Evolutionary and molecular genetics of regulatory alleles responsible for lactase persistence

Anke Liebert

Submitted for the Doctor of Philosophy degree at University College London

July 2014

Research Department of Genetics, Evolution and Environment
University College London (UCL)

I, Anke Liebert 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

Persistence of lactase into adulthood varies in frequency worldwide and is attributable to several different single nucleotide changes in an enhancer of the *LCT* gene. One of these is at particularly high frequencies in Europeans and several others have been found elsewhere. However, information about their worldwide distribution is patchy.

2056 DNA samples from populations of Europe, Asia and the Middle East were sequenced to examine the distribution of allelic variants of the *LCT* enhancer region. It was confirmed that *-13910*T* is also the predominant allele around the periphery of Europe, and that this allele extends as far as the South and East of the Arabian Peninsula. Other alleles appear to have spread out of Africa or Arabia and most variation was found in the Middle East. No new common alleles were found that were likely to be causal.

In previous studies four lactase persistence associated nucleotide substitutions *(-13910*T, -14010*C*, *-13915*G* and *-13907*G*) have been studied functionally. In this thesis four additional enhancer alleles were examined using enhancer/promoter construct transfections and electrophoretic mobility shift assays. Three enhancer variants alter transcription factor binding *in vitro* and/or reporter-gene expression. Bioinformatic tools and specific antibodies were used to assist in identifying the transcription factors involved. The results show that different mechanisms lead to a disruption of the normal down-regulation of lactase in adult life.

Known haplotypic markers were assessed on an overlapping set of samples and a greater number of flanking markers was typed in an extended region of 1.8 Mb around *LCT*, in order to chart the evolutionary relationships and extent of historic recombination of the chromosomes carrying the derived alleles, as well as those that do not. All of the functional alleles tested have longer extended haplotypes than their ancestral counterparts, but so also does the derived variant at position -958, a haplotype marker for the B haplotype. The finding of an extended region of high linkage disequilibrium in all populations, and an extended B haplotype, is discussed in relation to the methods to study selection.

Because phenotypic studies suggest missing functional variants, the immediate promoter and a part of intron 2 of *LC*T were also selected for sequencing. No serious candidates were found in intron 2. Two alleles in the immediate promoter were studied by transfections but there was little evidence for any *in vitro* effect of these variants.

Acknowledgements

First of all I would like to thank my supervisors Professor Dallas Swallow and Professor Mark Thomas for giving me insights in their quite complementary methods of approaching the field of genetics. A special thank you goes to Professor Dallas Swallow for her continuous support and inspiring discussions as well as her patience and trust throughout this thesis.

I am very grateful to our collaborator Professor Jesper Troelsen for making my time in Denmark an enjoyable experience, for sharing his expertise in the world of proteins and introducing me to the Danish mentality. I would also like to thank Tine Jensen, Thomas Danielsen, Anders Krüger Olsen, Lotte Laustsen, Lotte Bram, Mette Juel Riisager and Sisse Marie Schmidt for their help and good company in Copenhagen and Roskilde.

Thanks to all the present and past people from the Darwin first floor office(s): Nic Montalva (for the constant coffee supply), Rosemary Ekong (for her special pieces of advice), Mirna Kovacevic, Kate Brown, Anna Rudzinski, Adrian Timpson, Heather Elding (hmm, those cakes), Bryony Jones, Ru Bains, and Chris Plaster for being around and brighten up my days in the office, especially with the sometimes mad conversations, within and outside of UCL.

Thanks to the wonderful LeCHE group for all the experiences, thoughts and adventures and the nice boat trips. A special thank you goes to Pascale Gerbault for being who she is and all the things we have shared.

A big thank you goes to Mari-Wyn Burley and Ranji Arasaretnam for their support in the lab and to Pascale Gerbault, Winston Lau and Yuval Itan for their contributions of programming work for this thesis.

Funding for this thesis was provided by the EU Marie Curie FP7, and the Annals of Human Genetics. I would also like to thank the sample donors and collectors of the G.E.E. DNA collections made available by DS and MT as well as Neil Bradman, Rosemary Ekong and Andres Ruiz-Linares.

A special thank you to my friends Kathrin Büttner, Sabrina Dettmer, Anique Lauber, Johanna Kirchhoff, Alice Albers-Westphal, Madlen Marschka, Caro and Pierre Bernoth and many more for all the good times outside of academia and helping me fill up my energy tanks when needed.

Most of all, I wish to thank my family: My parents Hannelore and Gerd, for their continuous encouragement and understanding, and my two beloved men, Maxim and Ben. Maxim for his support and patience, especially through the last months of this PhD, and both for all the nice disruptions and for making me smile almost every day.

Abbreviations/Glossary

bp	base pair
BP	before present
cal BC	years before Christ, from calibrated C^{14} data
cDNA	complementary DNA
CEPH	Centre d'Etude du Polymorphisme Humain
ChIP	chromatin immunoprecipitation
CNV	copy-number variation
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
EHH	extended haplotype homozygosity
EMSA	electrophoretic mobility shift assay
GWAS	genome-wide association studies
HWE	Hardy-Weinberg equilibrium
HT	haplotype (combination of alleles at adjacent locations (loci) on a chromosome that are inherited together)
ID	identification
kb	kilobase
kDa	kilodalton
LCT	lactase gene
LD	linkage disequilibrium (allelic association)
LDU	linkage disequilibrium units
LeCHE	Lactase persistence in the early cultural history of Europe
LP	lactase persistence
LNP	lactase non-persistence
LPH	lactase-phlorizin hydrolase
Mb	megabase
MCM6	minichromosome maintenance complex component 6
mRNA	messenger RNA
mtDNA	mitochondrial DNA
OMIM	Online Mendelian Inheritance in Man
PBS	phosphate buffered saline

PHASE	software that determines the chromosomal phase (the maternal or paternal chromosome on which the alleles lie in heterozygous individuals) by Bayesian inference
PCO	principal co-ordinates
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
SDS	sodium dodecyl sulfate
SNP	single nucleotide polymorphism
STR	short tandem repeat
TF	transcription factor
UniProtKB	Protein knowledgebase of the Universal protein resource (UniProt)

Table of Contents

Contents of Tables

Contents of Figures

1 Introduction

Human influence on the global ecosystem has increased and has become substantial, especially in recent years, even leading to suggestions of naming a new era, the 'Anthropocene'. No doubt, human culture has modified environments, but in reverse it was the ability to adapt well to environmental changes, phenotypically and culturally, that has made Homo sapiens successful as a species, being able to populate most parts of the world. Human genetic diversity we see today is the result of past evolutionary processes and one example of how adaptation to a new diet and a different lifestyle of former populations has shaped present genetic variation is lactase persistence.
This thesis is concerned with the evolutionary, molecular and population genetics of the variation in regulatory sequences responsible for altering the expression of-human lactase, and leading to adult lactase persistence. This introduction reviews the relevant background, and includes information about past migrations and domestication processes, previous work related to lactase persistence and the lactase gene and an overview-of methods for studying genetic variation and detecting selection, followed by the aims and an overview of the thesis.

1.1 Neolithisation – a revolutionary cultural change

The European and Asian continents have been inhabited by modern humans since their dispersal out of Africa from about 50,000 years ago (reviewed in Stoneking and Krause 2011).

The Neolithic era, which marks the transition from a hunter-gatherer lifestyle to farming and hence food production, represents a remarkable change in life circumstances which is well described by the term 'Agricultural revolution' coined-by Childe (1936). It broadly involved the appearance of a package of innovations, also called the 'Neolithic package' within a few thousand years, which included the domestication of plants and animals, the use of polished stone tools, the manufacture of pottery and the appearance of permanent settlements. However, these innovations developed as gradual processes and at slightly different times across the world (Figure 1.1). The boundaries between Mesolithic and Neolithic lifestyles also often overlap, as there is evidence of hunting and gathering by Neolithic societies and typical Neolithic pottery that was found at Mesolithic

archaeological sites. The stepwise development of agriculture, where domesticated plants and animals like sheep, goat, pig and cattle increasingly built food resources, allowed the settlement of larger groups of people in permanent dwellings. This in turn allowed the development of new technologies and changed the social structures within societies, presumably accompanied by an alteration of mindset of its members (reviewed in Bellwood 2008).

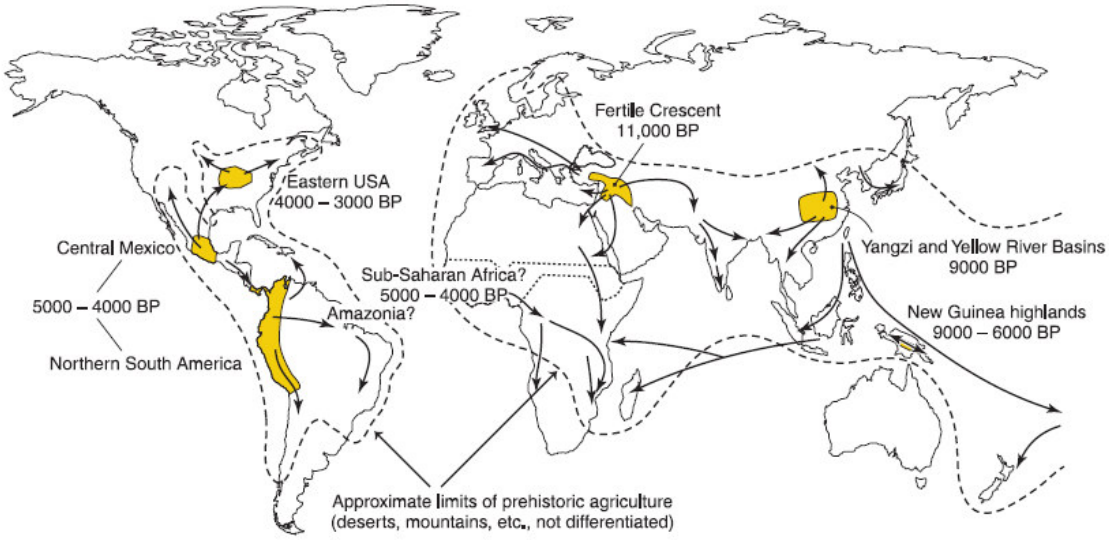


Figure 1.1: Origins and approximate expansion of agricultural systems and early farming cultural complexes from archaeological records (Diamond and Bellwood 2003).

1.1.1 The Neolithic in Europe

The process of Neolithisation relevant to Europe began around 15000-12000 years ago (Bellwood 2008) in the region of the so called 'Fertile Crescent' (Breasted 1916). This 'Neolithic core zone' extended from the Levant to Southwestern Iran via Southern Anatolia. Substantial climate changes took place at that time - the late Pleistocene to early Holocene. The temperature started to rise and the world climate became relatively stable, which would have allowed a better management of crops (reviewed in Bellwood 2008). Hunting and gathering was the only subsistence strategy across Europe before 8500 years before present (BP), until farming started to spread from the Neolithic core zone, as shown in Figure 1.2. Evidence suggests that the Neolithic culture started to spread from the Fertile Crescent to other parts of Asia and Europe and the North of Africa at approximately the same time. Two models have been suggested as to how farming could have spread into Europe. The 'cultural diffusion' or 'acculturation' model proposes that the Neolithic culture was adopted by local hunter-gatherer populations from their farming

neighbours, whereas under the 'demic diffusion' model the Neolithic lifestyle spread with the migration of farmers into new territories (Ammerman and Cavalli-Sforza 1984). In fact, it is likely that the Neolithisation process was more complex, with local differences (Gronenborn 2007). Palaeogenetic data support a demic diffusion model, but also showed the coexistence of hunter-gatherer and farmer populations over longer time periods (Bollongino et al. 2013; Bramanti et al. 2009; Pinhasi et al. 2012).

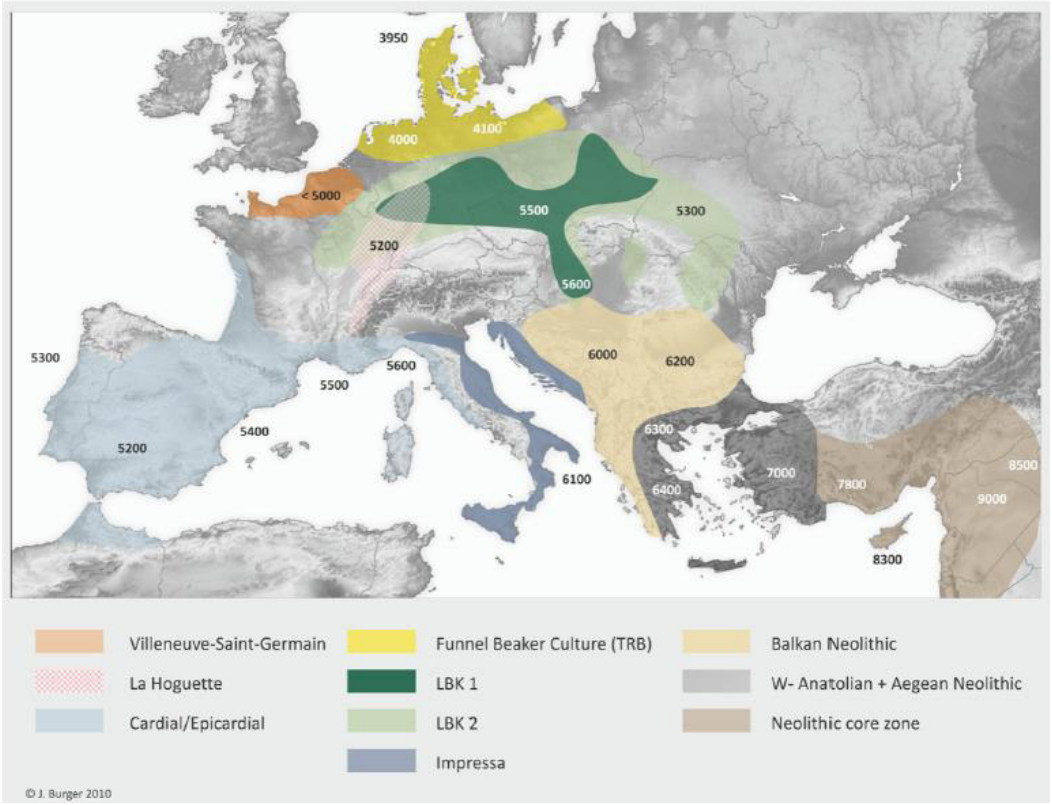


Figure 1.2: Approximate arrival dates and expansions of Neolithic cultures across Europe. Note that dates are given as years before Christ (cal BC). The picture was taken from Burger and Thomas (2011).

1.1.2 Domestication of animals

Bone assemblages from archaeological sites can reveal information about the domestication process. Besides the analyses for species affiliation, morphology, sex, age at death and cutting or cooking signs, bones can be directly radio-carbon dated and analysed for ancient DNA, which puts them in an archaeological context (Gerbault et al. 2011). Differences in morphology can be seen between the wild and domesticated forms of a species, which resulted from the break in gene flow between wild and domesticated animals and the deliberate breeding for certain traits (Vigne and Helmer 2007).

Archaeozoological evidence suggests that goat and sheep were first domesticated about 11000-10500 years ago in the Levant, followed by taurine cattle and pigs at about 10500-10000 BP. The domesticates spread rapidly to the areas of central Anatolia and Cyprus but are only found outside these areas after 8500 BP (Zeder 2008).
Archeological and archeozoological information indicated that the spread of the Neolithic culture and domesticated animals across Europe followed two routes (Figure 1.2): The first, the 'Mediterranean route' stretched along the Aegean and Adriatic Sea, to Italy, the South of France and the Iberian Peninsula. The second, the 'Danubian route' went from the Balkans via the South and West of Central Europe and further to the North. It is not clear where the routes have met but Greece, the Rhine valley and Northwest Europe are discussed as possible regions, before the Neolithic spread further to the British Isles around 6000 BP (Tresset and Vigne 2007).

The evidence for an introduction of domestic goat and sheep into Europe is relatively clear as no wild progenitors of these species have been reported. The Anatolian origin of the domesticated cattle of Central Europe is supported by ancient DNA analysis that found no evidence of mixture with the European wild cattle (Bollongino et al. 2008; Edwards et al. 2007; Troy et al. 2001). The introduced domestic pigs in contrast seem to have been substantially mixed with local wild boar as suggested by ancient DNA data (Larson et al. 2007).

1.1.3 Evidence for dairying

The management strategies of domestic herds are reflected in different patterns of slaughtering of animals (reviewed in Vigne and Helmer 2007), called 'kill off' profiles. The interpretation of these profiles can help to distinguish between culling strategies that more likely focussed on dairying or on meat production. A study on modern sheep herds in Turkey (Payne 1973) showed a massive killing of young and sub-adult males between 6-18 months, as required for meat production. Specialised dairy farming in contrast was characterised by slaughtering of very young lambs, aged below 2 months. This would of course only be true for sheep and goat since cows needed their calves to be present to stimulate milk production. Traditionally, many calves were kept until weaning in order that milk could be obtained from their mothers (reviewed in Vigne and Helmer 2007).
The earliest evidence from archaezoology for specialised breeding for milk use was found after 7500 BP in the Balkan region and 7000 BP in Central Europe. However an earlier practise of milking is very likely (Gerbault et al. 2011). Vigne and Helmer (2007) concluded from their research that milk procurement may have been at least a motivation

for the domestication of sheep, goat and cattle, which developed from stock-keeping by hunter-cultivators, to true Neolithic farmers. Because of the low milk yield at the beginning of dairying it could not be the only subsistence strategy; people must still have been hunting and gathering.

The archeometrical analysis of organic residues in pottery is another method of investigating prehistoric dairying activities. The isotopic composition of fatty acids can provide information about whether residues originated from milk or adipose fats but it cannot distinguish whether fresh or fermented milk was stored in the vessels examined (Copley et al. 2003; Dudd and Evershed 1998). With this approach, the earliest evidence for milk use was found from 8500 BP in Northwest Anatolia and Thrace (Evershed et al. 2008) and from further Neolithic sites in the Carpathian basin from around 7900-7500 BP (Craig et al. 2005), from Poland 7400-6800 BP (Salque et al. 2013), from Scotland from about 6100 BP (Copley et al. 2003) and probably Denmark around 6000 BP (Craig et al. 2011). Pot shards dated to 7200-5800 BP also revealed evidence for dairying practices in Sub Saharan Libya (Dunne et al. 2012).

However, ancient human DNA suggests that the lactase persistence was not common in early Neolithic Europe up to about 5070 BP (Burger et al. 2007; Lacan et al. 2011a; Lacan et al. 2011b; Malmstrom et al. 2010) which will be further discussed later in this thesis.

1.2 Lactase persistence (an example of gene-culture coevolution)

It has been known since Roman times (described by Hippocrates around 400 B.C.) that people show differences in their ability to digest milk. Some can handle large amounts of fresh milk without problem, others suffer from symptoms of intolerance, like abdominal pain, cramps, flatulence or diarrhoea. Much research has been done over the last 50 years, which enables us to explain these differences, although there are still unanswered questions. About a third of the world human population are capable of digesting the lactose in milk into adulthood (reviewed in Ingram et al. 2009a), a trait which is called lactase persistence (LP, OMIM #223100), as the result of a process which is seen as a remarkable example of coevolution: The interaction between genetics causing lactase persistence and the culture of milk drinking.

1.2.1 Lactase persistence status

Newborn mammals rely on milk as their first source of nutrition. Lactose is in most cases the main carbohydrate in milk, and the enzymatic activity of lactase is necessary to cleave lactose into its two absorbable component monosaccharides, glucose and galactose. Hence, lactase is present in the intestine of nearly every young eutherian mammal and its activity is high during the suckling period, except in the Pinnipedia, for example sea lions, where milk lactose is very low and no lactase is expressed (Jenness et al. 1964; Kretchmer 1971; Pilson and Kelly 1962). The enzyme is located in the brush border membrane of the small intestinal mucosa (see also 1.2.6).

After the weaning period the production of the enzyme is usually substantially reduced, to about 5-10%. This reduction is seen as the ancestral state, occurring in most humans and all other mammals tested (Ingram and Swallow 2009b; Sebastio et al. 1989). However, some of us escape this down-regulating mechanism and keep high lactase activity after childhood and throughout life, being called lactase persistent (reviewed in Ingram et al. 2009a). The mechanism for this is not fully understood yet but it was shown that lactase expression cannot be induced by prolonged intake of milk (Keusch et al. 1969).

Differences exist in the timing of the decline of human lactase. Decline can start at a young age from 1 to 5 as shown in British, Chinese and Thai children (reviewed and shown by Wang et al. 1998), whereas in others it takes up to 18-20 years to reach the lowest level of

activity as described in Finns (Sahi et al. 1983). It is not yet clear what the reason for this variation is, but it may either be due to environmental factors or other genetic influences or both.

1.2.2 The geographic distribution of the lactase persistence trait

From a European point of view, lactase non-persistence (LNP) was thought to be rare and an abnormality until the early 1960s when cases of adult hypolactasia were reported as a normal condition (Auricchio et al. 1963; Bayless and Rosensweig 1966; Dahlqvist et al. 1963). With the collection of more data on worldwide lactase persistence frequencies the picture became clearer, that lactase non-persistence is actually the more common phenotype (Flatz 1987; Swallow and Hollox 2000), in 65% of the world population (Ingram et al. 2009a).

Figure 1.3 shows the worldwide distribution of lactase persistence frequencies, which are highest in Northwest Europe, namely Denmark, Ireland and the British Isles. Lactase persistence frequency decreases as one moves towards the East and South from Northern Europe. A similar gradient has been observed in India with more people being lactase persistent in the North (Swallow 2003).

In Africa and the Middle East the picture is different, with a variable lactase persistence distribution. Some African and Arabian nomadic tribes that are milk dependent have higher lactase persistence frequencies than their neighbouring non-pastoralists (reviewed in Ingram and Swallow 2009a; Itan et al. 2010) such as the Beni Amer (pastoralists) and Donglawi (non-pastoralists) of Sudan with lactase persistence frequencies of 64% to 20% respectively (Bayoumi et al. 1982; Bayoumi et al. 1981; Holden and Mace 1997).

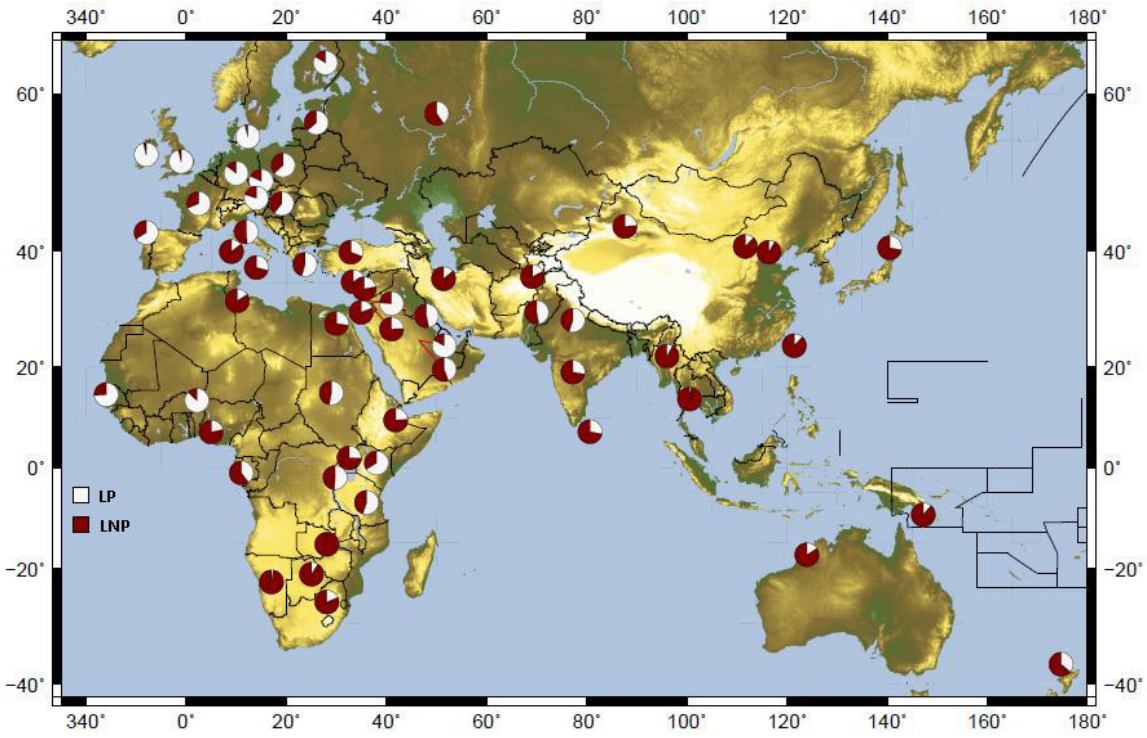


Figure 1.3: Geographic distribution of lactase persistence frequencies. White indicates the proportion of LP individuals in a population, red LNP individuals respectively. The overall frequency for each country is mostly comprised of different ethnic groups with different LP frequencies. Note that large differences in LP frequency are sometimes seen for example between Bedouin and non-Bedouin groups of Saudi Arabia and Jordan and North Indian and Indian populations. Data are taken from Itan et al. (2010).

1.2.3 Clinical definitions of lactase deficiencies

Lactase non-persistent individuals have a high probability of suffering from symptoms of lactose intolerance after consumption of a certain amount of lactose or milk, because the low level of intestinal lactase activity may not be sufficient to hydrolyse the quantity consumed. Symptoms can then occur when the undigested lactose passes into the small intestine, where it builds up an osmotic gradient leading to an inflow of water, causing diarrhoea. In the large intestine, colonic bacteria can ferment the lactose and this produces gases and fatty acids which can cause symptoms, mainly such as discomfort, flatulence and even cramps (reviewed in Hammer and Hammer 2012; Hammer et al. 1996; Hollox and Swallow 2002).

Lactose intolerance symptoms vary individually but if present they are usually recognisable within a few hours of the lactose consumption. Even quite large amounts of lactose can be in the tolerable range for some non-persistent people, as described below, while others claim to be able to take little or none. The reason for these inter-individual

differences in lactose tolerance are not completely clear but are likely be due in part to differences in gut flora composition or a different gastric emptying rate (Hertzler and Savaiano 1996; Lee and Krasinski 1998; Lomer et al. 2008; Szilagyi et al. 2004).

The consumption of fermented milk products like cheese and yoghurt can be a way of exploiting the nutrients in milk products and avoiding intolerance symptoms, as the lactose content of milk products decreases with natural fermentation (Food Standards Agency 2002; Swallow 2003). It should however be noted that some soft cheeses and commercial yoghurts contain substantial amounts of lactose. Interestingly, clinical trials have also shown that psychosomatic effects can influence the perception of symptoms (Briet et al. 1997; Peuhkuri et al. 2000).

The normal genetically determined trait of reducing the levels of lactase in adulthood (primary adult hypolactasia) is different from secondary loss of lactase due to other causes (secondary hypolactasia). Enzyme levels can be reduced when the intestinal epithelium is damaged, for example by inflammatory or autoimmune diseases like enteritis or coeliac disease. Lactase levels usually go back to normal after treatment of the primary disease and restoration of the intestinal epithelium, but this can also take longer if the patient is suffering from malnutrition, as shown for kwashiorkor ,and may in some of those cases even become permanent (reviewed in Dahlqvist and Lindquist 1971; Villako and Maaroos 1994).
Another very rare condition called 'congenital lactase deficiency' (CLD, OMIM #223000) describes the absence of lactase from birth. This autosomal recessive disorder can be fatal for a new born child if the condition is not identified straight after birth and is attributable to mutations in the coding sequence of the lactase gene (*LCT*, see also 1.2.5) (reviewed in Jarvela et al. 2009; Kuokkanen et al. 2006).

It has been suggested that lactase persistence status plays a part in a number of diseases of complex aetiology. The influence of lactase non-persistence on osteoporosis has been discussed, since an avoidance of milk products as a rich source of easy absorbable calcium, without the supplementation of calcium from other dietary sources, may contribute to a loss of bone mass and therefore an increased risk for osteoporosis (Lomer et al. 2008; Mattar et al. 2012). Several studies also have tried to relate milk consumption and lactase persistence/non-persistence to the risk of developing cancer (Larsson et al. 2006; Meloni et al. 1999; Shrier et al. 2008), metabolic syndromes (Almon et al. 2010; Corella et al. 2011; Enattah et al. 2004; Meloni et al. 2001) or gastrointestinal syndromes (Hollox and

Swallow 2002; Lomer et al. 2008; Villako and Maaroos 1994). However, a clear effect of lactase persistence/non-persistence could mostly not be confirmed and confounding effects such as mixed ancestry of the individuals tested are likely to have caused false positive associations.

1.2.4 Diagnosis of lactase persistence and non-persistence

Because lactase is restricted to the intestine, the only way of determining lactase persistence status directly is to take an intestinal biopsy to measure lactase enzyme activity. Ideally the measurement should be made in relation to another brush-border disaccharide such as sucrase or maltase, which together with histology will allow exclusion of secondary hypolactasia (reviewed in Swallow and Hollox 2000). The mRNA levels of the human proximal jejunum were shown to correlate well with the lactase persistence/non-persistence phenotype (Escher et al. 1992; Fajardo et al. 1994; Lloyd et al. 1992).

Population studies however require less invasive methods. Lactose tolerance tests are used to infer the levels of lactase indirectly as differences in lactose digestion, and volunteers can be categorised as digesters or non-digesters. Tests are carried out after an overnight fast and usually a lactose dose of 50 g is given to the volunteer. The physiological outcome is measured as difference between levels of blood glucose, urine galactose or breath hydrogen before and after the 'lactose drink' (reviewed in Ingram and Swallow 2009b). As glucose and galactose are cleavage products of lactose, an increase in blood glucose or urine galactose indicates a lactase persistent person.
All methods of testing are affected by a certain error rate but the most accurate test (Mulcare et al. 2004) is the breath hydrogen (H_2) test, which is-now most commonly used for large scale tolerance testing. Hydrogen is one of various gases produced by the colonic bacteria when they ferment lactose. Partly absorbed through the intestinal mucosa, transported through the bloodstream and exhaled by the lungs it can be measured in the breath. In lactose maldigesters an increase of hydrogen in the breath is observed, provided that appropriate hydrogen producing bacteria are present in the gut.

The term 'lactose intolerant' is often used both by the public and professionals for people with lactase non-persistence. This is misleading since lactase non-persistent people do not always show symptoms in their everyday life; note that a relatively high dose of lactose (equivalent to 1 litre of milk) is used for the tests to clearly distinguish digesters from non-

digesters (reviewed in Ingram et al. 2009a; Sahi 1994), but that even in these tests breath hydrogen production does not necessarily lead to significant symptoms.

1.2.5 Evidence for a genetic cause and *cis* regulation

A genetic cause for lactase persistence was first proposed by Bayless and Rosensweig (1966) who examined differences of lactose intolerance in black and white Americans. Several family studies followed (reviewed in Swallow 1993), in particular one, using lactose tolerance tests on a cohort of large Finnish families, showed the strongest evidence for an autosomal dominant inherited trait (Sahi 1974). Later twin studies confirmed this pattern of inheritance with the lactase persistence phenotypes of monozygotic twins in complete concordance and the phenotype distribution of dizygotic twins matching Hardy Weinberg expectations (Metneki et al. 1984).

A trimodal distribution of lactase activity, as expected to be seen phenotypically by the combination of two different alleles for hypolactasia and lactase persistence, was first suggested by Rosensweig et al. (1967). By further examining the enzyme activity ratios (lactase/sucrase or lactase/maltase) of biopsy material this 'gene-dosage effect' was confirmed, with the result of one group being interpreted as homozygous persistent (highest lactase activity), one homozygous non-persistent (low lactase activity), and one heterozygous (with intermediate activity) (Flatz 1984; Ho et al. 1982). This also suggested that only one copy of the lactase gene is fully expressed in heterozygotes suggesting a *cis*-acting mechanism (Flatz 1984; Ho et al. 1982) and indeed Wang et al. (1995) were able to verify this: The allelic variants of two known exonic SNPs within the lactase gene were used to reveal the chromosomal origin of the corresponding lactase mRNA transcripts. Lactase persistent individuals who were heterozygous for the SNP marker in most cases showed a difference in the amount of transcript from each of the chromosomes, which would not exist, had the influencing factor been *trans*-acting. The conclusion from this observation of monoallelic expression was that the causal element must indeed be *cis*-acting, being located within or adjacent to the gene.

The lactase coding gene itself, *LCT* was first characterised by the group of Mantei (1988) who investigated the cDNA sequences of the gene of humans and rabbits. They deduced a 5 domain structure of the enzyme and a primary translation product of 1927 or 1926 amino acids. They found 17 coding regions (exons) and further conclusions about the structure, evolution and processing of the gene were made, as described below. Boll et al.

(1991) further examined the fine structure of the gene and compared sequences of persistent and non-persistent individuals (see below).

LCT was mapped to chromosome 2 and more specifically to the region 2q21 (Harvey et al. 1993; Kruse et al. 1988) and additional information about the sequence of *LCT* was collected within the work of the International Human Genome Sequencing Consortium (Consortium 2004) and with subsequent annotations of chromosome 2 (for example Hillier et al. 2005) which describe the gene to have a length of about 49.3 kb and being transcribed from the reverse strand of the chromosome.

