included as they had previously been shown to be one of the few sub-Saharan African populations in which *–13910*T* was able to account for lactase persistence (Mulcare *et al.*, 2004). A new sample group of nomadic Fulani (collected in different, more northern locations in Cameroon) was collected for this study, as the nomadic Fulani are reported to have a higher frequency of lactase persistence than the sedentary urban Fulani (Kretchmer *et al.*, 1971).

A cohort of Sudanese samples with associated phenotype data was also collected for this investigation. Ninety-nine individuals who were self identified as belonging to the Jaali ethnic group were lactose tolerance tested using the breath hydrogen method. This population was specifically selected for this study, as they had been previously reported to have an intermediate lactose tolerance frequency of 0.53 (Bayoumi *et al.*, 1981), giving the study good power.

3.3 Marker selection

LCT gene markers were selected by consideration of the haplotypes which occurred in South African San & Bantu-speaking populations at a frequency of ≥0.05 (n = 7) (Hollox *et al.*, 2001). Four SNPs (*-942/3TC>ΔΔ*, *-678A>G*, *666G>A* and *5579T>C*) were chosen that gave maximum discrimination between the haplogroups present in the African populations. This approach to SNP selection was subsequently complimented by entering the SNP frequency data extracted from Hollox *et al.*, (2001) into a web-based SNP tagger program (Ke and Cardon, 2003), which gave a similar selection of markers. Figure 3.2 shows the location of these SNPs in relation to *LCT* and the combination of the four alleles in each distinguishable haplotype.

In their study of Old world populations, (Hollox *et al.*, 2001) defined 42 haplotypes by typing 11 SNPs. Only four of these SNPs were genotyped in this study, and so only a subset of the original diversity identified has been captured. Table 3.1 shows 'haplogroups' (subsequently referred to within this thesis as 'haplotypes') of haplotypes that cannot be distinguished from one another without genotyping further markers. The defining name of each haplogroup represents the haplotype within that group that was most common globally in the original study, though not necessarily the most common in Africa (Hollox *et al.*, 2001). A single exception is haplogroup **K**, in which haplotype **Q** is more frequent in the original report. The **K** haplotype was selected to represent the haplogroup in this case as it is the ancestral allele (from comparison with primates) and the root of the haplotype network (Hollox *et al.*, 2001). Figure 3.2 depicts the allelic compositions of the Hollox haplotypes and their observed frequencies in the African groups included in the study.

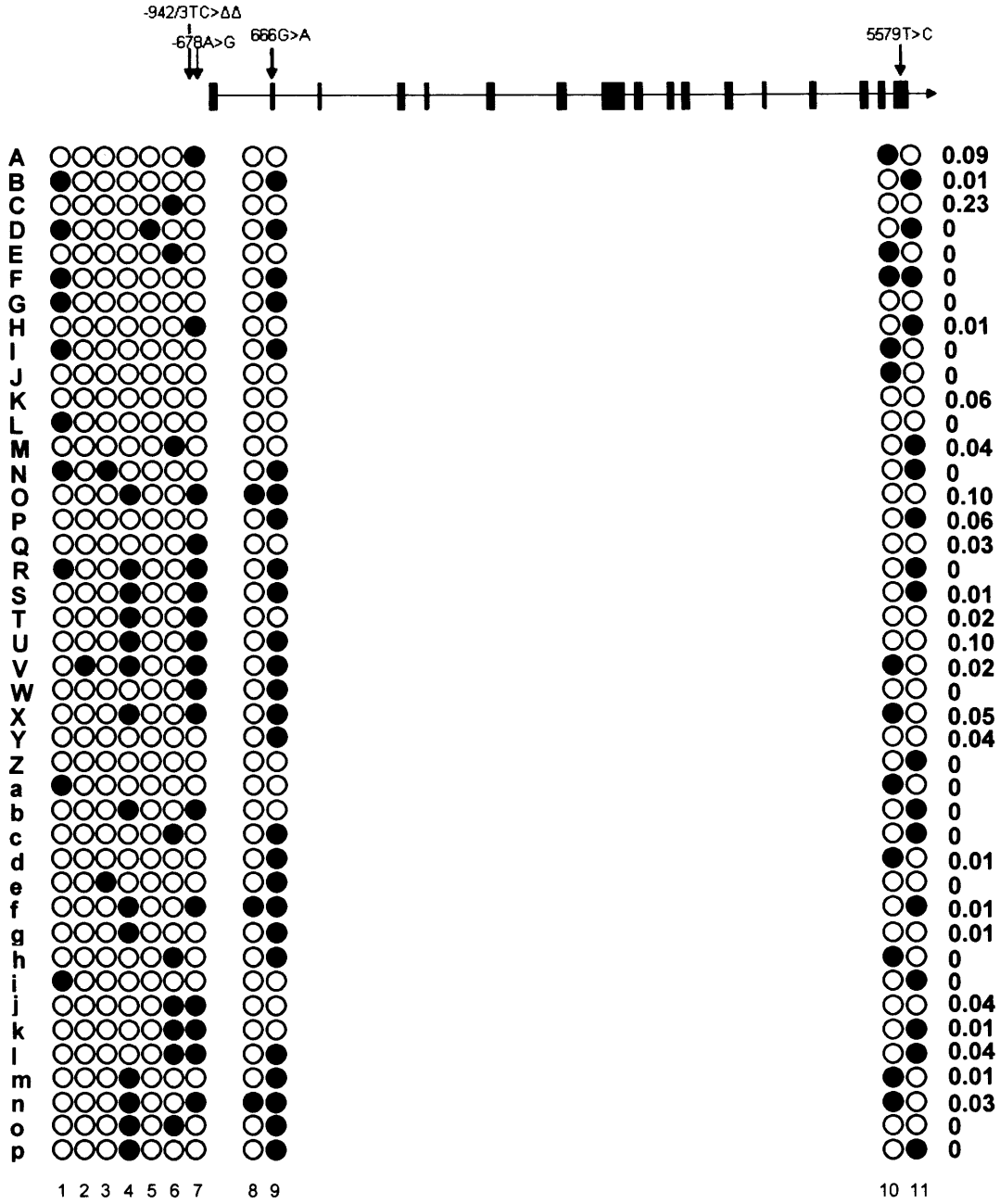


Figure 3.2 Location of SNPs selected for haplotype analysis depicted by arrows on the diagram of the *LCT* gene. Underneath are the possible allelic combinations of the 11 SNP haplotypes described in Hollox et al., 2001. The open circles indicate an ancestral allele and filled circles denote the derived allele at a locus. The circles with red outline (4, 6, 9 & 10) are the ones genotyped in this study (*-942/3TC>ΔΔ*, *-678A>G*, *666G>A* and *5579T>C* respectively). Haplotypes are named as they were in the original study and the given frequency of the haplotypes was obtained by pooling the African (San and Bantu) samples included in Hollox et al., (2001). Other markers (named according to Hollox et al 2001) are:1. *-958C>T*; 2. *-946A>G*; 3. *-942C>G*; 5. *-875G>A*; 7. -*552/559A$_8$A$_9$*; 8. *458CintT* and 11. *6236/7TG> ΔΔ*.

Haplogroup Name	Undiscriminated haplotypes within haplogroup
A	A, J, a
B	B, D, G, N, P, W, Y, e
C	C, M, j, k, l
E	E
F	F, I, d
h	h
K	K, H, L, Q, Z, i
U	U, O, R, S, f, g and p
T	T, b
V	V, m, n
X	X
c	c
o	o

Table 3.1 Haplotype names used in this thesis are given in the left hand column entitled 'Haplogroup name'. The right hand column lists the haplotypes which cannot be discriminated from one another by typing only the four selected markers (*-942/3TC>ΔΔ, -678A>G, 666G>A* and *5579T>C*).

3.4 Genotyping and Haplotype inference

-942/3TC>ΔΔ was genotyped using a tetra-primer ARMS PCR method (see Figure 3.3 and section 2.6.3), and *–13910C>T, -678A>*G, *666G>A & 5579T>C* were genotyped by PCR-RFLP (section 2.6.3).

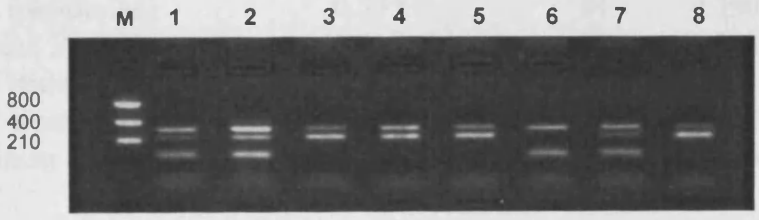


Figure 3.3 Example of electrophoresed *-942/3TC>ΔΔ* tetra-primer ARMS PCR products. The slowest mobility band (present in all sample lanes, 1-8) is the control PCR product. The intermediate mobility band (in all lanes except 6) is produced by amplification of the TC allele, and the lowest molecular weight band (present in lanes 1, 2, 6 & 7) is produced by amplification of the ΔΔ allele. M indicates the marker lane, with the molecular weight (in bp) indicated to the left of the gel image.

Haplotypes were inferred (either with or without *-13910*T* data) for each population individually using the computer programs PL-EM (Qin *et al.*, 2002) and Phase (Stephens *et al.*, 2001), which use an EM algorithm or a Bayesian approach respectively (see section 2.9.3 for a detailed description of both methods). The out-put phase calls of each program were compared for each individual and found to be identical.

3.5 Association of *-13910*T* with lactase persistence in east Africa and the Middle East.

In all the east African and Middle Eastern populations tested here, *-13910*T* was shown to be very rare, contrasting with the frequency (0.39) observed in the pastoralist Fulani group from Cameroon (Table 3.2). The computer program GenoPheno was used to test whether the discrepancy between the observed frequency of *-13910*T* and the expected frequency of the causal allele (calculated from reported lactose digester frequency in a matched group) were significantly different. In agreement with the previous findings of Mulcare *et al.*, (2004) the *-13910*T* allele fails to explain lactase persistence frequency in any of the populations tested with an inferred lactase persistence allele frequency above 0.2, apart from the Fulani (max p = 1.0x10-4, see Table 3.2).

Population	Reported Lac*P allele frequency of matched group*	n	*-13910*T* frequency	*p* value (GenoPheno)
Beni Amir	0.641	100	0.005	<0.0001
Saudi Bedouin	0.592	56	0.000	<0.0001
Fulani (pastoralist)	0.533	91	0.390	0.4828
Jordanian Bedouin	0.515	26	0.058	<0.0001
Israeli Bedouin	0.515	19	0.026	<0.0001
Israeli urban Arab	0.138	83	0.024	<0.0001
Palestinian Arab	0.138	19	0.026	0.8110

Table 3.2 Frequency of *-13910*T* in 7 new population groups in comparison with lactase persistence allele frequency *(LAC*P)* calculated from published lactose digester frequency ($p^2 + 2pq$). *Numbers taken from Holden & Mace (1997) who extracted adult data

3.6 Association of *–13910*T* and lactase persistence in a phenotyped Sudanese cohort

Lactose tolerance test data is available in appendix B. In all members of the cohort, the starting breath hydrogen was between 2 and 18ppm. Forty-five people were defined unambiguously as lactose digesters, with a rise in breath hydrogen of less than 20ppm (max. observed rise was 11ppm observed over 2 hours), and 49 individuals were clearly lactose mal-digesters, showing a sustained rise of greater than 20ppm within 2 hours of ingestion of the lactose load. Five individuals were considered intermediate as breath hydrogen fluctuated during the course of the test or did not show sufficient rise before the test was terminated. The researcher collecting these samples found it virtually impossible to detain volunteers for longer than 2 hours, and non-persistent individuals were allowed to leave after their status became evident in the breath hydrogen readings.

During the lactose tolerance test all volunteers completed a milk-consumption questionnaire. Lactose digester status and milk consumption behaviour were highly associated (p = 5.84x10-6, Table 3.3).

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

Table 3.3 Contingency table showing numbers of lactose digester and non-digester people reporting consumption of at least 500 mls milk per day. Five individuals were excluded from the analysis due to ambiguous lactose tolerance test results. (Note that the numbers are slightly larger than in subsequent tables because individuals are included from whom no (or poor quality) DNA was obtained.

Direct confirmation of the finding that *–13910*T* is not responsible for causing lactase persistence was obtained by genotyping the SNP in the Sudanese cohort. Of 94 individuals for whom clear lactose tolerance test results were obtained, 48% were

lactose digesters, in close agreement with the previously published frequency. *-13910*T* was found in only one person in the cohort, who was a lactose digester (presumed lactase persistent), clearly showing that in this population *–13910*T* cannot be the sole causal variant.

3.7 The A haplotype and lactase persistence in east Africa and the Middle East.

Since the findings in sections 3.5 & 3.6 provide conclusive evidence that *–13910*T* is not the major causal variant in the populations under investigation, the subsequent objective was to discriminate between two scenarios: a) that in Europe, *–13910*T* is only a highly associated marker of the true cause of lactase persistence, or b) that there are multiple evolutionary origins of the trait. To address this question, the association of the **A** haplotype (upon which *-13910*T* occurs) with lactase persistence was examined in the non-European groups.

3.7.1 Haplotype background of *–13910*T* in Africa

In the Fulani, as in Europeans, *-13910*T* is strongly associated with the **A** haplotype. Of 150 chromosomes, 50 carried the *-13910*T* allele and only 3 of these were found in individuals whose genotypes at other loci could not be resolved into an **A** haplotype. One of the *-13910*T* alleles was confidently identified on a **K** haplotype background since the individual is heterozygous only for *-13910*T*. The two other non-**A** haplotype *-13910*T* carrying individuals had inferred diplotypes of **B/K** and **B/h** (assignment of *-13910*T* to either one chromosome or the other would be arbitrary with such little information). Two Sudanese and four Middle Eastern individuals were also found to carry *-13910*T,* all of whose genotypes could be resolved into an **A** haplotype. These results are consistent with the findings of Myles *et al.*, (2005) who also found a low percentage of *-13910*T* alleles on non-**A** haplotypes in North African Berber populations in which the allele is frequent.

3.7.2 Haplotype distribution in east African and Middle Eastern populations

In east Africa and the Middle East the core *LCT* haplotypes show considerable inter group difference in frequency. The **A** haplotype is widespread, but is not at all correlated in frequency with published lactase persistence allele frequencies. The **C** haplotype is noted to be most frequent in three of the four other groups in which lactase persistence is reported to be high, and is sufficiently frequent to account for lactase persistence in the Saudi Bedouin population. Figure 3.4 shows the haplotype distribution in the pastoralist populations. Neighbouring urban populations are not included as the frequency of the lactase persistence allele is negligible, but haplotype distributions were not significantly different from the pastoralist groups (p = >0.05, Exact test of population differentiation; section 2.9.4). Fulani are included for comparison.

3.7.3 Haplotype association in the phenotyped Sudanese cohort

None of the 'core' *LCT* SNPs showed significant association with lactose digestion status (minimum p = 0.12, for Fishers Exact tests of 2x2 tables of allele counts) in the phenotyped cohort. Core haplotype distributions are shown in Figure 3.5. Although the **C** haplotype is more frequent in the lactose digester group, the difference in core haplotype distribution between the two groups was not significant, (p = 0.73, PHASE, case-control comparison) and this haplotype was not present in all lactase persistent individuals.

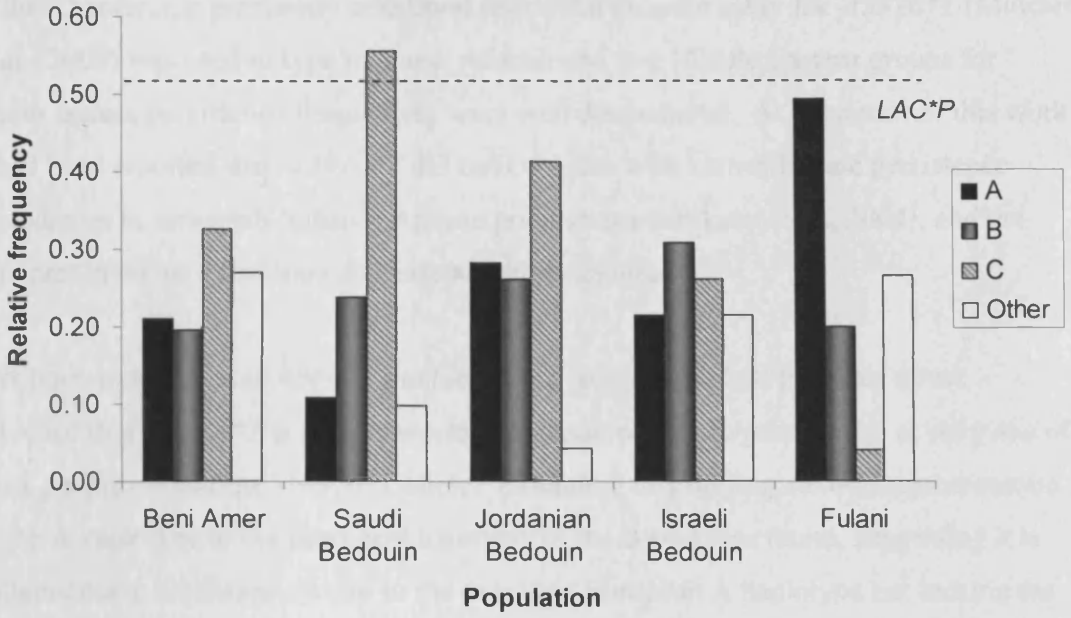


Figure 3.4 Core *LCT* haplotype frequencies in Middle Eastern & African pastoralist populations. Lactase persistence allele frequency *LAC*P*, is indicated at 0.51 – the lowest observed frequency in any of these populations (see Table 3.2).

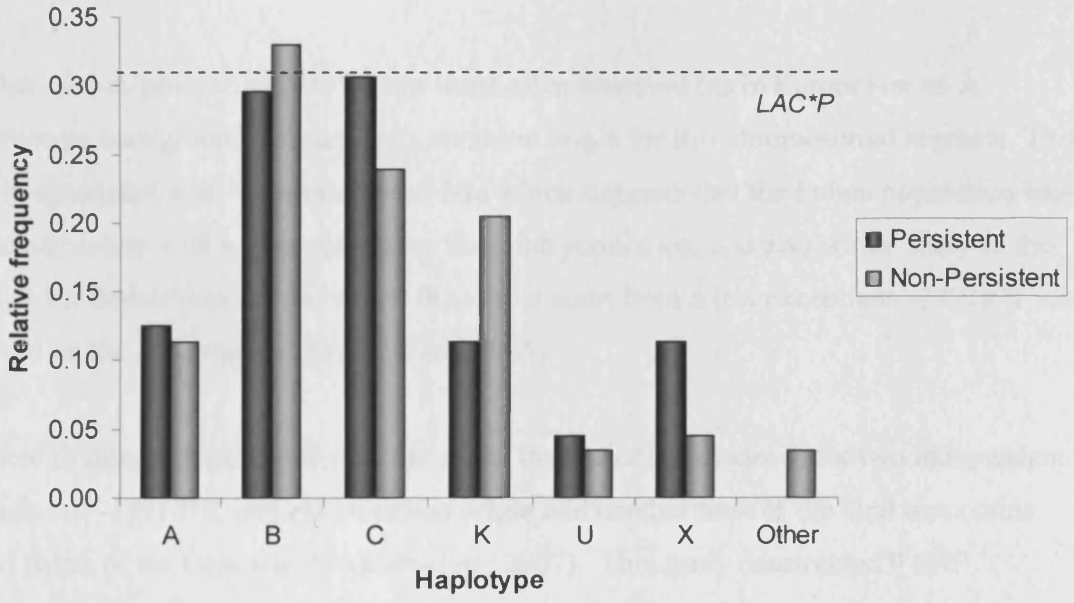


Figure 3.5 Comparison of core *LCT* haplotype distribution in persistent and non-persistent groups of the Sudanese Jaali cohort. Lactase persistence allele frequency *LAC*P* (calculated from lactase digester frequency, $p^2 + 2pq$), is indicated (0.31).

3.8 Discussion

In this chapter, the previously published restriction enzyme assay for *-13910*T* (Mulcare *et al.*, 2004) was used to type two new African and five Middle Eastern groups for whom lactase persistence frequencies were well documented. At the outset of this work it had been reported that *-13910*T* did not correlate with known lactase persistence frequencies in some sub-Saharan African populations (Mulcare *et al.*, 2004), and the data presented here confirms and extends those findings.

A cohort study in which 45/94 volunteers were lactose digesters provides direct evidence that *-13910*T* is not the worldwide cause of lactase persistence, as only one of these people carried the *-13910*T* allele. Extending this finding, no overrepresentation of the **A** haplotype in the persistent members of the cohort was found, suggesting it is unlikely that a haplotype similar to the extended European **A** haplotype but lacking the marker *-13910*T*, carries a single common cause of lactase persistence. In agreement, it is shown that neither the *-13910*T* allele nor the **A** haplotype are frequent enough to account for lactase persistence in other east African and Middle Eastern milk drinking groups.

Where it was present, *-13910*T* was most often observed (as in Europe) on an **A** haplotype background, suggesting a common origin for this chromosomal segment. This is in agreement with Y-chromosome data which suggests that the Fulani population has had admixture with a back-migrating Eurasian population, and also with a study of the Algerian Berber population, which found that apart from a few exceptions *-13910*T* was found on the **A** haplotype (Myles *et al.*, 2005).

These findings contrast with a recent paper that describes evidence for two independent origins of *-13910*T*, one of Caucasian origin and another west of the Ural mountains and North of the Caucasus (Enattah *et al.*, 2007). This study constructed 9 SNP haplotypes from intron 6 of *MCM6* to intron 1 of *LCT*, but did not cover the entire *LCT* gene region. Out of the 830 *-13910*T* carrying chromosomes tested, only 47 showed evidence of an independent origin of *-13910*T*, however 17 of these non-**A** haplotypes also carried *-22018*A*, known to be tightly associated with *-13910*T* in the 'European'

A haplotype. It would seem unlikely that two independent lineages gave rise to this combination twice and more probable that haplotype inference was incorrect.

Haplotype analysis showed no statistically significant difference in core *LCT* haplotype distribution between the lactose digesters and mal-digesters in the Jaali cohort, although it was noted that the **C** haplotype was slightly more frequent in the digesters than in the lactose non-digesters. These data point to another cause for lactase persistence in Africa and the Middle East, and even to the possibility that this may be trans-acting to *LCT*, although the lower *p* value (0.12) obtained for one SNP with lactase persistence (*-942/3TC>ΔΔ*) in a Fishers Exact test suggests a *cis*-acting affect should not be ruled out.

It is possible that the lack of association between lactose digester status and the *LCT* core haplotypes in the cohort study could in part reflect the shorter haplotype blocks seen in Africans (Gabriel *et al.*, 2002) or the somewhat reduced power resulting from errors in phenotyping (discussed in section 1.2.8). It proved very difficult to retain individuals for the entire three-hour test duration, and the researcher in charge of sample collection elected to reduce the test duration to two-hours, which still proved to be difficult in some cases. This modification is likely to have increased the error rate associated with breath hydrogen, with more false negatives (i.e. individuals incorrectly diagnosed as lactose tolerant), the most likely effect.

It can not be ruled out that a global cause of lactase persistence exists in the form of a shared **A**-haplotype block outside of the 60kb region tested here. There still remains a small possibility that a common up or downstream element exists in all populations. However, this should be countered with the observations that the most common haplotype observed in the non-European milk-drinking populations, with the exception of the Fulani, was the (**C**) haplotype. Furthermore, the observation that one of the core *LCT* SNPs (*-942/3TC>ΔΔ*) gave a decreased *p* value for association with lactase persistence, but is absent throughout Europe (Hollox *et al.*, 2001) also lent weight to the hypothesis of independent origins of lactase persistence.

4 Identification of novel variation associated with lactase persistence in Africa.

4.1 Introduction

As seen in chapter 3, data obtained from genotyping of *–13910*T* and analysis of core *LCT* haplotypes in east African and Middle Eastern populations provides strong evidence of an independent evolutionary origin of lactase persistence in sub-Saharan Africa. However, there were slight indications from haplotype distribution and association of marker loci that, in common with the European variation, the causal change in African populations may be *cis*-acting to the *LCT* gene. The aim of the work described in this chapter is to use the phenotyped Sudanese cohort to further investigate this possibility by resequencing a candidate functional region.

The region immediately surrounding *-13910C>T* was identified as the primary target for resequencing for two reasons;

- No association was observed between lactase persistence and *LCT* haplotype in the Jaali cohort, thus it was concluded that causal variation was most likely to reside outside the 60kb gene region
- *In vitro* evidence for the regulatory nature of *-13910C>T* with respect to *LCT* and the demonstration that even the ancestral sequence had enhancer function (Olds and Sibley, 2003; Troelsen *et al.*, 2003) made it a natural candidate.

4.2 Sequencing strategy

The original panel for resequencing consisted of eight samples (persistence status indicated following sample ID, P = persistent, NP = non-persistent): ATA (P), SD-J-01 (P), SD-J-02 (NP), SD-J-05 (P), SD-J-10 (NP), SD-J-12 (P), SD-J-16 (P), SD-J-46 (NP). The ATA sample was provided by an Ethiopian collaborator of Amharic extraction whose lactase persistence status had been verified by the breath hydrogen lactose tolerance test (under the supervision of Prof. Swallow). More DNA samples from persistent individuals were selected than from non-persistent to try to ensure even representation of persistent and non- persistent chromosomes (assuming dominant

inheritance, all non-persistent individuals will have two non-persistent chromosomes, but persistent individuals may carry only one).

PCR product was amplified using the primers MCM6778/LAC-C-L2 and MCM6i13 giving a 700/400bp product respectively. Attempts were made to amplify the larger fragment in all samples, however in some cases DNA quality was too poor, and in these cases the PCR yielding the smaller fragment was performed. All PCR products were sequenced with the MCM6i13 primer (details given in chapter 2.7).

4.3 Identification of novel variants

The original screen revealed two novel nucleotide variants. Samples SD-J-01, SD-J-05, and SD-J-12 were all T/G heterozygotes at the same nucleotide (-13915), 5bp upstream of *-13910C>T*. ATA was heterozygous C/G 3bp downstream (-13907). These individuals were all lactase persistent.

The entire Jaali cohort was therefore sequenced across the *MCM6* intron 13 enhancer region, and several of these, as well as other variant alleles were revealed. Figure 4.1 shows example chromatograms revealing the novel alleles located close to *-13910C>T*.

4.4 Evidence of association with lactase persistence

The most common of these new sequence variants *-13915*G*, is present at a frequency of 0.14 in the Jaali. Genotypic data for this SNP in both categories of lactase persistence status are shown in Table 4.1. *-13915*G* shows a significant association with lactose digester status ($p = 6.05 \times 10^{-3}$, for a Fishers Exact test of a 2x2 table of allele counts). However, the association is not 100%, and there are discrepancies in both directions.

Five lactose non-digester individuals carry the *-13915*G* allele, including one homozygote. None of these five reported any recent gastro-intestinal complaints when questioned, and four of the five reported drinking less than 500ml milk per day. Furthermore, the *-13915*G* allele was not frequent enough to explain all the lactase persistence observed in this population; 23/39 lactose digesters did not carry the allele

(Table 4.1). *-13910*T*, *-13913*C* and *-13907*G* were each present in a single lactose digester, leaving 20 with the ancestral allele in this region. 12 of these 20 reported drinking 0.5 litres or more of fresh milk per day.

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

Table 4.1 Contingency table showing numbers of each *-13915*G* genotype in the lactase digester and lactose non-digester categories. Data obtained by sequencing. Five genotyped individuals were excluded from the analysis due to ambiguous lactose tolerance test results

4.5 Distribution of the new alleles in non-phenotyped pastoralist groups.

The region around *-13910*T* was next sequenced in a further 434 individuals from various Middle Eastern and African groups, including pastoralists and non-pastoralists from each location, in order to better understand the distribution of the new alleles (Table 4.2). Of the populations tested, the *-13915*G* allele was found to be fairly widespread in eastern Africa and the Middle East, while *-13907*G* and *-13913*C* had more limited distributions. *-13915*G* was most common in the Saudi Bedouins and appeared to be more common in the milk-drinking pastoralists, though the frequencies are in most cases significantly lower than those published for lactose digester frequency (established using GenoPheno, see chapter 2.8.8). This is also the case if it is assumed that any variant allele at this locus is causative of lactase persistence (Table 4.2). Only one individual was found to carry two different variant alleles in the -13.9kb region. The majority of the Beni Amir, Afar and Somali tested reported drinking more than 500ml fresh milk per day and there was no association of milk drinking with carrier status for one, or any one, of the variant alleles (Fishers exact test, data not shown).

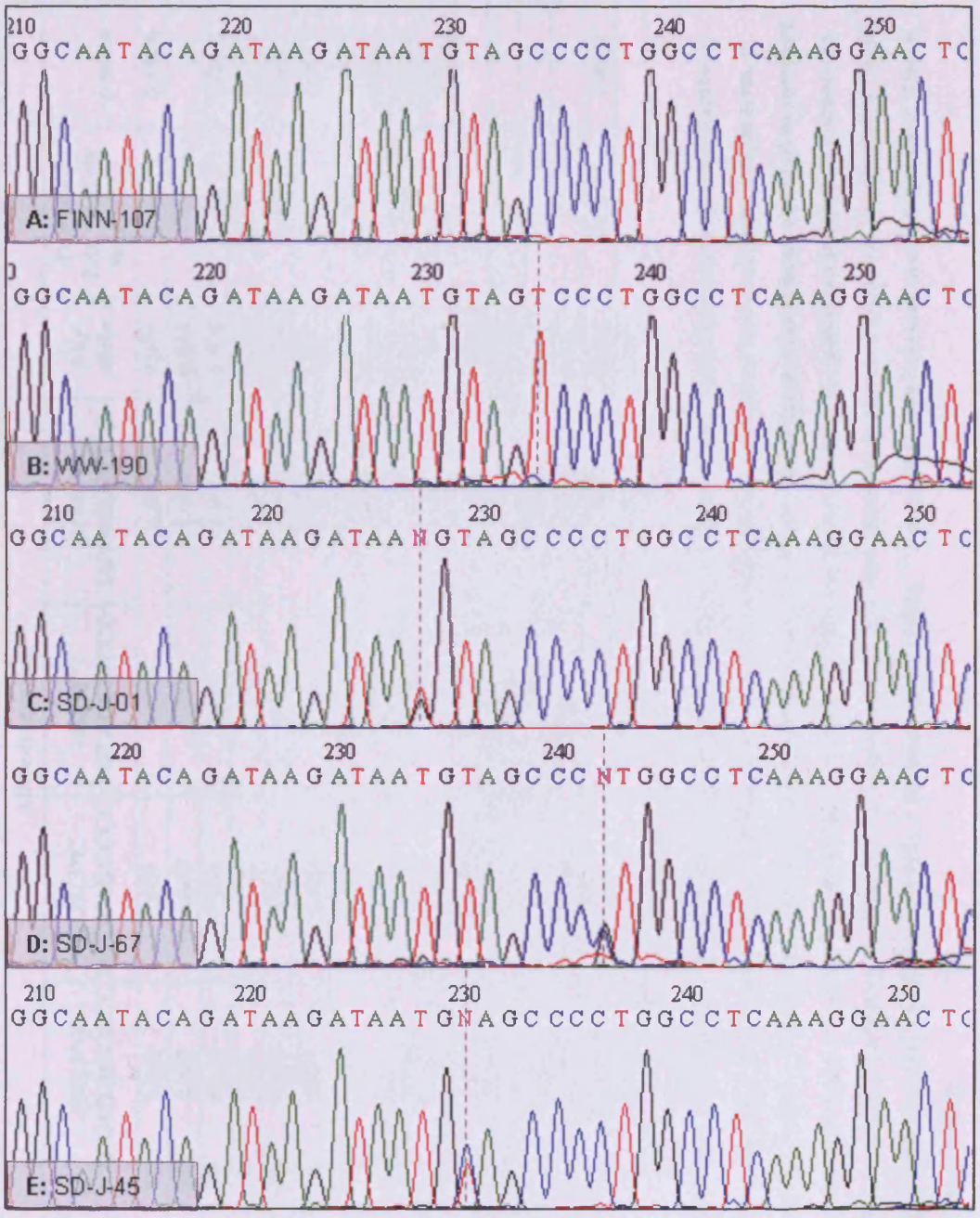


Figure 4.1 Example sequencing chromatograms of individuals carrying variant alleles within *MCM6* intron 13. Chromatograms A & B are from individuals of European descent and show homozygotes for the ancestral allele (*-13910*C*) and *-13910*T* respectively. Chromatograms C, D & E show sequences from three Sudanese individuals heterozygous at positions *-13915T>G*, *-13907C>G* and *-13913T>C* respectively.