1.2.6 The lactase-phlorizin hydrolase enzyme (LPH)

Lactase, also called lactase-phlorizin hydrolase, is a β-galactosidase and shows two distinct enzymatic activities working at different pH and temperature optima (Arribas et al. 2000; Naim et al. 1987; Skovbjerg et al. 1981; Wacker et al. 1992; Zecca et al. 1998). The first, lactase (EC 3.2.1.108), hydrolyses lactose and other β-galactosides such as cellobiose, -triose and -tetrose and has very low activity against cellulose. The second activity, phlorizin hydrolase (EC 3.2.1.62), cleaves the glucose from the calchone, phlorizin and other β-glycosides with large hydrophobic alkyl chains such as aryl- or alkyl-β-glycosides (reviewed in Skovbjerg et al. 1982; Troelsen 2005). Interestingly, studies revealed that lactase activity can be inhibited by phlorizin whereas phlorizin hydrolysis is unaffected by lactose presence (Leese and Semenza 1973; Skovbjerg et al. 1981). Phlorizin hydrolase is active in all vertebrates, whereas lactase seems to be specific for mammals.

The glycosyl and lactosyl ceramides of the fat globules in milk were shown to be substrates for phlorizin hydrolase and also cleaves the flavonoids-(Day et al. 2000; Nemeth et al. 2003). Dietary sources for flavonoids are mainly apples, onions, tea and red wine (Ehrenkranz et al. 2005).
In the human small intestine lactase is mainly expressed in the mid- to lower jejunum (Newcomer and McGill 1966; Skovbjerg et al. 1981). Lactase activity is generally not present in the stomach or colon, however, some mRNA was detectable in the fetal human colon and in the rat colon during the early postnatal period (Freund et al. 1990; Wang et al. 1994).

During cellular development from stem cells, epithelial cells migrate from the crypts up the villi and differentiate (as shown in Figure 1.5). In the enterocytes of the crypt/villus junction lactase mRNA expression but no protein is detected, which changes as the cells

migrate up the villus, where the protein is also present (Freund et al. 1995). It was earlier shown that lactase protein expression is highest at the middle of the villus, and decreases again towards the villus apex (Skovbjerg et al. 1981).

A number of post-transcriptional processes were shown to be necessary to form the final transmembrane protein from the LPH mRNA transcript: In humans, the mRNA of 6279 bp (Ensembl database, build 37) encoded by *LCT* is translated into a precursor protein of 1927 amino acids. This pre-pro LPH consists of a putative signal peptide of 19 amino acids, a large pro region of 849 amino acids, and an element of 1061 amino acids forming the main (extracellular) part with both active sites, a hydrophobic membrane anchor domain and a short hydrophilic cytoplasmatic tail (Figure 1.4). The pre-pro LPH shows an internal 4 fold homology which is suggested to have evolved in two gene duplication events, each being homologous to a β-glycosidase unit. The first two homologous domains (I and II) are found in the pro-protein segment, the other two (domains III and IV) in the mature protein (Mantei et al. 1988).

Several post-translational modifications take place during LPH maturation (see also Figure 1.4). First, the signal peptide necessary to translocate the protein through the membrane of the endoplasmatic reticulum (ER) is removed proteolytically. The resulting pro-LPH becomes N-glycosylated and forms homodimers (Grunberg and Sterchi 1995), an essential step for transport capacity and later enzyme activity (Naim and Naim 1996). The pro-region of the molecule seems to be crucial for folding and dimerisation of the protein and intracellular transport (Jacob et al. 2002; Naim et al. 1994; Panzer et al. 1998). The pro-region and domain III need to interact together for the correct folding of the 'lactase domain' IV, which in turn elevates the enzymatic activity of domain III (Behrendt et al. 2010). The pro-region is cleaved off in the Golgi apparatus after O-glycosylation has taken place (Naim and Lentze 1992; Naim et al. 1987). An extracellular cleavage of the mature transmembrane protein by luminal trypsin follows, to give rise to its final and active form (Jacob et al. 1996; Wuthrich et al. 1996).

The C-terminus of LPH is anchored to the surface of the enterocytes of the microvilli extending into the gut lumen. The two active sites are located in each of the domains of the mature 160 kDa enzyme, one for phlorizin hydrolase activity in domain III at position Glu1273 and the one for lactase activity at Glu1749 in domain IV (Arribas et al. 2000; Naim et al. 1987; Wacker et al. 1992; Zecca et al. 1998).

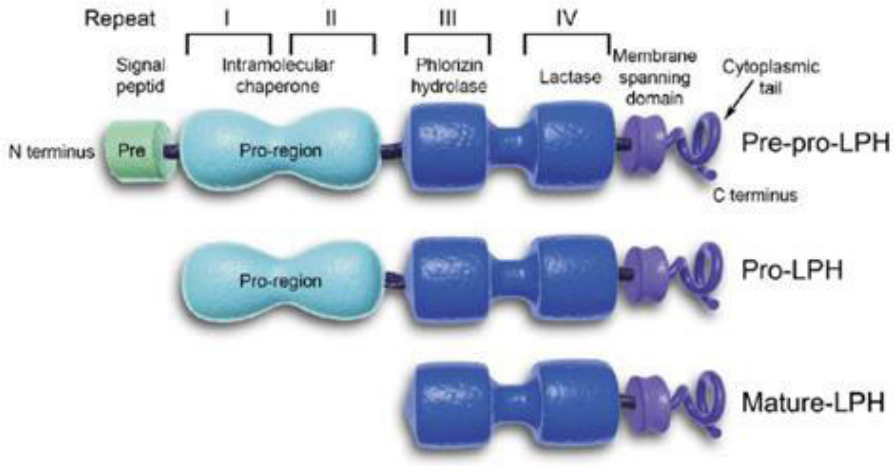


Figure 1.4: Structure of the LPH protein and its modification during the maturation process (Troelsen 2005). During post-translational processing the signal peptide and the pro-region, shown to have chaperone function, are cleaved off the 5 domain Pre-pro-LPH. The mature LPH protein consists of two domains, with phlorizin-hydrolase and lactase activities and is C-terminal anchored to the microvillus membrane.

1.2.7 Lactase expression during development - the ancestral state

Lactase down-regulation happens at a characteristic point of the mammalian development, depending on the species studied. The same seems to be true for the developmental rise of lactase activity. For example lactase mRNA expression is low in human foetuses but increases around the time of birth (Wang et al. 1998). Rodents, on the contrary, reach the maximum level of lactase activity around 3 days postnatally as their intestine is not completely matured at birth (Klein 1989).

During the post-weaning decline of lactase, a time when mammals switch their diet from milk as main food source to plants and meat, other enzymes necessary for plant digestion such as sucrase-isomaltase and maltase, are up-regulated in rodents, whereas these enzymes are already present in humans before birth (reviewed in Heitlinger et al. 1991; Troelsen 2005). This shows that although there are similarities in the developmental regulation of lactase expression in humans and other mammals, some patterns are different and the reasons for that are not fully understood.

1.2.8 Transcriptional regulation of the lactase promoter

Differences in a *cis*-acting element were suggested to be responsible for the inter-individual differences in lactase expression (as explained in section 1.2.5). Functional studies were performed to identify critical regulatory elements, first looking at the immediate promoter region of lactase. This region within the first 150 bp of the *LCT*

transcription initiation site was shown to be highly conserved between rat, mouse, pig and human, stressing its importance. Besides the TATA box, several *cis* elements in the form of binding sites for different transcription factors were identified in this region. HNF-1α, Cdx-2 and GATA factors were shown to influence the activity of the *LCT* promoter of all these species in Caco-2 cells (reviewed in Troelsen 2005). Caco-2 is a colonic carcinoma cell line and the only human cell lines known both to spontaneously differentiate in an enterocyte like manner (Pinto et al. 1983a) and to express lactase and other small intestinal proteins as normally expressed in the small intestine when grown to confluence (Chantret et al. 1988; Hauri et al. 1985).

Although this region of the promoter seems to be important, it only directs transcription at a relatively low level in Caco-2 cells and a region further upstream was suggested to be necessary for full lactase expression (Troelsen et al. 1992). *In vivo* studies in transgenic mice showed that promoter constructs containing 1 kb of the pig or 2 kb of the rat of *LCT* promoter were sufficient for full intestinal-specific expression and the regulation of the post-weaning decline (Krasinski et al. 1997; Lee et al. 2002; Troelsen et al. 1994). The 1 kb upstream part of the promoter is similar between the mammals studied, with the exception of a disruption by two tail to tail Alu elements in humans and other primate species (Hollox et al. 1999).

1.2.9 Transcription factors involved in *LCT* promoter activity

Transcription factors shown to be involved in lactase expression are part of the network of signalling pathways and transcription factors that play an important role in the development of the intestinal epithelium, as shown in Figure 1.5. The factors Cdx-2, HNF-1α and GATA-4 were shown to synergistically activate the lactase promoter (Mitchelmore et al. 2000; van Wering et al. 2004) as seen for sucrase-isomaltase (Boudreau et al. 2002).

A very important factor for the proliferation, differentiation and maintenance of the intestinal epithelium is Cdx-2 (Caudal-type homeobox protein 2). This transcription factor is exclusive to the postnatal intestine, mainly found in the distal small intestine and proximal colon (reviewed in Olsen et al. 2012; Silberg et al. 2000). It is involved in the transcription of many genes but the only intestinal specific factor involved in lactase expression (Troelsen 2005). As already mentioned, it binds to the immediate promoter of the human *LCT* and has been shown to up-regulate its activity when over-expressed in Caco-2 cells (Krasinski et al. 1997; Troelsen et al. 1997).

HNF-1α and HNF-1β (hepatocyte nuclear factor homologues 1α and 1β) were also shown to be important for lactase promoter activity of which HNF-1α seems to be more activating (10 fold higher than HNF-1β) and even more in combination with the Oct-1 transcription factor (Lewinsky et al. 2005). Both transcription factors are mainly expressed in the liver and pancreas but in other organs as well. In the gastrointestinal epithelia they are involved in the expression of several genes and essential for cell growth and differentiation (Bosse et al. 2006b; Troelsen 2005). HNF-1α is expressed throughout the intestine (Olsen et al. 2012).

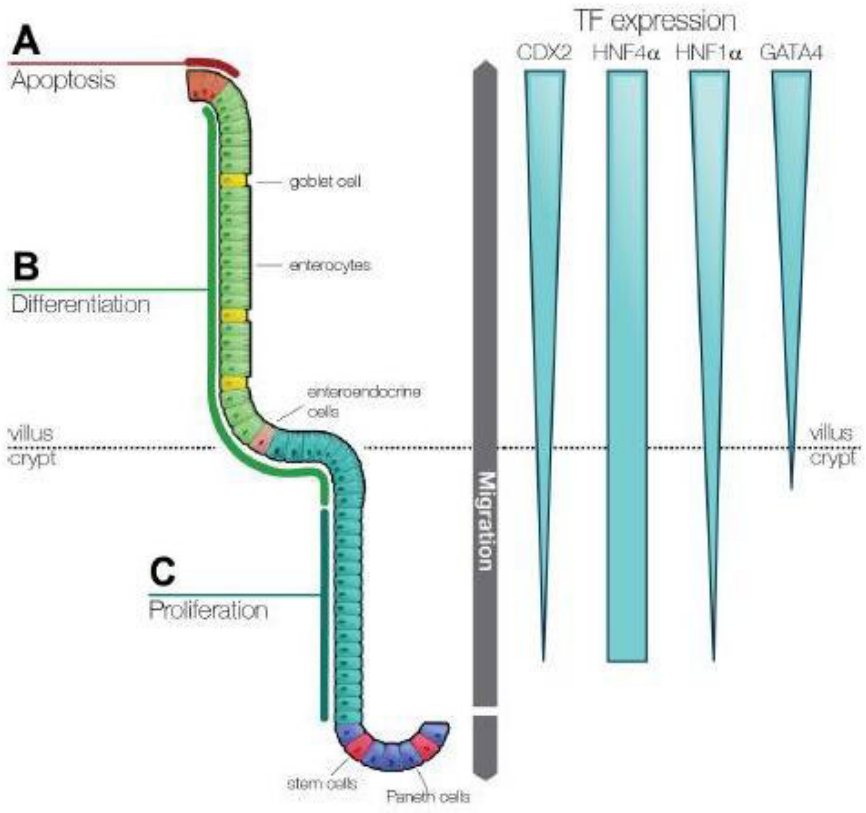


Figure 1.5: Schematic overview of the development of epithelial cells along the crypt/villus of the mammalian small intestine and important transcription factors involved (modified from Olsen et al. 2012). From stem cells, proliferating cells migrate either towards the base of the crypt or along the villus to the apoptotic zone (A). Cell proliferation (C) and differentiation (B) are mediated amongst other factors by differences in transcription factor (TF) expression: Cdx-2 and HNF-1α are highest in the villus but already present in the crypt, GATA-4 is detectable only above the crypt and is highest at the tip of the villus. HNF-4α expression is found equally along the crypt-villus axis.

Also GATA zinc finger family transcription factors are active in the intestine as well as in various other organs where they regulate gene expression, cell proliferation and differentiation. GATA-4, GATA-5 and GATA-6 factors have been shown to bind to the *LCT* immediate promoter (Fang et al. 2001; Fitzgerald et al. 1998) with similar affinity

(Krasinski et al. 2001). In contrast to GATA-6, highly expressed proliferating crypts of the villus, GATA-4 and -5 are mostly active in the differentiated enterocytes. Both were shown to interact with HNF-1α to synergistically activate the *LCT* promoter but GATA-4 can also act independently (van Wering et al. 2004; van Wering et al. 2002). GATA-4 was suggested to be the main factor involved in regulating lactase activity (van Wering et al. 2002) and its influence is necessary for *LCT* expression *in vivo* (Bosse et al. 2006a). Its expression decreases from the proximal part to the distal part of the mouse small intestine, which highly but not completely correlates with the regional expression of the lactase activity which is lower in the proximal small intestine (van Wering et al. 2004).

1.2.10 Lactase persistence caused by a different enzyme?

Much research has been conducted in order to explain the full mechanism underlying the differences between the enzyme activity of lactose digesters and non-digesters. The first suggestion, that the lactase enzyme produced in adults might differ from the one in infants was not supported experimentally (Lebenthal et al. 1974; Potter et al. 1985; Skovbjerg et al. 1978). Also the structure of the precursor and mature protein was indistinguishable for both persistent and non-persistent individuals (Sterchi et al. 1990). The groups of Sterchi (1990) and Witte (1990) both found that a decreased lactase biosynthesis accounts for low intestinal enzyme levels leading to non-persistence in humans, a similar process as previously been shown for the developmental loss of lactase in rats (Jonas et al. 1985). However, it was not clear whether the variation in lactase expression was controlled by transcription or translation until finally, by comparing mRNA levels with the levels of enzyme activity, it was interpreted that it is regulated at the transcriptional stage (Escher et al. 1992; Fajardo et al. 1994; Wang et al. 1994), although this did not exclude that other mechanisms, for example alterations in mRNA stability, may influence the final enzyme level.

1.2.11 The search for lactase persistence causal variation in and around the *LCT* gene

Research focussed on regions in and around the *LCT* gene to search for causal elements responsible for lactase persistence. However, cDNA sequencing of the *LCT* coding regions (including flanking intron-exon boundaries) and the *LCT* promoter, up to 1 kb upstream of the gene, did not show 100% association of any SNPs detected with lactase persistence or non-persistence (Boll et al. 1991; Lloyd et al. 1992) nor was there any evidence for alternative splicing (Boll et al. 1991). Nevertheless several polymorphisms were found and although not completely associated they were used as genetic markers in population

studies (Boll et al. 1991; Harvey et al. 1995a; Hollox et al. 1999). 11 of these variants in linkage disequilibrium across the gene were used to define the *LCT* 'core haplotypes' spanning 60 kb. Only four haplotypes A, B, C and U were common worldwide but a greater diversity was found in African populations (Hollox et al. 2001). The A haplotype showed an association with lactase persistence and is very frequent in Northern Europeans (Harvey et al. 1998; Hollox et al. 2001). The search for a causal allele within and outside the 60 kb haplotype region continued as the A haplotype was also shown to be present in non-persistent individuals (Harvey et al. 1998; Wang et al. 1998).

Sequencing of more distant regions around the *LCT* gene identified a SNP about 14 kb (-13910 bp) upstream of the transcription initiation site of *LCT*, that was completely associated with lactase persistence in Finns (Enattah et al. 2002). It is also highly associated in other Northern Europeans, occurring on a very extended A haplotype background of about 1 Mb (Bersaglieri et al. 2004; Poulter et al. 2003). This polymorphism (-13910 C>T) is located in intron 13 of the adjacent gene *MCM6* (minichromosome maintenance complex component 6, Figure 1.6), a gene with an important role in cell cycle regulation. In contrast to *LCT*, *MCM6* revealed no inter-individual differences in expression or tissue specificity (Harvey et al. 1996). Another polymorphism, -22018 G>A, was also associated with lactase persistence in the same study of Enattah and colleagues (2002). However this *-22018*A* always occurred with *-13910*T* but not vice versa and was also found in non-persistent individuals (Enattah 2002, Poulter 2003).

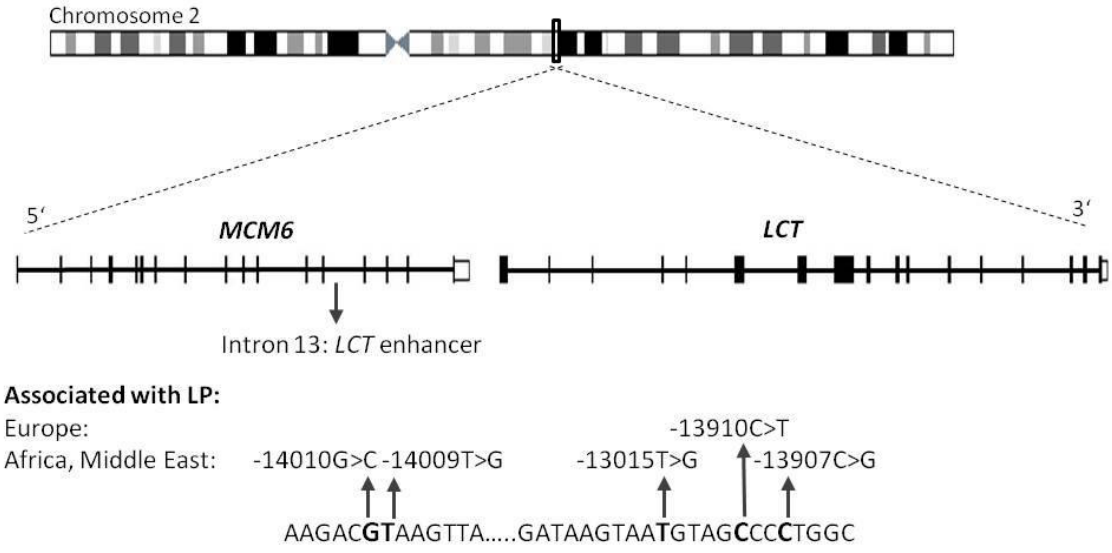


Figure 1.6: Location of the LP associated alleles in the '*LCT* enhancer' region in intron 13 of *MCM6*, upstream of *LCT*. Note that the genes are transcribed in - direction of the chromosome.

The *-13910*T* allele was seen as the likely cause (confirmed functionally later as described below) for lactase persistence in Europe and Asia but it did not explain the high lactase persistence frequencies of some sub-Saharan African populations or that of the Bedouins of the Arabian Peninsula where it is rare (Ingram et al. 2007; Mulcare et al. 2004).
In these geographic regions other genetic variants have been detected, located very near to -13910 C>T in the same intron of *MCM6,* within a segment which has been shown to influence enhancer function (see section 1.2.12). Three of them, *-13915*G*, *-14010*C and-13907*G,* showed clear association with lactase persistence (Enattah et al. 2008; Imtiaz et al. 2007; Ingram et al. 2007; Ingram et al. 2009b; Tishkoff et al. 2007), while a further allele, *-14009*G,* showed suggestive evidence of association (Ingram et al. 2009b).

Other rarer alleles in this region may or may not cause lactase persistence but even if all are taken into account some lactase persistent individuals carry no enhancer allele. Itan et al. (2010) tried to address this question at a worldwide level and compared phenotypic data with predicted phenotype frequencies from all four confirmed lactase persistence associated alleles. In some parts of the world a lack of correlation between the predicted and actual lactase persistence values was apparent (Figure 1.7). This was especially the case in parts of Africa, East and South Europe and Asia where genotype data were unable to explain lactase persistence frequencies. As most of the values were interpolated and sometimes only data for -13910 C>T were taken into account, this could simply be caused by sampling problems. Nevertheless, this pattern is striking in some regions and might suggest geographic regions under the influence of a novel causal allele resulting in lactase persistence occurring at high frequency. For example in the Wolof of Senegal, where 51% lactase persistence was reported, none of the four enhancer variants had been found so far (Ingram et al. 2009b; Mulcare et al. 2004).

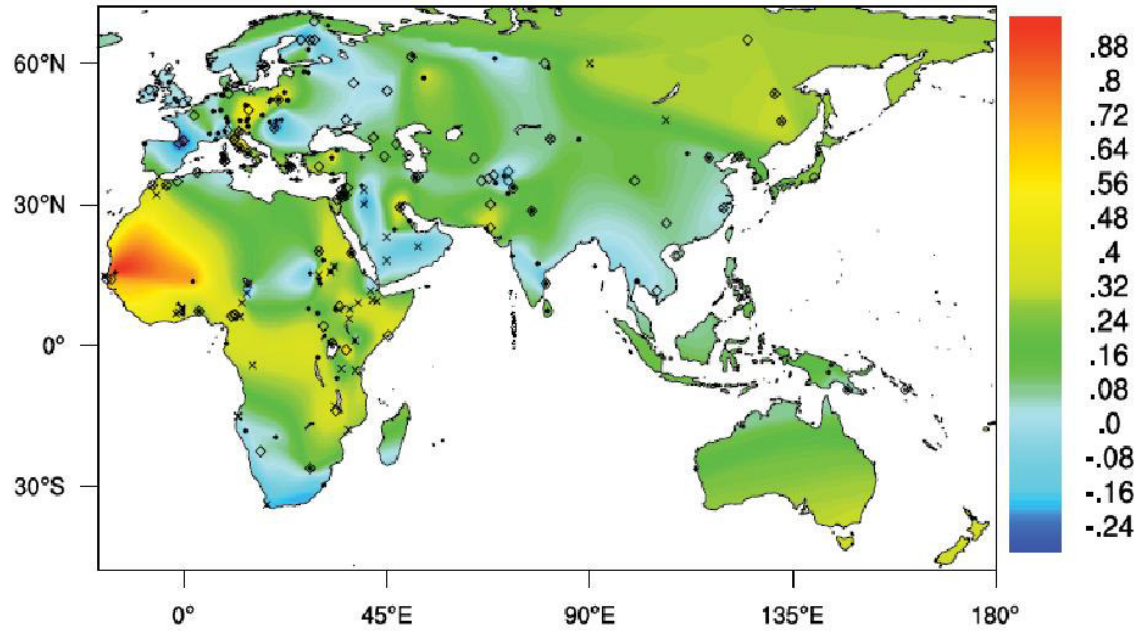


Figure 1.7: Lactase persistence genotype-phenotype correlation (taken from Itan et al. 2010). The map shows the quantitative differences between predicted LP phenotype frequencies from genotype data and observed phenotype data (both interpolated). Allele frequency data of alleles *-13910*T, -13915*G, -14010*C and -13907*G* were taken into account to predict the LP phenotype frequencies. For geographic regions where no phenotype or genotype data were available, values were interpolated and finally predicted LP frequency values subtracted from observed values. Colour key on the right: + values represent the genotype under-predicting phenotype at a certain location and - values over-predicting it, respectively. Data collection points: ◇*-13910*T* only, x all 4 alleles, ● LP phenotype.

1.2.12 Lactase persistence associated enhancer variants increase *LCT* expression

With the observation that *-13910*T* was completely associated with lactase persistence in a Finnish population (Enattah et al. 2002) a candidate functional allele had been found. And indeed it was subsequently shown in transfection experiments that the region around -13910 C>T has a transcriptional enhancer effect on the human and rat promoter of the *LCT* gene (Olds and Sibley 2003; Troelsen et al. 2003). The 450 bp human *LCT* enhancer was able to stimulate transcription about 4 to 7 fold in undifferentiated cells (2 days after transfection) and up to 50 fold in differentiated cells (9 days after transfection) compared to the promoter alone. The effect of *-13910*T* was much stronger than of *-13910*C* with a 3 to 6 times higher activity in differentiated Caco-2 cells (Troelsen et al. 2003).

One particular transcription factor, Oct-1, binds directly to the site at -13910 but much stronger to the **T* than the **C* variant (Lewinsky et al. 2005). The function of this allele has been supported recently *in vivo* (Fang et al. 2012).The post-weaning decline of transgene

luciferase expression in mice, shown previously to be mediated by a 2 kb rat lactase promoter (Lee et al. 2002), could be prevented in mice carrying *-13910*T* but not in those carrying the ancestral variant *-13910*C.*

Subsequently other enhancer variants (section 1.2.11) have been tested for transcription factor binding: *-13907*G* and *-13915*G* showed binding to Oct-1, *-13907*G* with a similar affinity as the*-13910*T* variant (Enattah et al. 2008) but *-13907*C* much less (Ingram et al. 2007). Each of these variants however increases the activity of the enhancer (Olds et al. 2011; Tishkoff et al. 2007) as does *-14010*C*, as will be described in more detail in chapter 4.

In addition to Oct-1, further binding sites for transcription factors, namely Cdx-2, GATA-6, HNF-4α and Fox have been identified within the 450 bp 'enhancer region' and mutation analyses revealed that all of them, except Cdx-2, are required for full enhancer activity. Further transfection experiments using co-transfection of expression plasmids for the different transcription factors themselves suggest that the interaction of Oct-1 with HNF-1α, shown to bind to the *LCT* promoter, enables a strong enhancement of reporter gene expression and mediates the further increasing effect of the enhancer carrying *-13910*T* compared to the ancestral variant (Lewinsky et al. 2005).

1.2.13 The cultural practice of milking and selection hypotheses for lactase persistence

It has been observed that milk drinking behaviour is correlated with high frequency of lactase persistence and it was suggested that this is due to strong positive selection (Aoki 1986; Holden and Mace 1997; McCracken 1971b; Simoons 1970a; Simoons 1978).The selective forces that drove the increase of lactase persistence in populations with dairying practices must have acted since the beginning of animal domestication about 5000-10000 years ago.

Several theories try to explain this process. Robert McCracken using a large literature review was able to classify three groups of human cultures according to their milk product dependence (McCracken 1971a; McCracken 1971b). The first has never used dairy animals and therefore did not include lactose in their diet after weaning (e.g. Aborigines, New Guinea natives, Eskimos and the other American Indians). The second group are pastoralists who keep animals but have never used much milk or milk products as part of their lifestyle (e.g. Chinese, Thai and Philippines). Thirdly there are populations who introduced the practise of milk drinking a long time ago and therefore have a large amount

of lactose in their diet (most Europeans, many African and Middle Eastern populations and e.g. North Indians).

The close correlation of the distribution of lactase persistence with the extent the populations investigated use milk as part of their culture lead to the 'culture historical hypothesis' independently proposed by Simoons and McCracken (McCracken 1971a; Simoons 1970a). It suggests that the genetically determined trait coevolved with the practise of dairying, since with the cultural adaption to animal breeding and milk consumption the dependence of adults to milk increased. The selective pressure would act in populations consuming lactose-rich forms of milk with otherwise marginal food availability and would have favoured those who were lactase persistent, as they benefit from the nutritional value of milk, whereas fresh milk consumption may have had adverse consequences for non-persistent people. A high selective pressure of this kind would not be present in populations that process milk to sour milk and cheeses which would also allow non-persistent people to benefit from the nutrients of these products. This might explain the presence of populations with a history of dairying but low lactase persistence frequencies, as seen for example in Southern Europe (Simoons 1978).

The 'reverse cause hypothesis' is based on the same observations but describes the opposite scenario (reviewed in Aoki 2001; first mentioned in McCracken 1971a). It suggests that dairying was only adopted by populations that were already largely lactase persistent, caused by another, unrelated process. This would mean that milk drinking would not necessarily have had any selective advantage. Nei and Saitou (1986) argued that lactase persistence is so recent that selection could not have been effective enough to cause these frequencies. Drift would have been the main factor, followed by later selection (reviewed in Ingram and Swallow 2009a).

The especially high rates of the lactase persistence phenotype in Northern Europe led other authors to suggest the 'calcium assimilation hypothesis' (Flatz and Rotthauwe 1973). This focuses on milk calcium as the nutritional benefit and also the improvement of calcium absorption by lactose. Calcium is an essential bone mineral and its absorption in the gut is facilitated by vitamin D_3 (cholecalciferol). Populations in northern latitudes with lower sunlight, especially during winter, have an increased risk of developing rickets and osteomalacia due to the lack of photochemically produced cholecalciferol in their skin. In contrast to hunter-gatherers, who could balance this with a vitamin D rich diet containing marine food, early agriculturalists in Northern Europe mainly relying on cereals might

have had a problem in reaching the necessary intake of this vitamin (Cordain et al. 2012; Richards et al. 2003).
Flatz and Rotthauwe (1973) argue that drinking milk could have been an advantage for lactase persistent individuals in the Neolithic to prevent these diseases. Milk is not only high in calcium it also contains small amounts of vitamin D, and milk proteins and it has been reported that lactose additionally promote vitamin D absorption (reviewed in Ingram and Swallow 2009a).
The lack of vitamin D would surely not have been a problem for African nomads or populations from the Middle East. According to the 'arid climate hypothesis' these groups could survive better in dry regions, with limited access to water and food, using fresh milk as relatively clean and uncontaminated fluid (Cook and al-Torki 1975). For lactase non-persistent people of course this could be extremely disadvantageous if they suffer from diarrhoea and dehydration (reviewed in Gerbault et al. 2011).

Holden and Mace (1997) statistically tested the culture historical, calcium absorption and arid climate hypotheses adjusting for genetic distances between populations. They found that pastoralism alone could explain most the lactase persistence distribution, while solar radiation or aridity could not. They suggested from their maximum likelihood approach that pastoralism must have been adopted before lactase persistence became frequent. This is also supported by palaeogenetic and archaeological data as mentioned in section 1.1. Ancient DNA samples from Meso- and Neolithic periods of Europe show the absence or very low frequency of *-13910*T* implying that it was very rare or absent in early Neolithic (Burger et al. 2007; Lacan et al. 2011a; Lacan et al. 2011b; Malmstrom et al. 2010). Pottery found in Western Turkey from as early as about 8500 BP contained preserved milk residues, which would also support the idea that milk was used before lactase persistence reached high enough frequencies (Evershed et al. 2008). However, the method used cannot distinguish residues from fresh or fermented milk.

There are however some groups of people who have relatively low lactase persistence frequencies like the Dinka and Nuer from Sudan and are traditional camel or cattle herders and drink fresh milk like the Somali in Ethiopia (Ingram 2008). Many of the non-persistent individuals drink more than 500ml of milk per day. It was proposed that an adaption of the bacterial gut flora within these people may allow them to drink large quantities of milk (Ingram et al. 2009a).

The potential role of holding cows as a 'status symbol' should be taken into account too when considering the relationship between pastoralism and lactase persistence. In many hierarchical cultures keeping livestock is restricted to a 'social elite' who were the ones who had access to milk. Reproductive behaviour is also influenced by social status in such cultures (Gillis 2003; Ihara 2011; Leonardi et al. 2012; Simoons 1970a). The effect of natural selection might therefore even been greater, favouring those who were lactase persistent in higher classes who in any case have greater fecundity (Heyer et al. 2005).

An alternative selection hypothesis was proposed that is linked to the negative effect of milk consumption, considering malaria as selective force for lactase non-persistence, and suggesting that lactase persistence was the ancestral state (Anderson and Vullo 1994; Meloni et al. 1996). Anderson and colleagues argue that without milk riboflavin a mild deficiency could develop in the red blood cells and as *Plasmodium falciparum* parasites are sensitive to that it would lead to milder malaria.
On the other hand an opposing hypothesis has been proposed, that selection for lactase persistence was driven by malaria. A milk diet would be protective against malaria in causing a deficiency of p-aminobenzoic acid, a pre-form of folate, and therefore impairing the folate metabolism of *Plasmodia* (reviewed in Cordain et al. 2012). However, other studies could not find evidence for any association with, the prevalence of lactase non-persistence and malaria incidence (Auricchio 1998; Meloni et al. 1998).

As will be seen below all studies using molecular data and results of modelling imply an extremely high positive selection coefficient. However, there is still an ongoing debate about the nature of these strong selective forces, which are probably different in Northern Europe from those in other parts of the world. Undoubtedly, milk is an edible source of protein and fat (also in its fermented forms) and the practice of dairying assures a constant food supply, especially in adverse situations such as crop failure. This may be particularly important for children after weaning. A selective pressure, even if normally weak could have become much stronger under extreme circumstances like famine and drought (Gerbault et al. 2011).

1.2.14 Genetic signatures for lactase persistence under selection

In summary lactase persistence was suggested to have reached its high frequency in some populations due to positive selection acting by providing an advantage to cultures who adapted pastoralism as early as the Neolithic (Aoki 1986; Holden and Mace 1997; McCracken 1971a; Simoons 1970a; Simoons 1978).