				Allele frequency						
Country	**Population group**	**Pastoralist /Milk drinkers [3]**	**N**	***-13915*G*** **T/GGTAGCCCC**	***-13913*C*** **TGT/CAGCCCC**	***-13910*T*** **TGTAGC/TCCC**	***-13907*G*** **TGTAGCCCC/G**	**Any allele**	**Published *LAC*P* allele frequency**	***p* value**
Israel	Urban Arabs	N	81	0.049	0.000	0.025	0.000	0.074	0.13[1]	**0.0030**
Israel	Druze	N	14	**0.107**	0.000	0.036	0.000	**0.143**	-	-
Israel	Bedouin	Y	19	**0.132**	0.000	0.026	0.000	**0.158**	0.51[1]	**0.0001**
Israel	Palestinians	N	18	0.000	0.000	0.028	0.000	0.028	0.13[1]	0.2791
Saudi Arabia	Bedouin	Y	46	**0.489**	0.011	0.000	0.000	**0.500**	0.59[1]	0.2507
Jordan	Bedouin	Y	23	**0.391**	0.000	0.065	0.000	**0.457**	0.51[1]	0.4073
Sudan	Beni Amir[$]	Y	82	**0.244**	0.000	0.006	0.061	**0.311**	0.64[1]	**0.0001**
Sudan	Shaigi	N	9	0.056	0.000	0.000	0.000	0.056	0.21[2]	0.2805
Sudan	Donglawi	N	6	0.000	0.000	0.000	0.083	0.083	0.10[2]	0.7670
Sudan	Jaali [$]	N	88	**0.142**	0.006	0.006	0.006	**0.159**	0.31[2]	**0.0032**
Ethiopia	Amharic	N	19	**0.132**	0.000	0.000	0.053	**0.184**	-	-
Ethiopia	Afar[$]	Y	10	**0.150**	0.000	0.000	**0.200**	**0.350**	-	-
Ethiopia	Somali [$]	Y	9	0.056	0.056	0.000	0.056	**0.167**	-	-
Cameroon	Mambila	N	38	0.000	0.000	0.000	0.000	0.000	-	-
Cameroon	Fulani [$]	Y	63	0.000	0.024	**0.389**	0.000	**0.413**	0.53[1]	0.6237
Cameroon	Shuwa Arabs	Y	16	0.063	0.000	0.000	0.000	0.063	-	-

Table 4.2 Allele frequencies of new *MCM6* intron 13 polymorphisms. Allele frequencies above 0.10 are shown in bold. No deviation from Hardy-Weinberg equilibrium was observed in any of the populations. Lactase persistence allele frequency (*LAC*P*) calculated from published lactose digester frequencies, is shown for matched groups for which published data are available. References (superscripts) are given below. The program GenoPheno was used to compare the expected lactose digester frequency, if carrying any one of the non-ancestral variants at -13.9kb was causative, with the published (observed) phenotypic frequencies for matched populations. In four cases there were significantly fewer expected digesters than observed in matched groups in the literature (bold). [$] Milk drinking data available. [1.] Numbers taken from Holden & Mace (1997). [2.] Bayoumi *et al.*, (1981). [3.] Pastoralist/ milk drinking status taken from Blench (1999), Murdock (1967) and Holden & Mace (1997).

4.6 OCT1 binding affinity for the new sequence variants

Two of the novel variants *(-13915T>G* and *-13913T>C)* overlap the previously reported OCT1 binding site (Figure 4.2). The *-13907C>G* variant is just three nucleotides outside the OCT1 binding site and was present in a single lactose digester in the Jaali cohort (SD-J-67) as well as the Ethiopian lactose digester.

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

Figure 4.2 Sequence comparisons of the ancestral and variant sequences within *MCM6* intron 13 at -13.9kb upstream of *LCT*. The OCT1 binding site (determined by TRANSFAC analysis; Lewinsky *et. al.*, 2005) is shaded, and the variant alleles are shown in bold underline.

Because of the incomplete association of *-13915*G* with lactase persistence it seemed important to determine whether the allele promoted binding of the transcription factor OCT1, as this is the mechanism by which *-13910*T* is thought to confer an increase in lactase expression in Europeans (Lewinsky *et al.*, 2005), and would therefore provide good evidence of function. Electrophoretic mobility shift assays were selected as the most straight-forward way to investigate binding of the new alleles to the OCT1 transcription factor. The OCT1, *-13910*T* and ancestral sequence probes were 25 nucleotides long, double stranded and identical to those used by (Lewinsky *et al.*, 2005), and the new sequence variant probes were identical to the ancestral probe apart from single nucleotide changes made at the relevant nucleotide positions. The TAATGARAT probe was newly designed for this assay, after a review of the literature concerning target binding sequences of OCT1 was made, and the probe was designed to be of equal length to the intron 13 probes. The non-specific probe was designed by using the reverse sequence of the intron 13 ancestral probe. The probes were incubated with previously prepared protein extract from CaCo2 cells (see section 2.8.5), a colon

carcinoma cell line, and one of the only cell lines known to express lactase (Pinto *et al.*, 1983).

Gel shift analysis confirmed binding of the *-13910*T* oligonucleotide to the same protein as bound by the OCT1 and TAATGARAT probes (Figure 4.3), which was inferred from the study of Lewinsky *et al.*, (2005) to be OCT1. Specificity of the interactions was demonstrated by adding an excess of either unspecific or OCT1 unlabelled competitor probes. Binding was not affected by addition of unspecific probe, but addition of unlabelled OCT1 probe completely displaced binding of the labelled *-13910*T* probe. Binding of the *-13913*C* and *-13915*G* sequence variants to the protein was not detected and the ancestral and *-13907*G* variants showed only very faint binding after long exposures (not shown). In reciprocal experiments the *-13910*T* oligonucleotide was shown to be the only variant to displace the OCT1 oligonucleotide (Figure 4.4).

4.7 Haplotype association of new alleles

Haplotype association of the new alleles was made by visual inspection of the haplotypes that had been previously inferred for each member of the cohort study and the Middle Eastern populations and Beni Amer (see section 3.8.3). For the new populations included in this chapter, core *LCT* SNPs were typed and haplotypes inferred in the same way. The newly observed *-13915*G* allele was found to occur exclusively on a **C**-haplotype background in the Jaali, and only ten out of 131 *-13915*G* alleles were found on non-**C** haplotypes in the other populations tested; three on an **E** (an **A/C** recombinant) haplotype background, four on a **K** haplotype, and three on different, rare haplotypes. These non **C** *-13915*G* carrying chromosomes were found in Beni Amir (n=3), Saudi Arabian Bedouin (n=2), Israeli urban (n=1) and Bedouin Arabs (n=2), and Jordanian Bedouin (n=2) groups. The other new SNPs (*-13907*G* and *-13913*C*) are not associated with the **C** haplotype, and were rare in the populations genotyped, but from the preliminary data it appears that *-13913*C* is associated with **B** haplotype and *-13907*G* with **A**. Haplotype backgrounds are pursued in more detail in chapter 6.

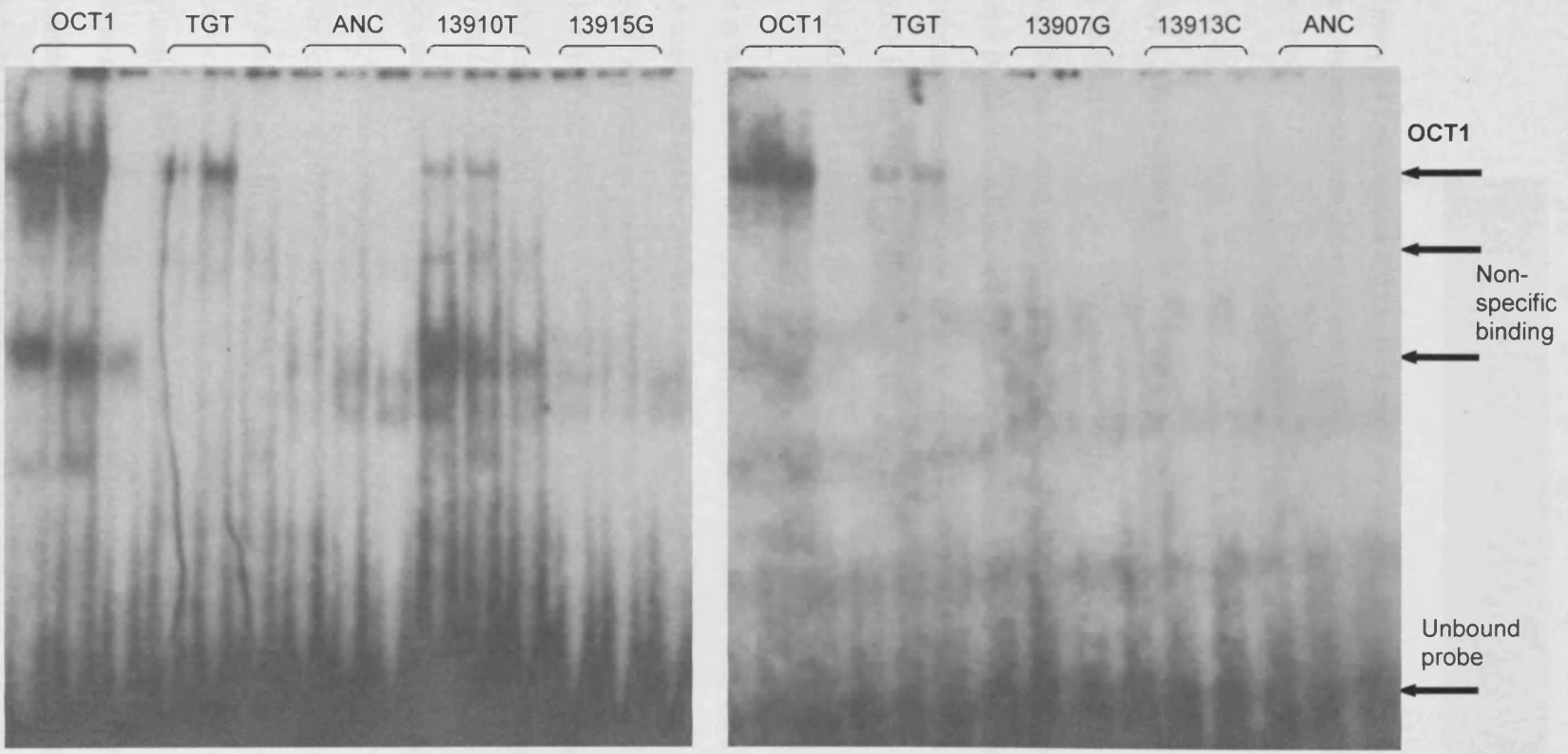


Figure 4.3 Electrophoretic Mobility Shift Assay (EMSA) of -13.9kb sequence variants. 32P labelled probes were incubated with nuclear protein extract. Unlabelled competitor probes were pre-incubated with the protein extract to demonstrate binding specificity. Above each set of three lanes the radioactively labelled probe is indicated (TGT = TAATGARAT, ANC = Ancestral). Competitor probe was present in the following order in each set of three lanes: none, unspecific and OCT1. N.B. Vertical lines of radioactivity in lanes are artefacts of the gel drying process.

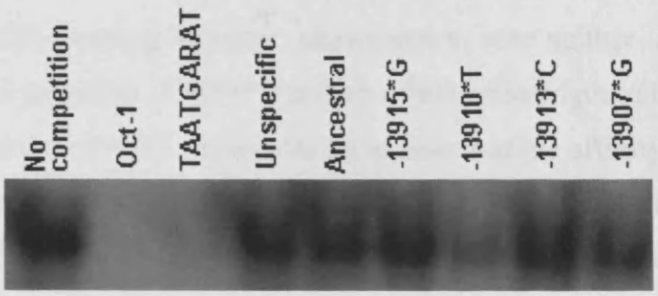


Figure 4.4 Competition EMSA: Radioactively labelled OCT1 probe was incubated with nuclear protein extract with either no competitor, OCT1, TAATGARAT, unspecific or intron 13 variant probes.

4.8 Discussion

The discovery of new SNPs in the vicinity of the OCT1 binding site, one of which is significantly associated with lactase persistence, strongly supports the notion that the trait is at least influenced by *cis*-acting variants in these populations, and points to a potential functional role, at least for *-13915*G*. However, despite the location of this SNP in the OCT1 binding site, the agreement with phenotype was not as tight as one would have expected for a causal SNP.

It is well known that lactose tolerance testing has an error rate by whichever method is used, and for breath hydrogen testing (which is considered most accurate) this has been reported as about 5% false positive (i.e. people found to be lactose intolerant by this, but not by other methods) and 8% false negative (Mulcare *et al.*, 2004). The identification of 23/39 individuals in the lactose digester group who do not carry *-13915*G* (and 20/39 who carry no variant allele) suggests an implausible false negative error rate of about 50%, if *-13915*G* is causal. There is also a high false positive rate (5/44). However it is possible that some or all of the five individuals who carry *-13915*G*, but are lactose non-digesters, were suffering from undiagnosed or undisclosed intestinal illness, which caused them to have secondary lactose intolerance. In the context of a field study it was not possible to do more than question the volunteers, and further lactose tolerance tests and other examinations were not feasible. Consequently, the possibility that *-13915*G* is one of several causes of lactase persistence in this group of people, along with other variants that are yet to be identified cannot be excluded. With this in mind it was important to ask whether this allele, like *-13910*T*, promoted binding to OCT1.

Analysis of OCT1 binding, however, shows that *in vitro* neither *-13915*G* nor any of the other alleles resemble *-13910*T* in their effect on strength of binding to the protein inferred to be OCT1. It is relevant to note that the affinity of this protein for the *-13910*T* containing sequence (TAATGTAGT) was similar to that seen with the TAATGARAT probe and significantly weaker than for the classic OCT1 motif (ATGCAAAT), and from sequence inspection it seems possible that OCT1 binds in its (OCTA-)TAATGARAT conformation (Cleary and Herr, 1995). If this is the case, the SNP (*-13915*G*) that associates with lactase persistence would be expected to disrupt the OCT1 binding site, since it converts a critical TAAT to TAAG of the PouH binding domain (Verrijzer *et al.*, 1992). Thus the identification of these new SNPs in the immediate vicinity of *-13910*T* casts serious doubt on the physiological role of this OCT1 binding site in the expression of *LCT*, although the site may be functional in relation to lactase persistence in some other way.

If the -13.9kb SNPs that are associated with phenotype (*-13910*T* and *-13915*G*) are not functional, one possibility is that they reside in a mutation hotspot (suggested by the occurrence of 4 SNPs in such close proximity) and that in different populations *-13910*T* and *-13915*G* happen to be associated with the true causal mutations. Interestingly, a fifth SNP immediately adjacent to *-13910*T* has been identified in the NIH Polymorphism Discovery Resource Panel (rs4988236, *-13908C>T)* (Entrez SNP database, NCBI, http://www.ncbi.nih.gov/), and a sixth variant, *-13914G>A* has been reported in two Austrian individuals (Tag *et al.*, 2007; Tag *et al.*, 2008)

*-13915*G* has not been reported so far in Europeans (Enattah *et al.*, 2002; Poulter *et al.*, 2003), and has not been reported in any of the public databases. Our preliminary studies on its distribution suggest that this allele may have originated in the Middle East, where it is seen at highest frequency in Bedouin groups. It is frequent in east Africa but hardly found in west Africa. It is possible that it was introduced in the last 1400 years as a result of the Arab expansion, which accompanied the spread of Islam. This is consistent with the presence of the allele in the Beni Amir and Jaali of eastern Sudan, as well as the Shuwa Arabs of Cameroon, all of whom are Muslim and claim some Arab ancestry (Robinson, 1927; Warburg, 1978; Levy and Holl, 2002; Vanhove, 2006). However if it is a marker of the principal cause of lactase persistence in Sudan, a longer history seems more likely since there is linguistic and

archaeological evidence for herding of cows in the Nile valley and eastern Africa at least 4-5000 years ago (Ehret, 1979; Smith, 1992). It is noteworthy that the frequency of *-13915*G* is low in the Israeli urban Arabs and Palestinians. The sharp differences in allele frequency in the Middle Eastern groups may reflect genetic drift magnified by the endogamous nature of these communities, or selection for lactase persistence in the nomadic groups.

*-13907*G* is widespread in Sudan and Ethiopia, being at highest frequency in the Afar (Table 4.2). It was not detected in the Middle East, suggesting a different origin for this allele. It would be of interest to estimate a date of origin for both of these SNPs using microsatellite markers, as has been done for *-13910*T* (Coelho *et al.*, 2005). The rarest of the new alleles, *-13913*C*, is seen at highest frequency (in the populations tested thus far) in the Fulani, who also carry the putative causative SNP *-13910*T* at high frequency.

It was tempting to speculate that a phenotype that may have had a great selective advantage is likely to have multiple causes. However, intuitively it seemed more likely that several different mutations would result in disruption, rather than gain of function, as attributed to *-13910*T*. Whilst it now seems less likely that OCT1 recruitment is essential for increased lactase expression *in vivo*, the association and proximity of *-13910*T* and *-13915*G* strongly suggest an important role for the locus in transcriptional regulation of the gene, perhaps in disrupting a repressor function associated with down-regulation.

5 Lactase persistence in Ethiopia: a Somali cohort study

5.1 Introduction

In chapter 4 it was shown that new alleles located in very close proximity to the original lactase persistence associated allele (*-13910*T*) were present in a Sudanese population, and that there is evidence for a positive association of one of these (*-13915*G*) with lactase persistence. Whilst that study was in progress, an independent group found another lactase persistence associated allele (*-14010*C*), located in the same intron of *MCM6*, but 100 base pairs upstream of the original SNP and situated far from the OCT1 binding site, lending support to the idea that the role of the locus with respect to lactase expression was not simple. This study nevertheless suggested a functional role for *-14010*C*, *-13915*G* and *-13907*G* as reporter/promoter construct assays showed an approximately 18-30% increase of transcription from constructs carrying these alleles compared to the ancestral allele. (Tishkoff *et al.*, 2007). Figure 5.1 shows the positions of known SNPs and transcription factor binding sites.

Figure 5.1 Position of reported SNPs in *MCM6* intron 13. Numbering of the nucleotide sequence shows only the last four digits (i.e. nucleotide 136,325,160 is labelled 5160). Depicted SNPs were reported in Ingram *et al.*, (2007); Tishkoff *et al.*, (2007); Tag *et al.*, (2007) and in public databases. Transcription factor binding sites for which there is experimental evidence are indicated by grey shading (Lewinsky *et al.*, 2005)

Importantly, the three lactase persistence associated variants were all located within a region of *MCM6* intron 13 that had been shown to have enhancer function *in vitro* (Troelsen *et al.*, 2003), suggesting that all variants may be functional and that complex regulatory mechanisms give rise to lactase persistence. The work described in this chapter reports another phenotyped cohort of Somali individuals from Ethiopia. In this study the strategy was to sequence the entire enhancer element.

The Somali population were selected because:

- The Somali are consistently described in the literature as a pastoralist population (Murdock, 1967; Blench, 1999), and at the time of sample collection, published lactase persistence frequency data was not thought to be available (subsequently a study of 244 Somali individuals from Somalia was found, which reported the lactose digester frequency to be 24% (Flatz, 1987).
- Population frequency data implied a Middle Eastern origin for *-13915*G*, but two other possibly functional variants that had been noted by ourselves and others and had been too rare (in the populations typed) to permit speculation regarding their origin. We had previously observed one of these alleles (*-13907*G*) in a lactase persistent Ethiopian, and this observation influenced our decision to study an Ethiopian pastoralist group.
- Our collaborative relationship with a group at Addis Ababa University also favoured collection of an Ethiopian group, and our collaborators advised us that the Somali region of Ethiopia would be safer than the locations for alternative population collections (e.g. Afar region) at that time.

5.2 Sequencing Strategy

Novel alleles showing an association with lactase persistence had previously been reported in *MCM6* intron 13 from positions -14010 to -13907 (numbered in bp upstream of the initiation codon of *LCT*) (chapter 4, Ingram *et al.*, 2007; Tishkoff *et al.*, 2007). A polymorphism with high minor allele frequency at -13730 that showed no association with either lactase persistence phenotype ($p = 0.55$) had also been noted in the earlier collected Sudanese cohort, and this provided a downstream boundary. We therefore sequenced from exon 13 of *MCM6* (coding regions are more highly constrained, reducing the possibility of SNPs under primers), and used a downstream primer that had already been used in many populations without any

reported mis-priming issues (MCM6778). The resulting amplicon encompassed the entire 'enhancer' region included in the most comprehensive promoter studies, Figure 5.2 (Troelsen *et al.*, 2003).

ATTTCCAAAGAGTCAGAGGACTTCATTGTGGAGCAATATAAACATCTCCGCCAGAGAGATGGTTCTGGAGTGACCAAGTC
TTCATGGAGGATTACAGTGCGACAGCTTGAGAGCATGATTCGTCTCTCTGAAGCTATGGCTCGGATGCACTGCTGTGATG
AGgtatcagagtcactttgatatgatgagagcagagataaacagatttgttgcatgtttttaatctttggtatgggacat
actagaattcactgcaaatacatt***tttatgtaactgttgaatgctcatacgaccatggaattcttccctttaaagagctt***
ggtaagcatttgagtgtagttgttagacggagacgatcacgtcatagtttatagagtgcataaagac[g/c]taagttac
catttaatacctttcattcaggaaaaatgtacttagaccctacaatgtactagtaggcctctgcgctggcaatacagata
agataa[t/g]gtag[c/t]ccctggcctcaaaggaactctcctccttaggttgcatttgtataatgtttgatttttaga
ttgttctttgagccctgcattccacgaggataggtcagtgggtattaacgaggtaaaaggggagtagtacgaaagggcat
tcaagcgtcccatcttcgcttcaaccaaagcagccctgc[t/g]ttttcctagttttattaataggtttgatgtaaggtc
gtctttgaaaaggggtttggctttttttacagtgtgactgaggtataatttataaaaagggaaatgtatggcatggtg
agtttttcacatacatccttgtgaatacccagctcaagatccaaaacatttccataatttcagaaagttccaaacccct
gcctcttttcagtcttagccctcttcccctgaagtaaccactgttccgacttcaatcactactttatcccacaggttaa

Figure 5.2 Genomic sequence from *MCM6* exon/intron 13 (chr2:136324660-136325603). Positions of PCR primers are indicated by grey highlight. Exon sequence is shown in upper case. The region demonstrated to have enhancer function (Troelsen *et. al.*, 2003) is indicated in bold italics. Four SNPs are indicated on the sequence, listed here in sequential order: *-14010G>C*; *-13915T>G*; *-13910C>T* and *-13730T>G*. The first three have all been shown to associate with lactase persistence, though not *-13730T>G*, which is the last known SNP to occur within the enhancer region.

5.3 Lactose tolerance test results

In total, 111 individuals of Somali ancestry were lactose tolerance tested in Shinile (near Dire Dawa, Ethiopia). Groups of 12-15 individuals were recruited on the day before the test and agreed to fast for a minimum of 8 hours prior to testing. Lactose tolerance test data is available in appendix C, and summarised in Table 5.1. We originally intended to include only individuals whose starting breath hydrogen was between 2 and 20ppm. However, it became apparent on the initial day of testing that having a breath hydrogen baseline reading of zero was far more frequent in this population than in the previously tested Sudanese population (chapter 3, and lactose tolerance test results included in appendix B). This may be due to differences in colonic bacteria and dietary habits between the two populations (Vogelsang *et al.*, 1988). Because of this observation, a strategic decision was made to lactose tolerance test all volunteers regardless of the lower limit of their starting breath hydrogen. A small number of individuals (n = ≤5) had starting breath hydrogen over 20 p.p.m. and in these cases testing did not proceed. All eligible (i.e. breath

hydrogen ≤ 20 p.p.m.) individuals were requested to stay for the entire test duration (3 hours) and due to the difficulty in retaining individuals experienced by M. Elamin in our previous study in Sudan, a locally recognised storyteller was employed to entertain the volunteers.

In only 8 cases was the test prematurely aborted. In six of these cases the volunteer was unambiguously classified as a lactose mal-digester, showing a breath hydrogen increase of greater than 20ppm for at least two consecutive readings. One individual was forced to abandon the test due to other commitments and could not be classified, and the remaining individual showed a breath hydrogen increase of 60p.p.m. and reported symptoms of lactose intolerance within one hour. In total 22 people were defined unambiguously as lactose digesters (presumed lactase persistent), with a rise in breath hydrogen of less than 15ppm (over 3 hours), and 68 individuals were clearly lactose mal-digesters (presumed lactase non-persistent), showing a sustained rise of greater than 25ppm.

12 individuals were classified as intermediate, because their breath hydrogen fluctuated during the course of the test or did not show sufficient rise within 3 hours (although sometimes rising by 19-21ppm in the final reading). 8 individuals did not produce hydrogen for the duration of the test, and could not therefore be classified with respect to lactose digestion status. These two groups are likely to be biased towards including more persistent individuals than non-persistent. Those classified as non-producers were not verified as such by way of a lactulose test, and may genuinely be lactase persistent. The intermediate category is most likely to include individuals heterozygous for high *LCT* expression who therefore have intermediate levels of intestinal lactase activity (Ho *et al.*, 1982).

	Lactose digesters	**Lactase mal-digesters**	**Intermediate**	**H_2 Non-producer**	**Unknown** (test aborted)
***n* individuals**	22	68	12	8	1
DNA (*n* individuals)	21	67	12	8	1
H_2 Minimum Rise	0	27	16	0	n/a
H_2 Maximum Rise	14	172	25	0	n/a

Table 5.1 Summary of phenotypes observed during lactose tolerance testing in the Somali cohort. H_2 minimum and maximum rises refer to breath hydrogen (BH) (p.p.m) over the three hour test duration. Individuals with 0 rise had starting BH $\neq 0$. The data in the DNA row indicates the number of buccal swabs from which high quality DNA yield was obtained.

During the lactose tolerance test all volunteers completed a milk-consumption questionnaire. Of the 111 individuals questioned, 79 (71%) reported their daily fresh milk consumption was greater than or equal to 0.5L. A visual aid of a 0.5L water bottle was used to confirm mutual understanding of the quantities being discussed between interviewer and respondent. Camel and cow's milk were predominant and equally common sources of milk whereas goat's milk consumption was far less common.

Interestingly, no statistically significant correlation was observed between lactose digester status and milk consumption (p = 1.00, Fishers Exact test) or symptoms associated with milk consumption in this population, in contrast to the findings in the Sudanese cohort (see section 3.7.2.1 and Table 3.2).

5.4 Sequencing results

In total, 109 DNA samples of good quality were made from the collected buccal swab samples. Individuals DD-109 (lactase non-persistent) and DD-111 (intermediate) are self-reported to be mother and daughter, and therefore only DD-109, who gave an unambiguous lactose tolerance test result, was included in the data analysis. All other individuals were reported to be unrelated at the grandparental level.

The DNA sequence from exon 13 up to position -14010 was completely invariant in the Somali cohort. Downstream of -14010 a number (n = 8) polymorphic sites were identified in the sequencing traces. Consistent with our findings in the Sudanese cohort, *-13730T>G* showed no association with persistence status (p = 0.53, Fishers exact test). Two loci were identified showing 2 (*-13806A>G*) and 1 (*-13779G>C*) alleles in only non-persistent individuals. Three different loci were variant only in the persistent group (*-14009T>G*, *-13910T>C* and *-13907C>G*), and at one locus the derived allele (*-13915*G*) was present at much higher frequency in the persistent (0.21) than in the non-persistent group (0.01). Three of the sites (*-14009T>G*, *-13806A>G* and *-13779G>C*) had not been previously reported. Table 5.2 shows the distribution of the SNPs in each category of lactase persistence status.

Phenotype	*n*	*-14010 G>C*	*-14009 T>G*	*-13915 T>G*	*-13910 C>T*	*-13907 C>G*	*-13806 A>G*	*-13779 G>C*	*-13730 T>G*
Persistent	42	0	2 (0.05)	9 (0.21)	1 (0.02)	6 (0.14)	0	0	2 (0.05)
Non-persistent	134	0	0	1 (0.01)	0	0	2 (0.01)	1 (0.01)	13 (0.10)
Intermediate	22	1 (0.05)	0	0	1 (0.05)	1 (0.05)	0	0	2 (0.09)
H_2 Non-producer	16	0	1 (0.06)	1 (0.06)	2 (0.13)	5 (0.31)	0	0	1 (0.06)
Unknown	2	0	0	0	0	1 (0.50)	0	0	0

Table 5.2 Summary of allele frequencies observed in Somali cohort. The table is categorised with respect to lactose tolerance phenotype. *n* = number of chromosomes sequenced. Numbers in brackets indicate allele frequency.

5.5 Statistical analysis

Using *-13730T>G* as a downstream boundary and exon 13 as the upstream boundary, simple counts of variant alleles occurring within the sequenced region in persistent and non-persistent individuals showed that the increased number of variant-carrying chromosomes observed in lactase persistent individuals is highly statistically significant ($p = 4.3 \times 10^{-6}$, including *-13730T>G*; Fishers exact test for a 2x2 table of variant/ancestral chromosomes and persistence status). This test depends on the correct phasing of chromosomes (because an individual carrying >1 allele may have a single variant chromosome carrying two variants, or alternatively two variant chromosomes), however, as only three individuals within the data set are doubly heterozygous this is very unlikely to affect the result of the test.

Although sample size is small, an intriguing property of these data is that three out of four of the people who carry more than one derived allele are lactase persistent (Table 5.3).

	No. of variant alleles (per person)		
	0	**1**	**2**
NP (*n* =67)	51	15	1
P (*n* =21)	4	14	3

Table 5.3 Table showing persistent and non-persistent people divided into categories that indicate the number of derived alleles carried.

Three SNPs (*-14009*G*, *-13915*G* and *-13907*G*) were present in the persistent group at a frequency ≥ 0.05, but were absent or present only at very low frequency (0.01) in the non-persistent group. Both *-13915*G* and *-13907*G* are also individually highly significantly associated with lactase persistence, ($p = 1x10^{-5}$ and $1x10^{-4}$ respectively for a 2x2 contingency table; Fishers Exact test, Table 5.4) and remain significant even after Bonferroni correction for 8 tests; significant p <= 0.006).

Another SNP of interest is *-14009T>G* which occurs in two lactase persistent individuals in the Somali, but not at all in the non-persistent group. This observation is of borderline statistical significance ($p = 0.056$, Fishers Exact test, Table 5.4) and curiously the SNP is immediately adjacent to the *-14010G>C* allele (Figure 5.3), previously shown to be associated with lactase persistence in Kenyan and Tanzanian populations (Tishkoff *et al.*, 2007).

	14010*C	***14009*G***	***13915*G***	***13910*T***	***13907*G***	***13806*G***	***13779*C***	***13730*G***
NP (n = 134)	0	0	1	0	0	2	1	13
P (n = 42)	0	2	9	1	6	0	0	2
***p* value**	n/a	0.056	**$1x10^{-5}$**	0.23	**$1 x10^{-4}$**	0.57	0.76	0.53

Table 5.4 Numbers observed of each type of derived allele in persistent and non-persistent categories. 2x2 tables of allele counts were tested for association with lactase persistence using Fishers Exact test. Alleles which show a significant association with lactase persistence (p ≤ 0.006, significant at the 5% level after Bonferroni correction for eight tests) are indicated in bold.

The discovery of *-14010G>C* and *-14009T>G* led us to re-examine the Sudanese cohort in which *-13915*G* was originally identified (chapter 4). *-14009*G* was found at a frequency of 0.06 (11/180 chromosomes). Eight of these eleven *-14009*G* carriers were lactase persistent, and 6 of the 8 carried no other candidate causal allele. Association of the allele with lactase persistence status was however not statistically significant ($p = 0.1295$, Fishers exact test).

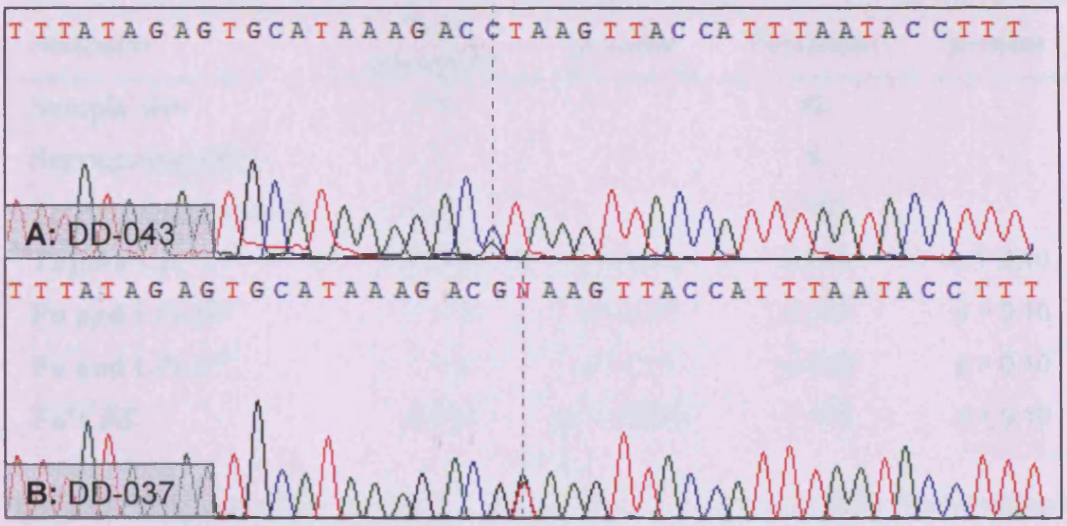


Figure 5.3 Sequencing chromatograms from members of the Somali cohort. Chromatogram A shows an individual heterozygous for *-14010G>C* (the derived *-14010*G* allele was reported to be associated with lactase persistence; Tishkoff *et al.*, 2007) and chromatogram B shows a heterozygote for the *-14009T>G* polymorphism, for which the *-14009*G* allele shows borderline association with lactase persistence within the Somali cohort.