At a molecular level, evidence for selection can be detected, for example by looking at the haplotype background of trait associated alleles (see also section 1.3.2). As already mentioned, the A haplotype is the most common in Northern Europe and shows an association with lactase persistence (Harvey et al. 1998). It was subsequently discovered that the lactase persistence causal polymorphism *-13910*T* occurs on the background of this A haplotype which extends up to 1 Mb. As this long haplotype is present at high frequency, it was thought likely to be a sign of a very recent selective process (Poulter et al. 2003).

Bersaglieri et al. (2004) further examined this signal of selection more formally, by looking at population differences of allele frequencies of 101 SNPs, spanning a region of about 3.2 Mb, and the patterns of LD as assessed by extended haplotype homozygosity. They detected a regional difference of allele frequencies, with SNPs in vicinity of *LCT* (over a 1.5 Mb spanning region) showing high inter population differences, and an unusually frequent and long haplotype of the *-13910*T* carrying chromosomes. They concluded that only strong positive selection could have led to this pattern and calculated a timeframe for it to have operated between 5000-10000 years and explained their observation with the model of allelic 'hitch-hiking' proposed by Braverman et al. (1995). Alleles surrounding the causal variant 'hitch-hike' with it to high frequency, creating a region of high LD (see also section 1.3 below). For *-13910*T* this mutation event must have been so recent that time was too short to decay the allelic associations in this region by recombination (reviewed in Sabeti et al. 2006).
Studies of microsatellite polymorphisms show similar results: a reduced diversity on the *-13910*T* carrying chromosomes (Coelho et al. 2005; Ingram et al. 2009a).

Estimations of the date when the *-13910*T* dispersion process started range from 7450-12300 years BP (Coelho et al. 2005) and 7475-10250 years BP (Mulcare 2006), both calculated by using closely linked microsatellite variation, to 2188-20650 years BP (Bersaglieri et al. 2004), estimated using the extended haplotype homozygosity (EHH) approach (see also section 1.3.2.2). The African allele *-14010*C* was reported to have occurred as an unusually extended haplotype, with the age of this variant estimated to be between 1200 and 23200 BP using EHH (Tishkoff et al. 2007). It has been observed that the variant *-13915*G* occurs mainly on a C haplotype background (Ingram et al. 2009b) and its age in a Saudi population was estimated using a coalescent approach to be about 4091 (± 2045) years (Enattah et al. 2008). No dating has been done so far for *-13907*G* but

it was shown to occur on an A haplotype background, as does-*13910*T* (Enattah et al. 2008; Ingram et al. 2009b). It is evident that lactase persistence evolved independently more than once with different causative alleles.

For alleles occurring at such high frequency in many populations, these age estimates are extremely young. With drift alone this would not have been possible, certainly not for -*13910*T* and-*14010*C*. However the combination of drift and very strong positive selection for lactase persistence could have left such a pattern and indeed the selection coefficients are estimated to be very high, between 0.8 and 19% for -*13910*T* (Bersaglieri et al. 2004; Gerbault et al. 2009; Itan et al. 2009) and between 1 and 15% for -*14010*C* (Tishkoff et al. 2007).

1.2.15 Simulation models for the spread of lactase persistence

From genetic data of modern populations alone it cannot be concluded where a variant originated or migrated from. This is especially the case because time and place a genetic variant arises does not necessarily correspond to the starting point of the selection process. To obtain further insight as to when lactase persistence first occurred and how it spread, computer simulations have been applied, since they provide a useful tool to test different scenarios. These models can take into account several factors, such as modern genetic, palaeogenetic, archaeological and geographic data to inform and to test statistical models simulating the migration.

As early as 1986, Aoki developed a model to investigate the coevolution of dairying and lactase persistence in Europe with the aim of explaining the incomplete correlation between the use of milk and the presence of lactase persistence seen today (Aoki 1986). He showed that this pattern could simply be a stochastic effect and is no evidence against the culture historical hypothesis. Although a quite simple model, for example not considering migration nor effective population sizes other than 100, he suggested that the selection coefficient must have been either constantly very high (> 5%) or the coevolution must have started much earlier than 6000 years ago.

Gerbault et al. (2009) aimed to test the calcium assimilation hypothesis and added a geographic component to their simulation models. Different selection pressures by latitude combined with two different diffusion models (Ammerman and Cavalli-Sforza 1984) were tested. The demic diffusion model assumes the spread of Neolithic over Europe and gene flow between the migrating farmers and hunter-gatherer populations. In

the cultural diffusion model, only the Neolithic cultural practice would have been exchanged between hunter-gatherers and their farming neighbours, without genetic admixture between them. Gerbault and colleagues (2009) simulated the selection process for a lactase persistence associated allele (present at 1% in LBK cultures) from the beginning of Neolithic (10000 years ago) and compared that with lactase persistence frequencies from European and Middle Eastern populations today. The results obtained would fit with the demic diffusion model and the calcium absorption hypothesis (Gerbault et al. 2009). They concluded that a positive selective process would indeed have been necessary to explain the high frequency of lactase persistence in Northern Europe. In Southern Europe on the other hand the effect of genetic drift alone could explain the lactase persistence frequencies observed today.

With a more complex model, Itan et al. (2009) simulated the spread of the lactase persistence *-13910*T* allele across Europe and the Middle East over the last 9000 years, allowing for gene flow between non-dairying and dairying farmers and hunter-gatherers. They concluded from their best fitting simulations that the *-13910*T* allele first underwent selection among dairying farmers between 6256-8683 years ago. It was also concluded that the coevolution process started between the central Balkans and Central Europe (Figure 1.8). Interestingly the combination of time and region would agree with the occurrence of Linearbandkeramik culture in Central Europe. In contrast to Gerbault et al. (2009) they suggest that a selective pressure due to latitude would not have been necessary to explain the high lactase persistence frequencies in Northern Europe, although they did not explicitly test for this.

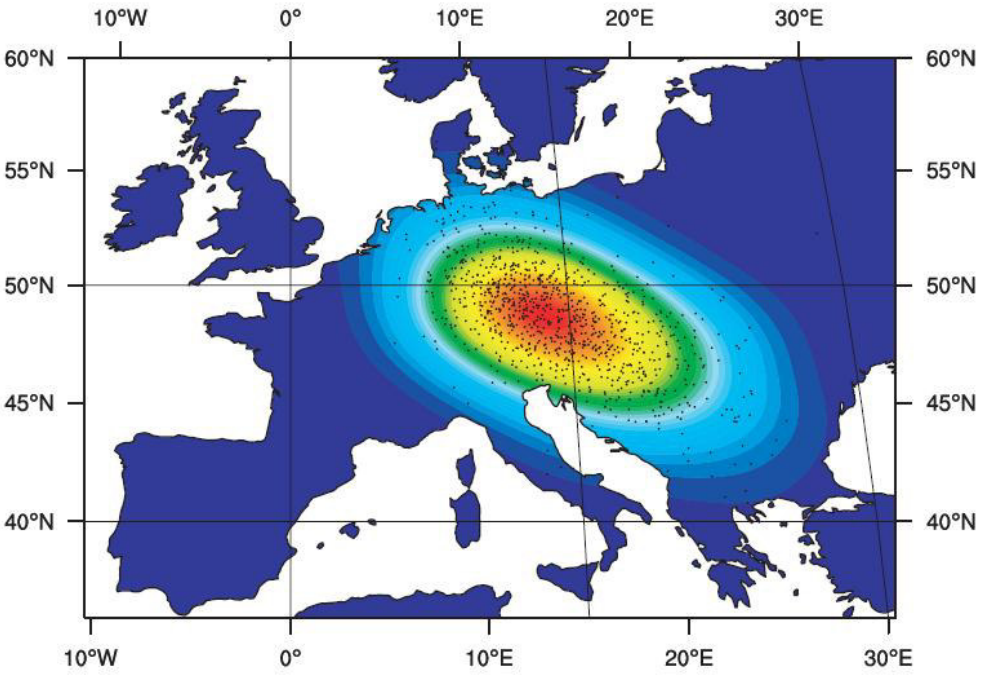


Figure 1.8: Simulated region of the origin for LP-dairying co-evolution. Best fit simulations (dots) are of highest density in the red coloured area between the Central Balkans and Central Europe (Itan et al. 2009).

1.3 Population genetics and methods to detect selection

1.3.1 Aspects of the population genetics theory

Population genetics is a research field that studies evolutionary processes leading to genetic diversity, by assessing inter and intra population differences of variation in allele and haplotype frequencies of populations. Population genetic models help to disentangle how different factors have acted over time and space to result in currently observed diversity patterns. Such factors are for example mutation, recombination, genetic drift, population subdivision, gene flow and selection. These issues are described in Jobling (2013) and presented in summary in the following sections.

New alleles are created by mutation and lead stepwise over time to other allelic forms (haplotypes). Mutation rates differ along the genome and depend on the type of mutation. Microsatellites for example have a higher mutation rate than SNPs. By knowing the mutation rate of a particular genomic region, the allele frequency can be calculated over generations with different models of DNA sequence evolution (see Jobling et al. 2013, chapter 5). Of course, only the heritable nucleotide changes in germ-cells are of evolutionary interest for these models.
In addition to mutation, meiotic recombination can generate variation and reduce the frequency of a haplotype. This process can be studied by looking at non-random correlations of alleles in populations, known as linkage disequilibrium (LD) using the *D* statistic. A new arising mutation is linked to all other alleles on that chromosome, being in complete LD with them (D_{max}) and forming one haplotype. Over time LD starts to decay as recombination breaks up the haplotype and creates new allelic combinations. One useful measure of LD is *D*' which is D/D_{max}. Closely linked loci on the same chromosome are less prone to recombination and selection can be detected by looking at their diversity (see below).

Whether an allele gets fixed or is lost within a population is mainly due to genetic drift. This source of variation is due to the stochastic process of sampling from one generation to the next. The extent of genetic drift relates to the size of the sampled population as described in the Wright-Fisher model (Wright 1931), if the population stays at a constant size, mates randomly and generations don't overlap. Of course, this is not a realistic model for human populations as generations do overlap, populations seldom stay at a constant in size and in large populations mating is not random. To take these influences into account and be able compare the degree of genetic drift between populations, the concept of

effective population size was developed. It specifies the size of a randomly mating population with the same degree of genetic drift as the actual population.
In most cases, new alleles will disappear from a population over time due to drift, when no selection or mutation is acting. The fixation of a neutral allele ~~only~~ depends on the population size. Considering the same mutation rate, an allele becomes fixed earlier in smaller than in larger populations and selection can accelerate this process. This means that even under neutral conditions past reductions in population size, as seen with population bottlenecks and founder effects, have a large influence on present day genetic variation.

As mentioned above, human mating may not be random because of mate choice not being independent from genetic relatedness, for example due to assortative mating (for example preference of phenotypic similar mates) or disassortative mating. Mating is also not random when a population is structured, which means it consists of partially isolated and smaller subpopulations. If isolated, members of each of these subpopulations are closer related than to individuals of all other subpopulations, due to partial genetic differentiation. If individuals migrate between these subpopulations they cause gene flow, which consequently reduces population differentiation. Population structure can be measured by the F_{ST} statistics, which shows the genetic distance between subpopulations (see also section 2.3.7).

The genetic variation of individuals defines their 'fitness', the ability to survive and reproduce in certain environments. Natural selection leads to variable reproduction of individuals and resulting in different genotypes in successive generations. Selection can operate on phenotypes at any stage of development: Survival to reproductive age (viability and mortality), success in mate attraction (sexual selection), the ability to fertilize (fertility and gamete selection), and to produce viable offspring (fecundity). These factors determine the relative fitness of an individual genotype/phenotype compared with others competing for the same resources, which can be measured as the selection coefficient (the proportion of fertile progeny relative to those without the genotype, or relative probability that a genotype will reproduce-i.e. positive selection w of 0.1 = 1-0.9, where progeny of 100% (1) for carriers and 90% (0.9) for non-carriers).

Selection can act in different ways. Mutations that increase the fitness of the carrier undergo positive selection and may rise in frequency in a population, whereas mutations that reduce fitness are subject to negative or purifying selection and are possibly

eliminated from the population. However, it is important to take the combination of mutations into account when considering their effect on the fitness of a genotype. Two alleles on a diploid locus for example can have a different impact on individual fitness. A novel allele may increase the fitness of a heterozygote compared to that of both homozygotes and possibly decrease their fitness, which is called heterozygote advantage. This results in balancing selection, which favours more than one allele, increasing diversity but preventing certain alleles from fixation. Another process creating balancing selection is frequency-dependent selection, where the frequency of a genotype is crucial for its fitness and leads to higher fitness at lower frequencies compared to other genotypes.

1.3.2 Statistical tests to detect selection

Of special interest for this thesis are methods that detect natural selection. Most of the statistical methods that aim to detect selection work in considering significant deviations of the observed genetic diversity to that expected under neutrality. The neutral theory of evolution, developed by Kimura (Kimura 1968; Kimura 1983), states that most variation within and between species is neutral, which means it does not affect the phenotype or fitness of individuals and is therefore not under selective pressure. Neutral mutations are fixed in a population randomly by genetic drift and most of them stay at low frequencies or are lost from a population.

A positive selective sweep reduces the variation within a population but does not lead to decreased variation between species. Divergence data build the background to identify positive selection between species and diversity data within species. Negative selection instead tends to reduce variability within and between species (Nielsen 2005). Population genetics distinguishes between hard selective sweeps, soft selective sweeps and polygenic adaptation (Jobling et al. 2013). A recent hard selective sweep (or classical selective sweep) leads to an increase in frequency of the haplotype carrying the advantageous mutation as shown in Figure 1.9. The following paragraphs review the methods that aim to detect this type of positive selection, based on their different approaches and Figure 1.9 summarises commonly used methods. Positive selection can also act over a longer time span and favour more than one variant (soft selective sweep) that may even lay within different genes (polygenic adaptation). These forms of selection are harder to detect.

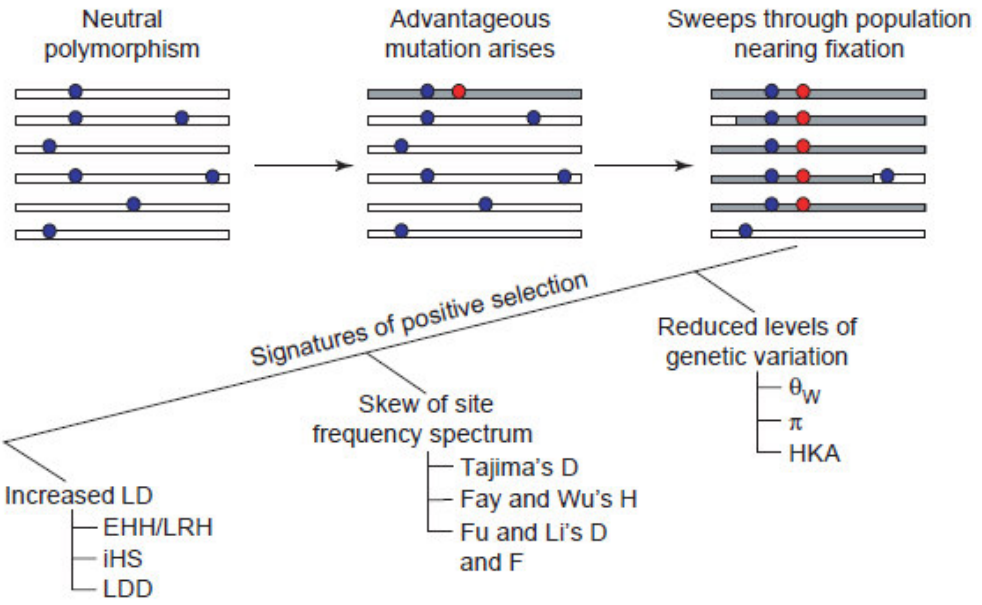


Figure 1.9: How to detect signatures of positive selection. The upper part of the figure shows how the pattern of neutral polymorphisms on different haplotypes changes when an advantageous allele (red) arises and increases in frequency and thereby drags along linked neutral polymorphisms. This process reduces variation levels surrounding the selected region, causes a skew in allele frequency distribution and increases linkage disequilibrium levels. Statistical methods commonly used to tests for these signatures are shown below, of which some are further described below. The figure was taken from Biswas and Akey (2006).

1.3.2.1 Proportion of function altering mutations

Variation in gene coding regions can be of two types: synonymous, which usually don't alter amino acids and therefore protein structure or their function and non-synonymous, which can have serious effects depending on which part of a protein they occur. By comparing the ratio of non-synonymous (d_N) to synonymous (d_S) mutations, selection can be detected. Under neutrality, a neutral non-synonymous change would increase in frequency in a population at the same rate as a synonymous change and d_N/d_S will equal one. A deleterious non-synonymous mutation will most probably be subject to purifying selection and the d_N/d_S ratio will be less than one whereas a positive selected non-synonymous change increases its frequency and d_N/d_S will be greater than one. The latter effect is expected to be stronger for species divergence data than within a species and can be tested using the McDonald-Kreitman test (McDonald and Kreitman 1991), which compares the d_N/d_S ratio at a specific locus within and between species. This test is useful in identifying selected function altering mutations in gene coding regions and it is robust against demographic factors (Nielsen 2005). However, it ignores selection of variation in gene regulatory elements in non-coding parts of the genome.

The application of a modified version of the Hudson-Kreitman-Aguade (HKA) test (Hudson et al. 1987), which compares the variation rate within and between species at multiple loci, and can be applied to non-coding sequences.

1.3.2.2 Haplotype based methods

A recent and strong positive sweep (hard sweep) can be detected when an allele under selection has increased in frequency and created a region of LD with neighbouring SNPs, which reduces genetic diversity of the neighbouring regions. Two test methods have been developed to detect this skewed allele frequency spectrums around a frequent mutation, the extended haplotype homozygosity test (EHH) (Sabeti et al. 2002) and the integrated haplotype score test (iHS) (Voight et al. 2006). EHH detects a combination of alleles that are in LD over a long region (core haplotype) and that are at high frequencies (i.e. common homozygosity). The method statistically evaluates the expansion of LD of this core haplotype. The extension of the core haplotype is compared to all other haplotypes of that region (relative EHH, rEHH) in comparing their EHH decay. The rEHH statistics also takes local recombination rates into account. A recently positive selected haplotype should show high EHH values at high frequencies, whereas a haplotype under less selective pressure or just reaching high frequencies due to genetic drift would have been more prone to recombination and mutation as this process takes longer, and would therefore show lower EHH values.

1.3.2.3 Allele frequency spectrum

Tests considering the site frequency spectrum or allele frequency spectrum take the allele frequencies at segregating sites into account. As described above, a positively selected allele tends to increase in frequency and causes a reduced frequency spectrum of other variants around it. This causes a skew in allele frequencies towards more alleles at low frequencies (Nielsen 2005). To detect this, Tajima's *D* (Tajima 1989) for example tests the excess of rare alleles compared to neutrality or the HKS test (Wright and Charlesworth 2004) the reduction in polymorphism levels. A further test takes the proportion of the variants at high frequencies into account that are as a result of this 'hitchhiking' process associated with each other. Fay and Wu's H (Fay and Wu 2000) test the excess of derived alleles at high frequencies.

1.3.2.4 Population differentiation

Population differentiation tests work under the assumption that selection operates differently in different populations, especially under different environmental conditions.

They can detect selection that occurred after subdivision of a population. Differences in allele frequencies might therefore be specific for a population and can be indicated by an increased level of genetic differentiation between populations, as measured for example by the F_{ST} statistic (Wright 1950), further described in chapter 2. However, inter-population differences can be highly influenced by demographic factors, such as population bottlenecks that reduce genetic diversity and F_{ST} tests should ideally be combined with other tests to get better evidence for selection.

1.3.3 Detection of different patterns of selection of *LCT*

Over the last years the availability of genome wide data and new or modified statistical approaches to scan for genomic signatures of adaptations in these datasets have identified a range of candidate genes or gene regions under positive selection. As described in above these methods mainly investigate variation patterns like allele frequency spectrums at certain regions or the reduction of haplotype diversity across populations. The *LCT* region mostly appears in these studies as strongly positive selected in European datasets (with chromosomes carrying *-13910*T*) as shown by long-range haplotype tests , haplotype diversity or population differentiation (Akey et al. 2002; Bersaglieri et al. 2004; Hofer et al. 2012; Sabeti et al. 2002; Sabeti et al. 2006; Voight et al. 2006). Extended haplotype homozygosity in combination with a high allele frequency was also reported for *-14010*C* in Kenyans, which supports the notion that a strong selective pressure has also acted on this variant (Tishkoff et al. 2007).

Recently, evidence for another type of selection acting on the *LCT* enhancer was detected in an Ethiopian dataset. Our group assessed the signatures of a soft selective sweep (Jones et al. 2013) which allows the accumulation of many rare advantageous alleles, causing the same phenotype (as described by Hermisson and Pennings 2005; Pennings and Hermisson 2006a; Pennings and Hermisson 2006b). This was inspired by the findings that several lactase persistence associated enhancer alleles were present in the Jaali and Somali groups previously investigated and that enhancer sequences of persistent people of these groups accumulated more derived alleles than those of non-persistent individuals (Ingram et al. 2009b). In taking neighbouring genetic regions of the *LCT* enhancer into account to control for demographic and mutational effects a comparison of haplotype and nucleotide diversity between lactose digesters and non-digesters revealed a clear pattern of reduction in variability only at the enhancer in the non-digesters (Jones et al. 2013).

1.3.4 Methods to study *LCT* regulation

The function of a gene is to a great extent defined by the protein it codes for. However, the regulation of gene expression 'fine-tune' the spatial and temporal expression of genes and can have a great impact on the phenotype in for example development processes of an organism and adaptation to environments. Inter-individual and inter-population differences in gene expression have been noticed (reviewed in Jones and Swallow 2011).

Factors that influence gene expression can be environmental, epigenetic and genetic, with elements of the latter acting *cis* and *trans* of the gene. Enhancers, silencers, promoters and insulators are such *cis* elements, which can have influence on transcription factor binding (*trans* elements) and mutations in these regions can have substantial effects on the expression of a gene. As mentioned earlier, it was shown using mRNA activities assigned to certain chromosomes that the transcription of *LCT* is regulated by a mechanism *cis*-acting to the gene (section 1.2.5) and a regulatory network of transcription factors is involved for the different lactase expression between lactase persistent and non-persistent individuals (see also section 1.2.9, 1.2.12).

Evidence for differences in transcription factor binding came from the analysis of DNA-protein interactions between proteins of the nuclear extract of the lactase expressing matured Caco-2 cells and radioactively labelled DNA oligonucleotides with sequence parts of the promoter or enhancer of *LCT* containing different mutations. The broader gene regions that are involved in the differential lactase expression were defined in transfection experiments in Caco-2 cells with plasmids containing different parts of upstream genetic regions of *LCT* (see also section 1.2.9 and 1.2.12). Details of the procedures are described in sections 2.2.6 and 2.2.7 of chapter 2. However further experimental procedures might identify other regions or elements that are involved in lactase expression, some of them are discussed in chapter 7.

1.4 Aims and overview of the thesis

The general aim of this thesis is to examine the genetic variation in and around the *LCT* gene in human populations with emphasis on Europe and the Middle East, to shed more light on the evolution of the lactase persistence. My studies also focus on how specific genetic variants can influence lactase expression.

The following specific aims were addressed in each chapter:

- To investigate the variation of *LCT* enhancer alleles in Eurasia and the Middle East (chapter 3).
- To study putative functional *LCT* enhancer alleles *in vitro* (chapter 4).
- To examine patterns of linkage disequilibrium and obtain insight into the demographic history and origin of these alleles from their haplotypic background (chapter 5).
- To explore new genetic regions for possible missing alleles for lactase persistence (chapter 6).

2 Material and Methods

2.1 DNA samples and population histories

All DNA samples analysed within this project were part of University College London, G.E.E. DNA collections, mainly of the former TCGA (The Centre for Genetic Anthropology) and of Dallas Swallow's group. The Spanish and some of the Catalan DNA samples were kindly provided by the laboratory of Professor Andres Ruiz-Linares.
Samples were collected anonymously with informed consent of the donors (or their legal guardian when under 18 years old for a collaborative project with the City of London school). The collections are covered by UCL/H ethics approvals (ULCH 99/0196, 01/0236 and UCL 2670/001) and consents were also obtained from the relevant authorities in the countries outside the UK.

The samples were obtained from unrelated individuals, being unrelated at the parental and at least grandpaternal level. Information about self declared ethnic background/cultural identity (also for parents and grandparents) and sampling locations were available in most cases. Individuals were grouped accordingly into population groups for analysis. Individual samples were classified to be part of a certain ethnicity group if their self declared cultural identity and those of both parents belong to the same group, otherwise as mixed within that country or other for that geographic region. Samples from individuals of very mixed ancestry were excluded from the study. Language classification of populations were done using the ethnologue online resource (http://www.ethnologue.com/) (Lewis et al. 2013).

2.1.1 Samples from Europe and Asia used for the geographic survey of chapter 3

A summary of the population samples used for chapter 3 of this thesis are shown in Table 2.1. Some of the populations, with a history particularly interesting for this thesis, are further described.

2.1.1.1 Han-Chinese

Han-Chinese do not have a long history of milk drinking; this only very recently became fashionable. According to Murdock (1967) Chinese rely on animal husbandry for up to 6-15% of their diet, but usually do not milk their animals. China was one of the centres of

domestication around 8000 years ago but the domesticated animals were non-milkable species like pigs and dogs (Cavalli-Sforza et al. 1994). Except for minority ethnic groups and regional influences of pastoralists groups from neighbouring countries, there were only two periods when the use of milk and milk products was introduced on a broader scale. Between 250 and 1000 AD Northern China was influenced by cattle-breeding societies and also 1280-1368 under Mongolian leadership (Simoons 1970b) during the Yuan Dynasty. The DNA samples were collected from Han-Chinese individuals living in Singapore. Chinese with its several forms belongs to the Sino-Tibetan languages, Mandarin is the official national language of China.

2.1.1.2 Mongols (Mongols and Khalka)

The Mongolian samples were collected in different parts of Mongolia. Dairy products are and have been of great importance to Mongols, especially kumiss, a fermented milk product.
Most of them belonged to different ethnic groups, who speak a language of the Altaic language family. Only the samples from the Khalka were enough to build an 'own population group' for analysis. The Khalka live to 76 - 85% from animal husbandry. They are pastoralists who hold cattle, water buffalo and yaks (Murdock 1967).

2.1.1.3 Tyroleans

The three Tyrolean groups sampled have a different historic background. Samples are from Bozen, Gadertal and Vinschgau and indviduals speak different languages or dialects with Italian, Ladin (a mixture of German and Italian) and German respectively. The Italian speakers only recently immigrated from the South of Italy (around 1920) from different places of the country. Ladins are suggested to be the closest related to native Tyroleans and 'they acquired the typical Romance Rhaetian language during the Roman invasion of the Alps' (from the 1st century BC to the 3rd century AD). The German speaking groups from South Tyrol are believed to have immigrated around the 5th century from the North (Valentina Coia, personal communication).

2.1.1.4 Sami

The Sami are traditional hunter-gatherers living in the far North of Scandinavia and Russia where they have been indigenous for about 2000 years (Cavalli-Sforza et al. 1994). The Sami subsisted mainly from hunting and fishing until they became reindeer herders in the recent centuries and many of them live in semi-settlements most of the year (Leonard and

Crawford 2002). Samples tested for this thesis originate in Sweden. Sami is part of the Uralic language family.

2.1.1.5 Yakuts

The Yakuts, or Sakha as they describe themselves, live in the northeastern part of the Russian Federation in the Sakha Republic (or Yakutia). They are thought to originate from the Lake Baikal region and were forced to move north by the Buryat-Mongols during the 13th and 14th centuries.

Yakuts were traditional nomads and Northern and Southern Yakuts are culturally quite distinct (Wixman 1984). Northern Yakuts mainly rely on fishing, hunting and reindeer breeding whereas Southern Yakuts are cattle and horse breeders. Reindeer and mare's milk are an important part of the Yakutian diet with koumiss, their traditional drink of fermented mare's milk (Zeder 2006).

The Yakut DNA samples were divided in two sub-groups of differing origin. Samples collected in the Abyysky and Allaikhovsky Districts were grouped together as Northern Yakuts; samples from the Nyurbinsky, Ust-Aldansky and Vilyuysky Districts and from Yakutsk (phenotyped samples) were grouped as Yakuts (South). Yakut is a northern Turkic language branch of the Altaic language family.

Table 2.1. Population samples used in chapter 3, their location and language family (according to: http://www.ethnologue.com/)**.**

Region	Country	Population	Language family
Northwest/	Germany	**Germans**	Indo-European
Central Europe		**Sorbs**	Indo-European
	Ireland	**Irish**	Indo-European
	Netherlands	**Frisians**	Indo-European
	Norway	**Norwegians**	Indo-European
	Slovakia	**Roma**	Indo-European
	Sweden	**Sami**	Uralic
		Swedes	Indo-European
	UK	**Ashkenazi-Jews**	Afro-Asiatic
		English	Indo-European
South Europe	Greece	**Greeks**	Indo-European
	Italy	**Tyroleans Bozen**	Indo-European
		Tyroleans Gadertal	Indo-European
		Tyroleans Vinschgau	Indo-European
	Portugal	**Portuguese**	Indo-European
	Spain	**Catalans**	Indo-European
		Spanish	Indo-European
East/	Belarus	**Belarusians**	Indo-European
Southeast	Macedonia	**Macedonians**	Indo-European
Europe	Romania	**Romanians**	Indo-European
	Russia	**Erzya**	Uralic
		Moksha	Uralic
		Russians Perm	Indo-European
	Ukraine	**Ukrainians**	Indo-European
Middle East	Cyprus	**Greek Cypriots**	Indo-European
	Iran	**Iranians**	Indo-European
	Kuwait	**Kuwaiti**	Afro-Asiatic
	Syria	**Syrians**	Afro-Asiatic
	Turkey	**Anatolian-Turks**	Altaic
	Yemen	**Yemeni Hadramaut**	Afro-Asiatic
		Yemeni Sena	Afro-Asiatic
West Asia	Armenia	**Armenians**	Indo-Eurpean
	Azerbaijan	**Azeri**	Altaic
	Georgia	**Georgians**	Kartvelian
Central/	Afghanistan	**Pashtuns/Afghans**	Indo-European
South Asia		**Tadjiks**	Indo-European
		Uzbeks	Indo-European
	India	**North Indians**	Indo-European
		South Indians	Dravidian
	Nepal	**Nepalese**	Indo-European
		Tharu	Indo-European
	Uzbekistan	**Uzbeks**	Altaic
Central/East/	Mongolia	**Khalka**	Altaic
Southeast		**Mongols**	Altaic
Asia	Russia	**Northern Yakuts**	Altaic
		Yakuts	Altaic
	Singapore	**Han-Chinese**	Sino-Tibetan
		Japanese	Japonic

2.1.2 African samples

Information about the African and Middle Eastern samples used in chapters 5 and 6 were extracted from the theses of Bryony Jones and Kate Ingram where they were studied in detail (Ingram 2008; Jones 2012). Table 2.2 summarises anthropological information such as broad geographic location, language families and lifestyle patterns of these populations.

Table 2.2: Summary of geography, language family and subsistence strategy of the African and Middle Eastern populations used for comparative analyses (compiled from Ingram 2008; Jones 2012)**.**

Populations	Country	Language family	Settlement pattern	Primary substistence method	Presence of milking	Predominant milked animal
Afar	Ethiopia	Afro-Asiatic	Nomadic	Pastoralism	Yes	Camels
Amhara	Ethiopia	Afro-Asiatic	Sedentary	Farming	Yes	Cattle
Asante	Ghana	Niger-Congo	Sedentary	Farming	No	-
Beni Amer	Sudan	Afro-Asiatic	Nomadic	Pastoralism	Yes	Camels
Chagga	Tanzania	Niger-Congo	Sedentary	Farming	Yes	Cattle
Chewa	Malawi	Niger-Congo	Sedentary	Farming	No	-
Jaali	Sudan	Afro-Asiatic	Sedentary	Pastoralism	Yes	Cattle
Mambila	Cameroon	Niger-Congo	Sedentary	Farming	No	-
Oromo	Ethiopia	Afro-Asiatic	Nomadic	Pastoralism	Yes	Cattle
Shabo	Ethiopia	Nilo-Saharan	Sedentary	Farming, hunting, gathering	No	-
Bedouin	Israel, Jordan, Saudi Arabia	Afro-Asiatic	Nomadic	Pastoralism	Yes	Camels
Israeli Arabs, Palestinians	Israel	Afro-Asiatic	Sedentary	Farming	Yes	Cattle

2.2 Experimental methods

2.2.1 DNA collection

DNA samples newly collected during the course of this thesis were obtained from buccal cell DNA. For this collection, cotton swabs were rubbed around on the inside of the cheek of the sample donors for at least 30 seconds and 10 of the swabs were placed in a 15ml Falcon tube containing 2.5ml 'Slagboom' buffer including proteinase K (see 2.5 for receipe). The tubes were stored in a dark environment at room temperature until extraction.

2.2.2 DNA extraction

The samples used in this thesis were extracted by a variety of different procedures, depending on the cellular source of the DNA and the laboratory from which they were obtained by different adaptations of the phenol-chloroform method or salting out precipitation methods.

2.2.2.1 Within group extraction method

The buccal cell samples newly collected for this thesis and other buccal samples prepared within the Swallow group were extracted by Ranji Araseretnam or Kate Ingram following an extraction protocol devised by Freeman et al. (2003).
Each collection tube containing Slagboom buffer and cotton swabs was incubated at 65°C for 2 hours. After removing the cap, the tube was inverted into a clean 50ml tube and both centrifuged for 10 minutes at 650x g. The original tube was discarded, the swabs were drained and all liquid was transferred into a new 15ml tube. 300µl Yeast Reagent 3 (Autogen Bioclear, diluted 1:1 with 100% ethanol) was added. This contains potassium acetate and phenol and chloroform to denature remaining proteins. The solution was vigorously mixed manually for 1 minute and then centrifuged for 25 minutes at 8000x g. A repeat of this step followed, after transfer of the supernatant to clean tube and further addition of 300µl Yeast Reagent 3. The supernatant containing the DNA was transferred again into a clean tube. 1.8ml 100% isopropanol was added and the tube gently inverted to mix, and precipitate the DNA. A centrifugation step followed at 8000x g for 25 minutes. The supernatant was discarded and the DNA containing pellet washed gently with 1ml 70% ethanol and centrifuged again at 8000x g for 10 minutes. The ethanol supernatant was then discarded and the tube left open to allow the pellet to air dry for 15 minutes. The DNA pellet was resuspended in 400µl Puregene DNA hydration solution (1mM EDTA,

10mM Tris-CI (pH 7.5), Gentra) and incubated overnight at 4°C on a rocking platform. Sample aliquots were stored ad +4°C and -70°C.
DNA from blood and cell line samples was extracted previously using Puregene DNA extraction kits, which involve salting out precipitation, following the manufacturer's instructions.

2.2.2.2 Extraction from dried blood

The DNA of 43 phenotyped Yakut samples was extracted from blood dried onto filter paper, using the QIAamp DNA Mini Kit (QIAGEN). The procedure followed the protocol 'DNA purification from dried blood spots':
Approximately 1cm^2 of paper was cut from each blood spot with-scissors, cut into 4 smaller pieces and placed in a 1.5ml eppendorf tube. After each sample the scissors and forceps used were washed with 1% Alconox solution (Sigma-Aldrich), rinsed with dH_2O and 70% ethanol and dried with a separate tissue to avoid cross contamination of the samples. 180µl of Buffer ATL was added and the tube and incubated at 85°C for 10 minutes. The addition of 20µl of proteinase K stock solution (600 mU/ml) followed, the mix was vortexed briefly and incubated at 56°C for 1 hour. 200µl lysate buffer (Buffer AL) was added to the sample, the solution mixed thoroughly by vortexing and incubated at 70°C for 10 minutes. 200µl 100% ethanol was added, the mix quickly vortexed and then transferred to a QiAamp mini spin column in a 2ml collection tube.
After centrifuging at 6000x g for 1 minute, the tube with supernatant containing proteins and other contaminants was discarded and the column transferred into a new collection tube. The DNA absorbed onto the silica membrane of the column was then washed in two steps to remove any remaining contaminants. First 500µl Buffer AW1 (96% ethanol) was added to the spin column and it was centrifuged at 6000x g for 1 minute. The collection tube with the filtrate was emptied and the second washing step followed with the addition of 500µl Buffer AW2 (96% ethanol) and centrifugation at 16000x g for 3 minutes. The spin column was placed in a 1.5ml eppendorf tube and 150µl Buffer AE (10mM Tris Cl, 0.5mM EDTA, pH 9.0) was added to the column and incubated for 1 minute. In the following centrifugation step at 6000x g for 1 minute the DNA was eluted from the membrane. The samples were stored at -20°C.

2.2.3 Lactase persistence data and lactose tolerance testing

As described in section 1.2.4 of the introduction, the classification of the lactase persistence status of an individual can be made from biopsy material in measuring lactase activity directly or indirectly via lactose tolerance tests. Most of the phenotype data used

in chapters 5 and 6 were obtained during the course of several other projects and were reassessed for this thesis. Test methods are briefly mentioned below, further details can be obtained from the corresponding publications.

The Northern European sample series was first analysed by Harvey et al. (1995b). Duodenal biopsy material from hospital patients (UCLH) was examined for lactase and sucrase enzyme activity, and checked by histology.
Phenotypes of the Italian samples were also determined from lactase activity of biopsy samples, collected from hospital patients in Naples who underwent jejunal surgery (Harvey et al. 1998; Maiuri et al. 1991).
The Finnish samples were lactose tolerance tested by Dr Riitta Korpela in Helsinki via urinary galactose and breath-hydrogen tests. This cohort does not represent a randomly chosen population sample as it was collected to contain approximately half lactase persistent and non-persistent female individuals inferred from lactose digester/non-digester status (Harvey et al. 1998; Hollox 2000), which is higher than the population average for lactase non-persistence and was therefore only used for analyses in chapter 6.
Lactose tolerance testing of the Yakut samples was done via blood glucose tests, conducted by Sardana Markova and Maria Khandy of the Yakutsk State University (Hollox 2000).

The volunteers from the African groups were tested by the breath hydrogen method and this was done by Mohamed Elamin from Khartoum, Tamiru Oljira from Addis Ababa University and Kate Ingram from UCL: The Jaali were from Sudan (Ingram et al. 2007; Ingram et al. 2009b) and the Somali camel herders, the Amhara and the Oromo groups were from Ethiopia (Jones 2012).

5 individuals from Northern Europe were also newly tested with this method, which can be briefly described as follows. The volunteers were given a lactose load of 50g dissolved in water after fasting overnight. Breath hydrogen was measured before and every 30 minutes (up to 3 hours) after drinking the lactose solution, with a MicroH2 meter (Micro Medical) (Peuhkuri et al. 1998). Volunteers were classified as lactose non-digesters if they showed and maintained an increase in breath hydrogen of at least 20ppm above the baseline, and as digesters with values below this. 2 individuals with a baseline and all test results lower than 2ppm were excluded as they could have been potential hydrogen non-producers.

2.2.4 Sequencing

All genomic regions of interest were firstly amplified via polymerase chain reaction (PCR) to be later analysed by sequencing: The *LCT* enhancer region in intron 13 of *MCM6;* for haplotype analysis a control region in intron 4 of *MCM6* and 3 sequence regions containing the *LCT* core haplotype defining SNPs (hapdef) previously described by Hollox et al. (2001), and the immediate promoter (IP) and parts of intron 2 of *LCT* (Figure 2.1).

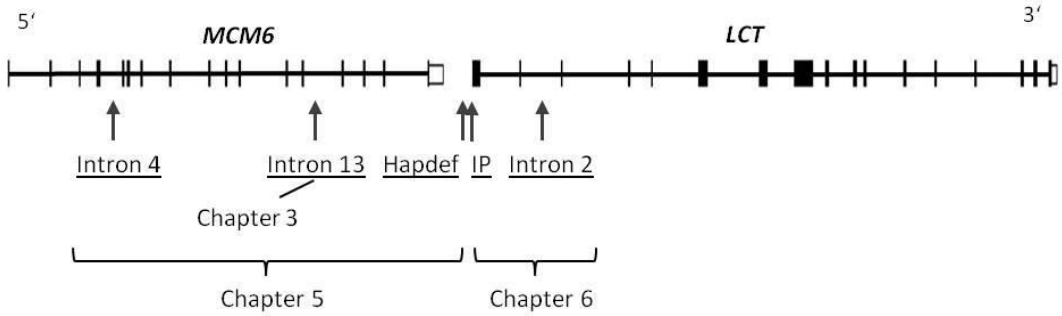


Figure 2.1: Overview about the sequenced regions for the different chapters of this thesis.

Some of the data collection was part of undergraduate projects. Most of the data for the *LCT* enhancer and immediate promoter region for South and East Asia (India, China and Japan) was collected by Anshua Gosh (Ghosh 2010). Some of the haplotype marker data for the Middle Eastern populations of Kuwait, Iran and Syria was accumulated by Lana-Mai Couzens (Couzens 2011).

The sequences of intron 4 of *MCM6*, the lactase enhancer and hapdef regions for the African populations were obtained from previous projects of Kate Ingram, Pawel Zmarz and Bryony Jones (Ingram 2008; Ingram et al. 2009b; Jones 2012; Zmarz 2010). Kate Ingram also performed the sequencing of the *LCT* enhancer and the genotyping of some haplotype defining SNPs for the Bedouin, Israeli and Palestinian populations and the European phenotyped samples (Ingram et al. 2007).
All of these sequences were re-checked and re-sequenced if necessary for chapters 5 and 6 of this project and all remaining samples were tested as part of this project.

2.2.4.1 Primer design

Primers for a multiplex PCR were designed to simultaneously amplify two of the haplotype marker regions of *LCT* exon 2/intron 1 and exon 17 of the haplotype defining regions. The sequences for the two regions were obtained from the human reference sequence, downloaded from the Ensembl database (build 71, GRCh37/hg19) and the FastPCR

software (version 5.4.8, Kalendar et al. 2011) was used for primer design. Primer specificity was checked with the 'BLAT search' of the UCSC browser (Kent et al. 2002). The primers for *LCT* intron 2 were also designed during the course of this PhD, by Bryony Jones.

2.2.4.2 PCR

Sampled DNA was PCR amplified using conditions optimized for each genomic region (temperature and cycling conditions). The multiplex PCR needed additional adjustment for optimal dNTP and $MgCl_2$ content. Primer sequences and reaction specific PCR cycling conditions are summarized in Table 2.3.

PCR reactions were carried out in 96 well plates each in a total volume of 10µl. Standard reactions contained 1x Reaction Buffer IV, 0.15mM $MgCl_2$ (0.25mM for Multiplex), 200µM dNTPs (400µM for Multiplex), 0.25U *Taq* DNA polymerase (all Thermo Fisher), 0.25-0.5µM of each primer (Sigma-Aldrich) and about 10-20ng genomic DNA. The basic method included an initial denaturation step at 95°C for 5 minutes, followed by a variable number of cycles of denaturation (95°C for 30sec), annealing (variable temperature for 30sec) and elongation (72°C for a variable time). A final elongation step followed for 5 minutes at 72°C.

2.2.4.3 Gel electrophoresis

All PCR products (after the PCR and after the PCR cleanup) were separated on 1 or 2% agarose gels in comparison with a DNA ladder (Stretch marker, see 2.5.1), to check their quantity and ensure the correct sizes. 2µl of the PCR product mixed with 3µl loading buffer was loaded on the gels that were stained with 50ng/ml ethidium bromide (Sigma) that was included in the gel. The gels were visualised and photographed under UV light for documentation.

2.2.4.4 PCR cleanup

To prepare the PCR products for the sequencing process they needed to be purified. The PCR cleanup protocol is essentially a polyethylene glycol (PEG) precipitation method: The precipitation solution (2/3 'Microclean', see section 2.5.1) was added to the PCR product in 3 fold volume. The plates were mixed thoroughly and centrifuged at 1500x g for 1 hour. The supernatant, containing primers, excess nucleotides and salts was then discarded by inverting the plate onto tissue and centrifuging at 20x g for 1 minute. A washing step followed to remove the remaining PEG and any contaminants with the addition of 150µl

70% ethanol, centrifugation for 10 minutes at 1500x g. The ethanol solution was discarded as before by inverting and centrifuging of the plate at very low speed (20x g) for 30 seconds. The DNA pellets were left at room temperature to air dry for at least 15 minutes and were then resuspended in 10µl dH_2O.

2.2.4.5 Sequencing

All sequencing reactions were performed by myself or carried out by Mari-Wyn Burley from 'The Centre for Comparative Genomics' at UCL using a standard protocol based on the 'Sanger sequencing' method (Sanger et al. 1977) and the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, ABI). For the reactions either PCR primers or additionally designed sequencing primers were used (Table 2.3).
Each sequencing reaction was carried out in a 10µl volume with 2-2.5µl cleaned up PCR product. The mix further contained 0.35µl BigDye Terminator v3.1, 1x sequencing buffer and 0.16µM primer. Thermo cycle sequencing was performed using the following parameters: After an initial denaturation at 96°C for 1 minute 25 cycles of 96°C for 1 minute, 50°C for 5 seconds and 60°C for 4 minutes followed.

Following the sequencing reaction, the DNA was purified by ethanol precipitation to remove contaminants. 2.5µl of 125mM EDTA and then 30µl 100% ethanol was added to each well. The plate was thoroughly mixed and centrifuged for 1 hour at 3870rpm. The supernatant was removed by inverting the plate onto tissue and centrifuging at 300rpm for a few seconds. Any remaining contaminants were then removed by the final ethanol wash: 30µl of 70% ethanol was added to each well and the plate centrifuged again at 3870rpm for 10 minutes. The ethanol was removed by inverting the plate onto tissue again and centrifugation at 300rpm for a few seconds. The plate with the DNA pellets was left at room temperature to air dry for 15 minutes.

To prepare the samples for the run on an ABI 3730xl DNA Analyzer, the pellets were resuspended in 10µl HiDi formamide (ABI), denatured at 95°C for 5 minutes and then immediately put on ice for at least 5 minutes before they were loaded in the machine.

2.2.4.6 Sequence analysis

Sequence chromatograms were analysed using the software ChromasPro (version 1.5, Technelysium Pty Ltd) and compared to a reference sequence (Ensembl database, GRCh37/hg19). The DNA was mostly sequenced in both directions. New variants were in

each case confirmed on both strands or sequencing of at least one additional and independent PCR.

The ancestral allelic state at each new locus was inferred by comparing the human reference sequence for each genetic region with those of other primate species (see Appendix A). The human reference sequence (Feb 2009, GRCh37/hg19) for each region was downloaded from the UCSC browser (Kent et al. 2002) using the flanking PCR primers for the 'In-Silico PCR' tool. The corresponding primate sequences were generated using the UCSC 'BLAT search' with the reference sequence assemblies for chimpanzee (Feb 2011, CGSC 2.1.4/panTro4), orang-utan (July 2007, WUGSC 2.0.2/ponAbe2), gorilla (May 2011, gorGor3.1/GorGor3) and rhesus macaque (Oct 2010, BGI CR_1.0/rheMac3). The BioEdit sequence alignment editor (version 7.0.9.0, Hall 1999) and the integrated CLUSTAL W (version 1.81) multiple sequence alignment tool (Thompson et al. 1994) was used to align the sequences.

2.2.5 Genotyping

2.2.5.1 KASP genotyping

Genotyping of 880 samples for analysis in chapter 5 was performed externally by LGC Genomics (KBioscience) using their KASP ('Kompetitive allele specific PCR') genotyping assay. Results were viewed with SNPviewer2 (version 3.2.2.16), provided by the company. Quality control checks were performed after the data were provided. For each population, the SNP calls were visually inspected to check they were assigned to the correct genotype. Uncertain calls (marked as '?') that could be clearly grouped with either one of the homozygous or the heterozygous genotypes were corrected manually. Finally the data were tested as to whether they agree with Hardy Weinberg expectations (as described in section 2.3.1).

Table 2.3: Primers used for PCR amplification and sequencing at UCL

Genetic region	Location (bp 5' of *LCT* start of transcription)	Primer name	Primer sequence (5'-3')	Product length	Annealing temperature (°C)	Elongation time (sec)	cycles
***LCT* enhancer**							
MCM6, intron 13	-14,397	MCM6ex13	ATTTCCAAAGAGTCAGAGGACTTC	940	58	60	39
		MCM6778	CCTGTGGGATAAAAGTAGTGATTG				
	-14,163	MCM6i13	GGACATACTAGAATTCACTGCAAATAC	706	58	45	38
		MCM6778	CCTGTGGGATAAAAGTAGTGATTG				
		MCM6ex13	ATTTCCAAAGAGTCAGAGGACTTC	661	58	45	38
		LAC-C-L2	CTGCTTTGGTTGAAGCGAAGAT				
LCT,	-527	LCT IP_f	TCTTCAGACATTTTCCGGGTTC	719	56	45	37
immediate promoter		LCT IP_r	TGTCTTTGTCCCCTGCTACA				
***LCT* haplotype marker**							
LCT, hapdef region	-1,162	LCT_far_prom_f	ATCCACATTCTACAGGTGACAA	701	59	60	38
		LCT_far_prom_r	GACCAACACAAAAACCTCAGAC				
LCT, exon 2	3,762	LCT_Ex2_F	CCTTGCCAACTCTCCAACTGC	502	59	45	38
		LCT_Ex2_R	ACACTTGACCAAGCAGGAGC				
LCT, exon 17	48,500	LCT_Ex17_F	GATGGGTACCTCCACCTCG	927			
		LCT_Ex17_R	TGGTGAGAAAGCTTAATCGGAGC				
LCT, intron 2	5,282	LCT_intron2_f	CTGGGAAGTGAACAGCTTTGG	839	58	60	39
		LCT_intron2_r	CCACCACATTAAGTCTTGGTCTCC				
MCM6, intron 4	30,461	MCM6_intron4f	ACCCTCAGATTTCAGCAGGAC	683	58	45	39
		MCM6_intron4r	ACTCCATGATGATTCAAGCAGC				

Other sequencing primers	Primer name	Primer sequence (5'-3')
LCT, inner promoter	LCT IP	AGATGGAGTCTCATTCTGTTGC
LCT, hapdef region	LCT-farprom_hap_seq_primer(r)	TTGGTGACCGGGTCTCACTCTG
MCM6, intron 4	intron4_R2	GACTCCAAATATCTGTTCCATGC

2.2.6 Electrophoretic mobility shift assays (EMSAs)

EMSAs were performed to detect nuclear proteins in the Caco-2 cell line that bind to DNA sequences of the *LCT* enhancer and distinguish possible differences in binding affinity to probes containing the derived or ancestral allelic variants.

2.2.6.1 Cell culture

Human Caco-2 cells (obtained from Martin Spiess, Basel, passage >80) were cultured in 180ml tissue culture flasks and kept in a humid environment at 37°C and 5% CO_2. The cells were grown in 30ml Dulbecco's modified eagle medium (DMEM, with 4.5 g/L Glucose and L-Glutamine, BioWhittaker-Lonza), supplemented by 100U/ml penicillin, 100µg/ml streptomycin and 10% foetal calf serum (Remal, all Fisher Scientific). At 80% confluence (after 3-4 days) the cells were split by trypsinisation. To do this the medium was discarded from the Caco-2 monolayer culture and the cells were washed with 5ml 0.05% Trypsin-EDTA solution (Gibco, Life Technologies) which was then all discarded with exception of a small amount just covering the cells. A short incubation followed for up to 5 minutes at 37°C until the cells could be detached from the flask surface by tapping the flask. 10ml of fresh DMEM 10% FCS, (full culture medium) were added (using the anti-trypsin in the FCS) to stop the reaction and suspend the cells. To start a new passage of cell culture 1ml of DMEM-containing cell solution was transferred to a new culture flask and 29ml culture medium added. The medium was changed at 2 day intervals.

2.2.6.2 Preparation of nuclear protein extract

Caco-2 cells were grown on 20 x 20cm cell culture plates. After 10 days of culture cells had grown to confluence and the formation of domes was visible by inspection under a microscope as described by Pinto and colleagues (Pinto et al. 1983b) indicating that they had become differentiated.

Nuclear protein extract was prepared from differentiated Caco-2 cells 13 days after seeding following the protocol described in Ausubel (2002). The medium was removed from the cells, they were washed twice in phosphate-buffered saline (PBS, pH 7.4, Gibco, Life technologies) and harvested by scraping them off the surface in 40ml PBS. A centrifugation step followed at 3000rpm for 10 minutes and the supernatant was discarded. The cell pellet was immediately resuspended in hypotonic buffer (see 2.5) of 5x the volume of the cell pellet and centrifuged for 5 minutes at 3000rpm to remove salt from the PBS solution for effective swelling in the next step. The supernatant was discarded and the cell pellet resuspended again in hypotonic buffer to a final volume of 3x the first pellet

volume and allowed to swell on ice for 10 minutes. In the following step the cells were homogenized in a glass Dounce homogenizer by 10 very slow up-and-down strokes with a type B pestle to lyse the cells and release the nuclei. Cell nuclei were collected by centrifuging at 3300x g for 15 minutes and after that resuspended in low-salt buffer of half the volume of the nuclear pellet. This was followed by application of the same volume (half the original nuclear pellet) of high-salt buffer (for recipes see 2.5.1) dropwise with continuous and gentle stirring and the nuclei were allowed to extract under further gentle mixing for 30 minutes. The lysed nuclei were pelleted by centrifugation at 20000x g for 30 minutes and the nuclear extract in form of the supernatant placed in dialysis tubing (Visking dialysis membrane, diameter size 6.3mm, volume 0.3ml/cm, molecular weight cut-off 12-14 kDa, Medicell) and left to dialyse against 50 volumes of dialysis buffer overnight. The extract of the dialysis bag was then centrifuged at 14500rpm for 20 minutes and the clean extract between the pellet and precipitates on the supernatant surface was removed, aliquoted and stored at -80°C.

The protein concentration of the nuclear extract was measured by the method of Bradford (1976) using the 'Bio-Rad Protein Assay' microassay procedure. 5 dilutions of the protein standard (Protein standard II, bovine serum albumin, Bio-Rad), to create a standard curve, and 2 dilutions of the nuclear extract with PBS in a 10µl volume were made in duplicate. 10µl of dialysis buffer was added and the mixture made up to 800µl with dH_2O. 200µl of dye (protein assay dye reagent concentrate, Bio-Rad) was added to each tube and after vortexing the tubes were incubated at room temperature for 5 minutes. The absorbance was measured at 595nm in a photometer and the protein concentration of the extract calculated from the standard curve.

2.2.6.3 EMSA oligonucleotide design

Gel shift assays were performed with labelled DNA probes and nuclear extract, competition experiments with the addition of unlabelled short DNA sequences and supershift experiments with the addition of antibodies specific for a certain transcription factor (see also chapter 4).
Two complementary oligonucleotides for each of the five enhancer variants tested (*-14028*C*, *-14011*T*, *-14009*G*, *-13779*C* and *14010*C* as control) and their corresponding ancestral sequences were designed by visual inspection of the sequence from the human reference sequence template (GRCh37/hg19). They were designed to be between 22 and 31 bp in length with the ancestral or mutated base near the centre and an extra 5' base on

both the sense and antisense oligonucleotide to create an overhang during later annealing, which is important for the labelling of the probe.
For EMSA competition experiments oligonucleotides with known transcription factor affinity sites were used, of which some had already been designed and tested by other lab members. Others were newly designed after bioinformatic analysis of the regions around the five enhancer variants (see section below and chapter 4). The core sequences consisting of matrices for predicted transcription factor binding sites were complemented with random bases at both sides to reach an oligonucleotide length of 25 bp. The oligonucleotide sequences were re-checked bioinformatically for binding specificity and the core flanking sequences modified accordingly. A 5'A base was added to both complementary oligonucleotides in case they would later be needed as probes. All double stranded oligunucleotide sequences used in EMSA experiments can be found in Table 2.4. A further set of oligonucleotides was ordered with sequences taken from technical information sheets of specific transcription factor probes from commercially available EMSA kits (Panomics, Affymetrix transcription factor EMSA kits, http://www.affymetrix.com/estore/).

2.2.6.4 Probe preparation: Radioactive labelling and purification

Sense and antisense oligonucleotides (MWG Operon) were annealed and used as unlabelled (cold) probe for competition experiments or labelled with radioactivity. Annealing reactions were carried out in a volume of 100µl containing 250pmol of each primer and 0.1M NaCl. After heating the mix at 95°C for 1 minute, the tube was left in the heating block on the bench to slowly cool down to room temperature.
2.5pmol of the double stranded probes were labelled using 5U T4 Polynucleotide Kinase (Fermentas) and 25-35µCi [γ^{32}P]ATP (Perkin Elmer) in Kinase reaction buffer A (forward, Fermentas). The reaction mix was incubated for 30 minutes at 37°C and 20µl 1x TE buffer were added afterwards. The probe was purified with a MicroSpin G-25 Column (Illustra, GE Healthcare) and the eluate filled up to 100µl total volume with TE buffer.

Table 2.4: Double stranded oligonucleotides used for EMSAs, variants indicated in bold.

Probe name	Sequence
Enhancer variants	
14028T (ancestral)	5'-ACGTCATAGTTTATAGAGTGCATAAA-3' 3'-GCAGTATCAAATATCTCACGTATTTC-5'
14028C	5'-ACGTCATAGTT**C**ATAGAGTGCATAAA-3' 3'-GCAGTATCAA**G**TATCTCACGTATTTC-3'
14011C (ancestral)	5'-TAGAGTGCATAAAGACGTAAGTTACCATTTA-3' 3'-TCTCACGTATTTCTGCATTCAATGGTAAATT-5'
14011T	5'-TAGAGTGCATAAAGA**T**GTAAGTTACCATTTA-3' 3'-TCTCACGTATTTCT**A**CATTCAATGGTAAATT-5'
14010C	5'-TAGAGTGCATAAAGAC**C**TAAGTTACCATTTA-3' 3'-TCTCACGTATTTCTG**G**ATTCAATGGTAAATT-5'
14009G	5'-TAGAGTGCATAAAGACG**G**AAGTTACCATTTA-3' 3'-TCTCACGTATTTCTGC**C**TTCAATGGTAAATT-5'
13779G (ancestral)	5'-AGTAGTACGAAAGGGCATTCAA-3' 3'-CATCATGCTTTCCCGTAAGTTC-5'
13779C	5'-AGTAGTAC**C**AAAGGGCATTCAA-3' 3'-CATCATG**G**TTTCCCGTAAGTTC-5'
TF binding competitors	
SIF1 24 (Cdx-2 binding site)	5'-TGGGTGCAATAAAACTTTATGAGTA-3' 3'-CCCACGTTATTTTGAAATACTCATT-5'
Oct control	5'-ATGTCGAATGCAAATCACTAGAA-3' 3'-ACAGCTTACGTTTAGTGATCTTA-5'
LPH-CE2c (HNF-1α binding site)	5'-ATAACCCAGTTAAATATTAAGTCTTAAT-3' 3'-TATTGGGTCAATTTATAATTCAGAATTA-5'
TREH's HNF4 site (HNF-4α binding site)	5'-CCTCAAAGGCTGGACTTTGGCCGACTTGG-3' 3'-CTCAAAGGCTGGACTTTGGCCGACTTGGA-5'
mut XbaI 13910 site (2 GATA binding sites)	5'-TACAGATAAGATAATTCTAGACCTGGCCTCAAAGGA-3' 3'-ATGTCTATTCTATTAAGATCTGGACCGGAGTTTCCT-5'
unspec24 (non-specific)	5'-AACGTAGCTGATCGAATCGGTTAC-3' 3'-TGCATCGACTAGCTTAGCCAATGA-5'
Newly designed TF binding competitors (see also Table 4.1)	
c-Ets-1	5'-AATCCTCTACCGGATGTAGGTCGAC-3' 3'-TAGGAGATGGCCTACATCCAGCTGA-5'
Ets/Tel-2	5'-ATCTTACTACTTCCTCGCTGACCGT-3' 3'-AGAATGATGAAGGAGCGACTGGCAA-5'
NF-kappaB	5'-AGTGGCGGGGAAAGTCCCCAGAATC-3' 3'-CACCGCCCCTTTCAGGGGTCTTAGA-5'
Pax	5'-AAGAAGTGGAACTCACGATCGTGCT-3' 3'-TCTTCACCTTGAGTGCTAGCACGAA-5'
Pax4-8	5'-ACGGCGTTCATGCGTGAGCGACCGT-3' 3'-GCCGCAAGTACGCACTCGCTGGCAA-5'
Pbx	5'-AGATGGGATTGATGGTAGCCGTATT-3' 3'-CTACCCTAACTACCATCGGCATAAA-5'

Probe name	Sequence
Competitors from Affymetrix EMSA kit	
Aff_c-Ets-1	5'-ACCAGGAAGCCAGGAAG-3' 3'-GGTCCTTCGGTCCTTCA-5'
Aff_ETS (1)	5'-AGGAGGAGGGCTGCTTGAGGAAGTATAAGAAT-3' 3'-CCTCCTCCCGACGAACTCCTTCATATTCTTAA-5'
Aff_ELF	5'-AGAGTCATCAGAAGAGGAAAAATGAAGGT-3' 3'-CTCAGTAGTCTTCTCCTTTTTACTTCCAA-5'
Aff_Elk-1	5'-ATTTGCAAAATGCAGGAATTGTTTTCACAGT-3' 3'-AAACGTTTTACGTCCTTAACAAAAGTGTCAA-5'
Aff_LEF1	5'-ACCCATTTCCATGACGTCATGGTTA-3' 3'-GGGTAAAGGTACTGCAGTACCAATA-5'
Aff_GKLF	5'-AATGCAGGAGGAGAAAGAAGGGCGTAGTATCTACTAG-3' 3'-TACGTCCTCCTCTTTCTTCCCGCATCATAGATGATCA-5'
Aff_AML1	5'-AATCTCTATGTGGGTGTGGGTGTGGGAATCAT-3' 3'-TAGAGATACACCCACACCCACACCCTTAGTAA-5'
Aff_GATA3	5'-AGTTATTTATCTCTTAGTTGTAGTTATTTATCTCTTAGTTGTA-3' 3'-CAATAAATAGAGAATCAACATCAATAAATAGAGAATCAACATA-5'
Aff_GATA4	5'-AGCCTAAGCCAAGTGATAAGCAGCCAGACAA-3' 3'-CGGATTCGGTTCACTATTCGTCGGTCTGTTA-5'

2.2.6.5 Protein binding reaction

A reaction mix, containing both the extract and radioactive end-labelled DNA fragments was prepared. Binding reactions for EMSA experiments were usually performed with 4.5-9µg differentiated Caco-2 nuclear extract, 4µl dialysis buffer, 10µl gelshift buffer. Each reaction also contained 0.25µg Poly (deoxyinosinic-deoxycytidylic) acid sodium salt (poly dI-dC, Sigma-Aldrich) and 2.5 pmol unlabelled oligonucleotides to prevent nonspecific DNA binding for clearer bands on the gel pictures.

Most reactions showed both specific and non-specific binding proteins. To specify further which protein of the extract is binding, competition and supershift experiments were performed. For competition assays, 2.5pmol of the unlabeled annealed competitor oligonucleotide was added to the reaction mix and for supershift assays 1-2µl of the antibodies Cdx-2 (BioGenex), Oct-1, HNF-1α, HNF-4α, Ets-1/2, Ets-1, GATA-6 (all Santa Cruz).

After incubation of the mixture for 10 minutes on ice 2.5 fmol of probe was added, mixed and incubated on ice for at least further 20 minutes.

2.2.6.6 Polyacrylamide (PAA) gel electrophoresis and analysis

For electrophoresis, 5% nondenaturing polyacrylamide gels (30% acrylamide:bisacrylamide 29:1, Sigma-Aldrich) were prepared. The PAA-gel was pre-run at 4°C for 1 hour using 0.5x TBE buffer before the mixes of the binding reactions, resolved

in 2µl EMSA loading buffer, were loaded and run at 4°C. Following electrophoresis the gel was transfered onto Whatman paper and vacuum dried for at least 2 hours at about 80°C. The gel-paper was exposed to a phosphoimaging screen for at least 2 hours which was then scanned with a Storm scanner and analysed using the Imagequant software (version 5.1, GE Healthcare).

2.2.7 Transfection studies

Transfection studies were performed to examine the effect of the different *LCT* enhancer and promoter variants, introduced in a reporter gene construct, on the expression of the reporter gene luciferase. Figure 2.2 summarises the cloning strategies described below.