5.6 Molecular diversity and neutrality tests

The DNA sequence from the beginning of intron 13 to -13684 (i.e. the entire enhancer) was analysed using the DnaSP suite of programs (chapter 2.10). Nucleotide diversity (π) was 0.4×10^{-3} in the lactose mal-digester group and 1.4×10^{-3} in the digesters. Nucleotide diversity was examined for departures from neutrality using Tajimas *D*, Fu and Li's *D** and *F** statistics and Fu's *FS* (Tajima, 1989; Fu and Li, 1993; Fu, 1997). These methods evaluate whether there is an excess of rare alleles in the sequenced region, a pattern associated with a selective sweep. All statistics showed a negative value in both groups, indicating an excess of rare alleles. However, only Fu's *FS* test revealed a statistically significant departure from neutrality (tested by generating random samples under the hypothesis of selective neutrality and population equilibrium using a coalescent simulation algorithm) in the non-persistent group (Table 5.5). While it is possible that this result reflects better power for Fu's *FS* to detect non-neutrality in the data, multiple testing may also have played a role. However, the Fu's *FS p* value remains significant even after Bonferroni correction for 4 tests (corrected $p <= 0.0125$)

Statistics	Non-persistent	*p*-value	Persistent	*p*-value
Sample size	134	-	42	-
Segregating sites	4	-	5	-
Pair wise differences	0.221	-	0.784	-
Tajima's *D*	-1.314	$p > 0.10$	-0.819	$p > 0.10$
Fu and Li's *D**	-1.576	$p > 0.10$	-0.757	$p > 0.10$
Fu and Li's *F**	-1.759	$p > 0.10$	-0.906	$p > 0.10$
Fu's *FS*	**-3.659**	**$p = <0.001$**	-1.978	$p > 0.10$

Table 5.5 The region -14236 to -13684 (552 basepairs) was analysed using the DnaSP software to calculate Tajima's *D* and Fu and Li's *D and *F** statistics. The Arelequin package was used to calculate Fu's *FS* for the same region. Only Fu's *FS* test showed a significant departure from neutrality and only in the non-persistent group.**

5.7 Haplotype Analysis

All samples were genotyped for 8 additional SNPs, which spanned intron 11 of *MCM6* to exon 17 of *LCT*, and included the four 'core' haplotype SNPs described in chapter 3 (*-942/3TC>ΔΔ*, *-678A>G*, *666G>A* and *5579T>C*; Figure 5.4). Haplotypes were constructed using all observed polymorphic sites (the 8 haplotyping SNPs plus the eight intron 13 variants) and inferred using a two-stage process. Initially, haplotypes were inferred using the Bayesian algorithm implemented in PHASE (section 2.9.3.1) for which the inferred haplotypes of 90/108 individuals had a posterior probability of >0.9. The remaining cases were individually inspected by eye, and those which had one favoured haplotype combination were checked and accepted. In some cases two haplotype combinations were equally likely for a given individual, and here haplotypes were adjusted manually to reflect knowledge of allelic combinations reported in the literature or observed in other populations. The second stage was to pool genotype data for the Somali cohort with a larger data set consisting of Jaali, Afar, Somali (from Jijiga), Fulani and Europeans, n = 358 and infer the haplotypes once again. This larger sample set gave greater power for inference of the haplotype background of the rarer alleles (Andres *et al.*, 2007) e.g. *-13910*T*, *-14009*G*. Finally the outputs of the two phasing processes were compared and the best haplotype pairs selected (haplotype pairs and posterior probabilities included in appendix D). It was important to use a two stage procedure

to phase haplotypes due to the hypothetical problems associated with pooling samples from different populations which may exhibit genetic heterogeneity.

In the Somali group, all 12 *-13907*G* alleles were inferred to be present on an **A** haplotype background (with a posterior probability >=0.74), as were all 4 *-13910*T* alleles (posterior probability >=0.42). *-13915*G* was always present on the **C** haplotype, consistent with the data obtained previously for the Jaali cohort. All three individuals (two persistent and one H_2 non-producer) who carried the *-14009*G* allele had an **X** haplotype (posterior probability >=0.99), which appears to be derived from a recombination between **U** and **A** haplotypes between the *666G>A* and *5579T>C* loci (Figure 5.4). The *-14010*C* allele was deduced to be present on the **B** haplotype, as the individual who carried this allele was heterozygous **B/C** for diplotype, and it had been stated in the literature that the *-14010*C* allele occurred on a distinct haplotype from *-13915*G* (Tishkoff *et al.*, 2007).

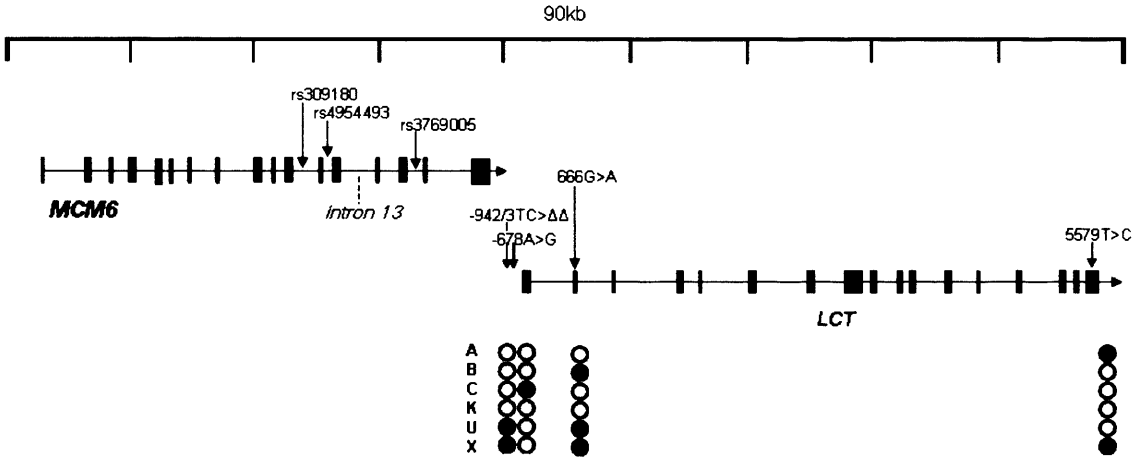


Figure 5.4 Location of genotyped SNPs with respect to *LCT* and *MCM6*. Combinations of the four 'core' SNPs that define *LCT* haplotypes are depicted by the shaded circles beneath the *LCT* gene (shading indicates the derived allele at that locus). Only the most common haplotypes observed in the previously typed populations (chapter 3) are shown.

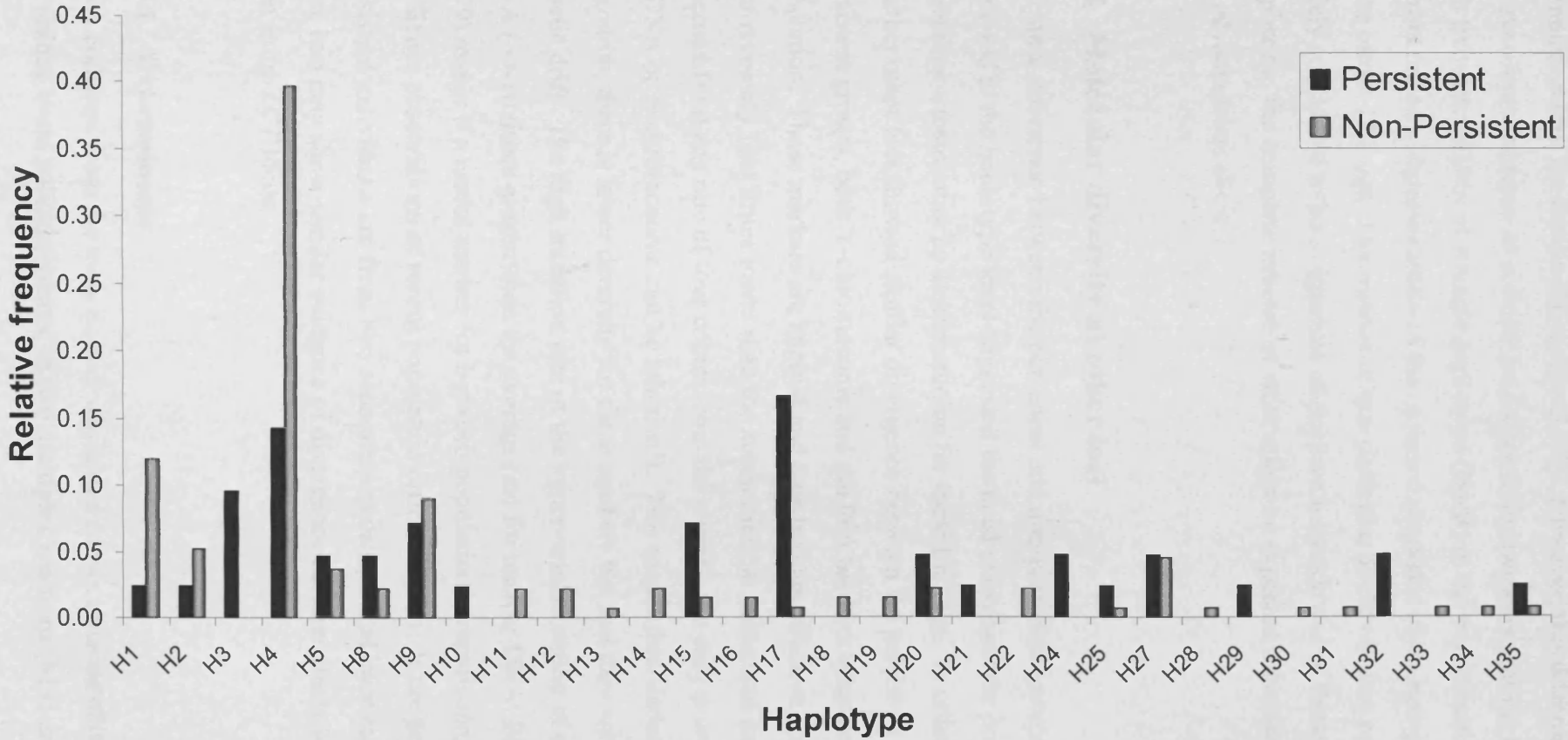


Figure 5.5 Haplotype distribution in persistent and non-persistent Somali. Haplotypes were inferred using the computer program Phase. The distribution was found to be significantly different between the two groups. $p = 0.001$ (case-control function). The most frequent haplotype (H17) in the persistent group corresponds to the C haplotype carrying the lactase persistent associated *-13915*G* allele. The most frequent haplotype (H4) in the non-persistent group corresponds to the B haplotype. H6, H7, H23 and H26 are not shown in this figure as they are not observed in the persistent/non-persistent groups.

Figure 5.5 shows the difference in haplotype distribution of the persistent and non-persistent groups, which is statistically significant ($p = 0.01$, Phase case-control permutation test, for details see section 2.9.3). However, this difference is not due to the over-representation of a single haplotype in the persistent group, but rather to the high frequency (0.40) of a single haplotype (H4/**B**) in the non-persistent group. This mirrors the over-representation of the ancestral sequence in the non-persistent group at the intron 13 locus. Dominance of one particular allele within a population is widely considered to be a signature of positive selection and so these findings are perplexing; the complete inverse of what might be expected if lactase persistence is the advantageous allele.

5.8 Molecular diversity at other loci

The stark difference between the persistent and non-persistent groups at the intron 13 locus and at the haplotype level suggested the need to exclude the possibility of population substructure as an explanation for these findings. In order to examine whether other loci showed similar divergence between the persistent and non-persistent groups, both Y-chromosome and mtDNA markers were genotyped in the population. These markers are haploid and thus have an effective population size approximately four times lower than the autosomes (an autosomal gene locus can be inherited from any one of four copies from the parents, but only a single copy of the mtDNA or Y-chromosome can be inherited). This means that, under most models of migration, there is lower diversity for these markers but that they are more prone to genetic drift. The high mutation rate of the hypervariable region of mitochondrial DNA (~5-10 times greater than the average rate for nuclear DNA; Brown *et al.*, 1979) makes it a useful marker for between population comparisons, as this facilitates observations of recent population differentiation. If the persistent and non-persistent individuals are from two separate populations that have recently mixed, these loci may show similar evidence of divergence between the two groups as is seen at the *LCT* locus.

5.8.1 Y chromosome

DNA collected from the male Somali was tested for six microsatellites markers and six unique event polymorphisms, in two multiplex reactions (MS1 and UEP1) as

described in (Thomas *et al.*, 1999) (for full details see section 2.6.4). Both UEP and microsatellite loci were used to form haplotypes. Approximately 50% of the Somali sample were male; 11/21 persistent and 38/67 non-persistent. Of these, 8 and 33 respectively were successful in the multiplex reactions. Three different haplotypes were observed in the persistent group, and ten in the non-persistent. Two haplotypes were common to both groups, and both groups shared a common modal (most frequent) haplotype.

5.8.2 mtDNA

Mitochondrial DNA was sequenced in both directions across 382bp between position 16019 and 16400 of the hyper-variable control region 1 (HVR1) (see section 2.7.2 for full details). Successful reads were obtained for 16/22 persistent and 59/68 non-persistent individuals. Mitochondrial DNA diversity was greater than that observed for the Y chromosome. In the persistent group 14 haplotypes were counted and 43 in the non-persistent (i.e. nearly every individual tested had a unique mtDNA sequence). Four haplotypes were found in common between the groups, although the modal haplotype was different in each case.

5.8.3 Comparison of genetic differentiation and diversity

Genetic diversity and differentiation of the lactase persistent and non-persistent subgroups of the Somali population were compared separately in each of the three genetic systems (mtDNA, Y-chromosome & *LCT*). Genetic differentiation was tested both by an exact test of population differentiation (Raymond and Rousset, 1995), and by permutation-based AMOVA of Wright's F_{ST} statistic (Wright, 1951; Excoffier *et al.*, 1992). A significant difference in genetic diversity was tested as described in (Thomas *et al.*, 2002). Descriptions of the methods can be found in chapter 2 (section 2.9). Table 5.6 shows the pairwise F_{ST} values between the persistent and non-persistent groups, along with the *p* value for F_{ST} and for an exact test of differentiation. Both F_{ST} and the test of differentiation show that the two populations differ significantly at only the *LCT* locus.

	LCT	Y	mtDNA
Pair wise F_{ST}	0.39	-0.03	0.01
F_{ST} *p* value	0.00 +/- 0.00	0.64 +/- 0.04	0.28 +/- 0.04
Non-differentiation (exact *p* value)	0.00 +/- 0.00	1.00 +/- 0.00	1.00 +/- 0.00

Table 5.6 F_{ST} and Exact test of population differentiation, calculated using Arlequin software. Nucleotide sequence from -14030 to –13803 of *LCT*, 16019 and 16400 of mtDNA and haplotypes formed of 6 STR and 6 UEPs on the Y-chromosome were entered into the program.

The set of functions TEST_h_DIFF were also used to test for significant differences in genetic diversity (*h*) (Nei, 1987) between the persistent and non-persistent Somali. The test looks for differences in haplotype diversity between the two populations and significance of the difference is obtained using both frequentist and Bayesian solutions (Thomas *et al.*, 2002). Results of this test for each locus are shown in Table 5.7. All tests found a significant divergence for *LCT* while only Bayes test B found evidence that genetic diversity was different in the persistent and non-persistent populations for the Y-chromosome markers. This difference may be due to the fact that one persistent individual carried a haplotype not observed in the non-persistent population.

Locus	*n*	*h*	*s.e*	*p* value	Bayes test A	Bayes test B
LCT						
Persistent	42	0.73	0.030	<0.00001	<0.00001	<0.00001
Non-persistent	134	0.09	0.007			
mtDNA						
Persistent	16	0.98	0.005	0.74	N/A	N/A
Non-persistent	59	0.99	0.001			
Y-chromosome						
Persistent	8	0.46	0.088	0.81	0.78	0.0018
Non-persistent	33	0.52	0.043			

Table 5.7 Test for differences in genetic diversity between the persistent and non-persistent Somali. Samples were tested using the TEST_h_DIFF method described in Thomas et al., 2002. Standard error was calculated using the formula $s = h(1-h)/\sqrt{n}$.

5.9 Discussion

In this chapter association of *-13915*G* with lactase persistence was replicated and association of *-13907*G* with lactase persistence was confirmed, although the most striking characteristic of the data set is the finding that the non-persistent Somali are much more homogeneous at the *LCT* locus than are the persistent Somali. Most models of selection are based on the principle that if under selection, an allele will rise to high frequency within a population. Therefore, consideration of both the haplotype distribution and the variation within the intron 13 enhancer give the impression that if any selection has occurred within the population, it was the non-persistent individuals who carried the advantage.

However, evidence that the intron 13 enhancer region does confer a selective advantage is presented here by the observation that the distribution of variant sites is not random with respect to lactase persistence status. A significant association of *-13915*G* with lactase persistence was observed, reproducing the association observed for this allele in the Sudanese cohort, and another SNP, *-13907*G* also shows a significant association with lactase persistence.

A fascinating property of this data set is the finding that lactose tolerance status is in no way correlated with milk consumption. This is contrary to our findings that individuals adapted their milk intake to reflect digestion status in the Sudanese cohort. Our lactose digester frequency agrees well with a large previously published study (Flatz, 1987), so that the low frequency of lactose tolerance is unlikely to be due to sampling error. It is possible that the gut flora has adapted to tolerate milk better in this population: The low breath hydrogen readings certainly reflect some difference compared to the Sudanese cohort, and may indicate increased colonic acidity which prohibits colonisation by some bacteria and favours colonisation by others. In retrospect we should have adjusted our questionnaire to find out whether 0.5L of milk was consumed in a single dose, or whether it was divided throughout the day. Whatever the nature of the adaptation (cultural or physiological) that allows the non-digesters to consume milk in large quantities, the observation may have implications for the genetic pattern observed. If lactose mal-digesters are not adversely affected by consuming large quantities of fresh milk, perhaps the selective pressure is not strong enough to drive a single allele to very high frequency within this population, but has allowed maintenance of several alleles introduced by genetic

drift or population movement. This perhaps implies considerable population admixture of migratory milk-drinking peoples.

The skewed distribution of variant alleles observed in the Somali cannot be easily explained by population stratification, as investigation of other loci (Y-chromosome and mtDNA) which are more likely to show evidence of recent population mixing did not indicate similar population divergence. The Y-chromosome haplotypes were quite homogeneous throughout the Somali population as a whole, and mtDNA diversity was similar in both groups. For closely related populations, the sensitivity of each genetic system to detect differentiation is dependent on diversity (*h*), thus the finding that no significant differentiation is observed in the mtDNA, which had higher *h* than *LCT*, is particularly noteworthy. Further confirmation of this at a number of autosomal loci would provide strong evidence that the population stratification observed within the Somali is peculiar to the *LCT* locus.

The clustering of variant alleles, and the evidence that at least four (*-13910*T*, *-13915*G*, *-14010*G* and *-13907*G*) of the enhancer SNPs may be functional with respect to lactase persistence status (Olds and Sibley, 2003; Troelsen *et al.*, 2003; Tishkoff *et al.*, 2007; Imtiaz *et al.*, 2007), seems to suggest that the 'enhancer' region is important in *LCT* expression in a complex manner, and is likely to consist of more than a single transcription factor binding site. There are now many known examples of regulatory regions such as insulators, locus control regions and enhancers which are often located a number of kilobases away from the gene whose expression they influence, and the mode of action of these regulatory elements is an active area of current research (for reviews see West and Fraser, 2005; Dean, 2006).

The sequence data obtained within this study meant that the most suitable analyses involved measuring departures from neutrality in nucleotide diversity. The tests used here evaluate whether an excess of rare alleles is present, indicating either selection or rapid population growth. Negative values indicating an excess of rare alleles, were observed in both the non-persistent and the persistent groups. Tajima's *D* did not significantly differ from neutrality, however it has been shown that the power of the test is weak when sample sizes and the number of segregating sites is small (Simonsen *et al.*, 1995). Fu's *FS* test did show a significant departure from neutrality, but only in the non-persistent group. The tests used here are modelled on

a selective sweep of a single allele, and in this case, as there may be more than one advantageous allele under selection, these traditional tests may not be applicable.

It is possible that the 'enhancer' region located 13.9kb upstream of *LCT* is a mutation hotspot, and that many of the observed alleles have no functional relevance. The genome average of SNP density within a population is approximately 4-5 SNPs per thousand base pairs (Brookes, 2005), and therefore, the 3-4 nucleotide substitutions observed in the Somali population within the enhancer region does not seem particularly high. Additionally, if the region was simply a mutation hotspot a similar level of nucleotide diversity would be expected to be observed in all populations, and this does not appear to be the case.

It seems likely from the data presented here and in other association studies (chapter 4, Tishkoff *et al.*, 2007) that the total spectrum of variation directly affecting lactase persistence has still not been fully characterised. In this study the breath hydrogen lactose tolerance testing was the most thoroughly conducted survey of lactase persistence within an African population, with full three hour breath hydrogen readings obtained for nearly all participants. Despite these strict procedures, a number of individuals were still found to have ambiguous test results. Furthermore, four Somali individuals were clear lactose digesters carried no variation in the entire sequenced region, and one more was heterozygous only at *-13730T>G*. This is also reflected in the Sudanese cohort, in which 14 persistent individuals with totally invariant intron 13 sequences are observed. All these observations suggest the presence of additional modifying factors in the lactase persistence phenotype, although whether these factors are genetic or not remains to be determined.

6 Geographic distribution and haplotype background of *MCM6* intron 13 variation

6.1 Introduction

Previous chapters have focussed on identifying putative causal SNP(s) of lactase persistence in non-European populations. A number of alleles have been identified within intron 13 of *MCM6*, some of which, in phenotyped cohorts, associate with lactase persistence and others which do not (chapters 4 & 5, Ingram *et al.*, 2007; Tishkoff *et al.*, 2007). The aim of this chapter is to collate the information regarding intron 13 derived allele frequency in all populations genotyped irrespective of phenotype to investigate their global distribution, possible origins and genealogy. To achieve this, a survey of the literature was also made to obtain additional information on the distribution of the novel *MCM6* intron 13 alleles reported in non-European populations.

For the populations typed within this project, allele frequencies for all *MCM6* intron 13 SNPs and also haplotype SNPs are used to measure pairwise F_{ST}s and produce principal co-ordinate (PCO) plots to display genetic similarities and differences between populations.

Haplotypes are inferred for the entire data set to identify haplotype backgrounds of all alleles and to compare linkage disequilibrium between them. Haplotype diversity of each of the derived alleles is focussed on as a possible indicator of both age and selection. Determination of haplotypes also allows reconstruction of the evolutionary relationships between alleles. To facilitate this, and to investigate conservation of regulatory regions, sequence alignment between human and primates is made in order to inform on likely ancestral alleles and also conservation between species. These data are used to construct a haplotype network based on the one previously constructed by Hollox *et al.*, (2001).

6.2 Total known distribution of *MCM6* enhancer alleles

At the outset of this thesis very little was known about the causes of lactase persistence in African populations. However, over the duration of this work a

number of research groups (including myself) have resequenced all or parts of *MCM6* intron 13 to reveal variation in African and Middle Eastern populations which is thought to affect *LCT* expression (Myles *et al.*, 2005; Tishkoff *et al.*, 2007; Imtiaz *et al.*, 2007; Enattah *et al.*, 2008).

During the course of this project the intron 13 enhancer region of *MCM6* was sequenced in a total of 747 individuals in 20 separate population samples (including two European groups as comparators) and in whom the four core *LCT* SNPs were also typed. Table 6.1 shows allele frequencies for all the SNPs identified in the enhancer region (that occurred more than once) and their allele frequencies in the populations genotyped within this thesis. All markers were tested for deviation from HWE using Arlequin software. Some significant deviations were observed: *-13915T>G* in the Saudi (p =0.003) and Israeli Bedouins (p = 0.01), and *-13910C>T* in the Fulani (p = 0.02), but none of these are significant after Bonferroni correction for multiple testing (of 8 markers across 20 population groups). Three novel variants (*-13779G>C, -13957A>G* and *-14028T>A*) were identified that occurred as singletons within this sample set and are not included in Table 6.1. The variants were confirmed by sequencing both strands of the PCR products. *-13957A>G* was identified in an individual of Amharic ancestry, *-13779G>C* was identified in a Somali individual, and *-14028T>A* was identified in a European individual previously shown to have two high expressing lactase alleles, but who was heterozygous for *-13910C>T* (Poulter *et al.*, 2003).

Considering the SNPs from left to right in Table 6.1, it can be seen that *-14025A>G* is at highest frequency in the Donglawi (although the sample size is small), and also at low frequency in two of the Middle Eastern Bedouin groups. The only population in which *-14010*C* is observed is the Somali (both in the phenotyped and unphenotyped populations), in which only 3 alleles were identified in total. *-14009*G* is most frequent in the Sudanese groups; the Beni Amer, Shaigi and Jaali, and is also present at lower frequency in some of the Ethiopian populations. *-13915*G* is the most frequent and also widespread of the intron 13 variants (outside Europe) and is present in a number of Ethiopian, Sudanese and Middle Eastern populations. It is observed at highest frequency in the Middle Eastern Bedouin groups. *-13913*C* is rare in all populations (<0.05), although at highest frequency in the Fulani.

Population group	*n*	P	*-14025A>G*	*-14010G>C*	*-14009T>G*	*-13915T>G*	*-13913T>C*	*-13910C>T*	*-13907C>G*	*-13806A>G*
Afar	37	**y**	0.000	0.000	0.013 ± 0.01	**0.122 ± 0.04**	0.013 ± 0.01	0.013 ± 0.01	**0.297 ± 0.05**	0.014 ± 0.01
Amharic	19	**n**	0.000	0.000	0.000	**0.132 ± 0.05**	0.000	0.000	**0.053 ± 0.04**	0.000
Beni Amer	81	**0.64**[1]	0.000	0.000	**0.105± 0.02**	**0.247 ± 0.03**	0.000	0.006 ± 0.01	0.006 ± 0.01	0.000
Bedouin - Israeli	19	**y**	0.000	0.000	0.000	**0.132 ± 0.05**	0.000	0.026 ± 0.03	0.000	0.000
Bedouin - Jordanian	23	**0.51**[1]	0.022 ± 0.02	0.000	0.000	**0.348 ± 0.07**	0.000	0.000	0.000	0.000
Bedouin – Saudi Arabian	47	**0.59**[1]	0.011 ± 0.01	0.000	0.000	**0.479 ± 0.05**	0.011 ± 0.01	0.000	0.000	0.000
Druze	14	**n**	0.000	0.000	0.000	**0.107 ± 0.06**	0.000	0.036 ± 0.04	0.000	0.000
Dunglawi	6	**n**	**0.083 ± 0.08**	0.000	0.000	0.000	0.000	0.000	**0.083 ± 0.08**	0.000
N. European	55	**0.80**[1]	0.000	0.000	0.000	0.000	0.000	**0.618 ± 0.05**	0.000	0.000
S. European	33	**0.18**[1]	0.000	0.000	0.000	0.000	0.000	**0.091 ± 0.04**	0.000	0.000
Fulani	55	**0.53**[1]	0.009 ± 0.01	0.000	0.000	0.000	0.036 ± 0.02	**0.391 ± 0.05**	0.000	0.000
Israeli non-Bedouin Arab	80	**n**	0.000	0.000	0.006 ± 0.01	**0.050 ± 0.02**	0.025 ± 0.01	0.000	0.000	0.000
Jaali	86	**0.27**[2]	0.000	0.000	**0.064 ± 0.02**	**0.134 ± 0.03**	0.006 ± 0.01	0.006 ± 0.01	0.006 ± 0.01	0.000
Mambila	37	**n**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Palestinians	18	**n**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Shaigi	9	**n**	0.000	0.000	**0.167 ± 0.09**	**0.056 ± 0.05**	0.000	0.000	0.000	0.000
Shuwa Arab	15	**y**	0.000	0.000	0.000	**0.133 ± 0.06**	0.000	0.000	0.000	0.000
Somali	37	**0.13**[2]	0.000	0.027 ± 0.02	0.013 ± 0.01	0.041 ± 0.02	0.013 ± 0.01	0.000	**0.095 ± 0.03**	0.019 ± 0.02
Somali (phenotyped)	109	**0.13**[2]	0.000	0.005 ± 0.01	0.014 ± 0.01	0.050 ± 0.01	0.000	0.018 ± 0.01	**0.055 ± 0.02**	0.009 ± 0.02
Wolof	59	**0.29**[3]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 6.1 Allele frequencies of all intron 13 variation observed in the populations genotyped. Standard error was calculated using the formula $s = \sqrt{(pq/2n)}$, where *n* is the number of individuals genotyped. The column entitled P indicates whether the population is considered to be pastoralist or not, or, if known, lactase persistence allele frequency (calculated from published phenotypic frequency: source references; [1]Holden & Mace (1997), [2] data presented in this thesis, [3] Arnold *et al.*, (1980). Allele frequencies >0.05 are indicated in bold. *-13730T>G* is not included here as complete data is not available in all populations.

Confirming the observations of (Mulcare *et al.*, 2004), *-13910*T* is only observed at appreciable frequencies in the European and Fulani populations. In all other populations it occurs at a frequency of less than 0.05, often represented as a singleton within the sample. *-13907*G* is observed at highest frequency in the Afar population, but is also present in other Ethiopian and Sudanese groups. *-13806*G* is very rare and has so far been observed only in the Ethiopian groups (Afar and Somali).

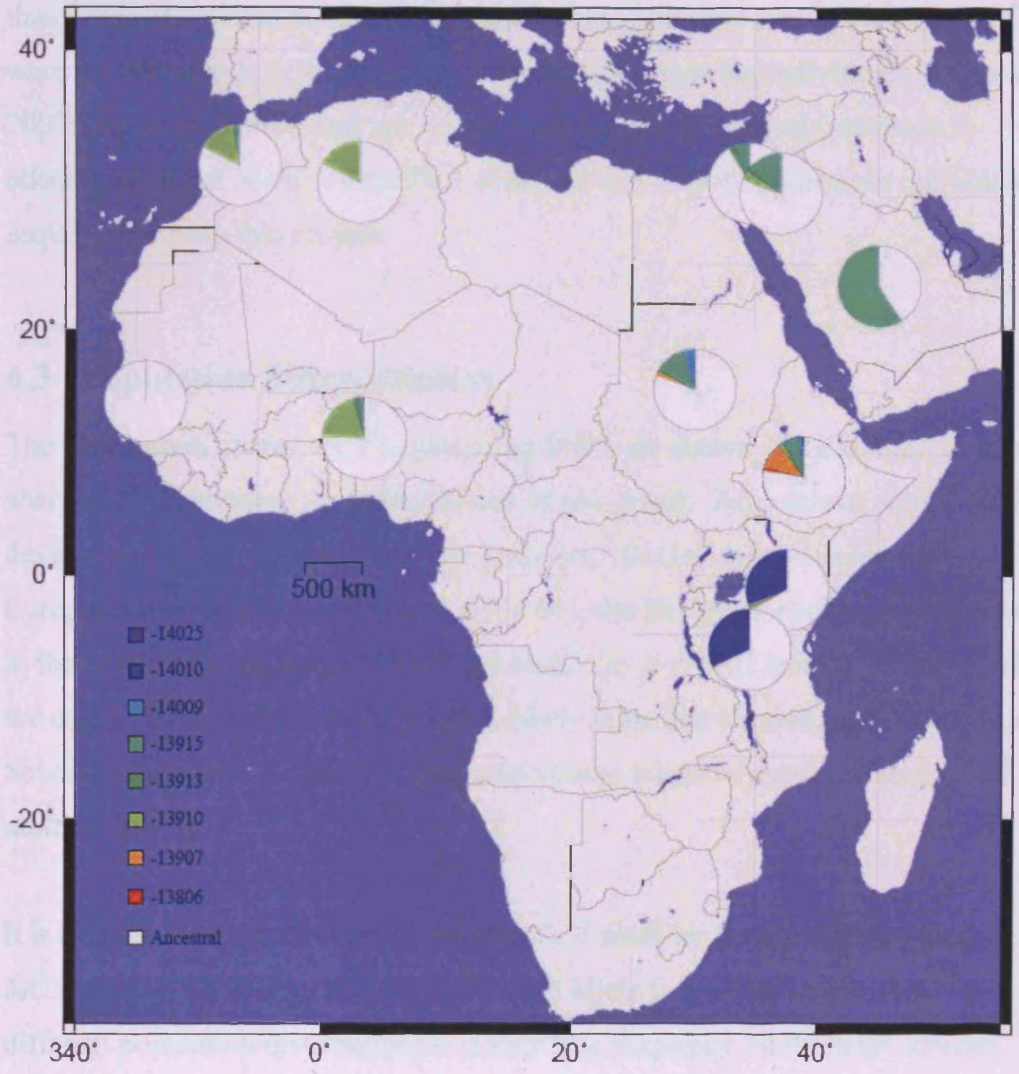


Figure 6.1 Overview of intron 13 allele frequencies in Africa and Arabia. All alleles occurring in the enhancer region (-14133 to -13684) were included with the exception of *-13730T>G*; see chapter 5. Data presented by country and compiled from Myles, *et al.*, (2005), Ingram *et al.*, (2007), Imtiaz, *et al.*, (2007); Tishkoff *et al.*, (2007), Enattah, *et al.*, (2008) and Ingram *et al.*, (this thesis).

A review of the literature to date finds that in total a further 2312 chromosomes have been sequenced in individuals from various non-European populations, bringing the total number of non-European chromosomes sequenced to ~4000. Figure 6.1 details the observed allele frequencies reported in different geographical regions, and depicts how many of the intron 13 alleles appear to be 'private'; mutations, occurring only in a single geographic location. Source data for this figure is available in appendix E. A further 2000 chromosomes have been sequenced across the region in European populations, and these are not included as the occurrence of an allele other than *-13910*T* within the intron 13 enhancer is extremely rare (n = 3). Two of these were occurrences of *-13914G>A* in heterozygous Austrian individuals (Tag *et al.*, 2007; Tag *et al.*, 2008), and one was the previously mentioned (section 6.2) occurrence of the novel *-14028T>A* observed in a northern European individual sequenced during this project.

6.3 Population differentiation

The distribution of core *LCT* haplotyping SNPs are shown in Table 6.2. In this analysis the Europeans were considered as one group. All markers were tested for deviation from HWE using Arlequin software. Deviations were observed only in the European population (at *-678A>G*, $p = 0.01$), the Shaigi (at *666G>A*, $p = 0.03$) and in the Saudi Bedouin (at *-678A>G* and *666G>A*; $p = <0.01$ and 0.03 respectively). In the case of the European sample this is likely to be due to pooling of Northern and Southern European groups, but again deviations might be expected because of multiple testing.

It is clear that the distribution of the core *LCT* markers is very different from the *MCM6* intron 13 SNPs. The *-942/3TC>ΔΔ* allele is the only one to show very different population distribution: It is seen at a frequency >0.05 in all African populations, at lower frequency in the Middle Eastern populations and is absent in all European samples genotyped here, in agreement with the observations of Hollox *et al.*, (2001). All other core SNPs were observed in all populations with the exception of *-678A>G*, for which the ancestral (A) allele frequency was 1.0 in the small Donglawi group.

Population group	*n*	*-942/3TC>ΔΔ*	*-678A>G*	*666G>A*	*5579T>C*
Afar	37	0.064 ± 0.03	0.115 ± 0.04	0.487 ± 0.06	0.397 ± 0.06
Amharic	19	0.053 ± 0.04	0.237 ± 0.07	0.579 ± 0.08	0.105 ± 0.05
Beni Amer	81	0.142 ± 0.03	0.364 ± 0.04	0.346 ± 0.04	0.370 ± 0.04
Druze	14	0.000	0.393 ± 0.09	0.214 ± 0.08	0.464 ± 0.09
Dunglawi	6	0.250± 0.13	1.000	0.500± 0.14	0.250 ± 0.13
European	89	0.000	0.86 ± 0.03	0.580± 0.04	0.640 ± 0.04
Fulani	55	0.109 ± 0.03	0.036 ± 0.02	0.300 ± 0.04	0.536 ± 0.05
Israeli urban Arab	80	0.012 ± 0.01	0.253 ± 0.03	0.457 ± 0.04	0.346 ± 0.04
Bedouin - Israeli	19	0.026 ± 0.03	0.316 ± 0.08	0.395 ± 0.08	0.289 ± 0.07
Bedouin - Jordanian	23	0.000	0.435 ± 0.07	0.283 ± 0.07	0.304 ± 0.07
Bedouin - Saudi	47	0.011 ± 0.01	0.564 ± 0.05	0.287 ± 0.05	0.213 ± 0.04
Jaali	86	0.116 ± 0.02	0.297 ± 0.03	0.424 ± 0.04	0.233 ± 0.03
Mambila	37	0.342 ± 0.08	0.316 ± 0.08	0.395 ± 0.08	0.211 ± 0.07
Palestinians	18	0.056 ± 0.04	0.250± 0.07	0.444 ± 0.08	0.333 ± 0.08
Shaigi	9	0.167 ± 0.09	0.111 ± 0.07	0.389 ± 0.11	0.611 ± 0.11
Shuwa Arab	15	0.167 ± 0.07	0.133 ± 0.06	0.267 ± 0.08	0.133 ± 0.06
Phenotyped Somali	109	0.078 ± 0.02	0.110± 0.02	0.573 ± 0.03	0.220 ± 0.03
Somali	37	0.054 ± 0.03	0.176 ± 0.04	0.649 ± 0.06	0.176 ± 0.04

Table 6.2 Allele frequencies of all core *LCT* markers in the populations genotyped. Standard error was calculated using the formula s = √(pq/2n), where n is the number of individuals.

The allele frequencies in Tables 6.1 and 6.2 were used to examine population differentiation by permutation-based AMOVA of Wright's F_{ST} statistic using Arlequin (Excoffier *et al.*, 1992; Wright *et al.*, 1951, see chapter 2.9) and these were subsequently used to generate PCO plots in order to visualise the genetic distances in two dimensions (Legendre and Legendre, 1998). These analyses were carried out using the core haplotype markers, the intron 13 alleles and both combined (Figure 6.2 a-c).

The intron 13 plots looked quite different from those including haplotype markers and it seems that inclusion of the intron 13 alleles increases population differentiation. However in all plots a central cluster is formed which includes most of the Middle Eastern, Sudanese and some of the Ethiopian populations. This seems to be a product of the rather similar allelic diversity observed in all these populations, and may reflect the close interactions between peoples of these regions over many generations.

In all PCO plots the Fulani population clusters with the Europeans, rather than with the other west African populations (Mambila and Wolof on the intron 13 plots), presumably because of the prevalence of both the **A** haplotype and *-13910*T* in both

populations. The Saudi Bedouin population are also consistent outliers in all plots, and are joined by the Jordanian Bedouin when only variation within intron 13 is compared. This is most likely to be caused by the high frequency of just one derived allele, *-13915*G*, which is particularly prevalent in the Bedouin groups. The Mambila are found in the central cluster in both plots which include core markers. However when only intron 13 variation is considered they become outliers, along with the Wolof. Both these populations show a dearth of variation in the intron 13 region (note that the Wolof are not included in the other plots as they have not been typed for the haplotypic markers).

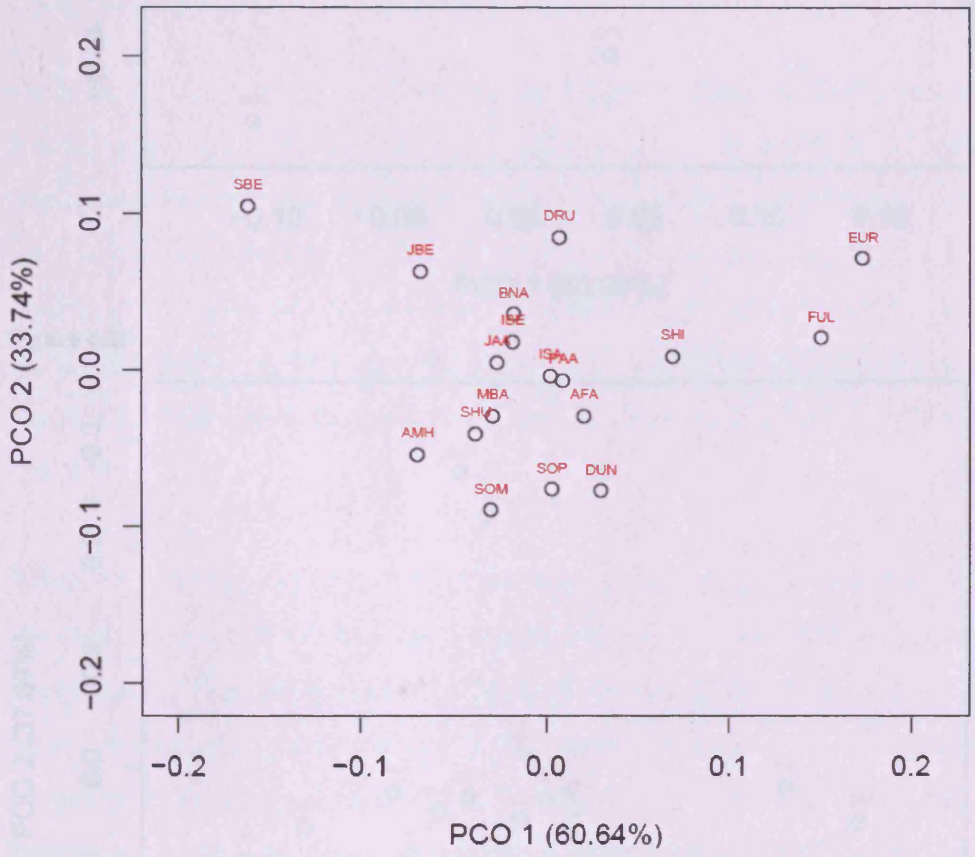


Figure 6.2a-c. Principal co-ordinate plot summarising pairwise F_{ST} values between populations. Axis labels indicate the percentage of information (genetic distance) captured by the first two principal axes. F_{ST}s were calculated from (a) allele frequencies of intron 13 and core *LCT* haplotyping SNPs (b) allele frequencies of core *LCT* haplotyping SNPs and (c) allele frequencies of intron 13 SNPs. Populations typed include: Afar (AFA), Amharic (AMH), Beni Amer (BNA), Druze (DRU), Dunglawi (DUN), European (EUR), Fulani (FUL), Israeli Bedouin (IBE), Israeli non-Bedouin Arab (ISA), Jaali (JAA), Jordanian Bedouin (JBE), Mambila (MBA), Palestinians (PAA), Saudi Arabian Bedouin (SBE), Shaigi (SHI), Shuwa (SHU), Somali (SOM), Somali – phenotyped (SOP) and Wolof (WOF).

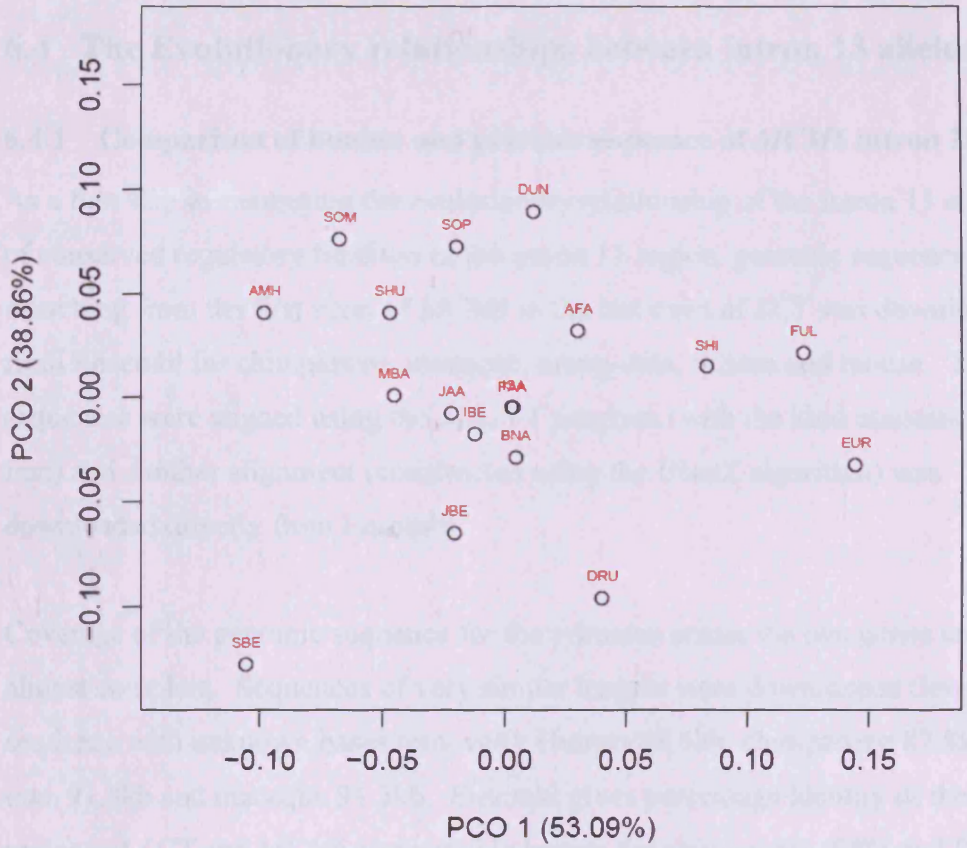


Figure 6.2b

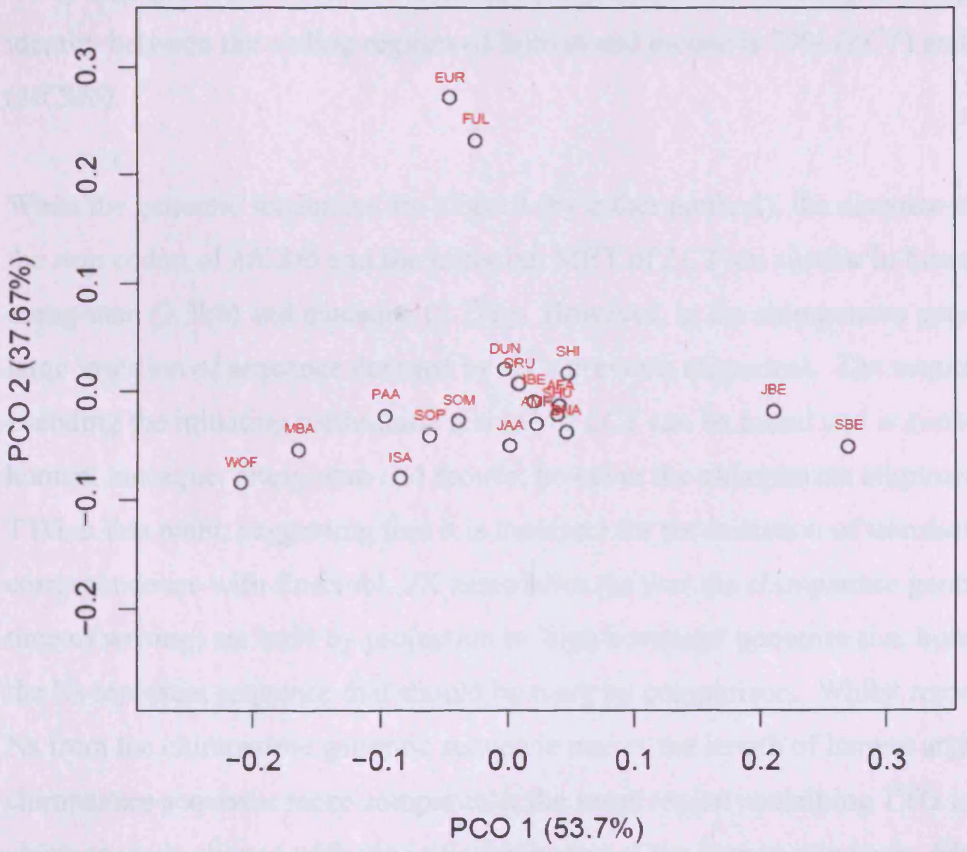


Figure 6.2c

6.4 The Evolutionary relationships between intron 13 alleles

6.4.1 Comparison of human and primate sequence of *MCM6* intron 13

As a first step in examining the evolutionary relationship of the intron 13 alleles and of conserved regulatory function of the intron 13 region, genomic sequence stretching from the first exon of *MCM6* to the last exon of *LCT* was downloaded from Ensembl for chimpanzee, macaque, orang-utan, human and mouse. The sequences were aligned using the MAFFT program (with the kind assistance of Y. Itan) and another alignment (constructed using the BlastZ algorithm) was downloaded directly from Ensembl.

Coverage of the genomic sequence for the primates across the two genes seems to be almost complete. Sequences of very similar lengths were downloaded (length of sequence with unknown bases removed): Human 88.6kb, chimpanzee 87.8kb, orang-utan 91.4kb and macaque 91.3kb. Ensembl gives percentage identity of the coding regions of *LCT* and *MCM6* compared to human for chimpanzee (98% and 99%), orang-utan (96% and 98%) and macaque (95% and 99%). In comparison, percentage identity between the coding regions of human and mouse is 79% (*LCT*) and 95% (*MCM6*).

When the genomic sequences are aligned (by either method), the distance between the stop codon of *MCM6* and the initiating MET of *LCT* are similar in human (3.6kb) orang-utan (3.3kb) and macaque (2.7kb). However, in the chimpanzee genome a large insertion of sequence denoted by 'N's prevents alignment. The sequence encoding the initiating methionine residue of *LCT* can be found and is conserved in human, macaque, orang-utan and mouse; however the chimpanzee alignment reads TTG at this point, suggesting that it is incorrect for the initiation of translation. From correspondence with Ensembl, 2X assemblies (as was the chimpanzee genome at the time of writing) are built by projection to 'high coverage' genomes (i.e. human), and the Ns represent sequence that should be there by comparison. Whilst removal of the Ns from the chimpanzee genomic sequence makes the length of human and chimpanzee sequence more comparable, the same region containing TTG in the chimpanzee is aligned with the initiation codon of the human sequence. Since the intergenic region is identifiable and quite similar in the other species (66% identity between human and macaque and 85% identity between human and orang-utan), the

mis-alignment of the chimpanzee does seem likely in part to be caused by missing sequence, but the addition of an arbitrary number of Ns does not seem to improve alignment.

Sequence alignment of the intron 13 region was less problematic. Chimpanzee, orang-utan and macaque sequence for this region is publicly available. Gorilla sequence was available but only from exon 13 of *MCM6* until just past the OCT1 binding site within intron 13. To get a more comprehensive comparison, two gorilla DNA samples (PRI-GOR-13 & 14), three orang-utan samples (PRI-ORA-01, 05 & 07) and two chimpanzee DNA samples (PRI-CHI-01 & 02) were sequenced using the primers MCM6i13, MCM6778 and LAC-C-L2 (for details see chapter 2.7). The Ensembl alignment for human, chimpanzee and macaque was downloaded and the gorilla and orang-utan sequences were added (Figure 6.3) after alignment to a human reference sequence in ChromasPro. The chimpanzee and orang-utan alignment obtained by sequencing agreed with that provided in Ensembl.

	Chimpanzee	Orang-utan	Macaque	Gorilla
Exon 13	99.4	98.8	99.4	
Intron 13	87.2	95.8	91.6	
Intron 13 enhancer	*94.2*	*95.8*	*94.2*	*95.6*
Exon 14	100	100	100	
Intron 14	98.4	94.7	91.9	

Table 6.3 Pairwise percentage identity calculated between human and various primates for regions of the *MCM6* gene. The intron 13 enhancer sequence includes nucleotides -14133 to -13684 with respect to *LCT*. Full gorilla sequence was not publicly available, and only the enhancer region was sequenced in this study.

Table 6.3 shows percentage identity between species for exons 13 and 14 and introns 13 and 14 of *MCM6* calculated using the EMBOSS pairwise alignment needle algorithm, which compares similarity over the entire length of both sequences. The algorithm was also used to compare sequence identity of the 450bp enhancer region (base pairs -14133 to -13684 upstream of *LCT*).

The intron 13 enhancer region appears quite similar between species (Figure 6.3), although the gorilla sequence has a small (7bp) deletion in intron 13, just upstream of the enhancer. The alignment shows that the ancestral version of the sequence corresponding to the OCT1 binding site (which contains the *-13910C>T*, and

-13915T>G lactase persistence associated SNPs in humans) is present and conserved in all five species. The small region surrounding another lactase persistence associated SNP (*-14010G>C*) in humans is also conserved between humans, chimpanzee, gorilla and orang-utan. However, the macaque sequence contains a transition to an A allele at -14010. The mouse alignment shows much less conservation of intron 13, but the OCT1 binding site can be aligned and contains two nucleotide substitutions (appendix F).

All SNPs typed within this thesis were compared with these primate alignments or (for those present in the databases) checked in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) to infer the ancestral allele by comparison with primates.

6.4.2 Haplotype association of intron 13 alleles

Haplotypes were inferred using Phase. All 747 individuals were pooled giving more power for inference of the rare intron 13 SNPs. Figure 6.4 shows the inferred haplotype association of all intron 13 alleles for which full analysis was possible. Nearly all the alleles have a dominant association with a single *LCT* core haplotype (*-14025*G* with **A**; *-14010*C* with **B**; *-14009*G* with **X**; *-13915*G* with **C**; *-13910*T* with **A**; *-13907*G* with **A**; *-13806*G* with **C**). The exception to this is *-13913*C* which is inferred to be present on both the **A** and **B** haplotypes in almost equal proportions.

Phase output gives haplotypes for both chromosomes and produces a probability value associated with the inferred haplotypes. Twelve *-13913*C* carrying chromosomes were typed in total. Three individuals were heterozygous only at *-13913*C* and their phase was unambiguous (two **A** haplotypes and one **B**), three had phase probabilities >0.98 (two **A**s and one **B**), one had phase probability of 0.80 (**B**), four had probabilities of >0.6 and one had a phase probability of 0.40. The inferred haplotypes of each *-13913*C* carrying individual were inspected by eye in order to investigate the possibility that the haplotypes were incorrectly assigned by the inference software: It was considered that given the genotypes, the program had made the most probable haplotype assignment.

Four of the *-13913*C* haplotypes were identified in Fulani individuals, five from Middle Eastern populations (four Israeli non-Bedouin Arabs and one Saudi Bedouin) and one each from the Somali, Jaali and Afar populations. All five of the Middle Eastern *-13913*C* carrying chromosomes were inferred to be **A** haplotypes (the other haplotype of this type was found in a single Fulani individual), and all other *-13913*C* alleles were inferred to be carried on **B** haplotypes.

It was also noticed that whilst *-13907*G* is most frequently inferred to be present on an **A** haplotype chromosome (78% of *-13907*G* carrying chromosomes, n = 58), a substantial proportion (14%) are inferred to occur on **F**-like haplotypes. Most of the phase calls (50/58) had a probability of ≥ 0.95, 4 had a probability of ≥ 0.80, three ≥ 0.60 (all **F** haplotypes) and the single **B** haplotype had a phase probability of 0.58.

6.4.3 Haplotype Network

A simple network of the observed haplotypes is depicted in Figure 6.5. This network is a simplified version of the one shown in Hollox *et al.*, (2001) but with the newly identified intron 13 alleles added on the haplotype with which they are most commonly associated. The network was constructed by connecting each haplotype that differed at only a single position.

This haplotype network is clearly over-simplified. For example, although from observation of the markers genotyped here haplotype **F** could be a **B/A** recombinant, typing of an extra downstream marker *TG6236/7ΔΔ* in the Hollox study indicated this is not the case. However, in the absence of genotype information for this and other markers, the **F** haplotype remains unresolved from **I** and **d**, of which **I** is likely to be a **B/A** recombinant (see chapter 3 and Figure 3.2 for a full description of haplotype composition). The network constructed here is based only on known information and does not attempt to incorporate unknown data.

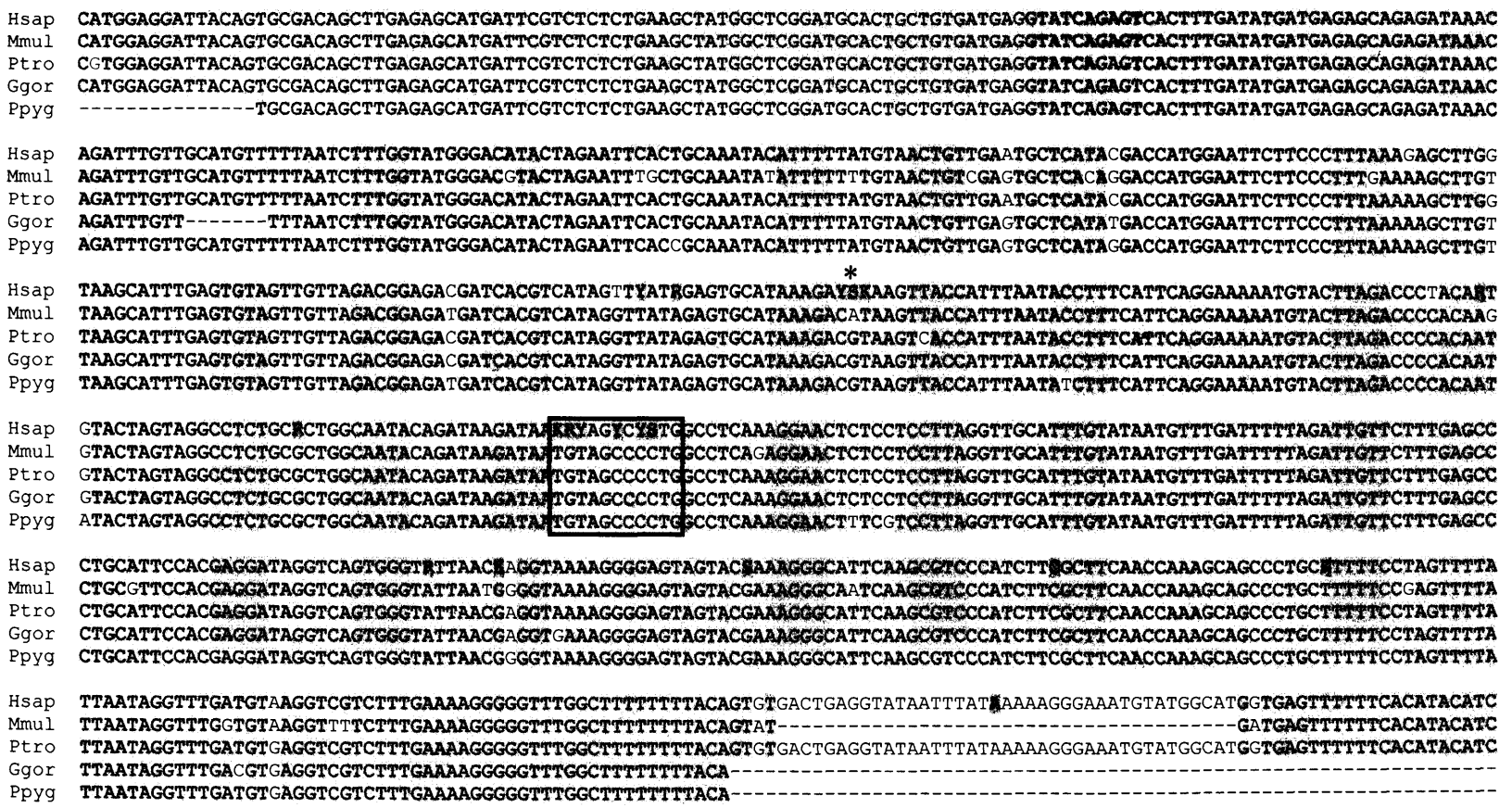


Figure 6.3 Alignment of the *MCM6* enhancer region. Exon 13 is indicated by red font. Conserved bases are highlighted in blue and known SNPs are highlighted in green with bases indicated according to the IUPAC Code (M = A or C, S = C or G, W = A or T, R = A or G, Y = C or T). Species aligned: Hsap, Human (Homo sapiens); Mmul, Macaque (Macaca mulatta); Ptro, Chimpanzee (Pan troglodytes); Ggor, Gorilla (Gorilla gorilla); Ppyg, Orang-utan (Pongo pygmaeus). The OCT1 binding site is outlined in dark blue, and position -14010 is indicated by an asterisk above the relevant base pair.

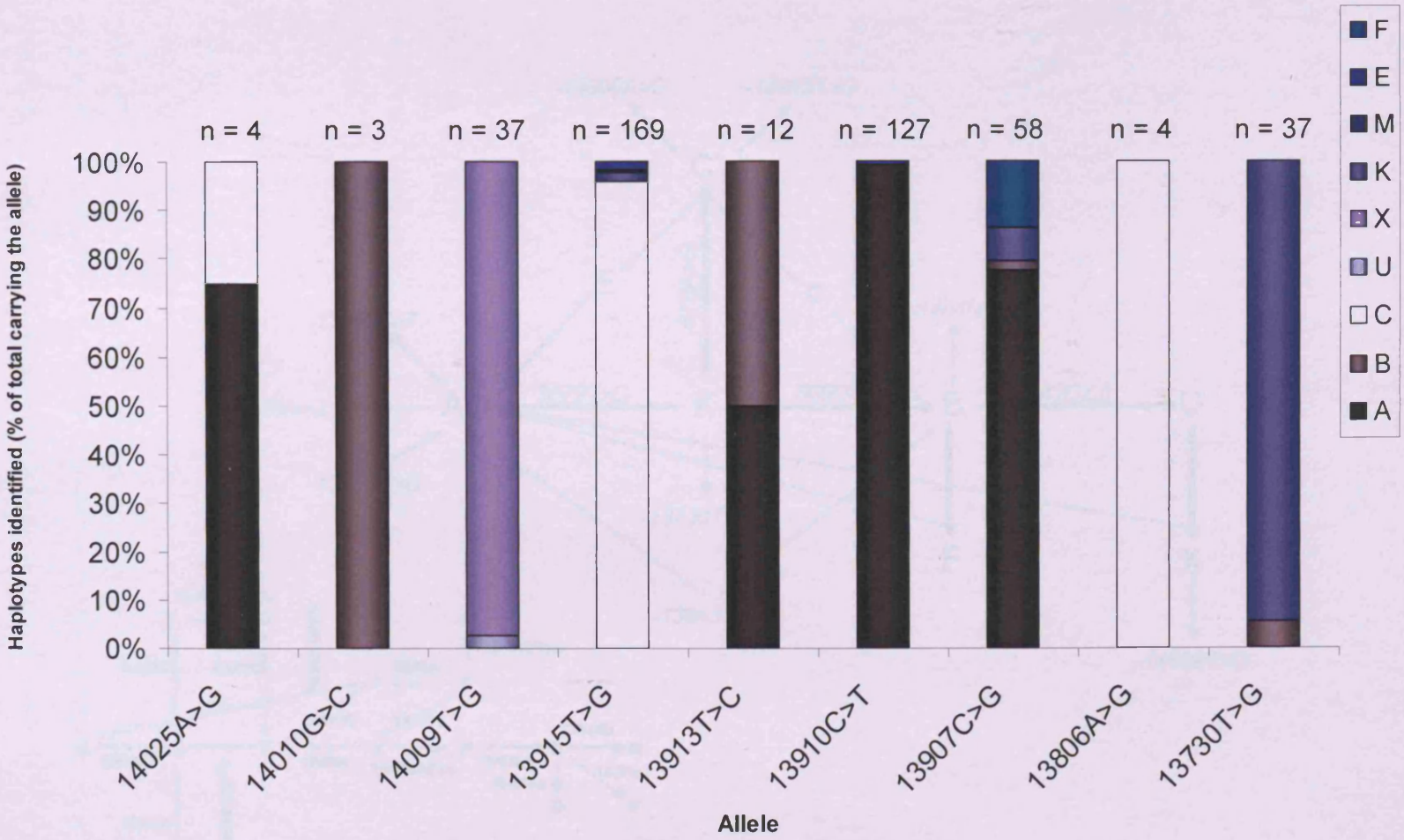


Figure 6.4 Haplotypic association of intron 13 alleles. Representation of the proportion of different haplotypes (different colours) that contribute to the total number of chromosomes observed carrying a particular intron 13 variant.. Number of chromosomes included is indicated above each bar.

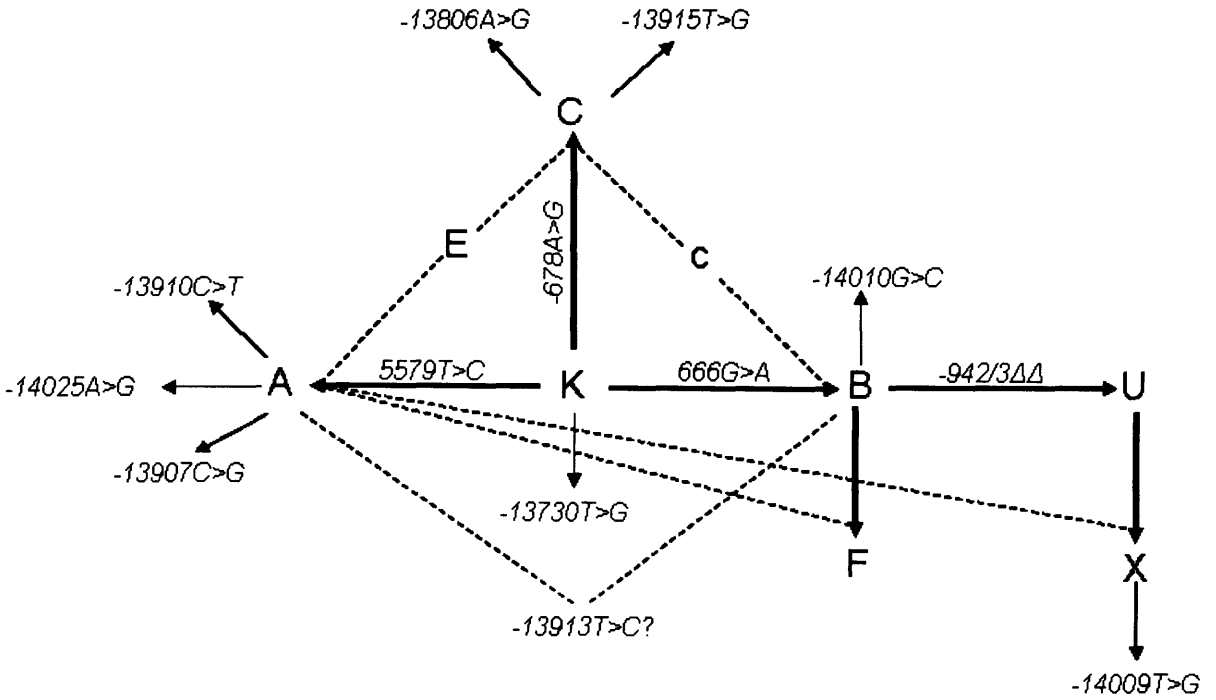


Figure 6.5 Haplotype network of observed *LCT* haplotypes. Solid lines indicate mutation, and dashed lines represent recombination between two haplotypes. Inset is the original haplotype network (Hollox *et al.*, 2001) which included more markers. *-13913*C* is indicated by dashed lines between haplotypes A & B, however the depiction of separate mutations from each haplotype would be equally valid.