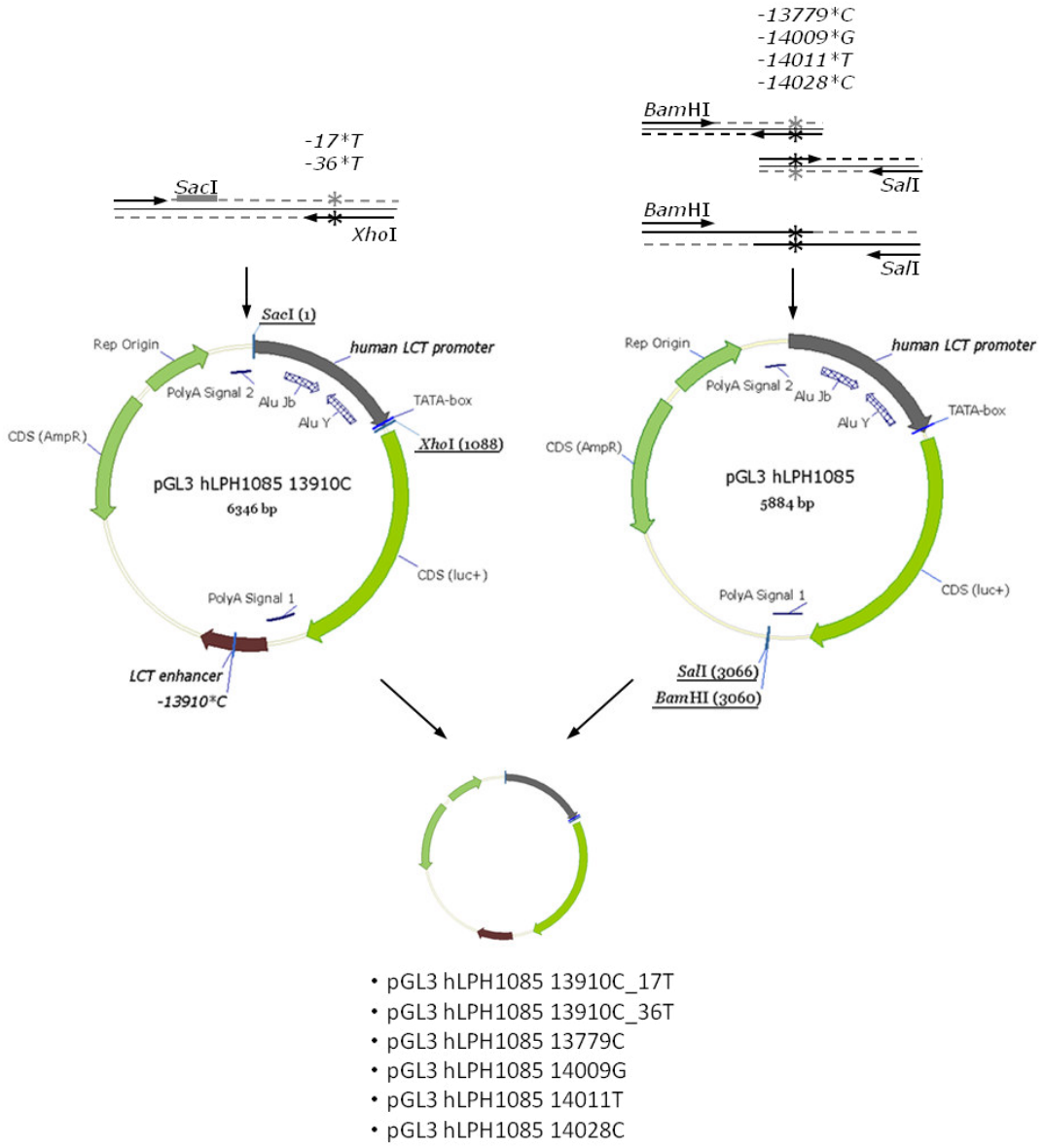


Figure 2.2: Summary of the different strategies for site directed mutagenesis and insertion of *LCT* promoter fragments (left) and enhancer fragments (right) into the luciferase reporter gene plasmids. *LCT* promoter variants were created in one PCR step and inserted via *Sac*I and *Xho*I into a plasmid, containing the ancestral enhancer region (pGL3 hLPH1085-13910*C (Troelsen et al. 2003)). *LCT* enhancer variants were generated in a two step PCR amplification and inserted via *Bam*HI and *Sal*I into the pGL3vector containing the ancestral promoter region (pGL3 hLPH1085 (Troelsen et al. 2003)). Vector maps were created using Vector NTI Advance software (version 11.5, Invitrogen).

2.2.7.1 Preparation of the LCT enhancer fragments

The *LCT* enhancer variants were generated via site directed mutagenesis in a two step PCR amplification of the pGL3 hLPH1085-13910C enhancer plasmid construct (Troelsen et al. 2003). Primers designed to add an upstream 5' *Bam*HI and downstream 3' *Sal*I restriction

site to the 450 bp enhancer fragment were used together with the overlapping primers designed for EMSA experiments to insert the variants (see Table 2.5).
First, the region was amplified in two separate PCRs to generate a left and right fragment (with the *Bam*HI site including primer and the reverse variant EMSA primer for the left fragment and the primer including the *Sal*I site and the forward variant EMSA primer for the right fragment). PCR reactions contained 1x *Pfu* Buffer with $MgSO_4$, 200µM dNTPs, 2.5U recombinant *Pfu* DNA polymerase, all Fermentas), 0.4µM of each primer (MWG Operon) and about 37ng of the pGL3 hLPH1085-13910*C plasmid as template. Cycling conditions are described below (2.2.7.3).

The two fragments, in form of 1µl of the purified left and right fragment, were then annealed to create an overlap at the mutated position with a denaturation step at 94°C for 1 minute and an annealing step at 72°C for 3 minutes prior the second PCR (for cycling conditions see 2.2.7.3). The flanking *Bam*HI and *Sal*I site containing primers were then used to amplify the whole enhancer fragment.

2.2.7.2 Preparation of the LCT promoter fragments

The *LCT* promoter variants were created in one PCR step with the mutations directly included in the 3' primers, also with the pGL3 hLPH1085-13910C plasmid as template. Primers containing the *-36*T* and *-17*T* variants and a 3' *Xho*I restriction site were designed and used together with a previous designed 5' primer to create a fragment containing a *Sac*I restriction site via PCR.

Table 2.5: **Primers used for PCRs to create fragments to be inserted into reporter gene plasmids** (mutated bases are indicated in bold).

Genetic region	Primer name	Primer sequence (5'-3')
***LCT* enhancer**	14028C_Fwd	ACGTCATAGTT**C**ATAGAGTGCATAAA
	14028C_Rev	CTTTATGCACTCTAT**G**AACTATGACG
	14011T_Fwd	TAGAGTGCATAAAGA**T**GTAAGTTACCATTTA
	14011T_Rev	TTAAATGGTAACTTAC**A**TCTTTATGCACTCT
	14009G_Fwd	TATAGTGCATAAAGACG**G**AAGTTACCATTTA
	14009G_Rev	TTAAATGGTAACTT**C**CGTCTTTATGCACTCT
	13779C_Fwd	AGTAGTAC**C**AAAGGGCATTCAA
	13779C_Rev	CTTGAATGCCCTTT**G**GTACTAC
	pcr14_5_131bp+ (BamHI)	GGATCCTTTATGTAACTGTTGAATGC
	pcr14_3_149bp- (SalI)	GTCGACTTTCAAAGACGACCTTACAT
***LCT* promoter**	5 pGL3 primer 3	CTAGCAAAATAGGCTGTCCC
	LPH_pro_36C>T_rev	CTCGAGATGTGGAACCCTTACTTTATAACT**A**CAACT
	LPH_pro_17C>T_rev	CTCGAGATGTG**A**AACCCTTACTTTATAACTGC

Sequencing primer	Primer name	Primer sequence (5'-3')
	lucsal	GTCATAAGTGCGGCGACGATAGT
	3pGL3 primer 1	CTTTATGTTTTTGGCGTCTTCCA
	5 pGL3 primer 3	CTAGCAAAATAGGCTGTCCC

2.2.7.3 PCR and gel purification

The PCR reactions to amplify the promoter fragments and the full enhancer (second PCR) were performed in a 25µl volume using the Advantage-GC cDNA PCR kit (Fermentas), which also creates a 5' overhang to the amplified product, important for later cloning. The reactions contained 1x cDNA PCR reaction buffer, 1x dNTP mix (200µM), 1x Advantage-GC cDNA polymerase mix, 0.2µM of each primer (MWG Operon) and about 37ng of the pGL3 hLPH1085-13910*C plasmid (or as template instead 2µl of the annealed enhancer fragment (section 2.2.7.1) or picked bacterial clones for colony PCR (section 2.2.7.5)).

The general procedure for the PCR included an initial denaturation step at 95°C for 1 minute followed by 25 cycles of denaturation (95°C for 30sec), annealing (55°C for 30sec) and elongation (72°C for 1min for enhancer or 2 min for promoter PCR) and a final elongation step for 5 minutes at 72°C.

The PCR products were run on a 1.5% agarose gel (as described in 2.2.4.3) with a marker lane (usually GeneRuler 100 bp, 100 bp Plus or 1 kb DNA Ladder, Fermentas). Bands of the right size were cut out of the gel and purified using the NucleoSpin Extract II kit (Macherey & Nagel) and the protocol for 'DNA extraction from agarose gels'. Briefly, the agarose gel slice and 200µl Buffer NT (high salt buffer) per 100mg gel were incubated in a tube at 50°C and vortexed every 2-3 minutes until the gel piece was dissolved completely. Then the mixture was placed on a prepared NucleoSpin Column and centrifuged for 1 minute at 11000x g to bind the plasmid DNA to the silica membrane of the column. In the following washing step all contaminants were removed. 600µl Buffer NT3 (96-100% ethanol) were added and the column centrifuged at 11000x g for 1 minute, after removing of the flow-through a further centrifugation step for 2 minutes followed. The column membrane was incubated with 15µl Buffer NE (5mM Tris-HCl, pH 8.5) for 1 minute and the DNA eluted in a new tube by centrifuging at 11000x g for 1 minute.

2.2.7.4 Cloning of PCR products

All purified enhancer and promoter fragments were TA-cloned into the pCR 2.1-TOPO plasmid using the TOPO-TA Cloning kit (Invitrogen). The 3'T overhang of the active plasmid vector was ligated to the 5'A of the PCR product with the help of Topoisomerase I, bound to the vector.
4µl of the purified PCR product and 1µl Salt Solution was added to 10ng of the pCR 2.1-TOPO plasmid DNA. After mixing and incubation at room temperature for 5 min, the mixture was transferred on ice.

The transformation into One Shot TOP10 *E.coli* competent cells (genotype: F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*araleu*) 7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G) was performed with 2µl of each TOPO cloning reaction added to a vial containing 25µl cells. After gentle mixing and incubation on ice for 30 minutes the cells were heat shocked for 30 seconds at 42°C to permeabilize the cell membrane for plasmid uptake, and transferred back on ice. 250µl Super Optimal broth with Catabolite Repression (S.O.C.) medium was added to the cells and the tubes placed on a shaking platform (170rpm) to incubate for 1 hour at 37°C.

Each cell solution was then divided and spread on two pre-warmed agar plates (20ml LB agar, Sigma-Aldrich), containing 50µg/ml ampicillin and 40µg/ml 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal, Invitrogen), and incubated at 37°C and 5% CO_2 over night. The plates were screened for blue and white colonies the next day. 5 white colonies

were picked per variant clone and incubated over night (37 °C, 5% CO_2) in 3ml LB medium (Lennox LB Broth, Sigma-Aldrich), containing 50µg/ml ampicillin.

2.2.7.5 Plasmid preparation (Miniprep)

1.5ml of each bacterial culture was used for a miniprep using the GenElute Plasmid Miniprep Kit (Sigma-Aldrich) following the protocol provided. All centrifugation steps were performed at 12000x g if not stated otherwise. Bacterial cells were pelleted for 1 minute and the supernatant discarded. The cells were completely resuspended in 200µl Resuspension solution and 200µl Lysis solution was added. The tube was inverted 6-8 times to mix gently. After the mixture became clear and viscous 350µl Neutralisation solution was added and gently mixed by inverting the tube 4-6 times to precipitate cell debris, proteins, lipids, salts and chromosomal DNA. By centrifuging for 10 minutes the debris was pelleted. The cleared lysate supernatant containing the plasmid DNA was transferred to a prepared binding column and centrifuged at 16000x g for 1 minute. The flow-through was removed and the column washed with 750µl Wash solution under centrifugation for 1 minute to remove contaminants. The flow-through was discarded and the columns dried in a further 1 minute centrifugation step. The purified plasmid DNA was eluted in a new collection tube with the addition of 100µl Elution solution to the column and a centrifugation for 1 minute.

The inserts of the plasmids were checked to be the right sized fragments on agarose gels subsequent to either restriction enzyme digest within the ligation preparation step (see below) or PCR of individual colonies. Agarose gel purification followed as described in section 2.2.7.3.

2.2.7.6 Enzyme digest, ligation and transformation of the pGL3 plasmids

The promoter fragments were inserted into the pGL3 hLPH1085-13910*C plasmid using *Sac*I and *Xho*I digestion and enhancer fragments were inserted into pGL3 hLPH1085 with *Sal*I and *Bam*HI digest. Both pGL3 plasmids needed to be opened prior to ligation and the mutated *LCT* enhancer and promoter fragments cut out of the plasmid via restriction enzyme digestion.
The optimal conditions for the digestion with two restriction enzymes were calculated using Fermentas' DoubleDigest web tool (June 2010, now Thermo Scientific).
Reactions with *Bam*HI and *Sal*I were carried out in a 20µl volume with 10µl 'miniprepped' plasmid DNA containing the mutated *LCT* enhancer, 1U *BamH*I and 2U *Sal*I in 1x Buffer *Bam*HI (all Fermentas) or in a 40µl reaction with ~2µg pGL3 hLPH1085 plasmid DNA and

2U *Bam*HI and 4U*Sal*I in 1x Buffer *Bam*HI respectively. The mixture was incubated at 37°C for at least 1 hour.
Restriction digests with *Sac*I and *Xho*I were carried out in 2 steps. First, 10µl 'miniprepped' plasmid DNA containing the mutated promoter fragments was digested with 2U *Sac*I in a reaction volume of 20µl 1x Tango Buffer or ~2µg of the pGL3 hLPH1085-13910*C with 4U *Sac*I in 40µl 1x Tango Buffer. After incubation for 1 hour at 37°C 1U *Xho*1 was added to the promoter plasmid digest reaction and 2U *Xho*I to the pGL3 hLPH1085-13910*C digest. Both reaction volumes were made up with Tango buffer to a 1x concentration. A further incubation step at 37°C for at least 1 hour followed.

The digested enhancer and promoter and linearised vector fragments were separated on gels and fragments of the right size cut out of the gel and purified as described above. The 5 digestion products (from the 5 different clones) for each mutated enhancer/promoter fragment were gel purified together.
The DNA concentration of the eluate was measured using a NanoDrop Spectrophotometer and was between 10.1 and 17.5ng/µl for the enhancer and promoter fragments and 51ng/µl for pGL3 hLPH1085-13910*C and 20ng/µl for pGL3 hLPH1085 for subsequent ligation and transformation.

A sticky end ligation of enhancer/promoter fragments with the linear vector DNA was performed with the help of T4 DNA ligase (Fermentas) in a total volume of 20µl 1x T4 DNA Ligase Buffer (Fermentas). The reactions were carried out with a 5:1 ratio of the enhancer/promoter fragment DNA over the linear vector DNA with 1U T4 DNA ligase. After 10 minutes incubation at room temperature the plasmids were either used for transformation or frozen at -20°C.

5µl of the plasmid ligation mix was transformed into 40µl (1:4 diluted) XL1-Blue *E.coli* competent cells (genotype: *recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F´ *proAB lacI*^{q}*ZΔM15* Tn*10* (Tet^{r})], Stratagene) and these spread on agar plates following the same protocol as for cloning of PCR products into TOP10 cells above (2.2.7.4).

As many as possible colonies were picked and checked for successful insertion of the right sized fragments by colony PCR (section 2.2.7.3). All clones containing inserts were chosen and grown in 3ml LB medium and 'miniprepped' as above or stored in 50% glycerol cultures at -80°C. Gel purified plasmid DNA was sent out for sequencing using the Eurofins

MWG Operon sequencing service. Sequences were verified in both directions using sequencing primers flanking the inserted regions (see Table 2.5).

2.2.7.7 Plasmid preparation (Maxiprep)

One clone for each mutated plasmid, with the right sequence, was chosen and 50µl of the bacterial glycerol culture was grown in 200ml LB medium (containing 1µl/ml ampicillin) and incubated on a shaking platform (170rpm) for 16 hours (37°C, 5% CO_2). Maxipreps were performed using the NucleoBond Xtra plasmid purification kit (Macherey & Nagel) following the corresponding protocol 'High-copy plasmid purification (Midi, Maxi)'.
The cells from each culture were transferred to a centrifuge flask and harvested in a centrifugation step at 6000x g and 4°C for 15 minutes. After discarding the supernatant, the cell pellet was completely re-suspended in 10ml re-suspension Buffer RES, containing RNase A to degrade RNA in the solution. 10ml lysis Buffer LYS (containing NaOH and SDS) was added to denature proteins, chromosomal and plasmid DNA. The solution was gently mixed and incubated for 5 minutes at room temperature. To precipitate salts, proteins, chromosomal DNA and put the plasmid DNA back into its supercoiled state in the solution, 10ml neutralisation Buffer NEU (containing potassium acetate) was added. Following gentle mixing, the lysate was applied to the inserted filter of the NucleoBond Xtra column, prepared with equilibration buffer (20ml Buffer EQU). The column was then washed by adding 10ml Buffer EQU to the filter rim and the filter containing unwanted contaminants discarded. The column containing silica resin was washed again with 20ml Buffer WASH and the plasmid DNA eluted from the column in 10ml elution Buffer ELU, collected in a 15ml falcon tube. The plasmid DNA was precipitated with 7ml room-temperature 100% isopropanol. The isopropanol-eluate mix was vortexed, incubated for 2 minutes and centrifuged at 15000x g for 30 minutes. The supernatant was carefully discarded and the pellet left to air dry for 10 minutes. Finally the pellet was re-dissolved in 200µl TE buffer. The DNA concentration of the mutated and control luciferase reporter plasmids was photometrically measured and was between 1.2 and 2.5µg/µl.

2.2.7.8 Transfections into Caco-2 cells

Caco-2 cells were grown to 80% confluence and plated in 24-well plates the day before transfection, with each well containing about $5x10^4$ Caco-2 cells. The medium was changed with 1ml fresh full culture medium before transfection.
Each transfection experiment was carried out in 4 or 8 fold repeats. A DNA mix was prepared for each well with 0.2µg luciferase reporter gene plasmid, 0.1µg pCMV-lacZ plasmid (for β-galactosidase expression, Promega), 0.9µg pBluescript SK^+ plasmid (to

adjust the total DNA amount to 0.3µg, Stratagene) in a 25µl volume containing 15mM NaCl. 25µl of transfection reagent containing 2µM polyethyleneimine (PEI) in form of Exgen 500 (22kDa, Fermentas) in 120mM NaCl solution or PEI25 (25kDa, Alfa Aesar) were added to the DNA mix and incubated at room temperature for 1 hour.
The mixture was slowly added to the cells and the plates centrifuged for 5 minutes at 1200rpm to settle the reagent on the cell surface. The plates were incubated at 37°C and 5% CO_2. The medium was changed the first day after transfection and later when necessary.

2.2.7.9 Reporter gene assay analysis

After 2 and 9 days of transfection the Caco-2 cells were harvested at an undifferentiated and differentiated stage respectively and luciferase and β-galactosidase activity were measured using the Dual-Light System chemiluminescent reporter gene assay (ABI, Bronstein et al. 1997; Martin et al. 1996) following the 'luciferase and β-galactosidase detection protocol'.
The wells with the Caco-2 cell cultures were rinsed with 1ml room temperature PBS, which was completely removed by inverting the plate onto a paper towel and pipetting the remaining PBS from the side of the well. 130µl Lysis solution (containing 0.5mM DTT) was added to each well to cover the cells and after incubation for 5 minutes on a shaking platform the plates were placed on ice. 10µl of the cell extract was transferred to a luminometer vial. 25µl Buffer A was added to the bottom of the vial and it was placed in the luminometer immediately afterwards. 100µl Buffer B (containing luciferin and 1:100 Tropix Galacton-Plus substrate) was automatically added by the luminometer and the luciferase signal read for 5 seconds after a 2 second delay. The signal from the β-galactosidase reaction is negligible at that stage due to the lack of turnover time and an enhancer and a non optimal pH. The tubes were further incubated at room temperature for 30-60 minutes and 100µl Accelerator-II was added to raise the pH of the solution and provide a luminescence enhancer for β-galactosidase. After automatically adding of the Accelarator-II solution by the luminometer the luciferase signal was read again with a 2 second delay for 5 seconds.

Luciferase activity was normalised against β-galactosidase activity by calculating relative luciferase/β-galactosidase ratios for each well and significance between the transfection results tested with a student unpaired *t* test in Excel.

2.3 Statistical methods/Bioinformatic analyses

2.3.1 Deviation from Hardy-Weinberg equilibrium (HWE)

In a randomly mating population the expected genotype proportions from a determined allele frequency should follow the Hardy-Weinberg-theorem as they remain constant over generations in a stable population. The HWE is described by the formula: $p^2+2pq+q^2=1$, where p represents the major allele frequency and the minor allele frequency can be calculated by $q=1-p$.
Deviations from HWE might be observed due to factors like non-random mating, population stratification, presence of selection (though unlikely to be detected in small sample sets), deletions and duplications (CNV), or even technical issues with collecting and analysing the data in the lab. In this project, tests for deviation from HWE were mainly used to check data for errors caused by allelic dropout or sampling problems.
All genotype data from enhancer and haplotype markers were tested for deviation from HWE with the software Arlequin (version 3.5.1.3, Excoffier and Lischer 2010), which uses a kind of Fisher's exact test for each locus separately (Guo and Thompson 1992).

2.3.2 Haplotype inference

As all samples typed for this project are unrelated individuals, a statistical approach is needed to reconstruct haplotypes. The software PHASE (Stephens and Donnelly 2003; Stephens et al. 2001) applies a Bayesian statistical method to infer haplotypes from population genotype data. The programme uses a Markov chain-Monte Carlo algorithm, called Gibbs sampling, to calculate an expected pattern of haplotype distribution taking into account observed information of the dataset. Known haplotypes, occurring in homozygous individuals for example are fully informative and can be used for the calculation of unresolved haplotype pairs and those that are even then not resolvable are assumed more likely to include ones most similar to known haplotypes. The haplotypes for each individual are then reconstructed as most likely haplotypes estimated from the calculated posterior distributions by constructing a Markov chain.
Version 2.1.1 of the PHASE package (Stephens and Scheet 2005), which was used in this study, can additionally take into account order and genetic distances between the genotyped markers and the resulting estimated decay of linkage disequilibrium. The PHASE output was checked by eye for possible anomalies. Default settings were used for most of the analysis, only for extended haplotype analysis (chapter 5) the number of iterations was reduced to 50.

2.3.3 Haplotype networks

Haplotype information analysed by PHASE was used to construct a haplotype network with the software Network (version 4.6.1.1, http://www.fluxus-engineering.com). Allelic composition of each haplotype occurring at more than 4 chromosomes and haplotype frequencies were entered into the program and manually rooted to an ancestral haplotype (ancestral alleles inferred as described above, 2.2.4.6). The median-joining algorithm (Bandelt et al. 1999) was used to calculate the networks. It can deal with all types of data and constructs rooted networks with the shortest possible trees and reduces unnecessary cross links.

2.3.4 Linkage disequilibrium

Haplotype information was also used to estimate the level of linkage disequilibrium between alleles, which describes the non-random association between them (see also chapter 1). Pairwise linkage disequilibrium D between two alleles at different loci (A and B) of a haplotype is calculated with the formula $D=p_{AB}-(p_{A}*p_{B})$, in subtracting the observed haplotype frequency from this expected under linkage equilibrium when $p_{AB}=p_{A}*p_{B}$ (whereas p is the frequency of one of the alleles at A and B). To be able to compare the values of two loci, the D parameter, which is highly dependent on allele frequencies, needs to be normalised. In dividing D by the theoretical maximum for the allele frequencies observed, D' can be calculated as proposed by Lewontin (1964). The values for D' lay between one (complete LD) and zero, values in between indicate a disruption of LD. Linkage disequilibrium was calculated using D' with PowerMarker (version 3.25, Liu and Muse 2005) and results were also checked in DnaSP (version 5.10.01, Librado and Rozas 2009). Both programs apply a chi-square test to assess the significance of associations between loci, whereas the Bonferroni corrected significance levels DnaSP calculates were used to mark significant values in the tables.

2.3.5 Linkage disequilibrium unit (LDU) maps

A high resolution LDU map for chromosome 2 was constructed with the help of Winston Lau using LDMAP software (Lau et al. 2007) with a method that was initially developed by Maniatis et al. (2002) and extended by Morton et al. (2007). The method is based on the Malecot model which describes the isolation by distance of two SNPs. The probability (ρ) of an association between two SNPs is given by the modified Malecot equation (Morton et al. 2001):

$$\rho = (1 - L)Me^{-\sum \varepsilon_i d_i} + L,$$

where L is the association level at large distance (LD asymptote), M the association level at no distance (initial value before LD decays) and ε_i the decline of association which happens exponentially with physical distance d_i, measured in kb between the ith pair of SNPs. LDU distances between SNPs are given by $\varepsilon_i d_i$ that is equivalent to θt, the product of recombination θ and t the time in generations this recombination has accumulated after population bottleneck(s). In summary εd provides a more useful measure for LD as it is more accurately known than θt. LDU can also reflect effects of population bottleneck events that can cause differences in LD patterns between populations but it is also influenced by evolutionary forces such as selection and mutation (Tapper 2007).
LD units for chromosome 2 were calculated by Winston Lau and values extracted for the region of interest. LDU were plotted as LDU/kb and LDU/Mb against distance (kb) in Excel.

2.3.6 Test for extended haplotype homozygosity (EHH)

Long range haplotype tests such as EHH tests use the principle that an allele under positive selection rises quickly in frequency and creates a region of extended homozygosity, as time would have been too short to break down LD of the region (details in chapter 1). The EHH test was applied in chapter 5 to compare haplotype data over a region of about 1.8 Mb, using Sweep software (version 1.1, Sabeti et al. 2002). EHH data for selected cores was exported in order to plot graphs of EHH decay using R, with a modified code from Pascale Gerbault.

2.3.7 Genetic distances between populations (F_{ST})

F_{ST} is a measure of genetic distances between subpopulations of a meta-population with divergence in time. It measures the deviation between the observed and expected heterozygote frequencies under the Hardy-Weinberg theorem. It takes into account that in a metapopulation the heterozygosity level is reduced with the divergence of subpopulations and in the subpopulations an excess of homozygotes appears, which is known as the Wahlund effect (Jobling et al. 2013). F_{ST} describes the differences between the average expected heterozygosity of sub-populations (H_S) and the expected total heterozygosity of the meta-population (H_T): $F_{ST}=(H_T-H_S)/H_T$. When high gene flow and little differentiation between the subpopulations are apparent, it is close to zero. It is close to one when the genetic diversity of the metapopulation is much higher than in any of the subpopulations, due to their high differentiation.

For comparison between two populations pairwise F_{ST} is calculated as $F_{ST}=V_p/p(1-p)$ with p and V being mean and variance of allele frequencies. Pairwise F_{ST} values were calculated with the Arlequin software (version 3.5.1.3, Excoffier and Lischer 2010).

2.3.8 Principal co-ordinates analysis (PCO)

The genetic distances between populations calculated as pairwise F_{ST} values were graphically represented in principal co-ordinates (PCO) plots (Gower 1966), in which multiple dimensions are reduced to be displayed on axes known as principal co-ordinates or eigenvectors.
These axes are extracted one after another to capture as much of the remaining variation as possible, with each axis being independent and orthogonal to the previous one. Plotted are the first two axes with the highest proportion of total variation that they account for. PCO plots were constructed using the R statistical environment (version 2.6.2., http://www.r-project.org/) using a code written by Christopher Plaster to convert the calculated pairwise F_{ST} values.

2.3.9 Diversity measures

Population diversity analyses were performed using DnaSP (version 5.10.01, Librado and Rozas 2009).
To prepare input files for these programs it was necessary to convert the haplotypes inferred by PHASE (and checked visually) to a full length haplotype sequences. Human reference sequences were downloaded as described in section 2.2.4.6. and manually edited and converted into FASTA format in BioEdit (version 7.0.9.0, Hall 1999) to only include the readable sequence length (559 bp) for the *LCT* enhancer region (chapter 3) and a combined sequence (1401 bp) of intron 4, enhancer and hapdef region (chapter 5). FASTA sequences for all haplotypes were created by assigning the nucleotides of a particular haplotype to their position on the FASTA reference sequences in using a Java code written by Maxim Scheremetjew. Individuals were assigned to their two haplotype sequences in using the most probable haplotype combination for each individual calculated by PHASE and connecting them to the haplotype FASTA sequences with Python using a code written by Mirna Kovacevic.

Haplotype diversity *H* (or gene diversity, heterozygosity) was measured using Nei's *H* (Nei 1987), which calculates the probability that two randomly taken sequences from a population are different from each other with the formula: $H = \frac{1}{n-1}\left(1 - \sum_{i=1}^{K} x^2 i\right)$.

Where n are the numbers of chromosomes (sample size), *K* are the number of different haplotypes and *x* the relative frequency of haplotype *i*. Values for *H* lie between zero and one, from no diversity to the most.

Another diversity measure calculated was the **nucleotide diversity π** (Nei, 1987). It is the probability that two randomly chosen sequences from a population differ at a certain site or simply the average number of nucleotide differences per site: $\pi = \frac{1}{n-1}\sum_{ij}^{K} x_i x_j d_{ij}$.

Where n is the numbers of chromosomes (sample size), *K* is the number of different haplotypes, x_i and x_j are the respective frequencies of the *i*th and *j*th sequences, d_{ij} is the number of nucleotide differences per nucleotide site between sequences *i* and *j*.
Both formulas calculate values that are corrected for sample size.

2.3.10 Tests of neutrality

Several statistics were applied to compare the observed diversity patterns with these expected under neutral evolution (see also chapter 1). Neutrality tests were conducted with DnaSP using phased haplotype data.

Tajima's *D* (Tajima 1989) calculates the genetic diversity within a population in comparing each pair of sequences in taking into account the number of segregating sites (S) and the average number of nucleotide diversity (π), and compares that to the expected level of diversity under neutrality. Significant negative Tajima's *D* values indicate positive selection or population growth and positive values could indicate population subdivision or balancing selection (Jobling et al. 2013). Tajimas' *D* should be zero under neutrality. *P*-values for of the significance of the Tajima'*D* values deviating from zero are also calculated by DnaSP.

Fu and Li's *D and *F*** statistics (Fu and Li 1993) were also applied, which test for an excess of rare variants in the populations. The *D** test measures the proportion of singletons compared to the total number of mutations in a population. The *F** statistics additionally compares the number of singletons to the average number of nucleotide differences between sequence pairs in a population. DnaSP indicates the statistical significance of both values in taking the critical values obtained by a two tailed test (Fu and Li 1993).

2.3.11 GenoPheno

To be able to compare the predicted phenotype data for the different populations in chapter 3, estimated from frequencies of lactase persistence associated alleles, with published lactase persistence phenotype data of matching populations, the GenoPheno test (Mulcare et al. 2004) was applied. GenoPheno, version 1.00, is an R script written by Dr. Mike Weale and it is accessible via the TCGA website (http://www.ucl.ac.uk/tcga/software/). This method takes sampling errors for genotype and phenotype data (summarised in Mulcare et al. 2004) into account and statistically tests the discrepancy of predicted and observed phenotype data.
To briefly describe the method, the predicted frequency of lactose digesters is calculated from the combined frequencies of the associated alleles $(p^2+2p(1-p))$ and corrected for different kinds of phenotyping error. Then a value for the number of lactose digesters is simulated from a Binomial distribution taking into account the numbers of predicted digesters and observed digesters. These steps are repeated 100,000 times to get a Monte Carlo sampling distribution of the numbers of observed digesters under the null hypothesis that genotype data and phenotyping error alone account for the predicted lactose digester frequencies. From this simulated distribution a two-tailed *p*-value is obtained.

To create the Old-World genotype-phenotype correlation map the approach of Itan et al. (2010) was followed. The program, written in the Python programming language, uses the statistical GenoPheno method described above and combines that with surface interpolation to cover areas with missing data points. Information about genotype data collection points by the group or taken from literature were converted into geographic co-ordinates in using Google Earth (http://www.google.com/earth/index.html). The genotype-phenotype correlation maps were plotted with the help of Pascale Gerbault and Yuval Itan using the modified Python code written by Yuval Itan (Itan et al. 2010).

2.3.12 Fisher's exact test

The significance of associations between loci of intron 2 of *LCT* and the lactase persistence phenotype (chapter 6) was tested with Fishers exact test. A 2x2 contingency table was generated with the observed allele counts at a locus and the existing phenotype data.
The test calculates the differences between the observed and expected values under the null hypothesis that values of rows and columns are not associated. The Graphpad online

software (http://graphpad.com/quickcalcs/contingency1.cfm) was used to calculate two tailed *p*-values for the Fishers exact tests.

2.3.13 Prediction of transcription factor binding sites

The bioinformatic search for potential transcription factor binding sites was done with several different software programs that use different mathematical algorithms to compare the input sequence to partly different libraries of known regulatory motifs. All results were checked for relevance in filtering for tissue and cell type specificity where possible.

The **MATCH** tool (Kel et al. 2003) uses the TRANSFAC database (TRANSFAC Professional version 12.1, BIOBASE Biological Databases) (Matys et al. 2006). With this software it is possible to compare DNA sequences with a library of mononucleotide weight matrices, matching possible transcription factors. The search can be additionally specified by species or tissue profiles. For the analyses done for this project, a vertebrate matrix was chosen with a cut-off value that reduces false positive and false negative results. The quality of a match is defined by the core similarity score (core match) and the matrix similarity score (matrix match). The threshold for the matrix score/match was set to 80% to avoid false negative results.

The freely online available tool **TFSEARCH** (version 1.3, http://www.cbrc.jp/research/db/TFSEARCH.html) also uses the TRANSFAC database, but with a slightly different algorithm to match the database entries. No information about the exact method was available. The program was used for comparison purposes.

MatInspector (version 8.0.5) accesses its own database (Cartharius et al. 2005). The software uses a similar scoring scheme as MATCH to compare a given sequence with the position weight matrices of the transcription factor binding sites of the database but additionally works with family matrices. The program calculates a matrix similarity score of a certain sequence to the matching matrices of the library consisting of information about the position weight matrix, a conservation profile and a core region for a transcription factor. The matrices of the program are also stored as family matrices to minimise redundant matches.

2.3.14 Further statistical tests or methods

Spearman Rank correlation tests were performed in Excel to analyse correlations between allele frequencies and geographic locations of populations. Therefore the distances between the assigned geographic coordinates for the population and 10°W longitude and 80°N latitude were calculated and converted in km using a converter online program (http://www.movable-type.co.uk/scripts/latlong.html).

Mann-Whitney U tests were performed online (http://elegans.som.vcu.edu/~leon/stats/utest.cgi). Mean lengths of haplotypes carrying derived and ancestral alleles were calculated in bp and tested with a two tailed test.

The geographic map showing the lactase persistence frequencies in chapter 1 was plotted with data from Itan et al. (Itan et al. 2010) by Mark Thomas using a Python scripting. The other geographic map showing enhancer alleles distribution in chapter 3 was generated in R (version 3.0.1) using a modified code written by Pascale Gerbault.

2.4 Web resources

Affymetrix: http://www.affymetrix.com/estore/

Arlequin: http://cmpg.unibe.ch/software/arlequin35/

BioEdit: http://www.mbio.ncsu.edu/bioedit/bioedit.html

ChromasPro: http://technelysium.com.au/

Centre for Comparative genomics: http://www.ucl.ac.uk/gee/research/CCG

CEPH: http://www.cephb.fr/en/hgdp/diversity.php

DnaSP: http://www.ub.edu/dnasp/

DoubleDigest: http://www.thermoscientificbio.com/webtools/doubledigest/

Ensembl: http://www.ensembl.org/index.html

Ethnographic Atlas: http://eclectic.ss.uci.edu/~drwhite/worldcul/atlas.htm

Ethnologue: http://www.ethnologue.com/

FastPCR: http://primerdigital.com/fastpcr.html

GLAD database: http://www.ucl.ac.uk/mace-lab/resources/glad

Google Earth: http://www.google.com/earth/index.html

Graphpad: http://www.graphpad.com/

HapMap: http://hapmap.ncbi.nlm.nih.gov/

MatInspector: http://www.genomatix.de/solutions/genomatix-software-suite.html

NCBI: http://www.ncbi.nlm.nih.gov/

Phase: http://stephenslab.uchicago.edu/software.html

Powermarker: http://statgen.ncsu.edu/powermarker/

R: http://www.r-project.org/

Sweep: http://www.broadinstitute.org/mpg/sweep/

SNPviewer: http://www.lgcgenomics.com/software

TCGA: http://www.ucl.ac.uk/tcga/software/

TFSEARCH: http://www.cbrc.jp/research/db/TFSEARCH.html

The intestinal transcription factor target database:
http://gastro.sund.ku.dk/chipchip/

TRANSFAC: http://www.biobase-international.com/product/transcription-factor-binding-sites

UCSC: http://genome.ucsc.edu/

UniProt: http://www.uniprot.org/

Vector NTI: http://www.lifetechnologies.com/uk/en/home/life-science/cloning/vector-nti-software.html

2.5 Laboratory chemicals and equipment

2.5.1 Labmade solutions

All pH values are at 25°C.

2/3 Microclean (HM-MC): 26.7% Polyethylene glycol (PEG 8000), 0.7M NaCl, 1.3mM Tris-HCl (pH 7.5), 0.13mM EDTA (pH 8.0), 2.3mM $MgCl_2$

TBE buffer (1x): 89mM Tris-HCl (pH8.0), 89mM boric acid, 2mM EDTA (pH 8.2-8.4)

TE buffer (1x): 10mM Tris-HCl (pH 8.0), 1mM EDTA (pH 8.0)

Agarose gel loading buffer: 15% Ficoll (PM400) in H_2O, traces (about 0.25%) of Xylene cyanol FF (Sigma) and Bromophenol blue (Bio-Rad)

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

Stretch marker: Agarose gel loading buffer containing about 50ng 200, 400 and 800 bp PCR products

Dialysis buffer: 20mM HEPES (pH 7.9), 20% glycerol, 1.5mM $MgCl_2$, 100mM KCl, 0.2mM EDTA, 0.5mM dithiothreitol (DTT)

EMSA loading buffer: 0.2% Bromophenol blue, 10% glycerol, 0.5xTBE buffer

Gel shift buffer: 25mM Tris-HCl (pH 7.8), 5mM $MgCl_2$, 6mM KCl, 0.5mM EDTA, 1mM DTT, 1µl/ml protease inhibitor cocktail (Sigma), 5%Ficoll (PM 400), 2.5% glycerol

High-salt buffer: 20mM HEPES (pH 7.9), 25% glycerol, 1.5mM $MgCl_2$, 1.2M KCl, 0.2mM EDTA, 0.5mM DTT, 1µl/ml protease inhibitor cocktail

Hypotonic buffer: 10mM HEPES (pH 7.9), 1.5mM $MgCl_2$, 10mM KCl, 0.2mM PMSF, 0.5mM DTT

Low-salt buffer: 20mM HEPES (pH 7.9), 25% glycerol, 1.5mM $MgCl_2$, 0.2M KCl, 0.2mM EDTA, 0.5mM DTT, 1µl/ml protease inhibitor cocktail

2.5.2 Commercial solutions

All pH values are at 25°C

LB Agar: 9.14g/L tryptone, 4.57g/L yeast extract, 4.57g/L NaCl, 13.72g/L agar, 1.6g/L inert tableting aids

LB medium: 10g/L enzymatic digest of casein, 5g/L yeast extract (low sodium), 5g/L sodium chloride, 2g/L inert agents, 22g/L total solids

Advantage-GC cDNA PCR Kit

*Advantage-GC cDNA Polymerase Mix, recombinant***:** KlenTaq-1 DNA polymerase and TaqStart Antibody (1.1µg/µl) in 1x buffer: 1.0%Glycerol, 0.8mM Tris-HCl (pH 7.5), 1.0mM

KCl, 0.5mM $(NH_4)_2SO_4$, 2.0 µM EDTA, 0.1mM β-mercaptoethanol, 0.005%Thesit, Deep $Vent_R$ as minor component

*1x GC cDNA PCR reaction buffer***:** 40mM Tricine-KOH (pH 9.2 at 25°C), 15mM KOAc, 3.5mM Mg(OAc), 2, 5% Dimethyl Sulfoxide (DMSO), 3.75 µg/ml µg/ml Bovine serum albumin

dNTP mix (1x): (10mM each of dATP, dCTP, dGTP, and dTTP; 1x concentration: 0.2mM each)

TOPO-TA Cloning kit

pCR 2.1-TOPO, 10 ng/µl plasmid DNA in: 50% glycerol, 50mM Tris-HCl (pH 7.4), 1mM EDTA, 1mM DTT, 0.1% Triton X-100, 100µg/ml BSA, phenol red

Salt Solution: 1.2M NaCl, 0.06M MgCl2

S.O.C. medium: 2% Tryptone, 0.5% Yeast Extract, 10mM NaCl, 2.5mM KCl, 10mM $MgCl_2$, 10mM MgSO4, 20mM glucose

Restriction enzymes and buffers

BamHI in 1x buffer: 10mM Tris-HCl (pH 7.4 at 25°C), 200mM NaCl, 1mM DTT, 0.1mM EDTA, 0.15% Triton X-100, 0.2mg/mL BSA, and 50% (v/v) glycerol

Buffer BamHI (1x): 10mM Tris-HCl (pH 8.0 at 37°C), 5mM $MgCl_2$, 100mM KCl, 0.02% Triton X-100, and 0.1mg/mL BSA

Pfu DNA polymerase, recombinant, in 1x buffer: 20mM Tris-HCl (pH 8.2), 1mM DTT, 0.1mM EDTA, 100mM KCl, 0.1% (v/v) Nonidet P40, 0.1% (v/v) Tween 20, and 50% (v/v) glycerol

Pfu Buffer with $MgSO_4$ (10x): 200mM Tris-HCl (pH 8.8), 100mM $(NH_4)_2SO_4$, 100mM KCl, 1mg/mL BSA, 1% (v/v) Triton X-100, 20mM $MgSO_4$.

T4 DNA ligase in 1x buffer: 20mM Tris-HCl (pH 7.5), 50mM KCl, 1mM DTT, 0.1mM EDTA and 50% (v/v) glycerol

*T4 DNA Ligase Buffer 10x:*400mM Tris-HCl, 100mM $MgCl_2$, 100mM DTT, 5mM ATP (pH 7.8)

T4 Polynucleotide Kinase in 1x buffer: 20mM Tris-HCl (pH7.5), 25mM KCl, 0.1mM EDTA, 2mM DTT and 50%(v/v) glycerol

Kinase reaction buffer A (10x): 500mM Tris-HCl (pH7.6), 100mM $MgCl_2$, 50mM DTT, 1mM spermidine

Taq DNA polymerase in 1x buffer: 100mM KCl, 20mM Tris-HCl, pH 8.0, 0.1mM EDTA, 1mM DTT, 0.5% Tween 20, 0.5% Nonidet P40, 50% (v/v) glycerol

Reaction Buffer IV (10x): 750mM Tris-HCl (pH 8.8), 200mM $(NH_4)_2SO_4$, 0.1% (v/v) Tween 20

SacI, SalI and XhoI in 1x buffer: 10mM Tris-HCl (pH 7.4), 100mM KCl, 1mM DTT, 1mM EDTA, 0.2mg/mL BSA, 50% (v/v) glycerol

Tango Buffer (1x): 33mM Tris-acetate (pH 7.9 at 37°C), 10mM Mg-acetate, 66mM K-acetate, 0.1mg/mL BSA

2.5.3 Equipment

Centrifuges, general

ALC Centrifuge PK120

Biofuge Pico, Heraeus

Capsule Tomy HF-120

Centrifuge 5430, Eppendorf

MSE-Europa 24M, SANJO

MSE MISTRAL 3000g, SANJO

Sigma 1-14

EMSA

Ole Dich Inmicrocentifuge 157 MP

SE 250Mighty small II Electrophoresis, Hoefer Scientifis Instr.

Multitemp II, Thermostatic circulator, Pharmacia LKB

SlabGEl Dryer SGD4050, Savant

GelPump GP100, Savant

Storage Phosphor Screen 20x25, Molecular dynamics, Amersham Biosciences

Image Eraser, Molecular dynamics, Amersham Biosciences

Storm Phosphorimager 820

Series 900 radiation mini-monitor G-M (Coderum) tube, Morgan Mini Instruments Ltd

Microbiological work

Centrifuges: Jouan CR 312, Sigma 3-18K, Sorovall SS-34

Heraeus UB 20 incubator

Lauda E100 Ecoline waterbath, VWR

Teche Hybridiser HB-ID

Berthold Lumat LB9501 Tube Luminometer

Other equipment

ABI GeneAmp PCR System 9700

ABI Veriti 96 well Thermal Cycler

GeneQuant II photometer, Pharma Biotech, Printer EPSON P-40

BioPhotometer 6131, Eppendorf

NanoDrop 1000 Spectrophotometer and -software, version 3.8.1

Luckham 4RT Rocking platform

G:BOX EF2, Syngene, Software GeneSnap and GeneTools, Syngene

UVP UV Transilluminator

2.5.4 Suppliers

AH diagnostics as (Stratagene): Runetoften 18, 8210 Aarhus V, DK

Autogen Bioclear UK Ltd: Holly Ditch Farm, Mile Elm, Calne SN11 0PY, UK

BioGenex: 49026 Milmont Drive, Fremont, California 94538, USA.

BioNordika Denmark A/S: Marielundvej 48, 2730 Herlev, DK

Bio-Rad Laboratories: Symbion Science Park, Fruebjergvej 3, 2100 København ø, DK

CareFusion Health UK 232 Ltd (Micro Medical Ltd): The Crescent, Jays Close, Basingstoke RG22 4BS, UK

Eurofins MWG Operon: Anzinger Str. 7a, 85560 Ebersberg, Germany

Fisher Scientific (Fermentas, Lonza Bioscience, Macherey-Nagel): Postboks 60, Industrivej 3, 3550 Slangerup, DK

Fisher Scientific UK Ltd: Bishop Meadow Road, Loughborough LE11 5RG, UK

GE Healthcare Europe GmbH (Amersham Biosciences): Park Allé 295, 2605 Brondby DK

Johnson Matthey A/S (Alfa Aesar): Frederikssundsvej 274 D, 2700 Bronshoj, DK

LGC Genomics Ltd (KBioscience): Units 1 & 2, Trident Industrial Estate, Pindar Road, Hoddesdon EN11 0WZ, UK

Life Technologies Ltd (Applied Biosystems): 3 Fountain Drive, Inchinnan Business Park, Paisley PA4 9RF, UK

Life Technologies Europe BV (Invitrogen): filial Danmark, PO Box 37, 2850 Naerum, DK

Medicell International Ltd: 239 Liverpool Road, London N1 1LX, UK

Perkin Elmer: Tonsbakken 16-18, Skovlunde 2740, DK

QIAGEN (Gentra): Skelton House, Lloyd Street North, Manchester M15 6SH, UK

Santa Cruz Biotechnology Inc: 2145 Delaware Avenue, Santa Cruz, California 95060, USA

Sigma-Aldrich Company Ltd: The Old Brickyard, New Road, Gillingham SP8 4XT, UK

Thermo Fisher Scientific (previous ABgene): Abgene House Blenheim Road, Epsom KT19 9AP, UK

VWR: Bie & Berntsen A/S: Transformervej 8, 2730 Herlev, DK

VWR International Ltd: Hunter Boulevard, Magna Park, Lutterworth LE17 4XN, U

3 Geographic distribution of LCT enhancer variation in Eurasian populations

3.1 Introduction

As described in chapter 1, the continued expression of lactase into adulthood is enabled by a mechanism *cis*-acting to *LCT* (Wang et al. 1995). A relevant regulatory region was found in intron 13 of the upstream gene *MCM6* that showed influence on *LCT* promoter activity with a transcriptional enhancing effect (Olds and Sibley 2003; Troelsen et al. 2003).

Within this sequence region, the allele *-13910*T* was shown to nearly completely associate with lactase persistence, in Northern Europe (Enattah et al. 2002; Enattah et al. 2007; Poulter et al. 2003). -13910C>T genotypes were also associated with a trimodal lactase expression pattern in a study on Finns, as one would expect from a *cis*-acting element (Kuokkanen et al. 2003). This general pattern was also seen by Poulter et al. (2003), although the pattern was not so clear-cut. This allele is also present at considerable frequencies in groups of European descent in America or African groups with a known history of admixture with Europeans or Asians (for example Friedrich et al. 2012; Myles et al. 2005).

The absence of *-13910*T* in sub-Saharan African and Middle Eastern populations where the lactase persistence phenotype is common (Ingram et al. 2007; Mulcare et al. 2004) led to further investigations involving re-sequencing the *LCT* enhancer region, which revealed more variation. Some of these alleles were too rare to assess for association but three were shown to associate with lactase persistence, for example - *13915*G,* which is frequent in the Middle East and East Africa (Enattah et al. 2008; Imtiaz et al. 2007; Ingram et al. 2007; Ingram et al. 2009b; Tishkoff et al. 2007). This allele is associated with lactase persistence in populations from Saudi Arabia (Imtiaz et al. 2007), Kenya (Tishkoff et al. 2007), in the Jaali of Sudan (Ingram et al. 2007) and Somali of Ethiopia (Ingram et al. 2009b).
In the same Somali, another allele, *-13907*G,* is also associated with lactase persistence (Ingram et al. 2009b) and was shown to be common in several East African populations (Tishkoff et al. 2007).

About 100 bp upstream and occurring at high frequencies in several populations from Kenya and Tanzania another allele, *-14010*C* showed association with lactase persistence too (Tishkoff et al. 2007). It is also relatively common in the South Africa Xhosa (Torniainen et al. 2009). A further allele, *-14009*G,* was associated in the Somali but did not reach statistical significance (Ingram et al. 2009b).

3.2 Chapter aims

Analysis of the published data on the frequencies of the four confirmed associated alleles did not fully explain the distribution of the lactase persistence phenotype.
The approach by Itan et al. (2010), mentioned in chapter 1 (see also Figure 1.7), illustrates this: The geographic mapping of the difference in phenotype distribution predicted from genotype information of all four alleles with published observed lactase persistence phenotype data showed that there are many geographic regions where the lactase persistence phenotype is underrepresented by the genetic data. Whether caused by yet undiscovered variants, data errors or simply due to a lack of sampling points, as the map shows interpolated values, could not be clarified. Of these regions, Eastern Europe and parts of West, Central and South Asia, are of special interest for this thesis.

Although much was known about the worldwide occurrence of *-13910*T* at the onset of this thesis, and even though reasoned to be causal for lactase persistence in Europe, it could not completely explain the phenotype distribution in parts of South and East Europe (Anagnostou et al. 2009; Itan et al. 2010). These regions were not well covered with information about enhancer SNP data and furthermore, if reported at all, the -13910 C>T SNP had been mainly studied via RFLP or other genotyping technologies. Only a few studies provided data from fully sequenced enhancer regions before the beginning of this thesis (Enattah et al. 2008; Imtiaz et al. 2007; Ingram et al. 2007; Ingram et al. 2009b).

Two further examples suggested that the information about lactase persistence alleles in Europe was not complete. One variant, -13914 G>A had been reported in two individuals, from Germany and Austria (Tag et al. 2008; Tag et al. 2007) and four family members from Northern Russia with partly Polish ancestry (Khabarova et al. 2010). Another European individual showed a chromosome with high lactase expression (Poulter et al. 2003) and carrying a new enhancer variant, *-14028*C*, but none of the others (Ingram 2008). The occurrence of this variant in particular, in other individuals, would be of great interest as it showed direct evidence of function.

The three other reported lactase persistence associated alleles *-13907*G, -13915*G* and *-14010*C* had mainly been described in African and Middle Eastern populations (Enattah et al. 2008; Imtiaz et al. 2007; Ingram et al. 2007; Ingram et al. 2009b; Tishkoff et al. 2007) and not much was known about their distribution elsewhere. Also with respect to other recently discovered variants (for example Ingram et al. 2009b; Tag et al. 2007) further investigations on Eurasian and Middle Eastern populations seemed to be necessary to complete the picture of the distribution of known *LCT* enhancer alleles and to identify new variants. It was thus of interest to obtain complete enhancer sequence data in these areas. Any novel variants could be subjected to genotype-phenotype association study if at sufficient frequency or could be tested *in vitro* for function.

This chapter aims to describe the variation of enhancer alleles and its distribution in Europe, the Middle East and other parts of Asia. In summary, the main questions to answer are:

- What is the distribution of the European *-13910*T* allele and what is the edge of this distribution in southern and southeastern parts of Europe and in the Middle East?
- Can other lactase persistence associated or derived alleles be found in the enhancer region of *LCT* in the available samples and what is their distribution?
- Can the distribution of lactase persistence associated *LCT* enhancer alleles now explain the phenotypic distribution of reported groups in the same geographic areas?

3.3 Sequencing strategy

To examine the variation in the *LCT* enhancer present in populations from all over Europe, the Middle East and other parts of Asia, DNA samples were sequenced with the classic Sanger method and analysed (see section 2.2.4). Several samples previously genotyped by our group for -13910 C>T (Mulcare 2006) or haplotype markers (Hollox et al. 2001; Poulter et al. 2003) were also re-analysed via sequencing to ascertain the whole enhancer. Of these, samples from India and Southeast Asia were sequenced as part of an undergraduate project (Ghosh 2010).

A combination of four primers successfully applied in previous studies (Ingram et al. 2007; Ingram et al. 2009b) was first used to amplify the 940 bp region in intron 13 of *MCM6* including the *LCT* enhancer (Chr2:136608194-136609133, GRCh37/hg19), and sequences were readable from bp position -14331 until -13489 from the start of transcription of *LCT*.

Depending on the quality of the DNA it was sometimes necessary to amplify two shorter overlapping fragments to get sufficient amounts of products. Details about primer sequences can be found in Table 2.3 of chapter 2.
During the later part of this project the focus was more on the smaller region 3' in *MCM6* (Chr2:136608194-136608899) which was typed for all samples. Of this 706 bp long PCR product, it was possible to analyse the enhancer sequences of most of the samples for a 559 bp region from position -14040 to -13482 from the start of transcription of *LCT*.

3.4 Grouping strategy

The definition of groups for population genetic analyses is always to some extent artificial as it depends on the information available. Nevertheless it is essential to aim as far as possible for uniformity of groups in the analyses. For this project, the individual samples were first grouped according to the geographic region, namely Northwest/Central Europe, South Europe, East/Southeast Europe, the Middle East, West Asia, Central/South Asia and Central/East/Southeast Asia, and the country they were collected. Samples that could not be allocated to one country of that geographic region were classified as 'other'. These samples were not included in the geographic maps (see below). A further differentiation into distinct ethnic groups with a minimum sample size of 10 followed, using self declared ethnic background if such information was available or in sub-groupings within countries with further information about the sample localization. Otherwise they were grouped by country only.

3.5 Variation of the *LCT* enhancer

The *LCT* enhancer region in intron 13 of *MCM6* was successfully sequenced for 2056 individuals from 52 populations within this project. A total of 25 SNPs was identified, of which 10 occurred more than once. Table 3.1 shows these enhancer variants and the allele frequencies of non-singleton SNPs for each population. 9 new variants were identified (-14145 T>G, -14062 G>A, -14010 G>A, -13964 C>A, -13926 A>C, -13950 A>G, -13771 A>G, -13744 C>G, -13693 G>A) which are mainly singletons. They were confirmed by sequencing of both strands of the products of at least two independent PCRs and the ancestral state of these SNPs was determined by sequence comparison with other primate species, which was the same as the common allele in humans (see also chapter 2, section 2.2.4.6). Sequencing chromatograms of some of the new variants are shown in Figure 3.1.

All SNPs were checked for deviation from HWE using the Arlequin software (see also chapter 2, section 2.3.1). A few significant deviations were observed but after Bonferroni

correction for multiple testing (51 tests) only one, -13910C>T in the Yakuts (p=0.0006) remained significant. An underrepresentation of heterozygotes at this position was observed which could be the result of population substructure within the group sampled. After reassessment of the sampling locations (see also chapter 2, section 2.1.1.5) the group was divided into Northern Yakuts and Yakuts which left both groups in HWE.

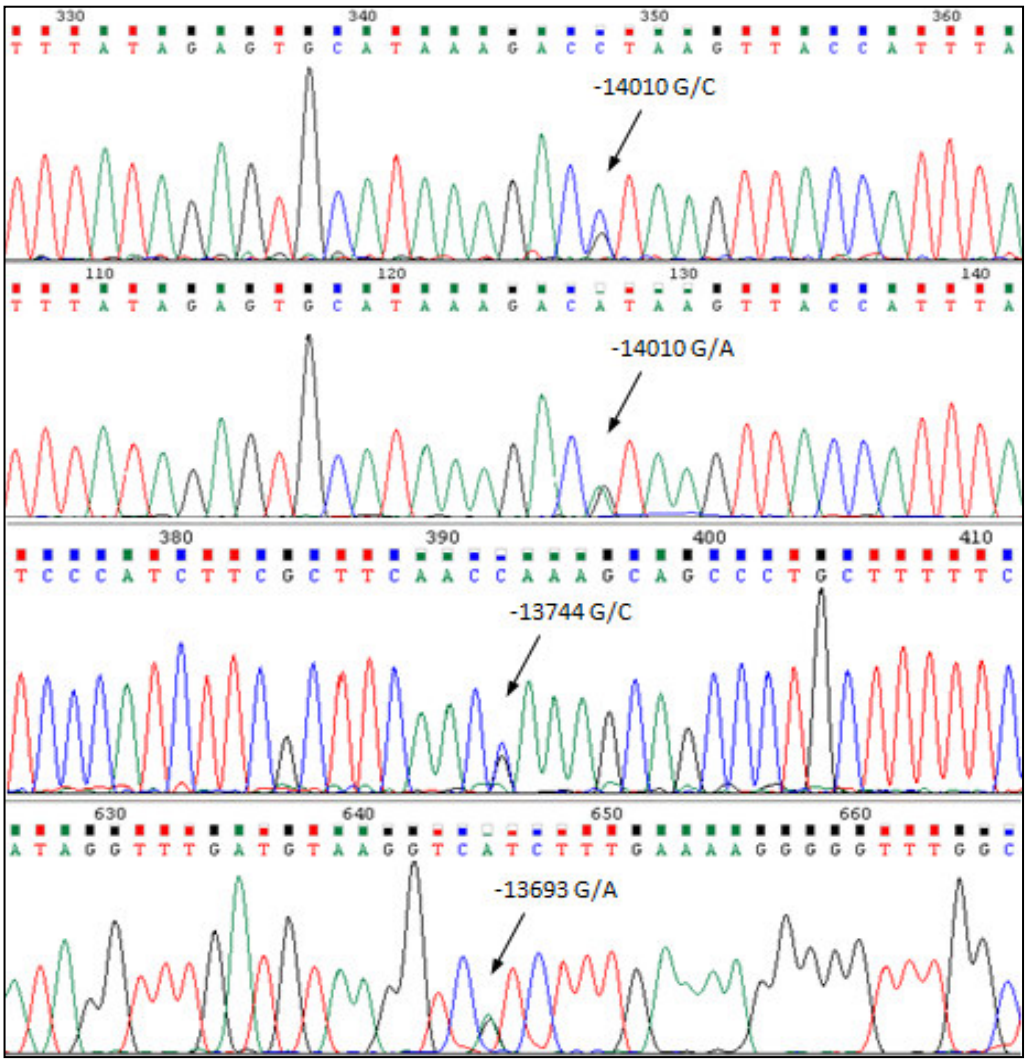


Figure 3.1: Sequence chromatograms showing examples of individuals heterozygous at various positions of the *LCT* enhancer. Note that the second chromatogram shows a third allele at the 14010 position which is known to be a G>C SNP (first chromatogram).

Table 3.1: *LCT* enhancer variation in the samples sequenced from 52 European and Asian population groups. Allele frequencies for all variants occurring more than once and the presence of singletons are indicated. N: number of chromosomes, data marked by * was not included in the geographic map (Figure 3.2).

Region	Country	Population	N	-14011 C>T	-13915 T>G	-13910 C>T	-13907 C>G	-13906 T>A	-13779 G>C	-13744 C>G	-13730 T>G	-13603 C>T	-13495 C>T	Singletons
Northwest/	Germany	**Germans**	60	-	-	0.550	-	-	-	-	-	-	0.617	
Central		**Sorbs**	64	-	-	0.500	-	-	-	-	-	-	0.617	
Europe	Ireland	**Irish**	68	-	-	0.956	-	-	-	-	-	-	0.969	
	Netherlands	**Frisians**	58	-	-	0.810	-	-	-	-	-	-	0.845	
	Norway	**Norwegians**	88	-	-	0.841	-	-	-	-	-	-	0.875	
	Slovakia	**Roma**	64	-	-	0.063	-	-	-	-	-	-	0.400	
	Sweden	**Sami**	60	0.033	-	0.293	-	-	-	-	-	-	0.750	
		Swedes	74	-	-	0.778	-	-	-	-	-	-	0.906	
	UK	**Ashkenazi-Jews**	38	-	-	0.105	-	-	-	-	-	-	0.158	-13914 G>A
		English	102	-	-	0.745	-	-	-	-	-	-	0.784	-14062 G>A, - 14028 T>C
	Other*	**other C/NW Europeans**	68	-	-	0.691	-	-	-	-	-	-	0.727	
South	Greece	**Greeks**	120	0.008	-	0.150	-	-	-	-	0.008	-	0.306	
Europe	Italy	**Tyroleans Bozen**	80	0.013	-	0.150	-	-	-	-	-	-	0.368	
		Tyroleans Gadertal	76	-	-	0.421	-	-	-	-	-	-	0.514	
		Tyroleans Vinschgau	102	-	-	0.363	-	-	-	-	-	-	0.521	
	Portugal	**Portuguese**	96	-	-	0.438	-	-	-	-	-	-	0.486	
	Spain	**Catalans**	58	-	-	0.259	-	-	-	-	-	-	0.364	
		Spanish	62	-	-	0.403	-	-	-	-	-	-	0.500	
East/	Belarus	**Belarusians**	100	0.010	-	0.230	-	-	-	-	-	-	0.477	
Southeast	Macedonia	**Macedonians**	100	-	-	0.240	-	-	-	-	-	-	0.398	
Europe	Romania	**Romanians**	118	-	-	0.169	-	-	-	-	-	-	0.278	-13753 C>T
	Russia	**Erzya**	42	-	-	0.143	-	-	-	-	-	-	0.333	
		Moksha	32	0.031	-	0.406	-	-	-	-	-	-	0.531	
		Russians Perm	46	-	-	0.239	-	-	-	-	-	-	0.395	
	Ukraine	**Ukrainians**	74	0.014	-	0.257	-	-	-	-	-	-	0.379	
	Other*	**Other E/SE Europeans**	38	-	-	0.289	-	-	-	-	-	-	0.444	
Middle East	Cyprus	**Greek Cypriots**	120	-	-	0.008	-	-	-	-	-	-	0.269	

	Iran	**Iranians**	154	0.006	0.006	0.032	-	-	-	-	-	-	0.234	
	Kuwait	**Kuwaiti**	66	-	0.250	0.030	-	-	-	-	0.015	0.015	0.141	-14009 T>G, -13693 G>A
	Syria	**Syrians**	140	-	0.029	0.021	-	-	0.007	-	-	-	0.231	-14145 T>G
	Turkey	**Anatolian-Turks**	116	-	-	0.078	-	-	-	-	-	-	0.225	
	Yemen	**Yemeni Hadramaut**	166	-	0.241	0.018	0.012	-	0.030	0.006	0.006	-	0.210	-14010 G>C, -13806 A>G
		Yemeni Sena	68	-	0.294	0.147	-	-	-	0.059	0.029	0.059	0.294	
	Other*	**Other Middle Easterns**	62	-	0.048	0.065	0.016	-	-	-	-	-	0.345	
West Asia	Armenia	**Armenians**	102	-	-	0.176	-	-	-	-	-	-	0.333	
	Azerbaijan	**Azeri**	80	-	-	0.013	-	-	-	-	-	-	0.171	
	Georgia	**Georgians**	108	-	-	0.075	-	-	-	-	-	-	0.293	
Central/	Afghanistan	**Pashtuns/Afghans**	32	-	-	0.063	-	-	-	-	-	-	0.375	
South Asia		**Tadjiks**	28	-	-	0.250	-	-	-	-	-	-	0.464	-13964 C>A
		Uzbeks	54	-	-	0.241	-	-	-	-	-	-	0.463	
	India	**North Indians**	120	-	-	0.200	-	-	-	-	-	-	0.410	
		South Indians	102	-	-	0.118	-	-	0.029	-	-	-	0.450	
	Nepal	**Nepalese**	38	-	-	0.342	-	0.026	-	-	-	-	0.526	
		Tharu	80	-	-	0.175	-	-	-	-	-	-	0.321	
	Uzbekistan	**Uzbeks**	76	-	-	0.039	-	-	-	-	-	-	0.269	
	Other*	**Other C/South Asians**	32	-	-	0.188	-	-	-	-	-	-	0.500	
Central/East/	Mongolia	**Khalka**	114	0.009	-	0.026	-	-	-	-	-	-	0.413	
Southeast		**Mongols**	52	-	-	0.038	-	-	-	-	-	-	0.417	
Asia	Russia	**Northern Yakuts**	22	-	-	0.136	-	-	-	-	-	-	0.500	
		Yakuts	110	-	-	0.055	-	-	-	-	-	-	0.443	
	Singapore	**Han-Chinese**	98	-	-	0.041	-	0.010	-	-	-	-	0.382	-13771 A>G
		Japanese	84	-	-	-	-	-	-	-	-	-	0.333	-14010 G>A, -13926 A>C

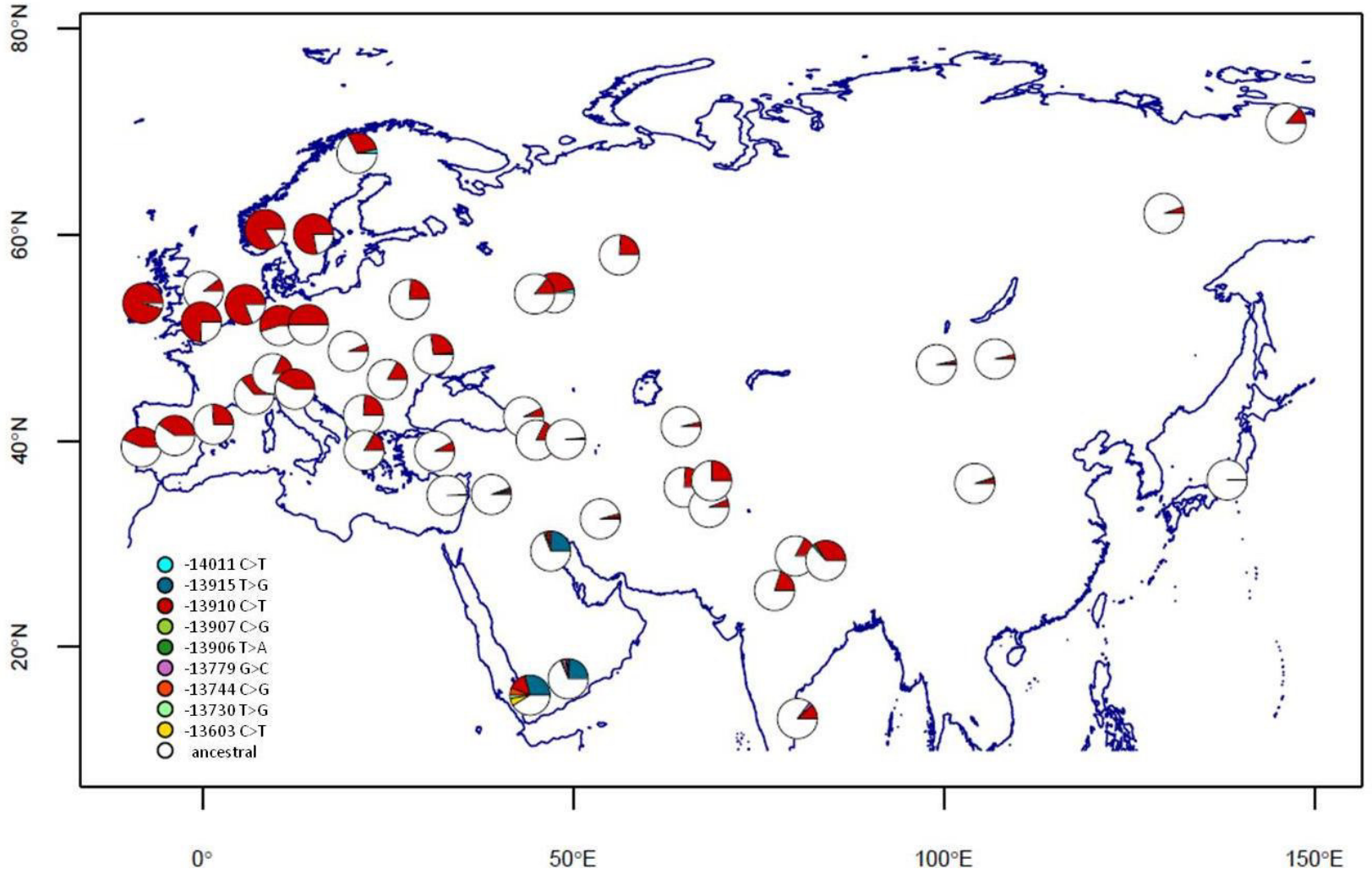


Figure 3.2: Geographic distribution of the common *LCT* enhancer variants (also shown in Table 3.1) in the 52 studied populations (-13495 C>T is not included).