6.5 Extending the haplotypes across intron 13

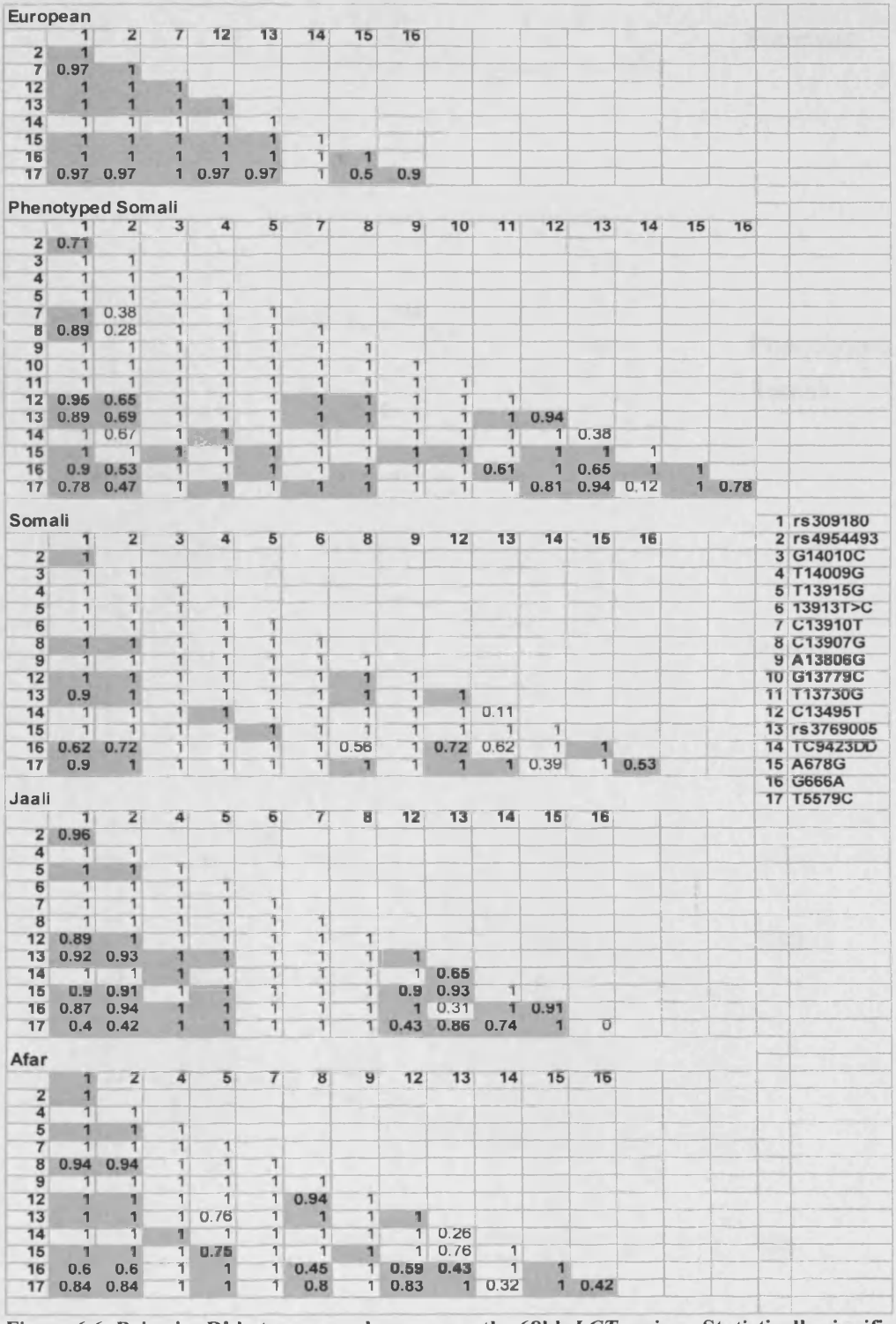
Prior to this thesis, SNPs used as markers for *LCT* haplotypes covered only the 60kb gene region. To extend the haplotypes so that they overlapped the intron 13 enhancer region we collected extra genotype information on two markers upstream (rs309180 and rs4954493) and two markers downstream (rs3769005 and rs4954490/*-13495C>T*) of the locus. These were located in introns 11, 12, 13 and 15 of *MCM6*. Inclusion of these markers extends the haplotypes to cover a distance of 68kb, from intron 11 of *MCM6* to exon 17 of *LCT*. The new markers were typed in a selection of the originally typed populations (European, Somali, Afar and Jaali, n = 358).

6.5.1 Linkage Disequilibrium

Pairwise linkage disequilibrium was calculated between the core *LCT* markers, the four extra markers and the *MCM6* intron 13 markers in all populations where this full data set was available (Figure 6.6) using DnaSP software (chapter 2.9).

The patterns of association appear to be consistent across populations. High D's are observed between the intron 13 SNPs (markers 3-12 in Figure 6.6) and other markers, although this is lower in the Afar group. rs4954490 (*-13495C>T*) and *5579T>C* (as well as other markers) show high D' with *-13907*G;* however *-13910*T* also has high D' with these SNPs. While there is often no evidence of recombination in the intron 13 region (D' = 1), the association between the intron 13 markers and their neighbours (Figure 6.7) or with more distant markers (Figure 6.6) is often not statistically significant because of low allele frequencies.

The neighbouring pair linkage disequilibrium plots (Figure 6.7) depict the slightly lower D' in the African populations than in the Europeans in the region of the core *LCT* markers, consistent with more recombination in these groups. Of note is the drop in D' and of significant association in all African populations between markers *-678A>G* and *-942/943TC>ΔΔ*, (markers 14 and 15, at ~15kb on Figure 6.8) due to the occurrence of all four gametes, as occurs in core haplotypes **A**, **B**, **C**, **U**, **X** and **o** (**U**, **X** and **o** not being found in Europeans).



European

	1	2	7	12	13	14	15	16
2	1							
7	0.97	1						
12	1	1	1					
13	1	1	1	1				
14	1	1	1	1	1			
15	1	1	1	1	1	1		
16	1	1	1	1	1	1	1	
17	0.97	0.97	1	0.97	0.97	1	0.5	0.9

Phenotyped Somali

	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
2	0.71														
3	1	1													
4	1	1	1												
5	1	1	1	1											
7	1	0.38	1	1	1										
8	0.89	0.28	1	1	1	1									
9	1	1	1	1	1	1	1								
10	1	1	1	1	1	1	1	1							
11	1	1	1	1	1	1	1	1	1						
12	0.95	0.65	1	1	1	1	1	1	1	1					
13	0.89	0.69	1	1	1	1	1	1	1	1	0.94				
14	1	0.67	1	1	1	1	1	1	1	1	1	0.38			
15	1	1	1	1	1	1	1	1	1	1	1	1	1		
16	0.9	0.53	1	1	1	1	1	1	1	0.61	1	0.65	1	1	
17	0.78	0.47	1	1	1	1	1	1	1	1	0.81	0.94	0.12	1	0.78

Somali

	1	2	3	4	5	6	8	9	12	13	14	15	16
2	1												
3	1	1											
4	1	1	1										
5	1	1	1	1									
6	1	1	1	1	1								
8	1	1	1	1	1	1							
9	1	1	1	1	1	1	1						
12	1	1	1	1	1	1	1	1					
13	0.9	1	1	1	1	1	1	1	1				
14	1	1	1	1	1	1	1	1	1	0.11			
15	1	1	1	1	1	1	1	1	1	1	1		
16	0.62	0.72	1	1	1	1	0.56	1	0.72	0.62	1	1	
17	0.9	1	1	1	1	1	1	1	1	1	0.39	1	0.53

1	rs309180
2	rs4954493
3	G14010C
4	T14009G
5	T13915G
6	13913T>C
7	C13910T
8	C13907G
9	A13806G
10	G13779C
11	T13730G
12	C13495T
13	rs3769005
14	TC9423DD
15	A678G
16	G666A
17	T5579C

Jaali

	1	2	4	5	6	7	8	12	13	14	15	16
2	0.96											
4	1	1										
5	1	1	1									
6	1	1	1	1								
7	1	1	1	1	1							
8	1	1	1	1	1	1						
12	0.89	1	1	1	1	1	1					
13	0.92	0.93	1	1	1	1	1	1				
14	1	1	1	1	1	1	1	1	0.65			
15	0.9	0.91	1	1	1	1	1	0.9	0.93	1		
16	0.87	0.94	1	1	1	1	1	1	0.31	1	0.91	
17	0.4	0.42	1	1	1	1	1	0.43	0.86	0.74	1	0

Afar

	1	2	4	5	7	8	9	12	13	14	15	16
2	1											
4	1	1										
5	1	1	1									
7	1	1	1	1								
8	0.94	0.94	1	1	1							
9	1	1	1	1	1	1						
12	1	1	1	1	1	0.94	1					
13	1	1	1	0.76	1	1	1	1				
14	1	1	1	1	1	1	1	1	0.26			
15	1	1	1	0.75	1	1	1	1	0.76	1		
16	0.6	0.6	1	1	1	0.45	1	0.59	0.43	1	1	
17	0.84	0.84	1	1	1	0.8	1	0.83	1	0.32	1	0.42

Figure 6.6 Pairwise D' between markers across the 68kb *LCT* region. Statistically significant values are indicated in bold and shaded in grey. A *p* value of 0.05 or lower was taken to be significant. Monomorphic markers were excluded from the analysis. Note that markers 14-17 are the core *LCT* haplotype markers. Pairwise D' between adjacent SNPs is shown on the diagonal, and this is depicted graphically in Figure 6.7.

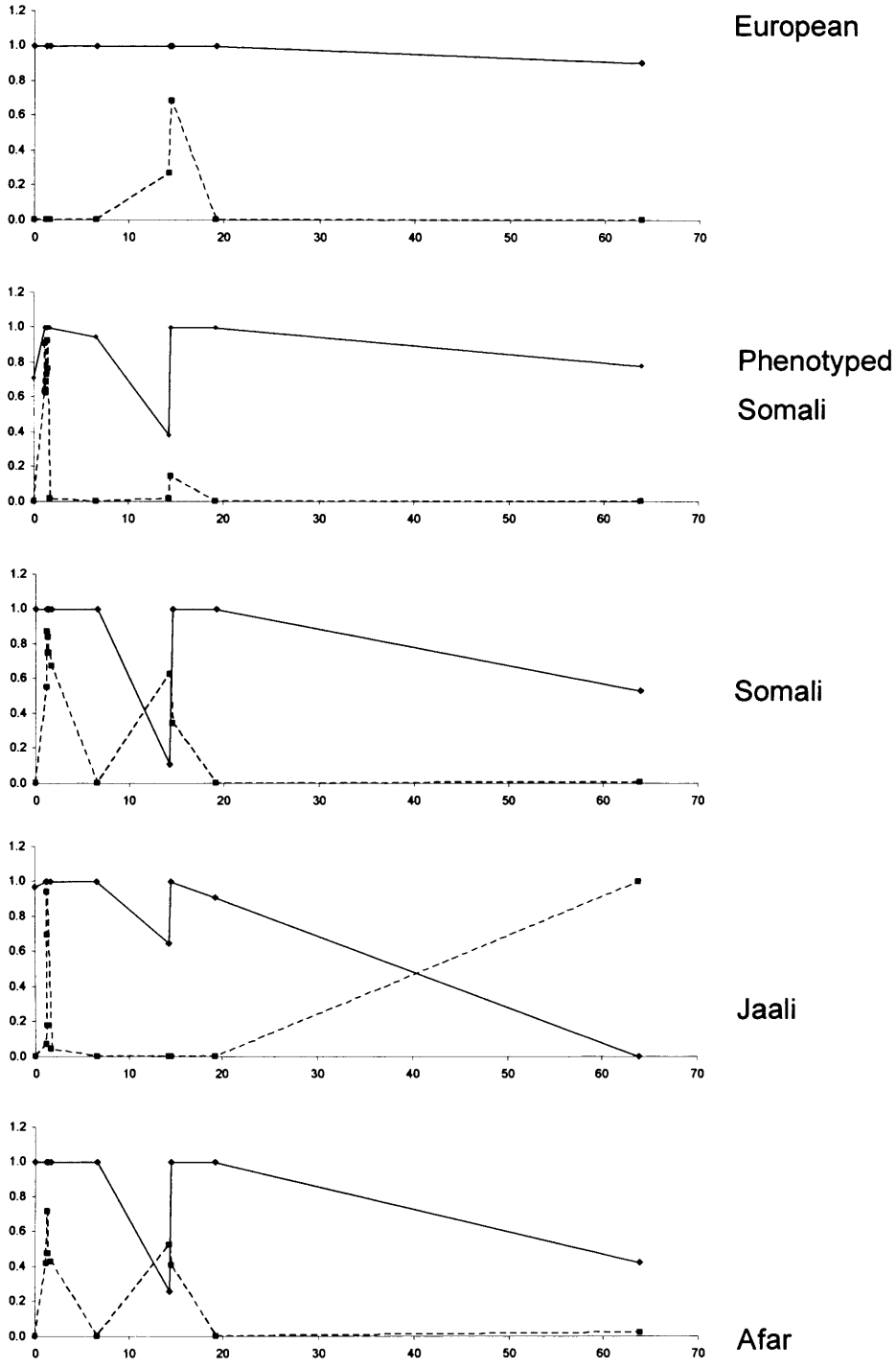


Figure 6.7 A graphical representation of the pairwise D' and *p* values for each marker with its neighbouring marker (the pairwise plotted D' values are given on the diagonal in Figure 6.6). D' values are shown by solid black lines, and corresponding *p* values are shown in dashed lines. The X axis shows distance between markers in kb.

As seen in Figure 6.6 the extra markers (1. rs309180, 2. rs4954493, 12. rs3769005 and 13. rs4954490) are significantly associated with the common *LCT* haplotype markers (14-17; *-942/943TC>ΔΔ, -678A>G, 666G>A, 5579T>C*. The allelic association of these extra markers with the derived intron 13 alleles and the core haplotype markers is shown in Figure 6.8. In most cases the derived alleles are all present on the same haplotype and where this is different it is mainly due to allelic decay at the ends of the haplotypes, possibly due to recombination.

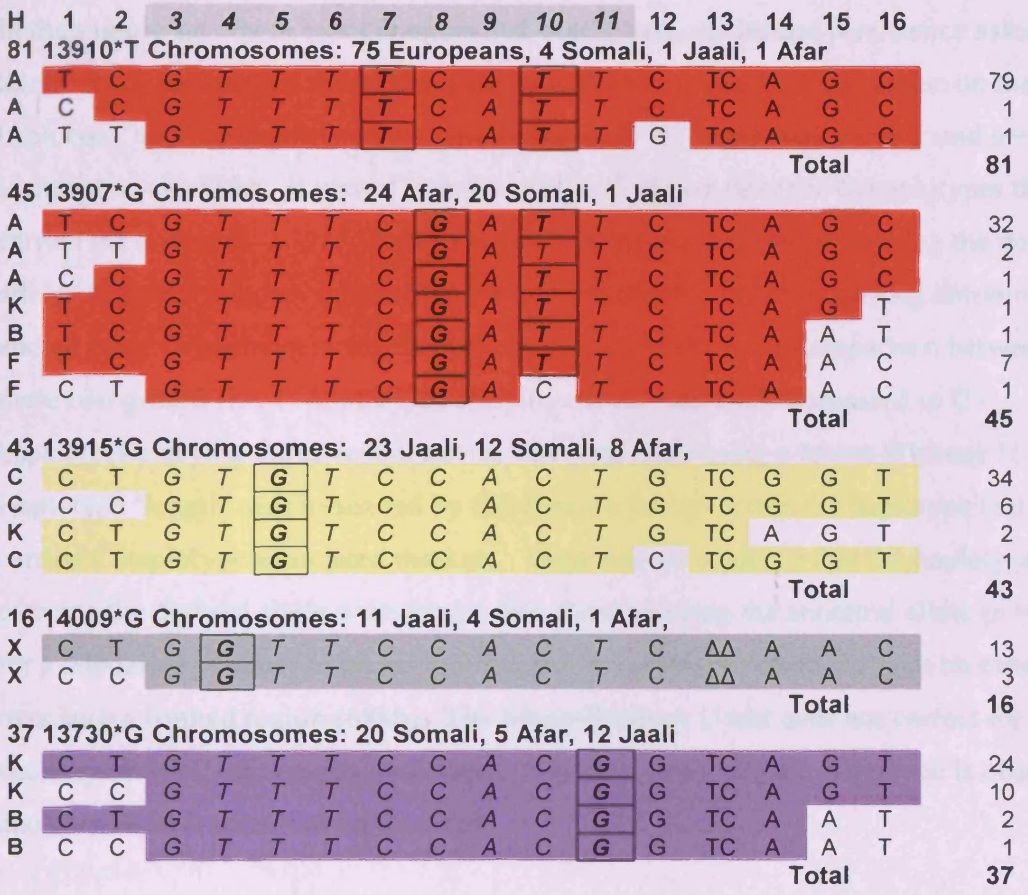


Figure 6.8 Shows inferred haplotypes (named according to the *LCT* core haplotypes) on which novel *MCM6* intron 13 alleles are carried. Only alleles for which full data existed for >10 chromosomes were included. The haplotypes were inferred using data for the following SNPs: 1 & 2; rs309180, rs4954493; 3-11; (intron 13 SNPs, indicated in italics and by grey shading), *-14010G>C, -14009T>G, -13915T>G, -13913T>C, -13910T>C, -13907C>G, -13806A>G, -13495C>T, -13730T>G*; 12; (located in intron 15 of *MCM6*) rs3769005 and markers 12-16 (which define core *LCT* haplotypes) *-942/943TC>ΔΔ, -678A>G, 666G>A, 5579T>C*. Outlined columns show derived alleles within intron 13.

6.5.2 Haplotype diversity of derived and non-derived alleles

When new alleles appear within a population they initially arise on a single chromosome and therefore haplotype. A neutral allele will rise to high frequency slowly, accruing mutation and undergoing recombination so that when it reaches high frequency it may be observed on a number of haplotypes. An advantageous allele will rise to high frequency rapidly and its associations with neighbouring alleles are more likely to stay intact, therefore the allele is likely to be associated with only a single haplotype.

In the sample set where extra markers had been typed, the lactase persistence associated allele which is observed most frequently is *-13915*G*. This SNP has arisen on the **C**-haplotype, and the majority of the chromosomes (79%) upon which it is found are identical across 68kb. It was of interest to investigate whether the **C**-haplotypes that carried the ancestral *-13915*T* allele were more diverse than those carrying the derived allele. Figure 6.9 shows a graphical comparison of all *-13915*G* carrying chromosomes and all other **C**-haplotypes which carry the *-13915*T* allele. A comparison between these two groups (i.e. **C**-haplotypes carrying the derived allele compared to **C**-haplotypes carrying the ancestral allele) was performed using a Mann-Whitney U test. Haplotype 'length' was measured by the distance (in bp) across the haplotype that carried **C** haplotype associated markers. There was no evidence that the haplotypes carrying the derived allele were longer than those carrying the ancestral allele ($p = 0.22$ for a one-tailed p value) however a difference in haplotype length may not be expected over such a limited region (68kb). The Mann-Whitney U test does not correct for shared ancestry of the **C**-haplotypes, however in this case no significant difference is observed and correction is therefore unnecessary.

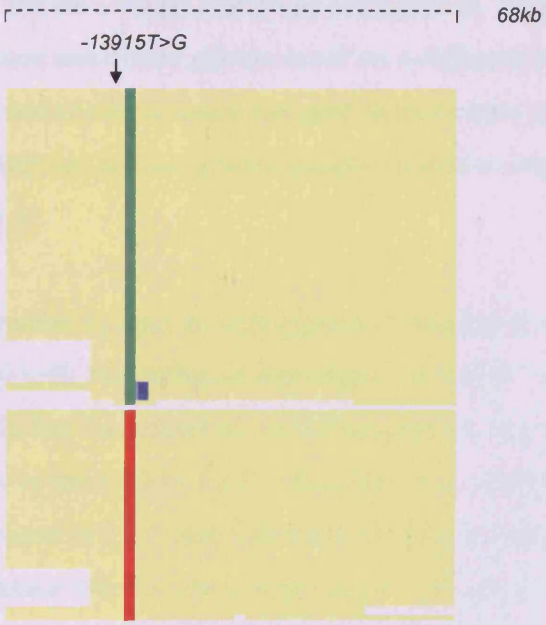


Figure 6.9 Graphical representation of C-haplotypes carrying either the ancestral *-13915*T* allele (green) or the derived *-13915*G* allele (red). Yellow colouring indicates identical alleles, and change of colour indicates the point at which haplotype identity breaks down. Blue indicates the *-13806*G* allele.

6.6 Discussion

In this chapter a survey of intron 13 allele frequencies was made from the available literature and my own data from which a geographic map was constructed by country. This gives an overview of the distribution of intron 13 alleles across Africa and Arabia and demonstrates the very different and sometimes isolated distributions of the alleles. For example *-14010*C* was observed in only 3/747 individuals included in this project, but was observed at frequencies up to 0.50 in several of the Kenyan and Tanzanian populations included in the original study (Tishkoff *et al.*, 2007). Visualisation of genetic distances between human populations (by principal co-ordinate analysis based on pairwise F_{ST}s) reveals that inclusion of the intron 13 alleles markedly increases population differentiation, whilst analysis using only *LCT* haplotype markers shows less clustering, and outlier populations are less obvious.

Investigation of the allelic associations of each of the intron 13 alleles shows that nearly all of them mainly fall on a single haplotype background. The observation that most of the lactase persistence associated alleles occur on a different haplotype implies that they are unlikely to be markers for a single untyped shared cause of lactase persistence, particularly as haplotypes are not genealogically related to one another in a simple way (shown in Figure 6.5).

*-13913*C* is an apparent exception with regards to haplotype association as it is inferred to associate equally with two different haplotypes; **A** and **B**. *-13913*C* was observed at highest frequency in the Cameroonian Fulani population, but was also observed in other groups, particularly in the Middle East. This allele was predominantly carried on a **B** haplotype chromosome in the Fulani and an **A** haplotype chromosome in the Middle East. It is hard to know whether this allele is observed on two separate haplotypes due to multiple mutations or because it is older and the haplotype association has been disrupted through a single recombination (between markers *-678A>G* and *666G>A*). Information for extra markers is not available for all *-13913*C* carrying chromosomes, so it is not known whether or not haplotypes upstream of *LCT* are the same as each other. Genotyping *-13495C>T* which occurs within intron 13 and is in strong association with **A** haplotype markers might aid resolution of this issue. West African populations (where it was thought *-13913*C* may be observed at higher frequencies) are not well represented within this study, and increased sample numbers may also help discriminate between the two possibilities. However ongoing research reveals that this allele is not observed at a frequency higher than 0.02 in a survey of 350 west African individuals comprising seven groups of Bantu-speaking peoples (B. Jones unpublished), so that both the likely origins and genealogy of the allele remain poorly understood.

To a lesser extent, *-13907*G* was also found to have association with two haplotypes. The allele is most commonly observed on the **A** haplotype, with identical allelic associations as observed for *-13910*T*, but it is also inferred to associate with the **F** haplotype chromosome in a proportion of people. Six of eight *-13907*G* **F** haplotype chromosomes occur in the Afar, the other two in the Somali. There are five other rare haplotypes, three are observed in the Afar, one in a Sudanese person and the other in an Amharic individual. These haplotypes differ only at the most 5' and 3' markers and so

appear most likely to be rare recombinants. By considering the allelic composition of haplotype **F** (see chapter 3, Figure 3.2) it can be seen that it may result from a single recombination between **A** and **X** haplotypes (between markers *-678A>G* and *666G>A*). Therefore a single historic recombination could explain the presence of this haplotype in the Ethiopian groups.

The observation that apparent recombinant *-13907*G* chromosomes occurs in Ethiopian populations may be a reflection of the higher genetic diversity present in this part of the world. In all the populations in which very extended haplotypes have been reported so far (*-13910*T* in Europe, (Poulter *et al.*, 2003; Bersaglieri *et al.*, 2004), *-13915*G* in the Middle East (Enattah *et al.*, 2008) and *-14010*C* in Kenya and Tanzania (Tishkoff *et al.*, 2007) the allele of interest has been present at a frequency of ≥ 0.5. It is possible that a combination of drift and selection have given rise to these extended haplotypes (Hollox *et al.*, 2001), and that the effect is heightened by the fact that recombination events are difficult to identify in a population which has low genetic diversity. Genotyping of distant markers may yet show the *-13907*G* haplotype to be extended in comparison to other haplotypes present at the locus in Ethiopian populations, if not in comparison to the other lactase persistence associated haplotypes.

Comparison of the 68kb haplotypes carrying the derived alleles showed that the majority of each haplotype (>70%) were identical across the region, consistent with findings of others who have reported extended haplotypes at the *MCM6* intron 13 locus (Tishkoff *et al.*, 2007; Enattah *et al.*, 2008). Interestingly, a slightly smaller percentage (65%) of the *-13730*G* carrying chromosomes seem to be conserved across the 68kb haplotype, compared to the haplotypes carrying the other novel alleles. This SNP is not a marker for lactase persistence, and it is located at the 3' end of the enhancer region. Furthermore, its occurrence at high frequency in a number of African populations (B. Jones personal communication) suggests that it is older than the other derived alleles.

Whilst the majority of *-13915*G* carrying chromosomes observed in this thesis are identical across the 68kb haplotype, it was found that this was not significantly different from **C** haplotypes carrying the ancestral *-13915*T* allele. Clearly the genomic distance tested is not sufficient to test for selection using haplotype homozygosity, but it does

indicate that the ancestral haplotypes might also be long. This was also observed in the Somali population (chapter 5), in which 40% of the chromosomes identified in the non-persistent group were identical across 68kb, and can also be observed in the haplotype homozygosity plots (for *-13907*G* and *-13915*G*) included in the supplementary information of Tishkoff *et al.*, (2007).

The linkage disequilibrium data presented here shows that *-13910*T* and *-13907*G* have strong allelic associations with the same markers but do not occur together. This indicates that they have arisen on the same haplotype lineage, consistent with the findings of Enattah *et al.*, (2008) and in contrast to previous claims that *-13907*G* was present on a distinct haplotype to the other intron 13 alleles (Tishkoff *et al.*, 2007). The finding that another intron 13 SNP (*-13495C>T*) is in strong linkage disequilibrium with both *-13907*G* and *-13910*T*, and the observation that *-13495*T* sometimes occurs with neither are good indicators that a haplotype carrying this allele was the common ancestor for both *-13907*G* and *-13910*T*.

The previously shown association of lactase persistence with some derived intron 13 alleles, and the very different population distributions and genealogies of these alleles strongly indicates that lactase persistence has evolved independently in these populations. However, the occurrence of these and other SNPs within such a small chromosomal region make it tempting to speculate that the mechanism by which the alleles confer lactase persistence is shared. Some of the observed alleles are extremely rare in the populations typed so far, making it difficult to conduct association studies. Therefore the region is an obvious candidate for functional studies which may further elucidate the significance of each variant with respect to *LCT* expression.

Alignment of human and primate genomic sequences across *LCT* and *MCM6* reveal that the sequence corresponding to the enhancer element is present in all species. However, the *MCM6* enhancer region is not particularly strongly conserved in comparison to other intron sequence from the same gene (Table 6.3), and as in other areas of the genome has lower percentage identity than between exons. However, experimentally identified protein binding sites within the enhancer region (Cdx-2, OCT1, HNF3α and HNF4α, Figure 5.1; Lewinsky *et al.*, 2005) are present in all species and are identical. It is

therefore possible that the region plays a similar role in regulation of *LCT* expression in both primates and in humans who have the ancestral allele. It is not unreasonable to suppose that disruption of this region due to the nucleotide substitutions that occur in some people may contribute to loss or alteration of function of the regulatory element, resulting in aberrant expression of *LCT* and lactase persistence. Information available regarding lactase persistence in primates is scant, and it would be of interest if this was better characterised, however one study reported 9/10 long-tailed or crab eating macaques (*Macaca fasicularis*) to be lactose digesters (ascertained by the blood glucose method), although all seven rhesus macaques (*Macaca mulatta)* tested were lactose mal-digesters (Wen *et al.*, 1973). This finding is particularly intriguing as a G>A transition is observed in rhesus macaque at position -14010, the site of the human lactase persistence associated *-14010*C* allele. A survey of persistence status in primates is unlikely to be available in the near future, but promoter-reporter construct assays (as conducted for many of the human variants (Olds and Sibley, 2003; Troelsen *et al.*, 2003; Tishkoff *et al.*, 2007) would at least inform on whether these two alleles (*-14010*C* and *-14010*A*) exert a similar effect on promoter function *in vitro*.

To investigate the possibility that the *MCM6* 'enhancer' region has regulatory function in primates as well as humans, conservation of the nucleotide sequence was examined (as discussed above). However spatial conservation between regulatory elements may also affect their efficacy (Vardhanabhuti 2007). The intergenic region between *MCM6* and *LCT* shows lower sequence similarity between species than either intron compared here (introns 13 and 14), with human and orang-utan percentage identity at 85%. However, alignment is possible and the distance between genes is similar in all species. Chimpanzee sequence for this region (which would be expected to be most similar to human) is incomplete, so comparison of the promoter region is not currently possible. However, the spatial similarity of *LCT* and *MCM6* between species and conservation of the genomic sequence across both genes does not provide any evidence to suggest that function of the putative regulatory region would not be conserved, at least in primates where it may originally have played a role in allowing expression over the critical period of weaning. In rodents (transgenic mice) there is evidence to suggest that normal down-regulation can be achieved with just 1kb of the pig, or 2kb of the rat promoter (Troelsen *et al.*, 1994; Lee *et al.*, 2002; see chapter 1.2.15).

General Discussion

At the outset of this project it was generally believed that lactase persistence was caused by a single nucleotide transition (Enattah *et al.*, 2002) occurring in a regulatory element (Troelsen *et al.*, 2003) 13kb upstream of the lactase gene, within an intron of the neighbouring gene. The mechanism by which this mutation was proposed to confer lactase persistence involved increased binding of the transcription factor OCT1 to an enhancer element (Lewinsky *et al.*, 2005). Although this provided a compact explanation, the reported absence of the putative causative allele in a number of African populations in which lactase persistence was common (Mulcare *et al.*, 2004), and the high mRNA expression of both *LCT* alleles in a heterozygote (Poulter *et al.*, 2003) indicated that something was amiss with the explanation.

One of the primary aims of this thesis was to use a phenotyped cohort from a sub-Saharan African population to investigate directly *-13910*T* status and also haplotype association with lactase persistence. *-13910*T* was observed in only one of 45 lactose digesters in a Sudanese cohort and no association of lactose digestion with the European persistence-associated haplotype was observed.

Haplotypic distribution was also investigated in a number of milk-drinking pastoralist populations from sub-Saharan Africa and the Middle East. A statistical procedure (Mulcare *et al.*, 2004) was applied to determine whether the frequency of the European associated haplotype was able to account for the published frequency of lactose digesters in matched groups. Despite taking into account phenotying and sampling errors, this analysis showed that there was insufficient representation of the European lactase persistence allele or haplotype to account for the observed frequency of lactose digesters. These findings led to the conclusion that *-13910*T* was not an associated marker for the lactase persistence haplotype and that lactase persistence had evolved independently in sub-Saharan Africa.

However, it was noted that in all of the groups included, the **C** haplotype was present at high frequencies, and that it was sufficient to explain lactase persistence in one of the pastoralist populations. It was found that this haplotype was also frequent enough to

account for the incidence of the trait, although it did not associate with lactase persistence in the Sudanese cohort. In fact, no haplotypic association was found with lactase persistence in the Sudanese cohort, but it was noted that one marker within the gene showed more association (although not statistically significant) with persistence than the others. These two findings (chapter 3 and Ingram *et al.*, 2007) – the high representation of the **C** haplotype in milk-drinkers and the slightly increased association of one of the haplotypic markers with lactase persistence in the phenotyped cohort, encouraged investigation for the existence of a *cis*-acting cause or marker of lactase persistence in the African and Middle Eastern groups.

The obvious candidate locus was the intron of *MCM6* in which the European variant had originally been found. Resequencing of this region in the phenotyped Sudanese cohort led to the identification of new alleles, one of which associated with lactase persistence. Functional analysis of these alleles using gel shifts showed that they disrupted rather than enhanced binding of the transcription factor OCT1 (chapter 4, Ingram *et al.*, 2007). This was recently confirmed by others (Enattah *et al.*, 2008), and implied that this transcription factor did not make a critical contribution to the sustained expression of *LCT*. In parallel work, other research groups reported identification of another allele (*-14010*C*) associated with lactase persistence in Tanzanian and Kenyan groups (Tishkoff *et al.*, 2007). Consistent with the assertion that increased OCT1 binding was not the key factor in determining persistence status, this allele was located 100bp upstream of the OCT1 recognition sequence. The occurrence of a persistence associated allele located at this distance from the other associated mutations which were more clustered (*-13910*T* and *-13915*G*) implied that the continued expression of lactase was not conferred by a single protein, and furthermore, indicated that the critical genomic region was at least 100bp in length. Notably *-14010*C* was contained within the 500bp segment shown to have enhancer effect *in vitro* (Troelsen *et al.*, 2003). With these findings in mind, along with the observation that other novel variants were present in the original cohort, DNA samples from a new cohort of phenotyped individuals collected from the Somali population in Ethiopia were sequenced across the entire enhancer region.