Figure 3.2 shows the geographic distribution of the common alleles within the (likely) functional region of the *LCT* enhancer in the populations studied and illustrates once more the predominance of *-13910*T* as main enhancer variant in the Europe. *-13910*T* was found in all populations tested except of the Japanese (Table 3.1). The results also broadly confirm the decline in allele frequency of *-13910*T* to the South and East of Europe. It is highest in the Irish and Norwegian populations with 96% and 84% respectively and lowest in the Tyroleans from Bozen and in Greeks (both at 15%).

Noticeable different frequencies of *-13910*T* in Northwest and Central Europe are observed in the Sami, the Ashkenazi-Jewish and Roma groups where the frequency of the **T* allele is comparatively low. The Sami are an indigenous population in the far North of Europe and are traditional reindeer herders, subsisting mainly from meat and fishing (Leonard and Crawford 2002). The Ashkenazi-Jewish and Roma have history of migration and were sampled in regions geographically distant from the living places of many of their recent ancestors. The recent Eastern European history of many of the Ashkenazi Jews and their complex history of migrations as close communities accounts for the overall genetic differentiation from the majority of the ethnic English population. The Roma originated in India and are likewise known to form a distinct ethnic-cultural minority mainly in Southeast and Middle Europe.

Differences in *-13910*T* frequencies of very closely located groups appear in the Spanish and the Catalans with a *-13910*T* frequency of 40% and 26% respectively and the Tyroleans between the German and Ladin speaking groups from Vinschgau and Gadertal and the Italian speaking population from Bozen with 36%, 42% and 15% *-13910*T* frequency respectively. The Catalans and Tyroleans are culturally and partly geographically quite isolated populations, which could explain these differences probably caused by decreased gene flow with neighbouring groups. The low *-13910*T* frequency of the Italian speaking Tyroleans, similar to other Italian regions from further South, reflects their recent immigration history into the Alps within the last century (Parteli 1988). The German Tyroleans were shown to be genetically influenced by the Ladins and both being highly genetically differentiated (Pichler et al. 2006).

The SNP -13495 C>T was polymorphic in all populations tested and has previously shown by members of the lab to be is too frequent to be causal of lactase persistence (unpublished data). This variant resides outside the functional enhancer examined by

Troelsen et al. (2003) and it was shown to be non-functional in reporter gene assays as it did not increase activity compared to the ancestral enhancer (Tishkoff et al. 2007).

Table 3.2 shows the distribution of the common enhancer alleles by geographic region. The previously observed pattern of a Northwest-Southeast decline in *-13910*T* frequency (Swallow 2003) is slightly disrupted across Asia. The prevalence of *-13910*T* is higher in Central and South Asia than Western Asia including the Middle East. Populations from Central and South Asia including Afghans, Indians and Nepalese have a similar *-13910*T* frequency to populations from East and Southeast Europe, for example the Macedonians and Belarus. The Georgians and Azeri of the West Asian Caucasus region have very low *-13910*T* frequencies comparable to those of their Middle Eastern neighbours.

Table 3.2: Frequency of the common enhancer alleles by geographic region (-13495 C>T included).

Region	N	-14011 C>T	-13915 T>G	-13910 C>T	-13907 C>G	-13906 T>A	-13779 G>C	-13744 C>G	-13730 T>G	-13603 C>T	-13495 C>T
NW/C Eur.	740	0.003	-	0.615	-	-	-	-	-	-	0.688
S Europe	594	0.003	-	0.305	-	-	-	-	0.002	-	0.359
E/SE Eur.	550	0.005	-	0.231	-	-	-	-	-	-	0.358
Middle East	892	0.001	0.094	0.041	0.003	-	0.007	0.006	0.004	0.006	0.219
W Asia	288	-	-	0.094	-	-	-	-	-	-	0.247
C/S Asia	562	-	-	0.167	-	0.002	0.005	-	-	-	0.370
C/E/SE Asia	480	0.002	-	0.038	-	0.002	-	-	-	-	0.360

*-13945*T* broadly follows the *-13910*T* distribution pattern. It is also particularly high in Northern European populations, even in the Sami as shown in Figure 3.3. To evaluate statistically whether the two alleles correlate with each other and follow the same Northwest to Southeast decline previously suggested for *-13910*T*, in the 48 geographically located populations of this study, Spearman Rank correlation tests were performed. The frequencies of the two markers correlate strongly with each other (Rho = 0.75, $p<0.01$). A significant negative correlation between *-13910*T* and *-13945*T* and longitude (distance calculated in km from 10°W) was observed which was less strong for *-13945*T* (Rho = -0.39, $p<0.01$) than for *-13910*T* (Rho = -0.65, $p<0.01$). The frequency for both alleles also significantly decreases with distance from latitude 80°N (in km). The negative correlation was slightly stronger for *-13945*T* (Rho = -0.52, $p<0.01$) than for *-13910*T* (Rho = -0.45, $p<0.01$). It should be noted that these results are consistent with the

finding that *-13945*T* forms part of the haplotype on which *-13910*T* arose as described in section 3.6 below.
The other 3 lactase persistence associated alleles (*-14010*C, -13915*G* and *-13907*G*) were only found in Middle Eastern populations. *-13915*G* is at high frequencies in Kuwait (25%) and Yemen (Hadramaut: 24%, Sena: 29%). *-14010*C* was rare, only occurring once in the Yemen Hadramaut, as well as *-13907*G,* found in two individuals from Hadramaut and in one other Middle Eastern sample. The two other alleles*-14028*C* and *-13914*A*, previously found in Europeans (Ingram 2008; Tag et al. 2008; Tag et al. 2007) occurred in one English individual and in an English person of Ashkenazi-Jewish background respectively.

Noticeable is the occurrence of another allele *-14011*T* previously described in populations from India, Estonia and Brazil (Friedrich et al. 2012; Gallego Romero et al. 2012; Lember et al. 2006). Although this allele was rare in any of the Eurasian groups tested, it was found scattered across Europe in individuals from Belorussia, Greece, Tyrol, Mordovia, Ukraine, Iran, Mongolia respectively and two Sami individuals and one other individual of mixed ancestry. It additionally occurred in one Saudi Arabian Bedouin sample and one from the phenotyped Italian cohort sequenced for chapter 5.

*-13779**C was identified in India, Syria and Yemen. This allele has been reported so far in the literature in one Somali individual who was diagnosed as a lactose maldigester (Ingram et al. 2009b) but is quite frequent in populations from India as shown in our collaborative project (Gallego Romero et al. 2012). Although it might not be associated with lactase persistence in these populations, its frequency in a milk drinking country and its location in the enhancer sequence part that shows a DNAse footprint (see chapter 4) could indicate some functional role.

An interesting case is an additional allelic variant at nucleotide position -14010. A Japanese individual carried an A at this position, the same location as the more frequent G>C SNP.
Another new variant is *-13744*G*, which was found in the Yemeni from Sena (4 individuals) and additionally in one individual of the Yemeni of Hadramaut. It occurs on a haplotype together with the lactase persistence allele *-13915*G* as shown in section 3.6 below.

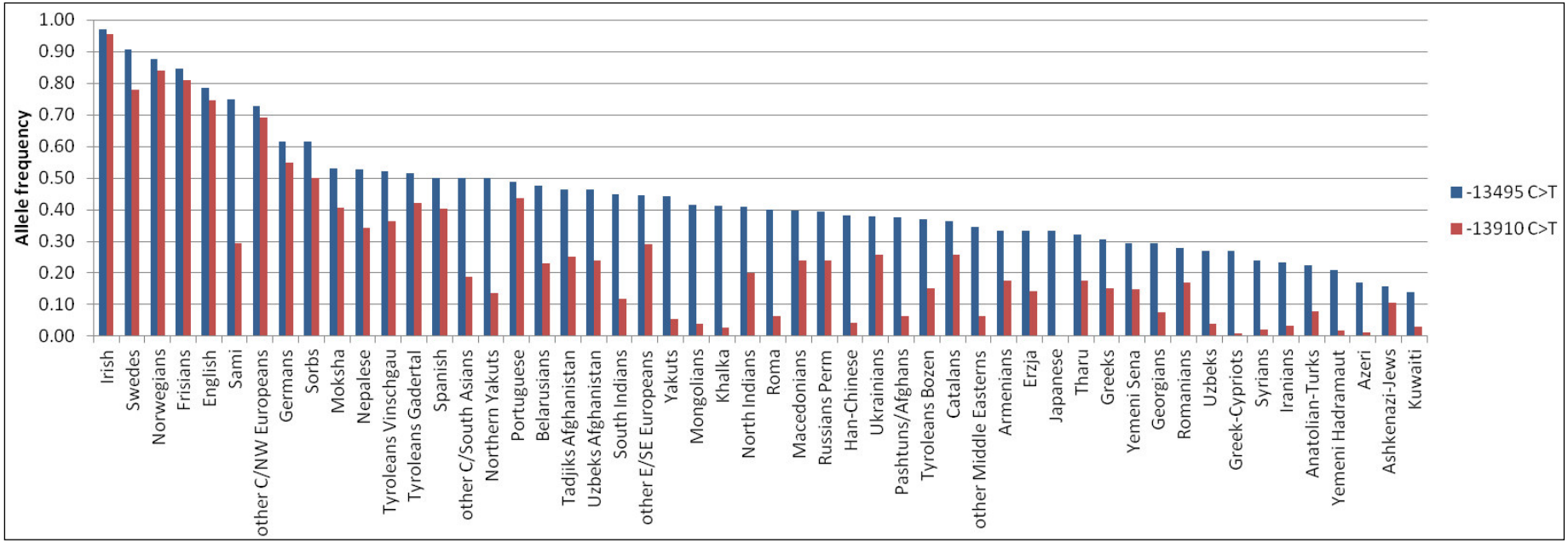


Figure 3.3: Allele frequency of *-13945*T* and *-13910*T* for the 52 populations tested, shown in order of *-13495*T* frequency.

3.6 Relationship of *-13495*T* to other enhancer alleles

To analyse the relationship of *-13495*T* to other enhancer alleles haplotype inference was done using PHASE. All samples with incomplete data at the -13945 nucleotide position were excluded which left 1888 samples for analysis. *-13495*T* occurred on the same haplotypes as four other enhancer variants: *-14011*T, 13910*T, -13907*G, -13603*T* and was also present on its own as a derived allele on the ancestral enhancer haplotype (Figure 3.4, haplotype no. 2). Chapter 5 will investigate this further in examining *-13495*T* association with the *LCT* core haplotypes.

Haplotype ID	-14028 T>C	-14011 C>T	-14010 G>C	-14009 T>G	-13915 T>G	-13910 C>T	-13907 C>G	-13906 T>A	-13806 A>G	-13779 G>C	-13744 C>G	-13730 T>G	-13603 C>T	-13495 C>T	N
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2079
9	.	.	.	.	.	T	.	.	.	.	.	.	.	T	860
2	.	.	.	.	.	.	.	.	.	.	.	.	.	T	716
10	.	.	.	.	G	.	.	.	.	.	.	.	.	.	77
5	.	.	.	.	.	.	.	.	.	C	.	.	.	.	9
15	.	T	.	.	.	.	.	.	.	.	.	.	.	T	9
3	.	.	.	.	.	.	.	.	.	.	.	.	T	T	5
4	.	.	.	.	.	.	.	.	.	.	.	G	.	.	5
11	.	.	.	.	G	.	.	.	.	.	G	.	.	.	5
7	.	.	.	.	.	.	G	.	.	.	.	.	.	T	3
6	.	.	.	.	.	.	.	A	.	.	.	.	.	.	2
8	.	.	.	.	.	T	.	.	.	.	.	.	.	.	2
12	.	.	.	.	.	.	.	.	G	.	.	.	.	.	1
13	.	.	.	G	.	.	.	.	.	.	.	.	.	.	1
14	.	.	C	.	.	.	.	.	.	.	.	.	.	.	1
16	C	.	.	.	.	.	.	.	.	.	.	.	.	.	1

Figure 3.4: Haplotype analysis of the *LCT* enhancer region, N: number of chromosomes.

3.7 Population differentiation across the *LCT* enhancer

To compare the genetic distance between the groups, pairwise F_{ST} values were calculated in Arlequin with the enhancer genotype data of the 52 populations studied. Not all populations are significantly different in pairwise F_{ST} when only taking the enhancer region into account (for a full results table, see Appendix D). The -13495 C>T SNP could not be included in this analysis because of too much missing data (9.6%). The F_{ST} values for inter-population differences were used to perform a Principal co-ordinate analysis

(PCO) in R and results are illustrated as PCO plot in Figure 3.5. Over 97% of the population differences are captured in the two first components. The first co-ordinate seems to be highly influenced by *-13910*T* as the Northwest European populations cluster on the right half of the plot and all other populations closely together. The Greek Cypriots are the furthest outstanding on the left side as they have the lowest **T* allele frequencies and none of the other rare derived enhancer alleles. The two Yemeni populations (51 and 52) can be seen at the upper edge of the point cloud, which may be explained by the numerous other enhancer alleles present in these groups.

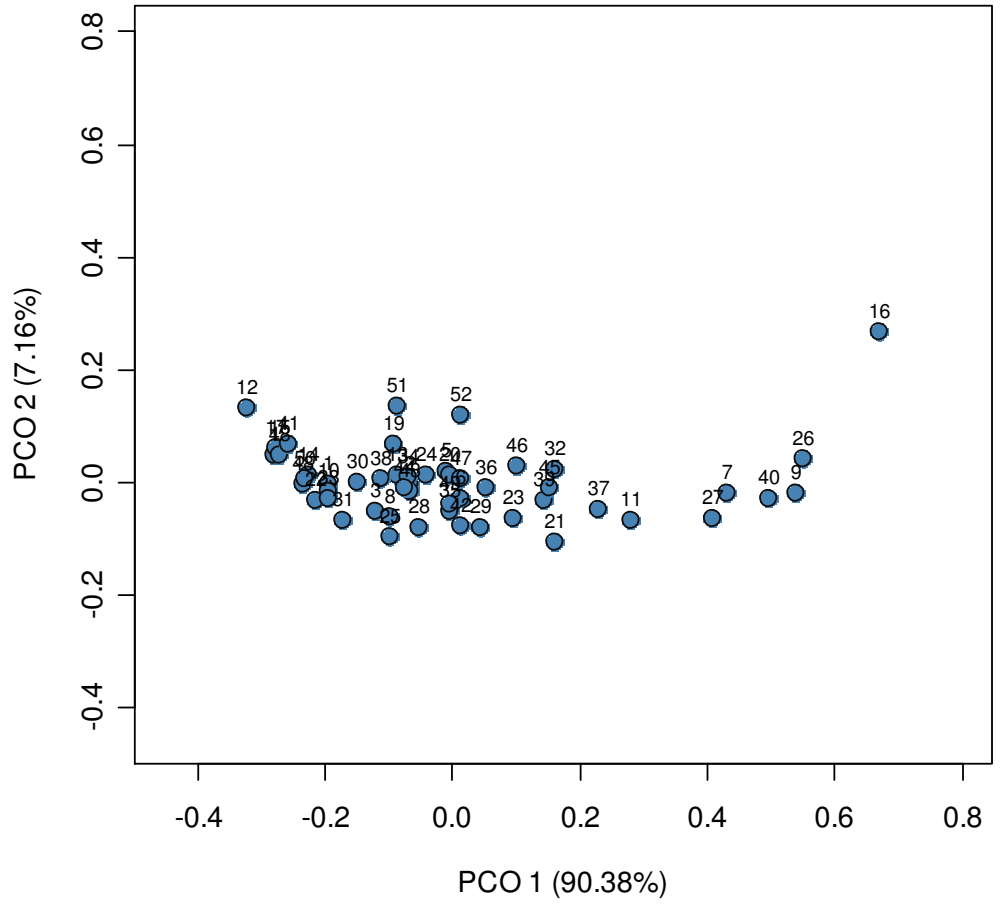


Figure 3.5: Principal co-ordinates plot of genetic distances between populations from pairwise F_{ST} values calculated from *LCT* enhancer genotype data. Populations included: 1:Anatolian-Turks, 2: Armenians, 3: Ashkenazi-Jews, 4: Azeri, 5: Belarusians, 6: Catalans, 7: English, 8: Erzja, 9: Frisians, 10: Georgians, 11: Germans, 12: Greek Cypriots, 13: Greeks, 14: Han-Chinese, 15: Iranians, 16: Irish, 17: Japanese, 18: Khalka, 19: Kuwaiti, 20: Macedonians, 21: Moksha, 22: Mongolians, 23: Nepalese, 24: North Indians, 25: Northern Yakuts, 26: Norwegians, 27: other C/NW Europeans, 28: other C/South Asians, 29: other E/SE Europeans, 30: other ME, 31: Pashtuns/Afghans, 32: Portuguese, 33: Roma, 34: Romanians, 35: Russians Perm, 36: Sami, 37: Sorbs, 38: South Indians, 39: Spanish, 40: Swedes, 41: Syrians, 42: Tadjiks Afghanistan, 43: Tharu, 44: Tyroleans Bozen, 45: Tyroleans Gadertal, 46: Tyroleans Vinschgau, 47: Ukrainians, 48: Uzbeks, 49: Uzbeks Afghanistan, 50: Yakuts, 51: Yemeni Hadramaut, 52: Yemeni Sena.

3.8 Tests of neutrality and diversity measures

Haplotypes of the enhancer region as inferred by PHASE were used with the DnaSP software to calculate haplotype and nucleotide diversity for each population. Several neutrality tests were also conducted to detect a possible excess of rare alleles, which might signify a selective sweep. Table 3.3 shows the values calculated for each population. The only significant deviation from neutrality was seen for Tajima's *D* in the case of the German and Ladin speaking groups and the Portuguese and Spanish. These showed a positive value, which is usually taken as a sign of balancing selection (alleles of equal frequency). However when excluding the -13495 C>T SNP from analysis none of them remained significant (data not shown).

3.9 Lactase persistence genotype-phenotype correlation

The frequencies of the lactase persistence associated alleles in the populations tested from Europe and Asia give an idea about how the lactase persistence phenotype is distributed. To see whether the predicted lactase persistence frequency from genotype data for the populations examined fit with observed published phenotype frequencies a GenoPheno test (Mulcare et al. 2004) was applied (for details see 1.3.10). Phenotype data from matching population samples were extracted from the 'Global Lactase persistence Association Database' (GLAD, http://www.ucl.ac.uk/mace-lab/resources/glad) compiled from different sources (Itan et al. 2010). Groups were selected as matching when they were from the same self-declared cultural identity living in the same or a directly neighbouring country.

Table 3.4 shows that for most of the groups studied, the frequencies of lactase persistence predicted from genotype data correlates well with published phenotype data. Significant differences were however observed in the Roma, English, Greek Cypriots and in Asia in the Uzbeks from Afghanistan, North Indians and Japanese, where the predicted frequencies were either higher or lower than that those that were reported from phenotypic testing. However, taking multiple testing into account (22 tests) only the Uzbeks, North Indians, and Greek Cypriots remain significant, with the more stringent *p*-value threshold of 0.002. These differences may in some cases result from sampling of slightly different populations for genotype and phenotype data and in others may reflect as yet undiscovered persistence alleles.

Table 3.3: Diversity and neutrality measures for all 52 tested populations. Significant values ($p<0.05$) are shaded, N: number of chromosomes.

Region	Population	N	Haplotype diversity (Neis's *H*) (±SD)	Nucleotide diversity per site (π)	Tajima's *D*	Fu and Li's *D**	Fu and Li's *F**
Northwest/ Central Europe	Germans	60	0.555 (0.001)	0.002	2.213	0.729	1.365
	Sorbs	64	0.603 (0.001)	0.002	2.235	0.723	1.370
	Irish	68	0.086 (0.002)	0.000	-1.091	0.718	0.201
	Frisians	58	0.324 (0.005)	0.001	0.588	0.733	0.802
	Norwegians	88	0.279 (0.003)	0.001	0.381	0.696	0.700
	Roma	64	0.482 (0.003)	0.001	0.574	0.723	0.790
	Sami	60	0.689 (0.001)	0.002	0.762	0.873	0.979
	Swedes	74	0.383 (0.004)	0.001	0.816	0.710	0.865
	Ashkenazi-Jews	38	0.284 (0.008)	0.001	-0.038	0.777	0.627
	English	102	0.405 (0.003)	0.001	0.519	-0.590	-0.282
	other C/NW Europeans	68	0.450 (0.003)	0.002	1.690	0.718	1.177
South Europe	Greeks	120	0.474 (0.002)	0.001	-0.101	-1.534	-1.259
	Tyroleans Bozen	80	0.526 (0.003)	0.001	0.427	-0.534	-0.274
	Tyroleans Gadertal	76	0.585 (0.001)	0.002	2.357	0.708	1.409
	Tyroleans Vinschgau	102	0.613 (0.001)	0.002	2.336	0.684	1.396
	Portuguese	96	0.561 (0.000)	0.002	2.440	0.689	1.435
	Catalans	58	0.525 (0.003)	0.002	1.707	0.733	1.190
	Spanish	62	0.559 (0.001)	0.002	2.268	0.726	1.383
East/ Southeast Europe	Belarusians	100	0.636 (0.001)	0.002	0.940	-0.586	-0.123
	Macedonians	100	0.553 (0.002)	0.002	1.842	0.685	1.219
	Romanians	118	0.444 (0.002)	0.001	1.259	0.673	1.002
	Erzya	42	0.511 (0.005)	0.001	0.968	0.766	0.956
	Moksha	32	0.625 (0.002)	0.002	1.023	-0.283	0.110
	Russians Perm	46	0.590 (0.004)	0.002	1.616	0.756	1.170
	Ukrainians	74	0.529 (0.003)	0.002	0.828	-0.515	-0.115
	Other E/SE Europeans	38	0.602 (0.003)	0.002	1.827	0.777	1.251
Middle East	Greek Cypriots	120	0.365 (0.002)	0.001	0.017	-1.113	-0.892
	Iranians	154	0.386 (0.002)	0.001	-0.682	-1.617	-1.548
	Kuwaiti	66	0.591 (0.003)	0.001	-0.953	-1.549	-1.595
	Syrians	140	0.402 (0.002)	0.001	-0.708	-0.337	-0.541
	Anatolian Turks	116	0.340 (0.003)	0.001	0.357	0.674	0.674
	Yemeni Hadramaut	166	0.661 (0.001)	0.002	-1.059	-1.943	-1.942
	Yemeni Sena	68	0.773 (0.001)	0.002	0.244	1.151	1.008
	Other Middle Easterns	62	0.548 (0.003)	0.001	-0.405	-0.153	-0.271
West Asia	Armenians	102	0.505 (0.002)	0.001	1.423	0.684	1.067
	Azeri	80	0.279 (0.003)	0.001	-0.411	-1.017	-0.971
	Georgians	108	0.394 (0.003)	0.001	0.505	0.680	0.732
Central/ South Asia	Pashtuns/Afghans	32	0.524 (0.004)	0.001	0.438	0.798	0.804
	Tadjiks Afghanistan	28	0.627 (0.004)	0.002	1.591	0.815	1.192
	Uzbeks Afghanistan	54	0.616 (0.002)	0.002	1.757	0.740	1.210
	North Indians	120	0.511 (0.002)	0.001	1.607	0.672	1.128
	South Indians	102	0.606 (0.001)	0.001	0.583	0.825	0.878
	Nepalese	38	0.666 (0.002)	0.002	0.968	-0.335	0.056
	Tharu	80	0.484 (0.003)	0.001	1.288	0.704	1.029
	Uzbeks	76	0.334 (0.004)	0.001	-0.040	0.708	0.561
	Other C/South Asians	32	0.605 (0.004)	0.001	1.319	0.798	1.093
Central/East/ Southeast Asia	Khalka	114	0.499 (0.001)	0.001	-0.068	-0.615	-0.519
	Mongols	52	0.520 (0.002)	0.001	0.494	0.743	0.778
	Northern Yakuts	22	0.619 (0.006)	0.001	0.845	0.851	0.976
	Yakuts	110	0.535 (0.001)	0.001	0.871	0.678	0.864
	Han-Chinese	98	0.456 (0.002)	0.001	-0.190	-0.581	-0.537
	Japanese	82	0.438 (0.001)	0.001	1.413	0.505	0.901

Table 3.4: GenoPheno analysis of studied populations where published lactase persistence or lactose digester frequency data were available.
Significantly different values are marked in bold, significant values after Bonferroni correction are shaded.

Population	No. of chromosomes	Frequency LP associated alleles	No. of individuals	reported LP frequency	GenoPheno *p*-value	Reference published data
Germans	60	0.550	221	0.864	0.094	Flatz, G., et al. (1982) Hum.Genet 62, 152.
Irish	68	0.956	50	0.960	0.553	Fielding et al. (1981) Ir J Med Sci. 150, 276.
Roma	64	0.063	113	0.442	**0.003**	Czeizel et al. (1983) HumGenet. 64, 398.
Sami	60	0.293	50	0.520	1.000	Kozlov (1998) Int J Circumpolar Health. 57, 18.
Swedes	74	0.778	156	0.827	0.311	Sahi (1974) Scand J Gastroenterol. 9, 303.
English	102	0.745	150	0.953	**0.020**	Ferguson et al. (1984) Gut. 25, 163.
Greeks	120	0.150	200	0.553	0.226	Kanaghinis et al. (1974) Am J Dig Dis. 19, 1021.
Spanish	62	0.403	338	0.660	0.806	Leis et al. (1997) J Pediatr Gastroenterol Nutr. 25, 296.
Russians Perm	46	0.239	112	0.500	0.629	Kozlov (1998) Int J Circumpolar Health. 57, 18.
Greek Cypriots	120	0.008	50	0.340	**0.002**	Kanaghinis et al. (1974) Am J Dig Dis. 19, 1021.
Iranians	154	0.039	21	0.143	1.000	Sadre et al. (1979) Am J Clin Nutr. 32, 1948.
Kuwaiti	66	0.295	70	0.529	0.920	Sanae et al. (2003) Med Princ Pract. 12, 160.
Anatolian-Turks	116	0.078	104	0.288	0.271	Flatz et al. (1986) Am J Hum Genet. 38, 515.
Yemeni Hadramaut	166	0.277	17	0.530	0.972	Dissanayake et al. (1990) Ann Saudi Med. 10, 598.
Yemeni Sena	68	0.441	17	0.530	0.357	Dissanayake et al. (1990) Ann Saudi Med. 10, 598.
Pashtuns/Afghans	32	0.063	71	0.211	0.971	Rahimi et al. (1976) Hum Genet. 34, 57.
Tadjiks Afghanistan	28	0.250	79	0.177	0.006	Rahimi et al. (1976) Hum Genet. 34, 57.
Uzbeks Afghanistan	54	0.241	16	0.000	**0.001**	Rahimi et al. (1976) Hum Genet. 34, 57.
North Indians	120	0.200	70	0.729	**0.000**	Gupta et al. (1971) J Trop Med Hyg. 74, 225.
South Indians	102	0.118	100	0.360	0.309	Desai et al. (1970) Indian J Med Sci. 24, 729.
Han-Chinese	98	0.041	248	0.077	0.052	Yongfa et al. (1984) Hum Genet. 67, 103.
Japanese	84	0.000	40	0.275	**0.028**	Yoshida et al. (1975) Gastroenterol Jpn. 10, 29.

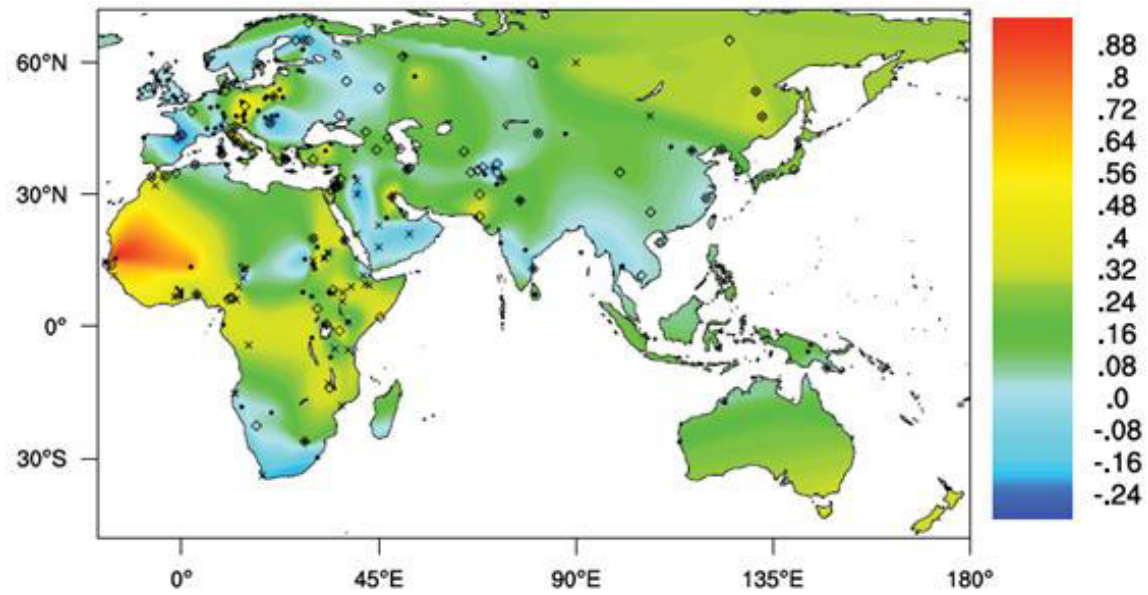
3.9.1 Worldwide genotype-phenotype correlation

Itan et al. (2010) devised a method for comparing lactase persistence genotype and phenotype data using interpolated values for regions with data lacking-one or the other type of information. Since the original publication of this approach in 2010, more publications giving genotype data for the *LCT* enhancer have become available and very recently a variant with previously reported borderline association, *-14009*G* (Ingram et al. 2009b), was shown by our group to be significantly associated with lactase persistence in a larger Ethiopians cohort (Jones et al. 2013).

A new interpolated lactase persistence genotype-phenotype correlation map of the Old World was constructed, with the help of Pascale Gerbault and Yuval Itan using the same Python code as in the original publication (Itan et al. 2010), to see whether the data collected in this thesis, the additional lactase persistence associated allele and the newly published data change the association patterns of the map. A literature review was necessary to update the GLAD database used for the original publication (Itan et al. 2010), this also included correction of coordinates of some data collection points. Tables with the full updated genotype and phenotype data are shown in Appendix B. Figure 3.6 and Figure 3.7 show the new lactase persistence genotype-phenotype correlation maps in comparison with the originally published maps by Itan et al. (2010).