Resequencing in this cohort (chapter 5) showed that a great deal of genetic diversity existed in the digesters, significantly more than that observed in the mal-digester group. Association of *-13915*G* with lactase persistence was confirmed, and a significant association of *-13907*G* was also observed. Several new alleles were identified, one of which (*-14009*G*) had a borderline association with lactase persistence and affected the nucleotide immediately adjacent to the Kenyan/Tanzanian associated *-14010*C* allele. Curiously this allele has not been reported in other studies.

The stark contrast in genetic diversity at the intron 13 locus between the persistent and non-persistent members of the Somali cohort was also reflected in the haplotype diversity observed in the population which was not attributable to population sub-structure.

As previously alluded to (chapters 1 and 5), other research groups have attempted to demonstrate functional activity of the lactase persistence associated alleles. The 450bp region (from -14133 to -13684) surrounding *-13910C>T* was cloned into one plasmid containing the SI promoter and another containing the LPH proximal promoter (1kb of the 5' flanking sequence). These constructs were transfected into the intestinal cell line Caco2. It was found that the 450bp region had enhancer function in both constructs: A 20/50-fold increase in transcription was observed upon addition of the 450bp ancestral sequence to the SI and LPH proximal promoters respectively, and it was shown that the constructs carrying *-13910*T* were a further 30-50% more active (Troelsen *et al.*, 2003). These types of experiments have also been carried out for the newly discovered alleles: Tishkoff *et al.*, (2007) made a basal promoter construct containing 3kb of the 5' flanking LPH promoter and a number of constructs which included the entire 2kb ancestral sequence of *MCM6* intron 13 which varied at particular nucleotides. Five constructs were made in total; one carrying the ancestral sequence, four that differed from this at only one position: *-14010*C*, *-13915*G*, *-13907*G* or *-13495*T* and a final construct carrying *-13907*G* and the associated *-13495*T* allele together. It was shown that addition of the intron 13 sequence increased transcription of the reporter gene by 20 fold, and that addition of any one of *-14010*C*, *-13915*G* or *-13907*G* significantly increased enhancer activity but only by another 18-30%. Addition of *-13495*T* did not increase transcription in comparison to the ancestral sequence.

Unfortunately, no single lab has tested transcription activity of all alleles, and since the assays have not been comparable between labs the only general conclusion that can be made is that all the lactase persistence associated alleles increase transcription to a lesser or greater extent *in vitro*. The data obtained here in the Somali cohort (chapter 5) found some alleles in this region at low frequency in lactase non-persistent individuals only, and one SNP *-13730T>G* which was frequent in both the Sudanese and Somali cohorts and that was not associated with lactase persistence in either. Inclusion of all the alleles observed within the 450bp region in a reporter promoter construct assay would inform on two things: differential effects of the alleles on transcription *in vitro* and its efficacy for indicating function (*-13730T>G* and *–13806A>G*) providing at least two negative controls).

By using different fragments of the ancestral sequence of intron 13 reporter promoter construct assays could also be investigated as a way of narrowing the boundaries of the functional region: From the sequencing data presented here it seems that 3' boundary of the enhancer could be located upstream of *-13806A>G* and is certainly located upstream of *-13730T>G*, because these nucleotide substitutions do not associate with lactase persistence (although they could simply be functionally neutral). However, the upstream boundary is unknown. The lactase persistence associated *-14010*C* allele does not seem to be the 5'-most variant affecting *LCT* expression. Another allele at -14011 is reported in dbSNP (rs4988236) and its location adjacent to *-14010*C* makes it tempting to speculate that it is also functional. In addition, resequencing of European samples during this study revealed another substitution at *-14028T>A* within a Cdx-2 binding site. The individual in which this allele was identified had been previously shown to have two high expressing *LCT* alleles (by mRNA extraction from biopsy material), but was heterozygous at *-13910C>T* (Poulter *et al.*, 2003). A current project resequencing unphenotyped west African populations for intron 13 variation reveals more SNPs lying upstream of this one (B. Jones, personal communication).

There is another approach to investigating likely function of alleles which is analogous to association studies but requires only a few samples to inform on relative expression of the alleles in relation to *LCT* expression *in vivo*. This involves testing mRNA from

intestinal biopsy material of heterozygous (variant/ancestral) individuals and performing quantitative allele specific PCR on exonic markers on the cDNA obtained. This method has previously been successfully used to demonstrate high expression of the *-13910*T* carrying **A** haplotype in Europeans (Wang *et al.*, 1995). In Saudi Arabian populations *-13915T>G* genotypes have been shown to significantly correlate with lactase activity of biopsy material, although measurement of mRNA expression of the different allelic transcripts was not performed (Imtiaz *et al.*, 2007). Whilst these assays (in the former) confirm and (in the latter) imply a *cis*-acting change alters *LCT* expression in these individuals they do not prove causality of the alleles, which could be attributable to a linked marker, and neither do they inform on the mechanism by which increased expression is achieved.

The mechanism by which the intron 13 enhancer influences *LCT* expression appears to be complex: Many nucleotide substitutions seem to allow *LCT* to escape down-regulation following weaning, presumably by a shared mechanism, however not simply through altered interaction of a single transcription factor. Furthermore, the substitutions only affect phenotype post-weaning and, despite the efforts of many research groups, the mechanism by which *LCT* is down-regulated is not well understood (although it is thought to be mainly transcriptional) (Escher *et al.*, 1992; Rossi *et al.*, 1997; Wang *et al.*, 1998). Characterisation of the mechanism by which the intron 13 alleles confer lactase persistence could therefore provide a focus for future research.

The location of *LCT*, adjacent to the replication licensing factor *MCM6* gene may have relevance to its expression profile. Down-regulation of genes by active control (i.e. continual regulation of the expression state of each gene) is thought to be the prevalent mechanism used by eukaryotes to control the expression of differentiation-specific genes (for review see Blau, 1992). *MCM6* is preferentially expressed in the intestinal crypts and in foetuses (Sykes and Weiser, 1995; Harvey *et al.*, 1996), i.e. where *LCT* is not expressed, and *MCM6* is down-regulated along the crypt-villus axis at the point where *LCT* is up-regulated. This suggests that a heterochromatin-euchromatin boundary (for reviews see Bi and Broach, 2001; Gaszner and Felsenfeld, 2006) exists between *LCT* and *MCM6* and that a possible role for the *MCM6* enhancer region is to overcome chromatin compaction at a critical point in the differentiation pathway. It is possible that

after weaning, alteration in expression of one or more DNA binding proteins (discussed in Troelsen, 2005) and possibly changes in the rate of migration along the crypt-villus axis (Smith *et al.*, 1984; Smith *et al.*, 1986) results in reduced efficiency in overcoming heterochromatinisation and ultimately silencing of *LCT* along with *MCM6*. It is not unreasonable to suppose that introduction of the various intron 13 mutations might alter the affinities of a protein binding complex and also the status of *LCT* transcription.

There are various methods which could be utilised to investigate this hypothesis. Chromatin immunoprecipitation (ChIP) techniques (Horak and Snyder, 2002) are now routinely used to characterise protein-DNA interactions *in vivo*. To investigate possible differences in protein binding at the intron 13 locus, immunoprecipitation using antibodies against candidate transcription factors could be carried out in biopsy samples from individuals heterozygous for variant and ancestral chromosomes. Candidate transcription factors would include those empirically known to affect *LCT* expression as well as those with binding sites identified bioinformatically. ChIP could also be carried out using antibodies against particular chromatin states to inform on differences in transcriptional status of *LCT* in persistent and non-persistent samples. Allele specific primer extension PCR of the immuno-precipitated DNA could be used to examine differences in representation of one allele over another, providing evidence of function *in vivo* and circumventing the need for large quantities of biopsy material required by other methods.

Micro-arrays could also be used to examine differences in chromatin compaction and RNA expression in the region surrounding *LCT* in persistent and non-persistent samples. Biopsy material could be extracted from intestinal tissue of individuals homozygous for lactase persistence genotype. Chromatin immmunoprecipitation micro-array analysis (ChIP-on-chip) on a custom-made array spanning 1Mb around *LCT* would allow very high resolution information on changes in protein binding to other regions in close proximity to *LCT*, including the proximal promoter. Again, candidate proteins for immunoprecipitation would be those already shown to have function with respect to *LCT* expression, as well as those identified bioinformatically, and proteins which inform on chromatin state. Expression studies using mRNA extracted from the same intestinal material could be used to identify differences in expression profiles of adjacent genes in

the region, and would also be able to detect and examine expression patterns of non-coding RNAs, which have recently been reported to play a functional role in recruiting transcription factors which confer a particular chromatin state (Rinn *et al.*, 2007).

Chromosome conformation capture (3C; Dekker *et al.*, 2002) is another technique which could conceivably be used to inform on function of the intron 13 enhancer. Over recent years a number of studies have revealed that enhancers, locus control elements and promoters are in contact with one another when genes are being transcribed. The 3C technique informs on whether two chromosomal segments are in close proximity within the nucleus. This method also utilises chemical cross linking of protein-DNA interactions, and tethers interacting chromosomal segments together via their shared interactions with a protein complex. Endonuclease digestion fragments the genomic DNA which is then treated with DNA ligase before reversing the cross-links. Locus specific primers are used to detect the ligation product by quantitative PCR. Ligation of cross-linked loci is strongly favoured over ligation of random fragments at very low DNA concentrations, and so will be more frequently represented in the quantitative PCR. This method could be carried out in undifferentiated and differentiated Caco2 cells as well as in biopsy samples from lactase persistent and non-persistent individuals (homozygotes) to examine whether there is interaction of the intron 13 enhancer with the *LCT* proximal promoter and whether the interaction changes with increased/decreased *LCT* expression.

The research presented within this thesis has gone some way in describing the current day distribution of lactase persistence alleles. The geographic location of the earliest *-13910*T* carrying populations and the mode and direction of spread of the allele remain to be fully understood, however a general consensus in the literature seems to be that the allele rose to high frequency in the Neolithic European populations, spreading throughout Europe alongside the cultural practise of dairying (Mulcare *et al.*, 2004; Burger *et al.*, 2007; Enattah *et al.*, 2007). The presence of *-13910*T* in the N. African Berber populations has been suggested to result from migration of Middle Eastern pastoralists (Myles *et al.*, 2005), however the population distribution data presented here suggests that the source of *-13910*T* in the Berber is more likely to have been European. The presence of *-13910*T* in the Fulani population

of Cameroon may be due to historic interactions between these two African populations, as documented in the oral traditions of the Fulani (Adebayo, 1991).

*-13915*G* was originally identified in Sudan. However, from population frequency data it seems likely that this allele originated in the Middle East, and its high frequency in the Bedouin suggests it was strongly selected for in the nomadic pastoralist groups. The allele appears to have spread south through Sudan, where it is observed in many of the groups who claim part-Arabic descent (Stanton, 1903; Robinson, 1927; Warburg, 1978). Interestingly, *-13915*G* is also observed in west Africa in the Shuwa Arabs of the Chad basin, who are said to have migrated across the Sahel corridor from Sudan in the 14th Century (Levy and Holl, 2002).

Another of the lactase persistence alleles seems to represent a localised 'private' adaptation to dairying. The *-14010*C* allele discovered at high frequencies in Kenyan and Tanzanian populations (Tishkoff *et al.*, 2007), was extremely rare in the east African populations typed within this thesis. *-13907*G* seems to be more widespread (reported in all African groups genotyped within this thesis and also present in Kenya and Tanzania (Tishkoff *et al.*, 2007) but has so far been observed at a highest frequency of 0.3 in Ethiopia. Future studies involving resequencing of a number of individuals from various populations across Africa should provide greater insight into the distribution and origins of these alleles and in conjunction with other genetic systems may perhaps inform on the origins and migration of, and gene-flow between, different pastoralists societies within Africa.

The evidence that lactase persistence has provided a significant adaptive advantage to some human populations is demonstrated here by the finding that multiple causes of the phenotype occur at high frequency in different populations. Work by others demonstrates that some of the alleles show a strong signature of positive selection, due to their occurrence on very extended invariant haplotypes juxtaposed with their high frequency (>0.5 within the population; Poulter *et al.*, 2003; Bersaglieri *et al.*, 2004; Tishkoff *et al.*, 2007; Enattah *et al.*, 2008). In the two phenotyped cohorts collected here a number of different lactase persistence alleles were observed. This was particularly marked in the Somali group who were selected as known milk-drinking pastoralists and

may reflect gene flow from other pastoralist groups (See chapter 2.1.12). Curiously lactose tolerance frequency was low in this group (0.24), although drinking fresh milk is a widespread practise (Murdock, 1967; Blench, 1999). This observation seems to imply that lactase persistence was not especially advantageous for individuals living in this pastoralist society. Therefore one of the fundamental questions remaining to be answered is the precise nature of the adaptive advantage lactase persistence provided to some pastoralists groups that was not particularly beneficial to others.

It is clear that cultural adaptation, i.e. fermentation of milk-products to reduce lactose content, or regulation of the quantity of milk consumed at one time may have reduced the selective advantage of lactase persistence in some pastoralist groups. It is also possible that in some populations the gut flora may adapt to enable consumption of fresh milk by non-persistent individuals with no adverse side-effects (such as diarrhoea and dehydration), reducing the selection for lactase persistence. This may be particularly pertinent to the Somali as no correlation between lactose tolerance status and milk drinking habits were observed in this cohort and the balance of hydrogen producing bacteria seemed to be quite different from the Sudanese cohort (in whom lactase persistence status and milk consumption were highly correlated). Alternatively, the Somali population may not have depended heavily enough on fresh milk historically to provide a strong selective advantage. Interdisciplinary research into the histories of different pastoralist populations and the past environmental pressures endured (e.g. drought or famines) along with anthropological investigation of milk production and consumption practises as well as measurement of lactose content of locally produced dairy products may aid clarification of the circumstances under which lactase persistence confers a strong selective advantage.

One of the other observations which can be made for the data presented here (and which is also true in the data sets presented by others (Tishkoff *et al.*, 2007)) is that a number of lactase persistent individuals in both cohorts carried no candidate causal allele in the intron 13 region. This observation was mirrored by the total lack of any intron 13 allele in the Senegal Wolof group sampled here, a population which has been previously reported to have a lactase persistence allele frequency of 0.29 (calculated from frequency of lactose digester phenotype; Arnold *et al.*, 1980). These observations imply

that there are unidentified factors affecting lactase persistence, and these may not necessarily all be due to *cis*-acting variation. There is experimental evidence that suggests that additional levels of control over expression of the enzyme exist (Rossi *et al.*, 1997), and heterogeneity of the lactase non-persistence phenotype has been reported by a number of research groups. Some studies observed individuals who show slower/abnormal processing (Sterchi *et al.*, 1990; Witte *et al.*, 1990) which may imply post-translational controls such as glycosylation and/or transportation, whilst others have made observations suggestive of epigenetic regulation. Although most non-persistent individuals show no immuno-histological staining for lactase in the jejunal biopsies of the small intestine (concordant with low lactase activity and transcriptional regulation of *LCT*), it was found by (Maiuri *et al.*, 1991) that some individuals show patchy expression of the enzyme in the intestinal epithelia. This mosaic expression pattern may result from somatic cell changes in methylation, or histone acetylation.

The data presented here shows that most of the alleles in *MCM6* intron 13 have occurred on independent haplotypes, indicating that they are unlikely to represent linked markers of a single cause of lactase persistence. However, two of the alleles *-13907*G* and *-13910*T* do represent recent mutations that have occurred independently on the same ancestral haplotype, and the allele which shows a borderline association with lactase persistence (*-14009*G*) is also present on a possible **A**-recombinant chromosome which becomes **A**-like somewhere within *LCT* (all upstream markers alleles are typical of **U**). This means that for these three alleles a common downstream cause of lactase persistence is conceivable. Genotyping of further haplotypic markers including distant markers and microsatellites would allow higher resolution of the haplotypes on which the alleles occur, allowing greater understanding of their evolutionary lineages and would also permit dating of the intron 13 alleles. Micro-satellite markers could also be used to inform on selection of the *LCT* haplotypes as has been previously carried out for *-13910*T* (Coelho *et al.*, 2005).

In summary, it has been shown that lactase persistence has evolved independently in different populations on numerous occasions, providing strong evidence of selection for the trait. This is analogous to G6PD deficiency, or thalassemia (see chapter 1) where different alleles that confer the same selective advantage have been selected for in

different locations. Some of the *LCT/MCM6* alleles show extended haplotype homozygosity (Tishkoff *et al.*, 2007; Enattah *et al.*, 2008), but genetic analysis for signatures of selection is complicated in populations in which several of the alleles co-exist. It is worth noting that analyses that involve measuring nucleotide diversity taken from sequence data are 'blind' to underlying haplotype differences.

Until recently SNPs that have been shown to cause functional differences have mainly been identified in the protein coding regions of genes, and lead to amino acid substitutions, as is the case with most of the examples in chapter 1, but it is now becoming clear that mutations in regulatory elements have played an important role in generating phenotypic differences e.g. expression of the Duffy antigen (chapter 1), or variation in eye colour (Eiberg *et al.*, 2008; Sturm *et al.*, 2008). Functional variants in regulatory elements are harder to identify than exonic changes, and often involve extensive resequencing; distant elements are harder to identify still. However, our understanding of the involvement of these elements in human disease is improving (Kleinjan and van Heyningen, 2005; Zeller and Zuniga, 2007), and their contribution to evolutionary adaptation and speciation may be widespread, although their importance is debated (Hoekstra and Coyne, 2007; Wray, 2007). Thus *LCT* provides a useful model in which to study the role of distant regulatory elements in human adaptation.

References

Adebayo, A.G. (1991). Of man and Cattle: A reconsideration of the traditions of origin of pastoral fulani of Nigeria. *History of Africa.* **18,** 1-21.

Allen, S.J., O'Donnell, A., Alexander, N.D., Alpers, M.P., Peto, T.E., Clegg, J.B. and Weatherall, D.J. (1997). alpha+-Thalassemia protects children against disease caused by other infections as well as malaria. *Proc. Natl. Acad. Sci U. S. A.* **94,** 14736-14741.

Allison, A.C. (1954). Notes on sickle-cell polymorphism. *Ann. Hum Genet.* **19,** 39-51.

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

Andres, A.M., Clark, A.G., Shimmin, L., Boerwinkle, E., Sing, C.F. and Hixson, J.E. (2007). Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet Epidemiol.* **31,** 659-671.

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

Arnold, J., Diop, M., Kodjovi, M. and Rozier, J. (1980). Lactose intolerance in adults in Senegal. *C R. Seances Soc. Biol Fil.* **174,** 983-992.

Arribas, J.C.D., Herrero, A.G., Martin-Lomas, M., Canada, F.J., He, S.M. and Withers, S.G. (2000). Differential mechanism-based labeling and unequivocal activity assignment of the two active sites of intestinal lactase/phlorizin hydrolase. *Eur. J. Biochem.* **267,** 6996-7005.

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

Barnwell, J.W., Nichols, M.E. and Rubinstein, P. (1989). In vitro evaluation of the role of the Duffy blood group in erythrocyte invasion by Plasmodium vivax. *J. Exp. Med.* **169,** 1795-1802.

Barreiro, L.B., Laval, G., Quach, H., Patin, E. and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40,** 340-345.

Barsh, G.S. (2003). What controls variation in human skin color? *Plos Biology.* **1,** 19-22.

Batzer, M.A., Arcot, S.S., Phinney, J.W., Alegria-Hartman, M., Kass, D.H., Milligan, S.M., Kimpton, C., Gill, P., Hochmeister, M., Ioannou, P.A. et. al., (1996). Genetic variation of recent Alu insertions in human populations. *J. Mol. Evol.* **42,** 22-29.

Bayoumi, R.A., Flatz, S.D., Kuhnau, W. and Flatz, G. (1982). Beja and Nilotes: nomadic pastoralist groups in the Sudan with opposite distributions of the adult lactase phenotypes. *Am. J. Phys. Anthropol.* **58,** 173-178.

Bayoumi, R.A., Saha, N., Salih, A.S., Bakkar, A.E. and Flatz, G. (1981). Distribution of the lactase phenotypes in the population of the Democratic Republic of the Sudan. *Hum. Genet.* **57,** 279-281.

Beck, F. (2004). The role of Cdx genes in the mammalian gut. *Gut.* **53,** 1394-1396.

Beja-Pereira, A., Luikart, G., England, P.R., Bradley, D.G., Jann, O.C., Bertorelle, G., Chamberlain, A.T., Nunes, T.P., Metodiev, S., Ferrand, N. and Erhardt, G. (2003). Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nat. Genet.* **35,** 311-313.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E. and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111-1120.

Besteman, C. (1993). Public History and Private Knowledge: On disputed History in Southern Somalia. *Ethnohistory.* **40,** 563-586.

Beutler, E. (1994). G6Pd Deficiency. *Blood.* **84,** 3613-3636.

Bi, X. and Broach, J.R. (2001). Chromosomal boundaries in S. cerevisiae. *Curr Opin. Genet Dev.* **11,** 199-204.

Birdsey, G.M., Lewin, J., Cunningham, A.A., Bruford, M.W. and Danpure, C.J. (2004). Differential enzyme targeting as an evolutionary adaptation to herbivory in carnivora. *Mol. Biol. Evol.* **21,** 632-646.

Blau, H.M. (1992). Differentiation requires continuous active control. *Annu. Rev. Biochem.* **61,** 1213-1230.

Blench R (1999). Why are there so many pastoral groups in eastern Africa? *In: Pastoralists under pressure? Fulbe societies confronting change in west Africa.* Azarya V, Breedveld A, De Bruijn M, Van Dijk H (eds). *Brill Press, Boston. U. S. A.*

Bloom, G. and Sherman, P.W. (2005). Dairying barriers affect the distribution of lactose malabsorption. *Evol. Hum. Behav.* **26,** 301-312.

Boll, W., Wagner, P. and 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.

Bond, J.H. and Levitt, M.D. (1976). Quantitative measurement of lactose absorption. *Gastroenterology.* **70,** 1058-1062.

Bosse, T., van Wering, H.M., Gielen, M., Dowling, L.N., Fialkovich, J.J., Piaseckyj, C.M., Gonzalez, F.J., Akiyama, T.E., Montgomery, R.K., Grand, R.J. and Krasinski, S.D. (2006). Hepatocyte nuclear factor-1alpha is required for expression but dispensable for histone acetylation of the lactase-phlorizin hydrolase gene in vivo. *Am. J. Physiol Gastrointest. Liver Physiol.* **290,** 1016-1024.

Boudreau, F., Rings, E.H.H.M., van Wering, H.M., Kim, R.K., Swain, G.P., Krasinski, S.D., Moffett, J., Grand, R.J., Suh, E.R. and Traber, P.G. (2002). Hepatocyte nuclear factor-1 alpha, GATA-4, and caudal related homeodomain protein Cdx2 interact functionally to modulate intestinal gene transcription - Implication for the developmental regulation of the sucrose-isomaltase gene. *J. Biol. Chem.* **277,** 31909-31917.

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

Brookes A.J (2005). Single Nucleotide Polymorphisms. *In: Nature Encyclopaedia of Life Sciences.* John Wiley & Sons, Ltd: Chichester http://www.els.net/ [doi: 10.1038/npg.els.0005006].

Brown, W.M., George, M., Jr. and Wilson, A.C. (1979). Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 1967-1971.

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

Burch, J.B. (2005). Regulation of GATA gene expression during vertebrate development. *Semin. Cell Dev. Biol.* **16,** 71-81.

Burger, J., Kirchner, M., Bramanti, B., Haak, W. and Thomas, M.G. (2007). Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl. Acad. Sci. USA.* **104,** 3736-3741.

Caldwell, E.F. (2005) *Molecular Evidence for Dietary Adaptation in Humans,* PhD. Thesis, University of London, London.

Caldwell, E.F., Mayor, L.R., Thomas, M.G. and Danpure, C.J. (2004). Diet and the frequency of the alanine:glyoxylate aminotransferase Pro11Leu polymorphism in different human populations. *Hum. Genet.* **115,** 504-509.

Cappellini, M.D. and Fiorelli, G. (2008). Glucose-6-phosphate dehydrogenase deficiency. *Lancet.* **371,** 64-74.

Carrington, M., Kissner, T., Gerrard, B., Ivanov, S., O'Brien, S.J. and Dean, M. (1997). Novel alleles of the chemokine-receptor gene CCR5. *Am. J. Hum. Genet.* **61,** 1261-1267.

Carter, R. and Mendis, K.N. (2002). Evolutionary and historical aspects of the burden of malaria. *Clin. Microbiol. Rev.* **15,** 564-594.

Cavalli-Sforza L.L., Menozzi P. and Piazza A. (1994) The History and Geography of Human Genes. *Princeton University Press, Princeton, New Jersey.*

Cerny, V., Hajek, M., Bromova, M., Cmejla, R., Diallo, I. and Brdicka, R. (2006). MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. *Hum Biol.* **78,** 9-27.

Christopher, N.L. and Bayless, T.M. (1971). Role of the small bowel and colon in lactose-induced diarrhea. *Gastroenterology.* **60,** 845-852.

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

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

Cole, D.P. (2003). Where Have the Bedouin Gone? *Anthropological Quarterly.* **76,** 235-267.

Columbo, V., Lorenz-Meyer, H. and Semenza, G. (1973). Small intestinal phlorizin hydrolase: the β-glycosidase complex. *Biochim. Biophys. Acta.* **327,** 412-424.

Cook, G.C. (1988). Human intestinal lactase and lamarckian evolution. *Lancet.* **2,** 1029.

Cook, G.C. and al-Torki, M.T. (1975). High intestinal lactase concentrations in adult Arabs in Saudi Arabia. *Br. Med. J.* **3,** 135-136.

Crisp, E.A., Czolij, R. and Messer, M. (1987). Absence of beta-galactosidase (lactase) activity from intestinal brush borders of suckling macropods: implications for mechanism of lactose absorption. *Comp Biochem Physiol B.* **88,** 923-927.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., stro-Bisol, G., Coia, V. et. al., (2002). A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am. J. Hum. Genet.* **70,** 1197-1214.

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

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

Dalby A. (1998) Dictionary of Languages. *Bloomsbury Publishing plc, London.*

Danpure, C.J., Fryer, P., Jennings, P.R., Allsop, J., Griffiths, S. and Cunningham, A. (1994). Evolution of alanine:glyoxylate aminotransferase 1 peroxisomal and mitochondrial targeting. A survey of its subcellular distribution in the livers of various representatives of the classes Mammalia, Aves and Amphibia. *Eur. J. Cell Biol.* **64,** 295-313.

Danpure, C.J., Guttridge, K.M., Fryer, P., Jennings, P.R., Allsop, J. and Purdue, P.E. (1990). Subcellular distribution of hepatic alanine:glyoxylate aminotransferase in various mammalian species. *J. Cell Sci.* **97 (Pt 4),** 669-678.

Day, A.J., Canada, F.J., Diaz, J.C., Kroon, P.A., Mclauchlan, R., Faulds, C.B., Plumb, G.W., Morgan, M.R.A. and Williamson, G. (2000). Dietary flavonoid and isoflavone glycosides are hydrolysed by the lactase site of lactase phlorizin hydrolase. *FEBS Lett.* **468,** 166-170.

de Silva, S.E. and Stumpf, M.P. (2004). HIV and the CCR5-Delta32 resistance allele. *FEMS Microbiol. Lett.* **241,** 1-12.

de Smith, A.J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N.A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R.J., Bruhn, L. et. al., (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet.* **16,** 2783-2794.

Dean, A. (2006). On a chromosome far, far away: LCRs and gene expression. *Trends Genet.* **22,** 38-45.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002). Capturing chromosome conformation. *Science.* **295,** 1306-1311.

Diamond, J. (2005). Evolutionary biology: geography and skin colour. *Nature.* **435,** 283-284.

Dissanyake, A.S., El-Munshid, H.A. and Al-Qurain, A. (1990). Prevalence of primary adult lactose malabsorption in the eastern provine of Saudi Arabia. *Ann. Saudi Med.* **10,** 598-601.

Eaton, S.B. and Konner, M. (1985). Paleolithic nutrition. A consideration of its nature and current implications. *N Engl J Med.* **312,** 283-289.

Ehret, C. (1979). Antiquity of Agriculture in Ethiopia. *Journal of African History.* **20,** 161-177.

Ehret, C. (2001). Bantu Expansions: Re-Envisioning a Central Problem of Early African History. *The International Journal of African Historical Studies.* **34,** 5-41.

Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K.W. and Hansen, L. (2008). Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123,** 177-187.

Enattah, N.S., Jensen, T.G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J.K., Alifrangis, M., Khalil, I.F. et. al., (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am. J. Hum. Genet.* **82,** 57-72.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. and Jarvela, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30,** 233-237.

Enattah, N.S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., Greco, L., Rossi, M., Lentze, M., Seo, J.K., Rahgozar, S. et. al., (2007). Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am. J Hum Genet.* **81,** 615-625.

Escher, J.C., de Koning, N.D., van Engen, C.G., Arora, S., Buller, H.A., Montgomery, R.K. and Grand, R.J. (1992). Molecular basis of lactase levels in adult humans. *J Clin. Invest.* **89,** 480-483.

Excoffier, Laval, L.G. and Schneider, L. (2005). Arlequin ver.3.0: An Integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online.* **1,** 47-50.

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics.* **131,** 479-491.

Faerman, M., Filon, D., Kahila, G., Greenblatt, C.L., Smith, P. and Oppenheim, A. (1995). Sex identification of archaeological human remains based on amplification of the X and Y amelogenin alleles. *Gene.* **167,** 327-332.

Fang, R., Olds, L.C., Santiago, N.A. and Sibley, E. (2001). GATA family transcription factors activate lactase gene promoter in intestinal Caco-2 cells. *Am. J. Physiol Gastrointest. Liver Physiol.* **280,** 58-67.

Fang, R., Santiago, N.A., Olds, L.C. and Sibley, E. (2000). The homeodomain protein Cdx2 regulates lactase gene promoter activity during enterocyte differentiation. *Gastroenterology.* **118,** 115-127.

Fay, J.C. and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. *Genetics.* **155,** 1405-1413.

Ferguson, A. and Maxwell, J. (1967). Genetic aetiology of lactose intolerance. *Lancet.* **2,** 188-190.

Fitzgerald, K., Bazar, L. and Avigan, M.I. (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 Rotthauwe, H.W. (1973). Lactose nutrition and natural selection. *Lancet.* **2,** 76-77.

Flint, J., Hill, A.V., Bowden, D.K., Oppenheimer, S.J., Sill, P.R., Serjeantson, S.W., Bana-Koiri, J., Bhatia, K., Alpers, M.P., and Boyce, A.J. (1986). High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature.* **321,** 744-750.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T. et. al., (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* **2,** E207.

Freeman, B., Smith, N., Curtis, C., Huckett, L., Mill, J. and Craig, I.W. (2003). DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behav. Genet.* **33,** 67-72.

Freund, J.N., Domon-Dell, C., Kedinger, M. and Duluc, I. (1998). The Cdx-1 and Cdx-2 homeobox genes in the intestine. *Biochem. Cell Biol.* **76,** 957-969.

Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics.* **147,** 915-925.

Fu, Y.X. and Li, W.H. (1993). Statistical tests of neutrality of mutations. *Genetics.* **133,** 693-709.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. et. al., (2002). The structure of haplotype blocks in the human genome. *Science.* **296,** 2225-2229.

Gabriel, S.E., Brigman, K.N., Koller, B.H., Boucher, R.C. and Stutts, M.J. (1994). Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science.* **266,** 107-109.

Galvani, A.P. and Slatkin, M. (2003). Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 15276-15279.

Gao, X.P., Sedgwick, T., Shi, Y.B. and Evans, T. (1998). Distinct functions are implicated for the GATA-4, -5, and -6 transcription factors in the regulation of intestine epithelial cell differentiation. *Mol. Cell. Biol.* **18,** 2901-2911.

Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev. Genet.* **7,** 703-713.

Gilat, T., Russo, S., Gelman-Malachi, E. and Aldor, T.A. (1972). Lactase in man: a nonadaptable enzyme. *Gastroenterology.* **62,** 1125-1127.

Groot, P.C., Bleeker, M.J., Pronk, J.C., Arwert, F., Mager, W.H., Planta, R.J., Eriksson, A.W. and Frants, R.R. (1989). The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. *Genomics.* **5,** 29-42.

Grunberg, J. and Sterchi, E.E. (1995). Human lactase-phlorizin hydrolase: evidence of dimerization in the endoplasmic reticulum. *Arch. Biochem. Biophys.* **323,** 367-372.

Gueguen, L. and Pointillart, A. (2000). The bioavailability of dietary calcium. *J. Am. Coll. Nutr.* **19,** 119S-136S.

Guo, S.W. and Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics.* **48,** 361-372.

Gutierrez, I., Espinosa, A., Garcia, J., Carabano, R. and De Blas, J.C. (2002). Effect of levels of starch, fiber, and lactose on digestion and growth performance of early-weaned rabbits. *J Anim Sci.* **80,** 1029-1037.

Haldane, J.B.S. (1949). Disease and evolution. *Ric. Sci. Suppl.* **Suppl. A 19:** 68-76.

Hamblin, M.T. and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J Hum Genet.* **66,** 1669-1679.

Hamblin, M.T., Thompson, E.E. and Di, R.A. (2002). Complex signatures of natural selection at the Duffy blood group locus. *Am. J Hum Genet.* **70,** 369-383.

Hammer, H.F., Petritsch, W., Pristautz, H. and Krejs, G.J. (1996). Evaluation of the pathogenesis of flatulence and abdominal cramps in patients with lactose malabsorption. *Wien. Klin. Wochenschr.* **108,** 175-179.

Hanchard, N., Elzein, A., Trafford, C., Rockett, K., Pinder, M., Jallow, M., Harding, R., Kwiatkowski, D. and McKenzie, C. (2007). Classical sickle beta-globin haplotypes exhibit a high degree of long-range haplotype similarity in African and Afro-Caribbean populations. *BMC. Genet.* **8,** 52.

Handley, L.J., Manica, A., Goudet, J. and Balloux, F. (2007). Going the distance: human population genetics in a clinal world. *Trends Genet.* **23,** 432-439.

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

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

Harvey, C.B., Pratt, W.S., Islam, I., Whitehouse, D.B. and 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., Darmoul, D., Phillips, A., Mantei, N. and Swallow, D.M. (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 Lett.* **398,** 135-140.

He, T., Priebe, M.G., Welling, G.W. and Vonk, R.J. (2006). Effect of lactose on oro-cecal transit in lactose digesters and maldigesters. *Eur J Clin Invest..* **36,** 737-742.

Hedrick, P.W. and Verrelli, B.C. (2006). "Ground truth" for selection on CCR5-Delta32. *Trends Genet.* **22,** 293-296.

Hertzler, S.R. and Savaiano, D.A. (1996). Colonic adaptation to daily lactose feeding in lactose maldigesters reduces lactose intolerance. *Am. J. Clin. Nutr.* **64,** 232-236.

Hertzler, S.R., Savaiano, D.A. and Levitt, M.D. (1997). Fecal hydrogen production and consumption measurements - Response to daily lactose ingestion by lactose maldigesters. *Dig. Dis. Sci.* **42,** 348-353.

Hijazi, S.S., Abulaban, A., Ammarin, Z. and Flatz, G. (1983). Distribution of Adult Lactase Phenotypes in Bedouins and in Urban and Agricultural Populations of Jordan. *Tropical and Geographical Medicine.* **35,** 157-161.

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

Hoekstra, H.E. and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution Int. J. Org. Evolution.* **61,** 995-1016.

Holbrook, J.D., Birdsey, G.M., Yang, Z., Bruford, M.W. and Danpure, C.J. (2000). Molecular adaptation of alanine:glyoxylate aminotransferase targeting in primates. *Mol. Biol. Evol.* **17,** 387-400.

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

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

Hollox, E.J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A.I. and Swallow, D.M. (2001). Lactase haplotype diversity in the Old World. *Am. J. Hum. Genet.* **68,** 160-172.

Horak, C.E. and Snyder, M. (2002). ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350,** 469-483.

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

Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B. and Oppermann, M. (2005). Detection of the CCR5-Delta32 HIV resistance gene in Bronze Age skeletons. *Genes Immun.* **6,** 371-374.

Imtiaz, F., Savilahti, E., Sarnesto, A., Trabzuni, D., Al-Kahtani, K., Kagevi, I., Rashed, M.S., Meyer, B.F. and Jarvela, I. (2007). The T/G 13915 variant upstream of the lactase gene (LCT) is the founder allele of lactase persistence in an urban Saudi population. *J. Med. Genet.* **44,** e89.

Ingram, C.J.E., Elamin, M.F., Mulcare, C.A., Weale, M.E., Tarekegn, A., Raga, T.O., Bekele, E., Elamin, F.M., Thomas, M.G., Bradman, N. and Swallow, D.M. (2007). A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum. Genet.* **120,** 779-788.

Jacob, F. and Monod, J. (1961). On regulation of gene activity. *Cold Spring Harb. Symp. Quant. Biol.* **26,** 193-211.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R. et. al., (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature.* **451,** 998-1003.

Jobling M.A., Hurles M.E. and Tyler-Smith C. (2004) *Human Evolutionary Genetics: Origins, Peoples, Disease*. Garland Science, Abingdon, UK.

Jorde L (2005) Human Genetic Diversity. *In: Nature Encyclopaedia of Life Sciences.* John Wiley & Sons, Ltd: Chichester http://www.els.net/ [doi: 10.1038/npg.els.0005079].

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T. and Batzer, M.A. (2000). The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66,** 979-988.

Ke, X. and Cardon, L.R. (2003). Efficient selective screening of haplotype tag SNPs. *Bioinformatics.* **19,** 287-288.

Keen, A.E. and Zeitlyn, D. (2007). Language, Diet, and Ethnicity in Mayo-Darle, Adamaoua, Cameroon. *Anthropos.* **102,** 213-219.

Keusch, G.T., Troncale, F.J., Thavaramara, B., Prinyanont, P., Anderson, P.R. and Bhamarapravathi, N. (1969). Lactase deficiency in Thailand: effect of prolonged lactose feeding. *Am. J. Clin. Nutr.* **22,** 638-641.

Khogali M.M. (1991). The Migration of The Danagla to Port Sudan. *GeoJournal.* **25,** 63-71.

Kleinjan, D.A. and van Heyningen, V (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76,** 8-32.

Kolho, K.L. and Jarvela, I. (2006). DNA test for hypolactasia premature: Authors' reply. *Gut.* **55,** 131-132.

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

Krasinski, S.D., van Wering, H.M., Tannemaat, M.R. and Grand, R.J. (2001). Differential activation of intestinal gene promoters: functional interactions between GATA-5 and HNF-1 alpha. *Am. J. Physiol Gastrointest. Liver Physiol.* **281,** 69-84.

Kretchmer, N., Ransome-Kuti, O., Hurwitz, R., Dungy, C. and Alakija, W. (1971). Intestinal absorption of lactose in Nigerian ethnic groups. *Lancet.* **2,** 392-395.

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

Kuokkanen, M., Kokkonen, J., Enattah, N.S., Ylisaukko-Oja, T., Komu, H., Varilo, T., Peltonen, L., Savilahti, E. and Jarvela, I. (2006). Mutations in the translated region of the lactase gene (LCT) underlie congenital lactase deficiency. *Am. J. Hum. Genet.* **78,** 339-344.

Labayen, I., Forga, L., González, A., Lenoir-Wijnkoop, I., Nutr, R. and Martínez, J.A. (2001). Relationship between lactose digestion, gastrointestinal transit time and symptoms in lactose malabsorbers after dairy consumption. *Aliment Pharmacol Ther.* **4,** 543-549.

Lacey, S.W., Naim, H.Y., Magness, R.R., Gething, M.-J. and Sambrook, J.F. (1994). Expression of lactase-phlorizin hydrolase in sheep is regulated at the RNA level. *Biochem. J.* **302,** 929-935.

Ladas, S., Papanikos, J. and Arapakis, G. (1982). Lactose malabsorption in Greek adults: correlation of small bowel transit time with the severity of lactose intolerance. *Gut.* **11,** 968-973.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et. al., (2001). Initial sequencing and analysis of the human genome. *Nature.* **409,** 860-921.

Lao, O., de Gruijter, J.M., van, D.K., Navarro, A. and Kayser, M. (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann. Hum Genet.* **71,** 354-369.

Lee, S.Y., Wang, Z., Lin, C.K., Contag, C.H., Olds, L.C., Cooper, A.D. and Sibley, E. (2002). Regulation of intestine-specific spatiotemporal expression by the rat lactase promoter. *J. Biol. Chem.* **277,** 13099-13105.

Leese, H.J. and Semenza, G. (1973). Identity between small intestinal enzymes phlorizin hydrolase and glycosylceramidase. *J. Biol. Chem.* **248,** 8170-8173.

Legendre, P. and Legendre, L. (1998). *Numerical Ecology*. Elsevier Science, Amsterdam, The Netherlands.

Leichter, J. (1973). Effect of dietary lactose on intestinal lactase activity in young rats. *J. Nutr.* **103,** 392-396.

Levy, T.E. and Holl, A.F.C. (2002). Migrations, ethnogenesis, and settlement dynamics: Israelites in Iron Age Canaan and Shuwa-Arabs in the Chad Basin. *Journal of Anthropological Archaeology.* **21,** 83-118.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov G. et. al., (2007). The diploid genome sequence of an individual human. *PLoS Biol.* 2007 **5,** e254.

Lewinsky, R.H., Jensen, T.G., Moller, J., Stensballe, A., Olsen, J. and Troelsen, J.T. (2005). T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Hum. Mol. Genet.* **14,** 3945-3953.

Lewis, H.S. (1966). The Origins of the Galla and Somali. *Journal of African History.* **7,** 27-46.

Lewis, I.M. (1960). The Somali Conquest of the Horn of Africa. *Journal of African History.* **1,** 213-229.

Lewontin, R.C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics.* **49,** 49-67.

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

Luzzatto, L., Usanga, F.A. and Reddy, S. (1969). Glucose-6-phosphate dehydrogenase deficient red cells: resistance to infection by malarial parasites. *Science.* **164,** 839-842.

Maiuri, L., Raia, V., Potter, J., Swallow, D.M., Ho, M.W., Fiocca, R., Finzi, G., Cornaggia, M., Capella, C., Quaroni, A. and 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. and 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.

Martin, T.W., Weisman, I.M., Zeballos, R.J. and Stephenson, S.R. (1989). Exercise and hypoxia increase sickling in venous blood from an exercising limb in individuals with sickle cell trait. *Am. J. Med.* **87,** 48-56.

Mason, P.J. and Vulliamy, T.J. (2005) Glucose-6-Phosphate Dehydrogenase (G6PD) Deficiency: Genetics. In: John Wiley & Sons, Ltd,

Mathers, C.D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* **3,** e442.

McCracken, R.D. (1971). Lactase Deficiency - Example of Dietary Evolution. *Curr Anthropol.* **12,** 479-517.

McDonald, J.H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature.* **351,** 652-654.

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

Mead, S., Stumpf, M.P., Whitfield, J., Beck, J.A., Poulter, M., Campbell, T., Uphill, J.B., Goldstein, D., Alpers, M., Fisher, E.M. and Collinge, J. (2003). Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics. *Science.* **300,** 640-643.

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

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

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

Mitchelmore, C., Troelsen, J.T., Spodsberg, N., Sjostrom, H. and Noren, O. (2000). Interaction between the homeodomain proteins Cdx2 and HNF1alpha mediates expression of the lactase-phlorizin hydrolase gene. *Biochem. J.* **346,** 529-535.

Mulcare, C.A. (2006) *The Evolution of the Lactase Persistence Phenotype,* PhD. Thesis, University of London, London.

Mulcare, C.A., Weale, M.E., Jones, A.L., Connell, B., Zeitlyn, D., Tarekegn, A., Swallow, D.M., Bradman, N. and Thomas, M.G. (2004). The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am. J. Hum. Genet.* **74,** 1102-1110.

Murdock G (1967). Ethnographic atlas. *University of Pittsburg Press, Pittsburgh. U. S. A.*

Murdock G.P (1959) Africa: Its peoples and their culture history. *McGraw-Hill, New York.*

Mutoh, H., Satoh, K., Kita, H., Sakamoto, H., Hayakawa, H., Yamamoto, H., Isoda, N., Tamada, K., Ido, K. and Sugano, K. (2005). Cdx2 specifies the differentiation of morphological as well as functional absorptive enterocytes of the small intestine. *Int. J. Dev. Biol.* **49,** 867-871.

Myles, S., Bouzekri, N., Haverfield, E., Cherkaoui, M., Dugoujon, J.M. and Ward, R. (2005). Genetic evidence in support of a shared Eurasian-North African dairying origin. *Hum. Genet.* **117,** 34-42.

Nagel, R.L. (2005) Sickle Cell Anaemia. *In: Nature Encyclopaedia of Life Sciences.* John Wiley & Sons, Ltd: Chichester http://www.els.net/ [doi: 10.1038/npg.els.0001454].

Naim, H.Y. and Naim, H. (1996). Dimerization of lactase-phlorizin hydrolase occurs in the endoplasmic reticulum, involves the putative membrane spanning domain and is required for an efficient transport of the enzyme to the cell surface. *Eur. J. Cell Biol.* **70,** 198-208.

Naim, H.Y., Sterchi, E.E. and Lentze, M.J. (1987). Biosynthesis and maturation of lactase-phlorizin hydrolase in the human intestinal epithelial cells. *Biochem. J.* **241,** 427-434.

Nei, M. (1987) Molecular evolutionary genetics. *Columbia University Press, New York.*

NIH/CEPH Collaborative Mapping Group (1992). A comprehensive genetic linkage map of the human genome. *Science.* **258,** 67-86.

Noguchi T (1987) Amino Acid Metabolism in animal peroxisomes. *In: Peroxisomes in biology and medicine.* Fahimi H.D. and Sies H. (eds), Springer-Verlag, Berlin, Heidelberg. pp 234-243

Norton, H.L., Kittles, R.A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V.A., Bradley, D.G., McEvoy, B. and Shriver, M.D. (2007). Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol Evol.* **24,** 710-722.

Novick, G.E., Batzer, M.A., Deininger, P.L. and Herrera, R.J. (2008). The Mobile Genetic Element "Alu" in the Human Genome. *BioScience.* **46,** 32-41.

Oesterreicher, T.J. and Henning, S.J. (2004). Rapid induction of GATA transcription factors in developing mouse intestine following glucocorticoid administration. *Am. J. Physiol Gastrointest. Liver Physiol.* **286,** 947-953.

Olds, L.C. and Sibley, E. (2003). Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum. Mol. Genet.* **12,** 2333-2340.

Panzer, P., Preuss, U., Joberty, G. and Naim, H.Y. (1998). Protein domains implicated in intracellular transport and sorting of lactase-phlorizin hydrolase. *J. Biol. Chem.* **273,** 13861-13869.

Parra, E.J. (2007). Human pigmentation variation: evolution, genetic basis, and implications for public health. *Am. J. Phys. Anthropol.* **Suppl 45,** 85-105.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R. et. al., (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet.* **39,** 1256-1260.

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

Peuhkuri K (2000) *Lactose, lactase and bowel disorders.* PhD. Thesis, University of Helsinki, Helsinki.

Pie, S., Lalles, J.P., Blazy, F., Laffitte, J., Seve, B. and Oswald, I.P. (2004). Weaning is associated with an upregulation of expression of inflammatory cytokines in the intestine of piglets. *J. Nutr.* **134,** 641-647.

Pier, G.B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S.S., Banting, G., Ratcliff, R., Evans, M.J. and Colledge, W.H. (1998). Salmonella typhi uses CFTR to enter intestinal epithelial cells. *Nature.* **393,** 79-82.

Pinto, M., Robine-Leon, S., Appay, M.D., Kedinger, M., Triadou, N., Dussaulx, E., Lacroix, B., Simon-Assmann, P., Haffen, K., Fogh, J. and Zweibaum, A. (1983).

Enterocyte-like differentiation and polarization of the human colon carcinoma cell line Caco-2 in culture. *Biol. Cell.* **47,** 323-330.

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.* **35,** 20-31.

Poulter, M., Hollox, E., Harvey, C.B., Mulcare, C., Peuhkuri, K., Kajander, K., Sarner, M., Korpela, R. and Swallow, D.M. (2003). The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann. Hum. Genet.* **67,** 298-311.

Qin, Z.S., Niu, T. and Liu, J.S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **71,** 1242-1247.

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

Raymond, M. and Rousset, F. (1995). An Exact Test for Population Differentiation. *Evolution.* **49,** 1280-1283.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. et. al., (2006). Global variation in copy number in the human genome. *Nature.* **444,** 444-454.

Riesman P. (1977) Freedom in fulani social life. *The University of Chicago Press, Ltd., London.*

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E. and Chang, H.Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell.* **129,** 1311-1323.

Robinson, A.E. (1927). Notes on the Gamuia Tribe, Sudan. *Journal of the Royal African Society.* **26,** 138-144.

Romeo, G., Devoto, M. and Galietta, L.J. (1989). Why is the cystic fibrosis gene so frequent? *Hum Genet.* **84,** 1-5.

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

Roth, E.F., Raventos, S.C., Rinaldi, A. and Nagel, R.L. (1983). The effect of X chromosome inactivation on the inhibition of Plasmodium falciparum malaria growth by glucose-6-phosphate-dehydrogenase-deficient red cells. *Blood.* **62,** 866-868.

Rousseau, K., Byrne, C., Griesinger, G., Leung, A., Chung, A., Hill, A.S. and Swallow, D.M. (2007). Allelic association and recombination hotspots in the mucin gene (MUC) complex on chromosome 11p15.5. *Ann. Hum. Genet.* **71,** 561-569.

Ruwende, C. and Hill, A. (1998). Glucose-6-phosphate dehydrogenase deficiency and malaria. *J Mol. Med.* **76,** 581-588.

Ruwende, C., Khoo, S.C., Snow, R.W., Yates, S.N., Kwiatkowski, D., Gupta, S., Warn, P., Allsopp, C.E., Gilbert, S.C., Peschu, N. and . (1995). Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature.* **376,** 246-249.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. et. al., (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature.* **419,** 832-837.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006). Positive natural selection in the human lineage. *Science.* **312,** 1614-1620.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. et. al., (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature.* **449,** 913-918.

Sabeti, P.C., Walsh, E., Schaffner, S.F., Varilly, P., Fry, B., Hutcheson, H.B., Cullen, M., Mikkelsen, T.S., Roy, J., Patterson, N. et. al., (2005). The case for selection at CCR5-Delta32. *PLoS Biol.* **3,** e378.

Sahi, T. (1974). The inheritance of selective adult-type lactose malabsorption. *Scand. J. Gastroenterol. Suppl.* **30,** 1-73.

Sahi, T. (1994). Hypolactasia and lactase persistence. Historical review and the terminology. *Scand. J. Gastroenterol. Suppl.* **202,** 1-6.

Saltzman, J.R., Russell, R.M., Golner, B., Barakat, S., Dallal, G.E. and Goldin, B.R. (1999). A randomized trial of *Lactobacillus acidophilus* BG2FO4 to treat lactose intolerance. *Am. J. Clin. Nutr.* **69,** 140-146.

Sanchez, J.J., Hallenberg, C., Borsting, C., Hernandez, A. and Morling, N. (2005). High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. *European Journal of Human Genetics.* **13,** 856-866.

Schliekelman, P., Garner, C. and Slatkin, M. (2001). Natural selection and resistance to HIV. *Nature.* **411,** 545-546.

Scozzari, R., Torroni, A., Semino, O., Sirugo, G., Brega, A. and Santachiara-Benerecetti, A.S. (1988). Genetic studies on the Senegal population. I. Mitochondrial DNA polymorphisms. *Am. J Hum Genet.* **43,** 534-544.

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

Semenza, G., Auricchio, S. and Mantei, N. (1999) Small-Intestinal Disaccharidases, in, *The Metabolic and Molecular Bases of Inherited Disease,* 8th edn., Vol. 1, C.R.Scriver, A.L.Beaudet, W.S.Sly and D.Valle eds., McGraw-Hill, New York, pp. 1623-1650

Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics.* **141,** 413-429.

Simoons, F.J. (1970). Primary lactose intolerance and the milking habit: a problem in biological and cultural interrelations, II. A culture historical hypothesis. *Am. J. Dig. Dis.* **15,** 695-710.

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

Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657-663.

Smith, A.B. (1992). Origins and Spread of Pastoralism in Africa. *Annual Review of Anthropology.* **21,** 125-141.

Smith, M.W., Patterson, J.Y. and Peacock, M.A. (1984). A comprehensive description of brush border membrane development applying to enterocytes taken from a wide variety of mammalian species. *Comp Biochem Physiol A.* 77, 655-662.

Smith, M.W., Peacock, M.A. and Lund, E.K. (1986). Testing the hypothesis that crypt size determines the rate of enterocyte development in neonatal mice. *Comp Biochem Physiol A.* **84,** 511-515.

Snook, C.R., Mahmoud, J.N. and Chang, W.P. (1976). Lactose tolerance in adult Jordanian Arabs. *Trop. Geogr. Med.* **28,** 333-335.

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

Spurr, N.K. and White, R. (1991). Report of the committee on the genetic constitution of chromosome 2. *Cytogenet Cell Genet.* **58,** 142-169.

Stanton, E.A.E. (1903). The Peoples of the Anglo-Egyptian Sudan. *Journal of the Royal African Society.* **2,** 121-123.

Stephens, J.C., Reich, D.E., Goldstein, D.B., Shin, H.D., Smith, M.W., Carrington, M., Winkler, C., Huttley, G.A., Allikmets, R., Schriml, L. et. al., (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am. J. Hum. Genet.* **62,** 1507-1515.

Stephens, M., Smith, N.J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978-989.

Sterchi, E., Mills, P., Fransen, J., Hauri, H., Lentze, M., Naim, H., Ginsel, L. and Bond, J. (1990). Biogenesis of intestinal lactase-phlorizin hydrolase in adults with lactose intolerance. Evidence for reduced biosynthesis and slowed-down maturation in enterocytes. *J. Clin. Invest.* **86,** 1329-1337.

Stover, P.J. (2006). Influence of human genetic variation on nutritional requirements. *Am. J Clin. Nutr.* **83,** 436S-442S.

Stryer L. (1995) Biochemistry. *W. H. Freeman and Company, New York.*

Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P., Stark, M.S., Hayward, N.K., Martin, N.G. and Montgomery, G.W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet.* **82,** 424-431.

Swallow, D.M. (2003). Genetic influences on carbohydrate digestion. *Nutrition Research Reviews.* **16,** 37-43.

Swallow, D.M. (2006). DNA test for hypolactasia premature. *Gut.* **55,** 131-132.

Swallow andHollox E.J. (2000) The genetic polymorphism of intestinal lactase activity in adult humans. *In: The metabolic and molecular basis of inherited disease*, 8th ed. Scriver C., Beaudet A., Sly W., and Valle D. (eds), McGraw-Hill, New York.

Swayd, S.S. (1998). The Druzes: One Thousand Years of Tradition and Reform. *International Studies and Overseas Programs.* **21,** 1-4.

Sykes, D.E. and Weiser, M.M. (1995). Rat intestinal crypt-cell replication factor with homology to early S-phase proteins required for cell division. *Gene.* **163,** 243-247.

Tag, C.G., Oberkanins, C., Kriegshauser, G., Ingram, C.J., Swallow, D.M., Gressner, A.M., Ledochowski, M. and Weiskirchen, R. (2008). Evaluation of a novel reverse-hybridization StripAssay for typing DNA variants useful in diagnosis of adult-type hypolactasia. *Clin. Chim. Acta.* **392,** 58-62.

Tag, C.G., Schifflers, M.C., Mohnen, M., Gressner, A.M. and Weiskirchen, R. (2007). A novel proximal -13914G > A base replacement in the vicinity of the common -13910T/C lactase gene variation results in an atypical LightCycler melting curve in testing with the MutaREAL lactase test. *Clinical Chemistry.* **53,** 146-148.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* **123,** 585-595.

Takayama, T., Fujita, K., Suzuki, K., Sakaguchi, M., Fujie, M., Nagai, E., Watanabe, S., Ichiyama, A. and Ogawa, Y. (2003). Control of oxalate formation from L-hydroxyproline in liver mitochondria. *J. Am. Soc. Nephrol.* **14,** 939-946.

Templeton, A. (2002). Out of Africa again and again. *Nature.* **416,** 45-51.

Thacher, T.D., Fischer, P.R., Pettifor, J.M., Lawson, J.O., Isichei, C.O., Reading, J.C. and Chan, G.M. (1999). A comparison of calcium, vitamin D, or both for nutritional rickets in Nigerian children. *N Engl J Med.* **341,** 563-568.

The International HapMap Consortium (2005). A Haplotype Map of the Human Genome. *Nature.* **437,** 1299-1320.

Thomas, M.G., Bradman, N. and Flinn, H.M. (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum Genet.* **105,** 577-581.

Thomas, M.G., Weale, M.E., Jones, A.L., Richards, M., Smith, A., Redhead, N., Torroni, A., Scozzari, R., Gratrix, F., Tarekegn, A. et. al., (2002). Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am. J Hum Genet.* **70,** 1411-1420.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M. et. al., (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39,** 31-40.

Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., stro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J. et. al., (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science.* **293,** 455-462.

Tournamille, C., Colin, Y., Cartron, J.P. and Le Van, K.C. (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet.* **10,** 224-228.

Troelsen, J., Olsen, J., Noren, O. and Sjostrom, H. (1992). A novel intestinal trans factor (NF-LPH1) interacts with the lactase phlorizin hydrolase promotor and co-varies with the enzymic activity. *J. Biol. Chem.* **267,** 20407-20411.

Troelsen, J.T. (2005). Adult-type hypolactasia and regulation of lactase expression. *Biochim. Biophys. Acta.* **1723,** 19-32.

Troelsen, J.T., Mehlum, A., Olsen, J., Spodsberg, N., Hansen, G.H., Prydz, H., Noren, O. and Sjostrom, H. (1994). 1 Kb of the lactase-phlorhizin hydrolase promoter directs post-weaning decline and small intestinal-specific expression in transgenic mice. *FEBS Lett.* **342,** 291-296.

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

Troelsen, J.T., Olsen, J., Moller, J. and Sjostrom, H. (2003). An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology.* **125,** 1686-1694.

van Wering, H.M., Bosse, T., Musters, A., de Jong, E., de Jong, N., Hogen Esch, C.E., Boudreau, F., Swain, G.P., Dowling, L.N., Montgomery, R.K. et. al., (2004). Complex regulation of the lactase-phlorizin hydrolase promoter by GATA-4. *Am. J. Physiol. Gastrointest. Liver Physiol.* **287,** 899-909.

Vanhove, M. (2006). The Beja language today in Sudan: The state of the art in linguistics. *Proceedings of the 7th International Sudan Studies Conference.*

Vardhanabhuti, S., Wang, J. and Hannenhalli, S. (2007). Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.* **35,** 3203-3213.

Verrelli, B.C., McDonald, J.H., Argyropoulos, G., stro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A.N., Loiselet, J. and Tishkoff, S.A. (2002). Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am. J Hum Genet.* **71,** 1112-1128.

Verrijzer, C.P., Alkema, M.J., van Weperen, W.W., Van Leeuwen, H.C., Strating, M.J. and van, d., V (1992). The DNA binding specificity of the bipartite POU domain and its subdomains. *EMBO J.* **11,** 4993-5003.

Villako, K. and Maaroos, H. (1994). Clinical picture of hypolactasia and lactose intolerance. *Scand. J. Gastroenterol. Suppl.* **202,** 36-54.

Vogelsang, H., Ferenci, P., Frotz, S., Meryn, S. and Gangl, A. (1988). Acidic colonic microclimate--possible reason for false negative hydrogen breath tests. *Gut.* **29,** 21-26.

Wacker, H., Keller, P., Falchetto, R., Legler, G. and Semenza, G. (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.

Wambua, S., Mwangi, T.W., Kortok, M., Uyoga, S.M., Macharia, A.W., Mwacharo, J.K., Weatherall, D.J., Snow, R.W., Marsh, K. and Williams, T.N. (2006). The effect of alpha(+)-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya. *Plos Medicine.* **3,** 643-651.

Wang, Y., Harvey, C.B., Hollox, E.J., Phillips, A.D., Poulter, M., Clay, P., Walker-Smith, J.A. and 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. and Swallow, D.M. (1995). The lactase persistence/non-persistence polymorphism is controlled by a *cis*-acting element. *Hum. Mol. Genet.* **4,** 657-662.

Wang, Z., Maravelias, C. and Sibley, E. (2006). Lactase gene promoter fragments mediate differential spatial and temporal expression patterns in transgenic mice. *DNA Cell Biol.* **25,** 215-222.

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

Weale, M.E. (2006). DNA test for hypolactasia premature. *Gut.* **55,** 131-132.

Weatherall, D.J. (2004). Thalassaemia: the long road from bedside to genome. *Nature Reviews Genetics.* **5,** 625-6U5.

Weatherall, D.J., Miller, L.H., Baruch, D.I., Marsh, K., Doumbo, O.K., Casals-Pascual, C. and Roberts, D.J. (2002). Malaria and the red cell. *Hematology. Am. Soc. Hematol. Educ. Program.*35-57.

Wen, C.P., Antonowicz, I., Tovar, E., McGandy, R.B. and Gershoff, S.N. (1973). Lactose feeding in lactose-intolerant monkeys. *Am. J. Clin. Nutr.* **26,** 1224-1228.

West, A.G. and Fraser, P. (2005). Remote control of gene transcription. *Hum. Mol. Genet.* **14 Spec No 1,** R101-R111.

Wilson, G.M., Flibotte, S., Missirlis, P.I., Marra, M.A., Jones, S., Thornton, K., Clark, A.G. and Holt, R.A. (2006). Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Res.* **16,** 173-181.

Witte, J., Lloyd, M., Lorenzsonn, V., Korsmo, H. and Olsen, W. (1990). The biosynthetic basis of adult lactase deficiency. *J. Clin. Invest.* **86,** 1338-1342.

Wiuf, C. (2001). Do delta F508 heterozygotes have a selective advantage? *Genet. Res.* **78,** 41-47.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8,** 206-216.

Wright, E.M., Hirayama, B.A. and Loo, D.F. (2007). Active sugar transport in health and disease. *J. Intern. Med.* **261,** 32-43.

Wright, S. (1951). The genetical structure of populations. *Annals Eugenics.* **15,** 323-354.

Ye, S., Dhillon, S., Ke, X., Collins, A.R. and Day, I.N. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.* **29,** E88.

Zeitlyn, D. and Connell, B. (2003). Ethnogenesis and fractal history on an African frontier: Mambila-Njerep-Mandulu. *Journal of African History.* **44,** 117-138.

Zeller, R. and Zuniga, A. (2007). Shh and Gremlin1 chromosomal landscapes in development and disease. *Curr. Opin. Genet Dev.* **17,** 428-434.

Appendix A: Example milk drinking questionnaire.

Sample number:_____ Age:_____ Sex: F □ M □

LACTOSE TOLERANCE TEST – RELEVANT INFORMATION

1. Have you recently been diagnosed with a gastro-intestinal disease? Y □ N □

Details:_____

2. Have you taken antibiotics in the last six weeks?

ETHNIC IDENTITY – RELEVANT INFORMATION

3. Tribe:

---

4. Current residence:

---

5. Place of birth:

---

6. Please list your first (and second) languages:

MILK DRINKING HABITS:

7. Which of the following descriptions best reflects the amount of milk you drink? Please tick all that apply:

A 1/2 litre of milk a day or more □

I drink milk infrequently (less than 1/2 litre a week) □

I actively avoid drinking milk □

I feel unwell after drinking milk □

8. What kind of milk do you usually drink? (e.g. camel, cow's milk etc):

---

9. Does any member of your family, to the best of your knowledge, have an intolerance or dislike of drinking milk? Please give brief details.

GENERAL:
10. Is this individual related to any of the other participants? If so, how? Eg father etc (prefer not to test cousins)

Comments:

Family Background:

Father

Place of birth:	
First language	
Second language, if any:	
Tribal group:	
Any known intolerance to milk?	

Father's father

Place of birth:	
First language	
Second language, if any:	
Tribal group:	

Father's mother

Place of birth:	
First language	
Second language, if any:	
Tribal group:	

Mother

Place of birth:	
First language	
Second language, if any:	
Tribal group:	
Any known intolerance to milk?	

Mother's father

Place of birth:	
First language	
Second language, if any:	
Tribal group:	

Mother's mother

Place of birth:	
First language	
Second language, if any:	
Tribal group	

Appendix B: Lactose Tolerance data for Sudanese Jaali Cohort

Sample ID	LTT: Baseline (ppm)	LTT: T=30	LTT: T = 60	LTT: T = 90	LTT: T = 120	LTT: RISE	LTT: STATUS	COMMENTS
SD-J-01	11	7	9	18	18	7	P	
SD-J-02	9	20	45			36	NP	
SD-J-03	12	10	31			19	I	
SD-J-04	7	13	29			22	NP	
SD-J-05	11	14	17	20	19	8	P	
SD-J-06	10	25	63			53	NP	
SD-J-07	3	8	13	10	10	7	P	
SD-J-08	13	20	30	37		24	NP	
SD-J-09	8	9	19	17	15	7	P	
SD-J-10	14	30	45			31	NP	
SD-J-11	6	18	60			54	NP	
SD-J-12	4	8	10	13	11	7	P	
SD-J-13	15	20	25	23	23	8	P	
SD-J-14	6	14	15	15	15	9	P	
SD-J-15	14	35				21	NP	
SD-J-16	3	8	4	13	14	11	P	
SD-J-17	2	9	24	36		34	NP	
SD-J-18	10	17	34			24	NP	
SD-J-19	5	6	14	48		43	NP	
SD-J-20	2	11	36			34	NP	
SD-J-21	3	4	10	4	5	2	P	
SD-J-22	5	41	48			43	NP	
SD-J-23	2	3	2	2	2	0	P	
SD-J-24	8	25	28	37		29	NP	
SD-J-25	3	9	9	9	9	6	P	
SD-J-26	5	5	4	5	5	0	P	
SD-J-27	17	16	13	15		-2	P	
SD-J-28	2	19	12	22		20	I	
SD-J-29	7	7	38			31	NP	
SD-J-31	14	4	8	12	15	1	P	
SD-J-32	16	21	49			33	NP	
SD-J-33	8	3	5	7		-1	P	
SD-J-34	5	8	54			49	NP	
SD-J-35	13	31	39			26	NP	
SD-J-36	10	13	25	40		30	NP	
SD-J-37	5	4	16	17	13	8	P	
SD-J-38	15	14	15	20		5	P	
SD-J-39	2	5	29			27	NP	
SD-J-40	11	12	23	25	24	13	I	
SD-J-41	9	10	6	6	8	-1	P	
SD-J-42	2	3	5	5		3	P	
SD-J-43	3	10	26			23	NP	
SD-J-44	4	18	25	40		36	NP	
SD-J-45	5	2	3	7		2	P	
SD-J-46	18	56				38	NP	
SD-J-47	7	5	3	9		2	P	
SD-J-48	14	22	37			23	NP	
SD-J-49	18	18	14	17	17	-1	P	
SD-J-50	2	8	48			46	NP	

Sample ID	LTT: Baseline (ppm)	LTT: T=30	LTT: T = 60	LTT: T = 90	LTT: T = 120	LTT: RISE	LTT: STATUS	COMMENTS
SD-J-51	4	18	53			49	NP	
SD-J-52	12	17	20	19	18	6	P	Poor Quality DNA
SD-J-53	11	20	39			28	NP	
SD-J-54	18	15	16	18	18	0	P	
SD-J-55	7	22	41			34	NP	Poor Quality DNA
SD-J-56	8	33				25	NP	Poor Quality DNA
SD-J-57	9	18	24	23	25	16	I	Poor Quality DNA
SD-J-58	14	13	17	13		-1	P	Poor Quality DNA
SD-J-59	3	9	40			37	NP	
SD-J-60	17	24	20	15	16	-1	P	
SD-J-61	13	12	25	37		24	NP	
SD-J-62	15	15	14	18	13	-2	P	
SD-J-63	7	12	8	7		0	P	
SD-J-64	13	20	21	21	21	8	P	
SD-J-65	12	18	20	21	19	7	P	
SD-J-66	17	45				28	NP	
SD-J-67	3	7	7	6		3	P	
SD-J-68	11	12	14	17	20	9	P	
SD-J-69	17	15	18	19	19	2	P	Poor Quality DNA
SD-J-70	4	6	12	11		7	P	
SD-J-71	18	20	39	67		49	NP	
SD-J-72	13	17	12	13	13	0	P	
SD-J-73	8	15	38			30	NP	Poor Quality DNA
SD-J-74	2	3	2	7	5	3	P	
SD-J-75	15	18	29	41	50	35	NP	Poor Quality DNA
SD-J-76	7	20	23	29		22	NP	
SD-J-77	17	45				28	NP	
SD-J-78	13	19	28	36		23	NP	
SD-J-79	11	10	13	13		2	P	
SD-J-80	7	20	12	15		8	P	
SD-J-81	12	18	17	47		35	NP	Poor Quality DNA
SD-J-82	8	11	20	29		21	NP	
SD-J-83	3	20	37			34	NP	
SD-J-84	9	32				23	NP	
SD-J-85	12	17	17	18	19	7	P	
SD-J-86	4	7	3	4		0	P	
SD-J-87	18	19	15	18		0	P	
SD-J-88	12	47				35	NP	
SD-J-89	9	15	18	13		4	P	
SD-J-90	15	30	39			24	NP	
SD-J-91	1	17	15	13		12	I	Poor Quality DNA
SD-J-92	8	13	10	12		4	P	
SD-J-93	14	14	18	20	22	8	P	Poor Quality DNA
SD-J-94	8	9	19	30		22	NP	
SD-J-95	11	20	12	11		0	P	
SD-J-96	7	19	37			30	NP	Poor Quality DNA
SD-J-97	6	18	23	35		29	NP	Poor Quality DNA
SD-J-98	2	18	55			53	NP	Poor Quality DNA
SD-J-99	7	8	18	39		32	NP	Poor Quality DNA
SD-J-100	11	27	38			27	NP	Poor Quality DNA

Appendix C: Lactose Tolerance data for Ethiopian Somali Cohort

Sample ID	LTT: Baseline (ppm)	LTT: T=30	LTT: T = 60	LTT: T = 90	LTT: T = 120	LTT: T = 150	LTT: T = 180	LTT: RISE	LTT: STATUS	COMMENTS
DD-001	8	16	68					60	NP	
DD-002	12	9	7					-3		
DD-003	11	31	50					39	NP	
DD-004	5	0	0	0	31	66	66	61	NP	
DD-005	0	0	0	0	0	0		0		
DD-006	0	0	0	0	15	42	57	57	NP	
DD-007	5	9	58	87	90	84	80	85	NP	
DD-008	0	6	68	103	84	86	147	147	NP	
DD-009	0	0	0	0	0	0	0	0		
DD-010	17	15	12	25	34	54	57	40	NP	
DD-011	0	5	66	68	47	46	52	68	NP	
DD-012	8	5	8	43	42	55		47	NP	
DD-013	10	19	8	6	7	6	10	9	P	
DD-014	0	13	37	33	74	57	55	57	NP	
DD-015	0	0	0	0	0	0	9	9	P	
DD-016	7	9	0	22	77	106	93	99	NP	
DD-017	0	0	0	53	43	39	36	43	NP	
DD-018	0	0	0	10	15	27	34	34	NP	
DD-019	6	7	0	0	0	0	5	-1	P	
DD-020	0	6	5	21	53	60	47	60	NP	
DD-021	0	0	0	0	63	127	121	127	NP	
DD-022	0	0	0	6	10	12	16	16	I	
DD-023	13	13	13	74	41	62	58	61	NP	
DD-024	0	0	0	0	14	127	133	133	NP	
DD-025	6	0	36	42	83	102		96	NP	
DD-026	0	0	0	0	0	0	0	0		
DD-027	5	5	5	0	10	16	15	11	P	
DD-028	0	0	0	18	18	19	19	19	I	
DD-029	18	24	25	49	77	138	105	120	NP	
DD-030	5	0	9	117	127	114	72	122	NP	
DD-031	10	7	6	6	0	0	5	-5	P	
DD-032	0	5	18	41	66	89	87	89	NP	
DD-033	0	0	0	0	0	0	0	0		
DD-034	0	0	15	38	35	30		38	NP	
DD-035	0	0	37	80	113			113	NP	
DD-036	7	25	60	58	76			69	NP	
DD-037	0	0	0	0	0	0	0	0		
DD-038	16	20	64	79	71	86	78	70	NP	
DD-039	5	0	6	39	49	51	47	46	NP	
DD-040	0	6	5	0	0	0	0	6	P	
DD-041	7	7	7	7	17	23	23	16	I	
DD-042	9	10	6	5	5	8	23	14	P	
DD-043	0	0	0	0	9	18	19	19	I	
DD-044	7	6	5	15	36	43	73	66	NP	
DD-045	16	30	34	22	25	40	32	24	I	
DD-046	8	12	11	7	10	8	39	31	NP	
DD-047	5	5	0	0	0	0	0	-5	P	
DD-048	0	0	0	0	0	33	57	57	NP	
DD-049	10	14	38	68	66	33	34	58	NP	
DD-050	0	5	0	0	0	8	9	9	P	

Sample ID	LTT: Baseline (ppm)	LTT: T=30	LTT: T = 60	LTT: T = 90	LTT: T = 120	LTT: T = 150	LTT: T = 180	LTT: RISE	LTT: STATUS	COMMENTS
DD-051	0	0	0	0	0	0	6	6	P	
DD-052	0	51	112	162	158	172	127	172	NP	
DD-053	0	0	0	8	12	18	17	18	I	
DD-054	0	0	25	53	24	60	39	60	NP	
DD-055	6	5	0	12	11	14	13	8	P	
DD-056	0	0	0	0	0	12	44	44	NP	
DD-057	11	16	7	6	34	31	55	44	NP	
DD-058	0	13	51	76	69	70	53	76	NP	
DD-059	9	7	0	0	0	0	7	-2	P	
DD-060	7	6	6	0	0	0	0	-7	P	Poor Quality DNA
DD-061	14	10	7	6	15	47	67	53	NP	
DD-062	0	6	0	5	5	0	39	39	NP	
DD-063	0	0	11	132	108	106	92	132	NP	
DD-064	0	8	48	84	72	99	97	99	NP	
DD-065	0	0	0	47	66	72	64	72	NP	
DD-066	0	6	58	53	80	77	74	80	NP	
DD-067	8	13	6	6	0	0	5	5	P	
DD-068	7	7	10	14	29	136	136	129	NP	
DD-069	5	0	0	0	0	0	5	0	P	
DD-070	0	0	0	0	0	0	0	0	[redacted]	
DD-071	0	5	9	54	68	65	48	68	NP	
DD-072	0	0	0	0	0	0	0	0	[redacted]	
DD-073	10	23	17	13	10	12	29	19	I	
DD-074	0	0	0	6	7	9	11	11	P	
DD-075	7	19	40	63	51	61	60	56	NP	
DD-076	0	35	43	55	50	43	57	57	NP	
DD-077	0	0	0	14	22	23	27	27	NP	
DD-078	17	12	146	154	153	185	123	168	NP	
DD-079	0	0	0	10	54	86	120	120	NP	
DD-080	0	58	110	168	99	146	91	168	NP	
DD-081	0	0	0	60	126	97	79	126	NP	
DD-082	6	13	16	11	12	21	27	21	I	
DD-083	0	0	0	0	0	5	40	40	NP	
DD-084	0	0	0	0	0	0	5	5	P	
DD-085	10	17	19	38	38	45	53	43	NP	
DD-086	5	23	36	62	84	88	68	83	NP	
DD-087	12	12	10	11	15	29	55	43	NP	
DD-088	10	11	34	64	97	120	90	110	NP	
DD-089	9	9	8	7	11	16	11	7	P	
DD-090	21	28	66	92	90	102	85	81	NP	
DD-091	0	0	0	0	7	62	49	62	NP	
DD-092	0	0	0	0	0	0	0	0	[redacted]	
DD-093	6	6	5	0	0	0	0	-6	P	
DD-094	19	9	5	0	0	0	6	-13	P	
DD-095	0	0	0	0	0	8	29	29	NP	
DD-096	10	0	19	65	73	95	80	85	NP	
DD-097	0	12	13	12	47	59	47	59	NP	
DD-098	0	0	8	8	20	22	13	22	I	
DD-099	5	12	8	10	27	75	67	70	NP	
DD-100	12	19	14	19	13	14	9	7	P	
DD-101	0	0	21	47	67	66	54	67	NP	
DD-102	0	0	0	5	10	34	92	92	NP	

Sample ID	LTT: Baseline (ppm)	LTT: T=30	LTT: T = 60	LTT: T = 90	LTT: T = 120	LTT: T = 150	LTT: T = 180	LTT: RISE	LTT: STATUS	COMMENTS
DD-103	0	13	12	12	24	23	12	24	**I**	
DD-104	1	13	54	37	36	86	71	85	**NP**	
DD-105	0	6	0	33	25	23	24	33	**NP**	Poor Quality DNA
DD-106	8	5	0	0	0	0	0	-3	**P**	
DD-107	11	11	20	59	61	48	80	69	**NP**	
DD-108	0	0	0	9	11	13	25	25	**I**	
DD-109	0	6	14	15	39	27	68	68	**NP**	
DD-110	15	13	11	14	40	61	35	46	**NP**	
DD-111	6	0	7	22	17	25	16	19	**I**	

Appendix D: Somali Haplotypes inferred using Phase (best pairs).
SNPs typed: **1.** rs309180; **2.** rs4954493; **3.** *-14010G>C*; **4.** *-14009T>G*; **5.** *-13915T>G*; **6.** *-13910C>T*; **7.** *-13907C>G*; **8.** *-13806A>G*; **9.** rs4954492; **10.** rs4954490; **11.** rs3769005; **12.** *-942/943TC>ΔΔ*; **13.** *-678A>G*; **14.** 666G>A; **15.** 5579T>C. Posterior probabilities are given for haplotype inference of just the Somali group alone (A) and haplotypes inferred with other populations, B (Afar, Jaali, European, Fulani and unphenotyped Somali, n = 358). Nucleotides in bold red are inferred.

																	Posterior probability	
Sample ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	H	A	B
DD-001a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1	0.039	0.962
DD-001b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2		
DD-002a	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	0.992	0.958
DD-002b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-003a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.992	1
DD-003b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5		
DD-004a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-004b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-005a	C	C	G	T	T	C	C	A	T	T	G	TC	A	G	T	H6		
DD-005b	C	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H7	0.735	0.66
DD-006a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-006b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-007a	T	T	G	T	T	C	C	A	T	T	C	TC	A	G	C	H8		
DD-007b	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1	0.986	0.99
DD-008a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9		
DD-008b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.972	0.979
DD-009a	T	C	G	T	T	T	C	A	T	T	C	TC	A	G	C	H10		
DD-009b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.994	0.998
DD-010a	C	T	G	T	T	C	C	A	G	C	G	TC	A	G	T	H11		
DD-010b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.017	0.818
DD-011a	C	C	G	T	T	C	C	A	G	C	G	TC	A	A	T	H12		
DD-011b	T	C	G	T	T	C	C	A	T	T	G	TC	A	G	C	H13	0.589	0.852
DD-012a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-012b	T	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H14	0.061	0.677
DD-013a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9		
DD-013b	C	C	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H15	0.88	0.958
DD-014a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-014b	C	C	G	T	T	C	C	G	T	C	G	TC	G	G	T	H16	0.935	0.998
DD-015a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9		
DD-015b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.973	0.979
DD-016a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-016b	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1	1	1
DD-017a	T	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H14		
DD-017b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-018a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-018b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.029	0.932

DD-019a	T	C	G	T	T	T	C	A	T	T	C	TC	A	G	C	H10		
DD-019b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	0.993	0.999
DD-020a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-020b	C	C	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H15	1	1
DD-021a	T	C	G	T	T	C	C	A	T	C	C	TC	A	G	T	H18		
DD-021b	T	C	G	T	T	C	C	A	T	C	C	TC	A	G	T	H18	1	1
DD-022a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-022b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-023a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-023b	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1	1	1
DD-024a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2		
DD-024b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	1	1
DD-025a	C	C	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H15		
DD-025b	C	C	G	T	T	C	C	A	T	C	G	TC	A	G	T	H19	0.534	0.378
DD-026a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9		
DD-026b	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	1	1
DD-027a	C	C	G	T	T	C	C	A	T	C	C	TC	A	A	T	H20		
DD-027b	C	C	G	T	G	C	C	A	T	C	C	TC	G	G	T	H21	0.929	0.855
DD-028a	C	C	G	T	T	C	C	A	T	C	C	ΔΔ	A	A	T	H22		
DD-028b	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1	0.623	0.91
DD-029a	C	C	G	T	T	C	C	A	T	C	C	TC	A	A	T	H20		
DD-029b	C	T	G	T	T	C	C	A	G	C	G	TC	A	G	T	H11	0.177	0.781
DD-030a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-030b	C	C	G	T	T	C	C	A	T	C	C	TC	A	A	T	H20	0.408	0.985
DD-031a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-031b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	0.995	0.999
DD-032a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-032b	C	C	G	T	T	C	C	G	T	C	G	TC	G	G	T	H16	0.926	0.998
DD-033a	T	T	G	T	T	T	C	A	T	T	G	TC	A	G	C	H23		
DD-033b	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1	0.799	0.839
DD-034a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-034b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.029	0.932
DD-035a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2		
DD-035b	C	C	G	T	T	C	C	A	G	C	G	TC	A	A	T	H12	0.793	0.996
DD-036a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-036b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-037a	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5		
DD-037b	C	C	G	G	T	C	C	A	T	C	C	ΔΔ	A	A	C	H24	0.998	0.868
DD-038a	T	T	G	T	T	C	C	A	T	T	C	TC	A	G	C	H8		
DD-038b	C	T	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H25	0.838	0.921
DD-039a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-039b	T	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H14	0.06	0.659
DD-040a	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3		
DD-040b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.97	0.902
DD-041a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-041b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-042a	T	T	G	T	T	C	C	A	T	T	C	TC	A	G	C	H8		
DD-042b	T	T	G	T	T	C	C	A	T	T	C	TC	A	G	C	H8	1	1

DD-043a	C	T	C	T	T	C	C	A	T	C	G	TC	A	A	T	H26		
DD-043b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	0.463	0.88
DD-044a	C	C	G	T	T	C	C	A	T	C	C	TC	A	G	C	H28		
DD-044b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.918	0.663
DD-045a	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3		
DD-045b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.992	0.957
DD-046a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-046b	C	C	G	T	T	C	C	A	T	C	G	TC	A	G	T	H19	1	1
DD-047a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2		
DD-047b	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	0.943	0.857
DD-048a	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9		
DD-048b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5	0.986	0.978
DD-049a	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27		
DD-049b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	1	1
DD-050a	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5		
DD-050b	C	C	G	G	T	C	C	A	T	C	C	ΔΔ	A	A	C	H24	0.998	0.868
DD-051a	C	T	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H25		
DD-051b	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	0.177	0.943
DD-052a	C	T	G	T	T	C	C	A	G	C	G	TC	A	G	T	H11		
DD-052b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2	0.023	0.958
DD-053a	C	T	G	T	T	C	C	A	G	C	G	TC	A	G	T	H11		
DD-053b	T	C	G	T	T	T	C	A	T	T	C	TC	A	G	C	H10	0.211	0.998
DD-054a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-054b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5	0.992	1
DD-055a	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27		
DD-055b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5	0.998	1
DD-056a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-056b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-057a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-057b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-058a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-058b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	1	0.999
DD-059a	C	T	G	T	G	C	C	A	T	C	G	TC	G	G	T	H29		
DD-059b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	N/A	0.973
DD-061a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-061b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-062a	C	C	G	T	T	C	C	A	G	C	G	TC	A	A	T	H4		
DD-062b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	0.938	0.998
DD-063a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-063b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	1	0.999
DD-064a	T	T	G	T	T	C	C	A	T	T	C	TC	A	G	T	H30		
DD-064b	T	T	G	T	T	C	C	A	T	T	C	TC	A	G	C	H8	1	1
DD-065a	C	C	G	T	T	C	C	A	T	C	C	ΔΔ	A	A	T	H22		
DD-065b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.92	0.929
DD-066a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-066b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-067a	C	C	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H15		
DD-067b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	0.998	1

DD-068a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-068b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.029	0.932
DD-069a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-069b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	0.995	0.998
DD-070a	C	T	G	T	T	C	C	A	T	C	G	TC	G	G	T	H31		
DD-070b	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	0.012	0.969
DD-071a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-071b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-072a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-072b	C	C	G	T	T	C	C	A	T	C	G	TC	A	G	T	H19	1	1
DD-073a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-073b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-074a	T	T	G	T	T	C	G	A	T	T	C	TC	A	G	C	H32		
DD-074b	T	T	G	T	T	C	G	A	T	T	C	TC	A	G	C	H32	1	1
DD-075a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-075b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.029	0.932
DD-076a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-076b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	C	H33	1	1
DD-077a	C	T	G	T	T	C	C	A	G	C	G	TC	A	A	T	H34		
DD-077b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2	N/A	0.961
DD-078a	C	C	G	T	T	C	C	A	T	C	C	ΔΔ	A	A	T	H22		
DD-078b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.917	0.952
DD-079a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-079b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-080a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-080b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-081a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-081b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-082a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-082b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.029	0.932
DD-083a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-083b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	0.029	0.932
DD-084a	C	C	G	T	T	C	C	A	T	C	G	ΔΔ	A	A	T	H15		
DD-084b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	0.302	0.226
DD-085a	C	T	G	T	T	C	C	A	T	C	G	TC	G	G	T	H31		
DD-085b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2	N/A	0.959
DD-086a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-086b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-087a	C	C	G	T	T	C	C	A	T	C	C	ΔΔ	A	A	C	H35		
DD-087b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	0.836	0.819
DD-088a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-088b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5	0.992	1
DD-089a	C	C	G	T	T	C	C	A	T	C	C	TC	A	A	T	H20		
DD-089b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	0.834	0.973
DD-090a	C	C	G	T	T	C	C	A	T	C	C	TC	A	A	T	H20		
DD-090b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	1	1
DD-091a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-091b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5	0.992	1

DD-092a	C	T	G	T	G	C	C	A	T	C	G	TC	G	G	T	H29		
DD-092b	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	**N/A**	**0.981**
DD-093a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-093b	C	C	G	G	T	C	C	A	T	C	C	ΔΔ	A	A	C	H24	**0.998**	**0.993**
DD-094a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-094b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	**0.995**	**0.998**
DD-095a	C	C	G	T	T	C	C	A	T	C	C	ΔΔ	A	A	T	H22		
DD-095b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	**0.939**	**0.995**
DD-096a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-096b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	**1**	**1**
DD-097a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-097b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	**0.029**	**0.933**
DD-098a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-098b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	**1**	**1**
DD-099a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-099b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	**1**	**1**
DD-100a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-100b	T	C	G	T	T	C	G	A	T	T	C	TC	A	G	C	H3	**0.068**	**0.925**
DD-101a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-101b	C	C	G	T	T	C	C	A	T	C	G	TC	G	G	T	H27	**1**	**0.999**
DD-102a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-102b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	**0.029**	**0.932**
DD-103a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-103b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	**0.029**	**0.932**
DD-104a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-104b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	**1**	**1**
DD-106a	C	C	G	T	T	C	C	A	T	C	C	ΔΔ	A	A	C	H35		
DD-106b	C	C	G	T	G	C	C	A	T	C	G	TC	G	G	T	H17	**0.911**	**0.892**
DD-107a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-107b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	T	H2	**N/A**	**0.905**
DD-108a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-108b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	**1**	**1**
DD-109a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-109b	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4	**1**	**1**
DD-110a	C	T	G	T	T	C	C	A	T	C	G	TC	A	A	T	H1		
DD-110b	T	C	G	T	T	C	C	A	T	T	C	TC	A	G	C	H9	**0.04**	**0.977**
DD-111a	C	C	G	T	T	C	C	A	T	C	G	TC	A	A	T	H4		
DD-111b	C	C	G	T	T	C	C	A	G	C	G	TC	A	G	T	H5	**0.992**	**1**

Appendix E: Frequencies of MCM6 intron 13 alleles. Data compiled from publications which sequenced the intron 13 region. SNPs included in this table are those that would have been included in sequencing fragments for all studies -14025 to -13806. Source references are: Myles, *et al.*, Hum. Genet. **117** (1), 34 (2005); Tishkoff, *et al.*, Nat. Genet. **39** (1), 31 (2007); Enattah, *et al.*, Am. J. Hum. Genet. **82** (1), 57 (2008); Imtiaz, *et al.*, J. Med. Genet. **44** (10), e89 (2007) and data included within this thesis. Samples only included in final geographic map if split into identity by country. Data not included in geographic map indicated by *.

Continent	Country	Ethnic group	*n*	*14025 A>G*	*14010 C>G*	*14009 T>G*	*13915 T>G*	*13913 T>C*	*13910 C>T*	*13907 C>G*	*13806 A>G*	Source Reference
Africa	Algeria	Berber Mzab	66	0.00	0.00	0.00	0.00	0.00	**0.17**	0.00	0.00	Myles 2005
Africa	Cameroon	Fulani	110	0.01	0.00	0.00	0.00	0.04	**0.39**	0.00	0.00	Ingram 2008
Africa	Cameroon	Mambila	74	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Ingram 2008
Africa	Cameroon	Shuwa Arab	30	0.00	0.00	0.00	**0.13**	0.00	0.00	0.00	0.00	Ingram 2008
Africa	Ethiopia	Afar	74	0.00	0.00	0.01	**0.12**	0.01	0.01	**0.30**	0.01	Ingram 2008
Africa	Ethiopia	Amharic	38	0.00	0.00	0.00	**0.13**	0.00	0.00	**0.05**	0.00	Ingram 2008
Africa	Ethiopia	Phenotyped Somali	218	0.00	0.01	0.01	0.05	0.00	0.02	**0.06**	0.01	Ingram 2008
Africa	Ethiopia	Somali	74	0.00	0.03	0.01	0.04	0.01	0.00	**0.10**	0.02	Ingram 2008
Africa	Kenya	Maasai	64	0.00	**0.58**	0.00	0.00	0.00	0.00	0.03	0.00	Tishkoff 2007
Africa	Kenya	Marakwet	14	0.00	**0.36**	0.00	**0.07**	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Nandi	8	0.00	**0.25**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Ogiek	22	0.00	**0.36**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Pokot	28	0.00	**0.29**	0.00	0.04	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Sabaot	12	0.00	**0.17**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Samburu	18	0.00	**0.28**	0.00	**0.06**	0.00	0.00	**0.06**	0.00	Tishkoff 2007
Africa	Kenya	Sengwer	32	0.00	**0.06**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Tugen	32	0.00	**0.19**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Turkana	26	0.00	**0.21**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Kikuyu	4	0.00	**0.75**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Kenya	Burji	16	0.00	**0.06**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Morocco		24	0.00	0.00	0.00	**0.08**	0.00	**0.21**	0.00	0.00	Enattah 2008

Continent	Country	Ethnic group	*n*	*14025 A>G*	*14010 C>G*	*14009 T>G*	*13915 T>G*	*13913 T>C*	*13910 C>T*	*13907 C>G*	*13806 A>G*	Source Reference
Africa	Morocco (High-Atlas)	Amizmiz	78	0.00	0.00	0.00	0.00	0.00	**0.14**	0.00	0.00	Myles 2005
Africa	Morocco (Mid-Atlas)	Berber Moyen-Atlas	66	0.00	0.00	0.00	0.00	0.00	**0.16**	0.00	0.00	Myles 2005
Africa		Saharawi*	22	0.00	0.00	0.00	**0.18**	0.00	**0.23**	0.00	0.00	Enattah 2008
Africa	Senegal	Wolof	118	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Ingram 2008
Africa	Sudan	Mahas	30	0.00	0.00	0.00	**0.17**	0.00	0.00	0.00	0.00	Enattah 2008
Africa	Sudan	Gaali	20	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	Enattah 2008
Africa	Sudan	Beni Amer	162	0.00	0.00	**0.11**	**0.25**	0.00	0.01	0.01	0.00	Ingram 2008
Africa	Sudan	Dunglawi	12	**0.08**	0.00	0.00	0.00	0.00	0.00	**0.08**	0.00	Ingram 2008
Africa	Sudan	Jaali	172	0.00	0.00	**0.06**	**0.13**	0.01	0.01	0.01	0.00	Ingram 2008
Africa	Sudan	Shaigi	18	0.00	0.00	**0.17**	**0.06**	0.00	0.00	0.00	0.00	Ingram 2008
Africa	Sudan	Beja (Banuamir)	12	0.00	0.00	0.00	**0.17**	0.00	0.00	**0.25**	0.00	Tishkoff 2007
Africa	Sudan	Beja (Hadandawa)	22	0.00	0.00	0.00	**0.09**	0.00	0.00	**0.18**	0.00	Tishkoff 2007
Africa	Sudan	Dinka	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Sudan	Koalib	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Sudan	Liguri/Logorik	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Sudan	Masalit	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Sudan	Nuer	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Sudan	Ama	4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Sudan	Shilook	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Burunge	36	0.00	**0.38**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Iraqw	78	0.00	**0.58**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Mbugu	60	0.00	**0.31**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Fiome	24	0.00	**0.55**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Akie	28	0.00	**0.25**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Datog	8	0.00	**0.63**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Dorobo	20	0.00	**0.40**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Maasai	38	0.00	**0.45**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Mbugwe	26	0.00	**0.27**	0.00	0.04	0.00	0.00	0.00	0.00	Tishkoff 2007

Continent	Country	Ethnic group	n	*14025 A>G*	*14010 C>G*	*14009 T>G*	*13915 T>G*	*13913 T>C*	*13910 C>T*	*13907 C>G*	*13806 A>G*	Source Reference
Africa	Tanzania	Pare	20	0.00	**0.10**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Rangi	**70**	0.00	**0.27**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Samba'a	**6**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Hadza	**36**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Africa	Tanzania	Sandawe	**62**	0.00	**0.13**	0.00	0.00	0.00	0.00	0.00	0.00	Tishkoff 2007
Asia	Iran	Iranians*	**42**	0.00	0.00	0.00	0.00	0.00	**0.10**	0.00	0.00	Enattah 2008
Asia	Israel	Druze	**28**	0.00	0.00	0.00	**0.11**	0.00	0.04	0.00	0.00	Ingram 2008
Asia	Israel	Israeli Arab	**160**	0.00	0.00	0.01	0.05	0.03	0.00	0.00	0.00	Ingram 2008
Asia	Israel	Israeli Bedouin	**38**	0.00	0.00	0.00	**0.13**	0.00	0.03	0.00	0.00	Ingram 2008
Asia	Israel/PAA	Palestinian Arab	**36**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Ingram 2008
Asia	Jordan		**112**	0.00	0.00	0.00	**0.05**	0.01	**0.05**	0.00	0.00	Enattah 2008
Asia	Jordan	Jordanian Bedouin	**46**	0.02	0.00	0.00	**0.35**	0.00	0.00	0.00	0.00	Ingram 2008
Asia	Saudi Arabia	Bedouin	**94**	0.01	0.00	0.00	**0.48**	0.01	0.00	0.00	0.00	Ingram 2008
Asia	Saudi Arabia	Central	**180**	0.00	0.00	0.00	**0.61**	0.00	0.00	0.00	0.00	Imtiaz 2007
Asia	Saudi Arabia	Eastern	**164**	0.00	0.00	0.00	**0.62**	0.00	0.00	0.00	0.00	Imtiaz 2007
Asia	Saudi Arabia	Western	**172**	0.00	0.00	0.00	**0.65**	0.00	0.01	0.00	0.00	Imtiaz 2007
Asia	Saudi Arabia	Southern	**184**	0.00	0.00	0.00	**0.58**	0.00	0.00	0.00	0.00	Imtiaz 2007
Asia	Saudi Arabia	Northern	**164**	0.00	0.00	0.00	**0.52**	0.00	0.01	0.00	0.00	Imtiaz 2007
Asia	Saudi Arabia		**248**	0.00	0.00	0.00	**0.57**	0.00	0.00	0.01	0.00	Enattah 2008
Asia	Syria, Iraq, Lebanon, Palestine	Arabs*	**40**	0.00	0.00	0.00	**0.11**	0.00	**0.13**	0.00	0.00	Enattah 2008
Europe	Finland	Finns*	**1876**	0.00	0.00	0.00	0.00	0.00	**0.58**	0.00	0.00	Enattah 2008
Europe	Italy	S. European*	**66**	0.00	0.00	0.00	0.00	0.00	**0.09**	0.00	0.00	Ingram 2008
Europe	Mixed	N. European*	**110**	0.00	0.00	0.00	0.00	0.00	**0.62**	0.00	0.00	Ingram 2008
Total chromosomes sequenced = 6154												

Appendix F. Alignment of the MCM6 enhancer region. Exon 13 is indicated by red font. Conserved bases are highlighted in blue and known SNPs are highlighted in green with bases indicated according to the IUPAC Code (M = A or C, S = C or G, W = A or T, R = A or G, Y = C or T). Species aligned: Hsap, Human (Homo sapiens); Mmul, Macaque (Macaca mulatta); Mmus, Mouse (Mus musculus); Ptro, Chimpanzee (Pan troglodytes). The OCT1 binding site is outlined in dark blue, and position 14010 is indicated by an asterisk above the relevant base pair.

Appendix G: Publications by the candidate arising from work relating to this thesis:

1. Ingram,C.J.E., Elamin,M.F., Mulcare,C.A., Weale,M.E., Tarekegn,A., Raga,T.O., Bekele,E., Elamin,F.M., Thomas,M.G., Bradman,N. and Swallow,D.M. (2007). A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum. Genet.* **120,** 779-788.

2. Ingram, C.J.E. and Swallow, D.M. (2007) Population Genetics of Lactase Persistence and Lactose Intolerance. In: Encyclopedia Of Life Sciences. John Wiley & Sons, Ltd: Chichester http://www.els.net/ [doi: 10.1002/9780470015902.a0020855].

3. Weiskirchen,R., Tag,C.G., Mengsteab,S., Gressner,A.M., Ingram,C.J.E. and Swallow,D.M. (2007). Pitfalls in LightCycler diagnosis of the single-nucleotide polymorphism 13.9 kb upstream of the lactase gene that is associated with adult-type hypolactasia. *Clin. Chim. Acta.* **384,** 93-98.

4. Tag,C.G., Oberkanins,C., Kriegshauser,G., Ingram,C.J., Swallow,D.M., Gressner,A.M., Ledochowski,M. and Weiskirchen,R. (2008). Evaluation of a novel reverse-hybridization StripAssay for typing DNA variants useful in diagnosis of adult-type hypolactasia. *Clin. Chim. Acta.* **392,** 58-62.

Hum Genet
DOI 10.1007/s00439-006-0291-1

ORIGINAL INVESTIGATION

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

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

Received: 20 September 2006 / Accepted: 25 October 2006
© Springer-Verlag 2006