The new data has contributed considerably to the density of genotype data points across Europe, South Asia and East Africa. The new genotype data for Oman (Al-Abri et al. 2012) has added to the areas where the lactase persistence phenotype was overpredicted by genetics. Areas where observed lactase persistence frequencies were not fully predicted from genotype data were slightly reduced but could not completely be resolved, such as in West Africa, the centre of Italy, some Eastern European areas and South Asian areas around Pakistan and North India. However, information about lactase persistence phenotype and genotype data is still rare in North and East Asia, the North and other parts of Africa and Australia and South East Asia

a)



b)

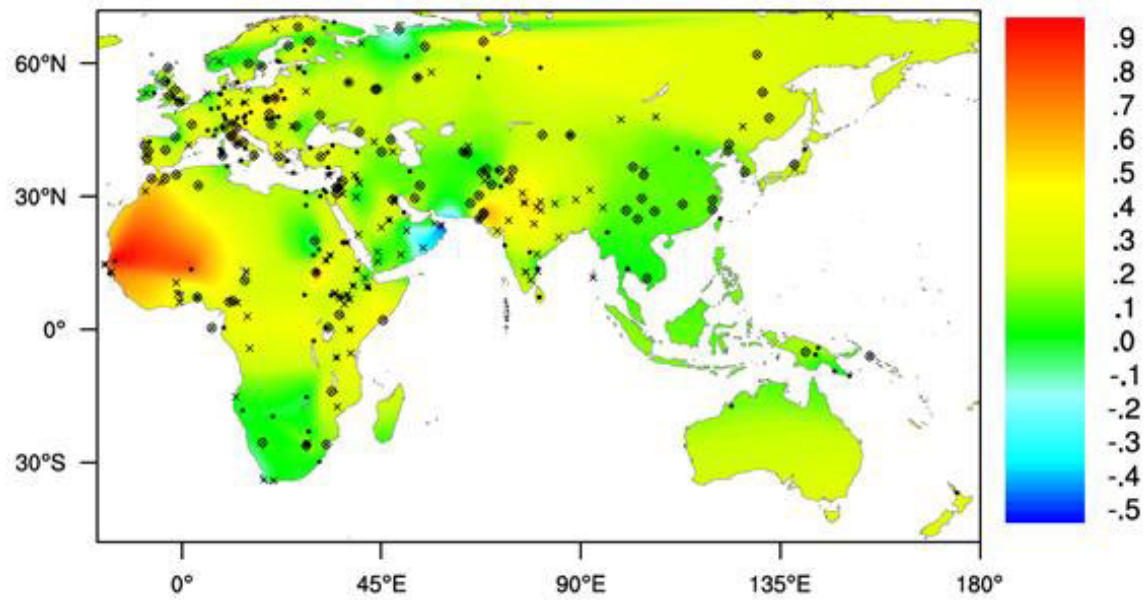
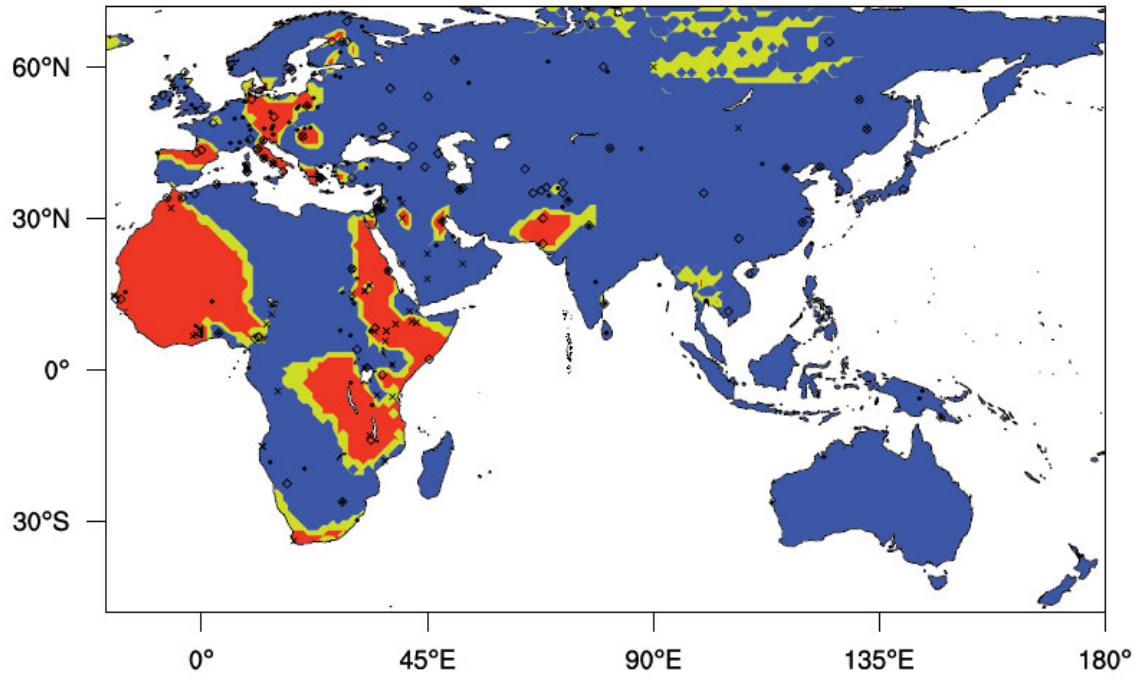


Figure 3.6: LP genotype-phenotype substractive map of the Old World. The map shows quantitative differences between observed LP phenotype data and these predicted from genotype data (both interpolated). Allele frequencies of the alleles *-14010*C, -14009*G, -13915*G, -13910*T and -13907*G* were taken into account to predict LP phenotype frequencies, which were subtracted from the observed LP phenotype frequencies. For geographic regions where no phenotype or genotype data were available, values were interpolated. a) shows the map originally published in Itan et al. (2010) and b) the newly calculated map. Colour key on the right: + values represent the genotype under-predicting phenotype at a certain location and - values over-predicting it, respectively. Note that the colour codes as shown in the scales on the right differ slightly between the maps, because of the greater range in the second map. Data collection points: ◇*-13910*T*, x all 4 or 5 LP alleles, ● LP phenotype.

a)



b)

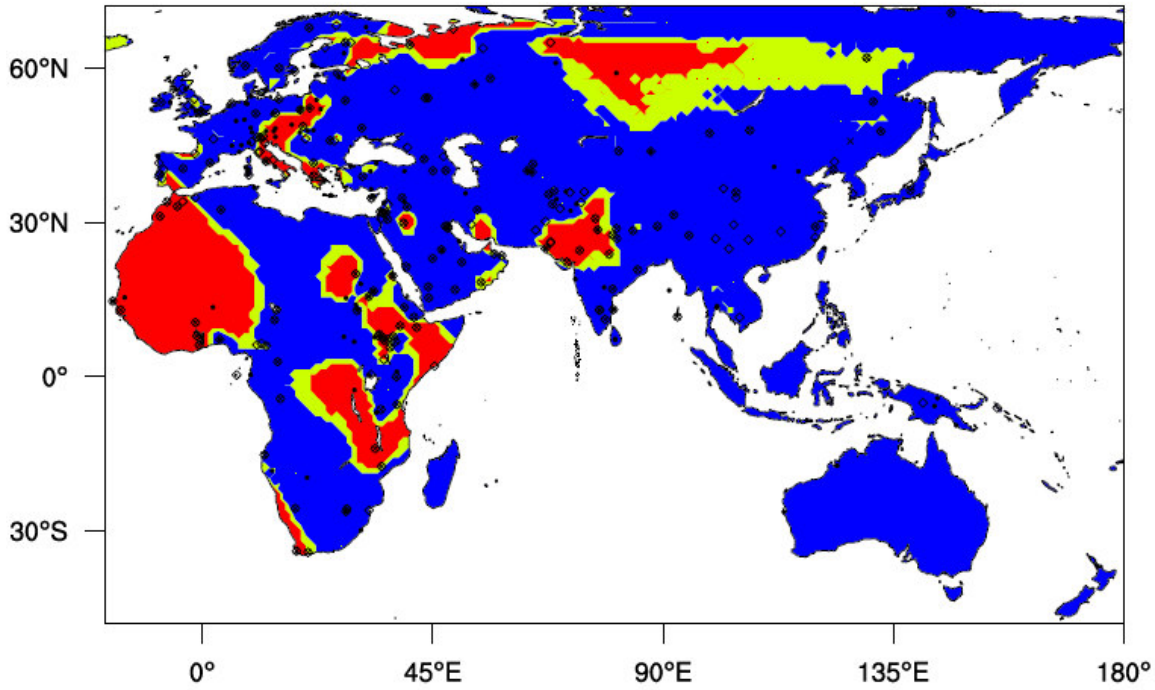


Figure 3.7: *P*-value contour map corresponding to the LP genotype-phenotype subtractive map of the Old World, shown in Figure 3.6. The colours represent the *p*-values obtained by the GenoPheno test: A highly significant difference is shown in red ($p<0.01$), yellow shows *p*-values of $0.01 \leq 0.05$ and blue indicates $p \geq 0.05$. Data collection points: ◇ *-13910*T* only, x all 4 or 5 LP alleles, ● LP phenotype.

3.10 Discussion

At the outset of this project European and Asian populations had mainly been typed specifically for the -13910 C>T SNP and only a few studies provided full enhancer sequencing data. Although more data are now available (see Appendix B for a full literature survey), this study represents one of the widest geographic surveys to date of the *LCT* enhancer sequence variation across Eurasia obtained via sequencing.

The results show that *-13910*T* is present in all populations tested except the Japanese, even in the Middle East at low frequencies. The allele is most prevalent in Northwest Europe and declines in frequency to the South and East as previously shown (Swallow 2003). However, deviating from this general pattern some of the Central and South Asian populations show similar *-13910*T* frequencies to Southern Europeans.
The other lactase persistence associated alleles on the other hand were only found in the Middle Eastern populations and of these mainly *-13915*G*, the others being surprisingly rare. *-14010*T* only occurred as a singleton as did *-14009*G.*

In contrast to this *-13495*T* was present in all populations at considerable frequencies. It broadly follows the Northwest to Southeast frequency decline of *-13910*T*, which could be explained by the fact that it occurs on the same haplotype background as *-13910*T*. It also occurs in combination with *-13907*G* and *-14011*T*, which might suggest it to be an evolutionary older variant. It was previously reported on a much smaller dataset to be in strong LD with *-13910*T* and *-13907*G* (Ingram 2008) and in association with the A haplotype but it was too frequent to be causal of lactase persistence. Further study of its pattern of association with the classical *LCT* haplotype markers in a wider–ranging set of populations, including some African groups should give better insight into the evolutionary relationship of this allele to the known haplotypes and is described in chapter 5.

Most of the newly discovered variants are rare and only one case of each of*-14028*C* and *-13914*A* were present in the European data. *-13914*A* was recently reported in one individual with high lactase activity in biopsy material (Khabarova et al. 2010) which did not carry other enhancer alleles thus could be, like *14028*C,* a candidate for causing lactase persistence.

*-14011*T* is rare-as well, but can be found in various European populations, which is of interest. Further study of the haplotype background of this allele might give insight into its origins, and this is also described in chapter 5.

None of the novel alleles occurred at frequencies high enough in any particular population to make new association studies a sensible option, but several of the alleles could be tested for function *in vitro* (chapter 4).

Several tests were conducted to investigate possible signs of selection in the enhancer region of the populations studied. Diversity measures as haplotype and nucleotide diversity as well as test for deviation from neutrality did not reveal any results that pointed to positive selection. It is most likely that the 559 bp enhancer region is too short to detect selection with these methods. The only clear genetic differentiation of the populations, as measured by pairwise F_{ST}, highlighted Northwestern Europeans as different, most probably because of their high *-13910*T* frequency.

Previously collected lactase persistence data, which was updated with new published information, and genetic data collected for this chapter were used to construct new interpolated maps showing the difference between reported lactase persistence and lactase persistence predicted from genotype, and of statistical significance of these differences calculated by the GenoPheno approach. Compared with the original publication (Itan et al. 2010) new data considerably increase the density of the data points and therefore the quality of the information that can be obtained by such an approach. However it did not decrease the areas with largest discrepancies nor change very much. The weaknesses of the interpolated GenoPheno approach were discussed extensively in Itan et al. (2010). The main issue in the case of lactase seems to be sampling from neighbouring populations with completely different lactase persistence frequencies. The interpolation approach is nevertheless a powerful tool to compare geno- and phenotypic information where it is not possible to collect both in the same populations, and the more data there are the more the effect of heterogeneity is reduced.

The new sequence data obtained in thesis has added considerably to the knowledge of the lactase persistence associated alleles in Eastern Europe and Asia. It was also shown that lacking information for lactase persistence enhancer alleles other than *-13910*T* was not the reason for the under-predicted phenotype from genotype data in the areas studied. There are still parts of East and South Europe and Asia, namely the regions

of West Russia, Italy and Pakistan and parts of Asia, where genotype data cannot explain the frequency of lactase persistence, which leads to speculations about other genetic influences on the lactase persistence trait. It is possible that further alleles have a functional effect. The next chapter will look further at some of the promising candidate alleles as they were tested in *in vitro* assays.

4 Functional studies of enhancer variation

4.1 Introduction

As discussed in chapter 3, a number of variants have by now been described within the sequence shown by Troelsen et al. (for example 2003) to have enhancer function. Several of these have been subjected to functional studies and showed differences in protein binding affinity and/or expression levels. These variants and the transcription factor binding sites identified so far are shown in Figure 4.1.

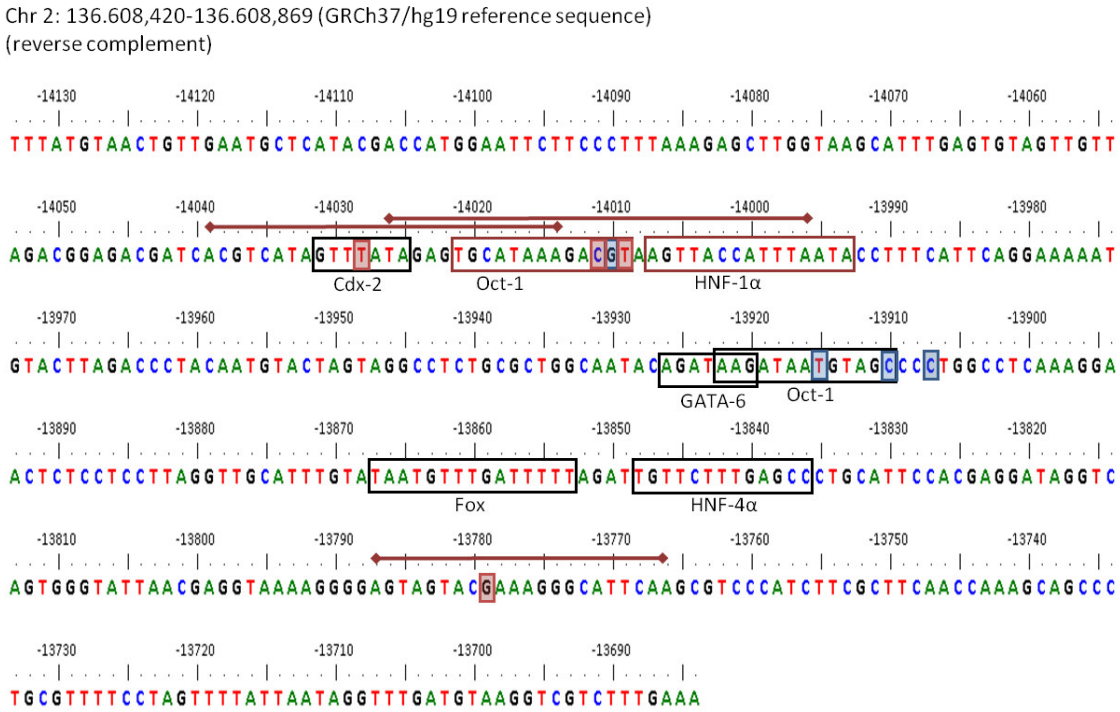


Figure 4.1: Positions in the 450 bp *LCT* enhancer of transcription factor binding sites and SNPs studied functionally. Indicated in boxes: black: transcription factor binding sites as shown in experiments before the outset of this thesis (Lewinsky et al. 2005); red: transcription factor binding sites as shown in experiments during the time of this thesis (Jensen et al. 2011); blue shadowed: SNPs previously examined functionally (Enattah et al. 2008; Ingram et al. 2007; Olds et al. 2011; Tishkoff et al. 2007); red shadowed: SNPs examined functionally within this thesis. EMSA probes used in this thesis are indicated with red lines.

The 'European' allele *-13910*T* was the first proposed to influence the level of enzymatic activity and to prevent the post-weaning decline of lactase, by preventing down-regulation of transcription, with evidence from reporter gene experiments and with the findings of increased Oct-1 transcription factor binding in EMSA studies (Ingram et al. 2007;

Lewinsky et al. 2005; Olds and Sibley 2003; Troelsen et al. 2003) (see also chapter 1). The causal role of *-13910*T* for lactase persistence is now even more evident as *in vitro* data were supported by the demonstration of its function in mice (Fang et al. 2012). Fang and colleagues established three transgenic mouse lines carrying a pGL3 luciferase reporter gene plasmid DNA, which included 218 bp of the *LCT* enhancer with either *-13910*T* or **C*, fused to a 2 kb rat *LCT* promoter. As demonstrated previously (Lee et al. 2002) the construct carrying the ancestral variant was able to create the typical lactase spatial and temporal expression pattern in mice, shown by *in vivo* bioluminescent detection and luciferase expression measurements. The two other mouse lines carrying different copy numbers of the *-13910*T* construct showed continued expression into adulthood (Fang et al. 2012).

Several other alleles in close proximity to *-13910*T* have also been studied since their location in and near the same Oct-1 binding site (see Figure 4.1) suggested a similar functional effect.
The lactase persistence associated alleles *-13915*G* and *-13907*G*, as well as *-13913*C*, which has only recently been shown to not be associated with lactose digester status (Jones et al. 2013), were tested in gel shift assays (Enattah et al. 2008; Ingram et al. 2007; Olds et al. 2011). The results of various investigators were slightly different. Compared with *-13910*T,* only minor Oct-1 protein binding of oligonucleotides carrying *-13907*G* was detected by Ingram et al. (2007), and was similar to that of the ancestral allele, whereas no binding at all was seen for *-13915*G* and *-13913*C*. Enattah and colleagues (2008) also showed no or very weak binding to *-13915*G* and *-13913*C* but their experiments revealed Oct-1 binding to the *-13907*G* probe at a level similar to *-13910*T*. Yet another study claimed to show binding of *-13915*G* to proteins of the nuclear extract, identified as partly being Oct-1 with competition and supershift experiments (Olds et al. 2011).

Transcription factor binding capacity *in vitro* is clearly complex so that gel shifts assays only give hints of differences between the alleles but results of transfection experiments provide further functional evidence. *-13915*G and -13907*G* alleles were shown clearly by several groups to increase reporter gene expression compared to the ancestral constructs (Enattah et al. 2008; Olds et al. 2011; Tishkoff et al. 2007). The Tishkoff study also included the lactase persistence associated allele *-14010*C*, which was shown to drive a higher *LCT* promoter expression too (Tishkoff et al. 2007). However, the influence of

transcription factor binding of the -14010 G>C SNP had only been tested in preliminary experiments at the outset of this thesis.
The *-14010**C variant was further examined in a project I was involved in, lead by Tine Jensen of the Troelsen group in Copenhagen. It was shown that the derived allele at -14010 G>C increased reporter gene expression compared to the ancestral variant. Gel shift assays located the SNP between an Oct-1 and HNF-1α transcription factor binding site and stronger binding of *-14010*C* to Oct-1 was reported (Jensen et al. 2011), see Appendix G. The functional region of the 450 bp enhancer described previously (Lewinsky et al. 2005; Troelsen et al. 2003) was investigated once more and it was shown that the 5' part of the enhancer region (-14133 to -13990) negatively influenced enhancer function, since the deletion of this part increased the activity of the remaining enhancer.

4.2 Choice of candidate functional variants

The focus in chapter 3 was on the search for new *LCT* enhancer variant alleles in European, Middle Eastern and other Asian populations that could tentatively be involved in lactase persistence and which would be further studied *in vitro*. Although some new alleles were found, they are quite rare, so it was not possible as a consequence to collect phenotypic data to test for association. Therefore four variants (*-14009*G*, *-14011*T*, *-14028*C*, *-13779*G*) were instead chosen to conduct *in vitro* functional assays, after a review of other studies in combination with the information in chapter 3.

-14009*G

*-14009*G* was found only once in the samples studied in chapter 3. However the association of *-14009*G* with lactase persistence had previously been shown to be of borderline significance in the Somali of Ethiopia (Ingram et al. 2009b). This allele is quite common in Ethiopia and Sudan and was noteworthy for being adjacent to a known lactase persistence variant, *-14010*C*. As described in the previous chapter, *-14009*G* has recently been confirmed to be associated with lactose digester status in a larger Ethiopian cohort (Jones et al. 2013).

-14011*T

*-14011*T* is rare but appears in several Eurasian groups as well as in a populations of Iran, Ethiopia and Brazil (Friedrich et al. 2012; Jones et al. 2013; Lember et al. 2006 and several groups in chapter 3 of this thesis) and like *-14009*G* is adjacent to a reported lactase persistence allele, which lead to speculations about possible function.

-14028*C

Another particularly interesting variant, -14*028*C,* was found in a European individual with two high lactase expressing alleles but only one chromosome containing *-13910*T* (Poulter et al. 2003), and a further individual of English descent in this study. Lewinsky et al. (2005) had found a Cdx-2 binding site overlapping with this position (see Figure 4.1), which was another reason to investigate the influence of this SNP more in detail.

-13779*G

In the present study a further allele, *-13779*G,* was found in India, Syria and Yemen (see chapter 3) and it was quite frequent in other populations from India as concurrently shown in a collaborative project (Gallego Romero et al. 2012). The preliminary suggestion that its frequency was higher in certain pastoralist than non-pastoralist Indian groups (Gallego Romero, personal communication 2009) motivated the choice of this allele but in fact this could not confirmed with the final data of the paper (Gallego Romero et al. 2012). However, additional evidence made this SNP interesting to examine.
Figure 4.2 shows a photograph of a DNase I footprint analysis of the *LCT* enhancer. Around nucleotide position -13800 a footprint of an unknown factor can be seen, suggesting some transcription factor binding there, which could be influenced by the -13779 C>G SNP.

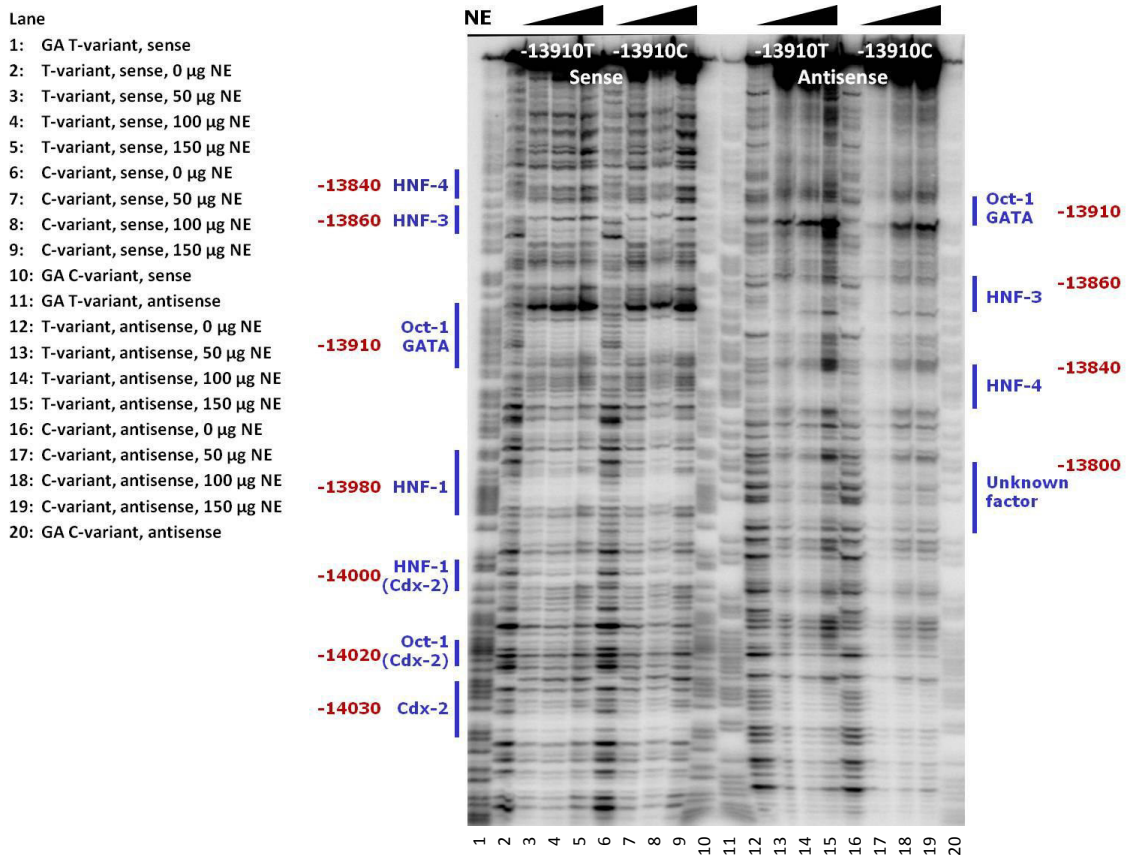


Figure 4.2: Autoradiograph picture of a DNase I footprint analysis done by denaturating polyacrylamide gel electrophoresis (PAGE), kindly provided by Jesper Troelsen. Blue bars indicate footprints with already identified transcription factor binding sites or unknown factors binding. Bp positions upstream of the transcription start of *LCT* are shown in red. With an increase of nuclear extract (NE) footprints get visible as light shaded areas, caused by binding of NE-proteins to the DNA and inhibition of DNase I cleavage.

4.3 Chapter aims

In summary, four *LCT* enhancer variants, *-14028*C, -13779*G, -14011*T* and *-14009*G* were selected for *in vitro* functional studies. These would be tested in comparison with the ancestral alleles and the lactase persistence alleles *-13910*T* and *-14010*C.* The combination of results from electrophoretic mobility shift essays and luciferase reporter gene assays should provide evidence as to whether these alleles are likely to be involved in the alteration of function of the *LCT* enhancer.

This chapter aims to investigate the influence of the four chosen allelic variants bioinformatically and *in vitro* by:

- Predicting transcription factors binding to the variants and the surrounding DNA regions, especially differences that occur for each allele of a SNP, by using different software tools,
- Conducting EMSA experiments to evaluate differences in binding of proteins of the nuclear extract to DNA sequences containing the variants and test specifically for binding of the predicted transcription factors,
- Transfecting Caco-2 cells with mutated plasmid DNA and measuring differences in expression of a reporter gene.

All experiments were conducted in collaboration with Jesper Troelsen and colleagues from Roskilde University and the University of Copenhagen in Denmark.

4.4 Prediction of transcription factor binding

A first step before conducting functional experiments was to analyse *in silico* the genetic region including 10 bp either side of a SNP or the full length sequence of the designed oligonucleotides for possible transcription factors binding. This was done by using different programs to compare these sequences with matrices of preferred targets for transcription factor binding as curated from literature (for more details see chapter 2, section 2.3.13). An important aspect was the matrix specificity to vertebrate species as considered in MATCH and TFSEARCH, both using the TRANSFAC database but with slightly different statistical algorithms and MatInspector.

As expected, the predicted results overlap to a great extent between the three programs, results are shown in the summary table in Appendix E. Candidate transcription factors

were those with different matrix matches/scores for the ancestral and derived variant and oligonucleotides were chosen to be tested in EMSAs if associated with expression in the digestive system (information from literature or MatInspector) and/or Caco-2 cells (information from 'The intestinal transcription factor target database' of the Copenhagen University, http://gastro.sund.ku.dk/chipchip/, with Caco-2 expression data from GeneChip analysis). As time for the experiments was limited, the transcription factors known to be involved in lactase expression, such as Cdx-2, Oct-1 and HNF-1α were tested first.

Matching competitor oligonucleotides were designed using information about binding site motifs from of the TRANSFAC database (see Table 4.1 below). For some of the predicted transcription factors, information about matching competitor oligonucleotides was available from the Affymetrix EMSA kit and these used to prepare oligonucleotides for competition experiments as described in chapter 2.

4.5 Transcription factor binding affinity of *LCT* enhancer variants

EMSAs were performed as described in section 2.2.6. For all *LCT* enhancer variant probes examined, binding specificity to proteins of the Caco-2 nuclear extract was confirmed if the DNA-protein complex could be competed with an excess of unlabelled oligonucleotides with the same sequences. The effect of competitor oligonucleotides for Cdx-2, Oct-1 and HNF-1α was in most cases tested, since these TFs have been shown to play an important role in the function of the *LCT* enhancer, as well as other specific probes selected from the *in silico* predictions. By adding target antibodies and specific competitors with a confirmed transcription factor binding sequence, several different factors were identified to bind to the variant probes (as summarised in Table 4.2 and Appendix E).

4.5.1 The influence of -14028 T>C on transcription factor binding

Both alleles of the -14028 T>C SNP bind similarly strongly to proteins of the nuclear extract. This binding could in both cases be competed with unlabelled oligonucleotides containing Oct-1 and HNF-1α binding sites (not shown). The binding of Oct-1 was expected as the probe sequence overlaps 3'with an Oct-1 site recently reported (Jensen et al. 2011)(see also Figure 4.1). The competition with the HNF-1α competitor suggests that this is probably another factor involved in the function of the -14133 to -14020 gene-regulatory region, which has been shown to decrease enhancer function in reporter gene

assays (Jensen et al. 2011). However, binding of these transcription factors could not be shown in supershift experiments.

The location of the -14028 position in a Cdx-2 binding site as proposed by Lewinsky et al. (2005) was confirmed. Both allelic probes of the SNP were competed with the same unlabelled Cdx-2 oligonucleotide but the ancestral variant much more strongly than the derived one. A clear supershift could be seen for 14028T with a Cdx-2 antibody (Figure 4.3a).

DNA-protein bands with 14028C were decreased with the addition of an HNF-4α competitor (Figure 4.3b). A supershift with the HNF-4α antibody confirmed that this transcription factor is binding to the 14028C probe.

-14028 T>C

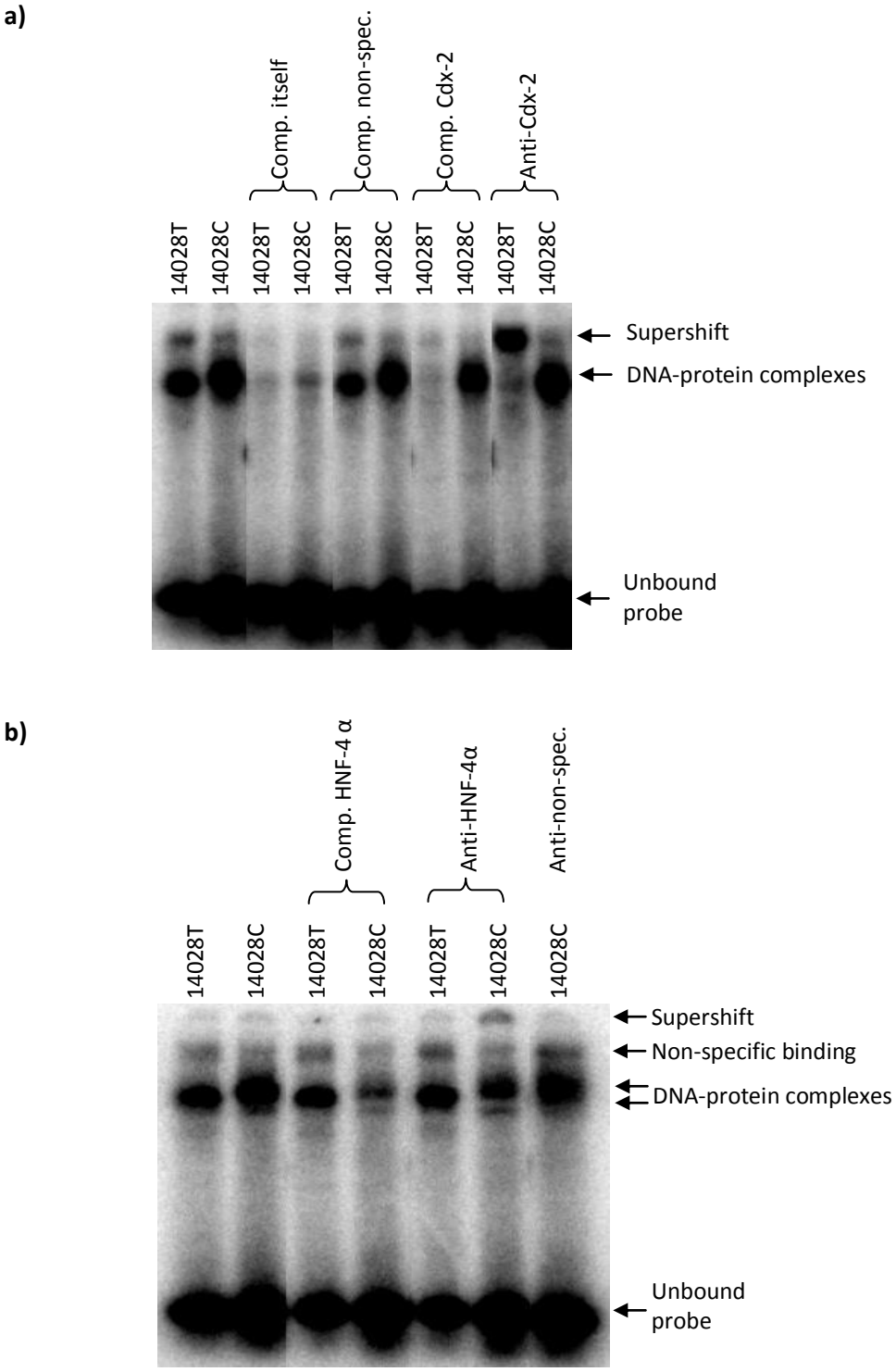


Figure 4.3: EMSAs of both variants of the -14028 T>C SNP in direct comparison. The phosphoimaging pictures of PAA gelshift assays show the binding of the labelled probes to proteins of the Caco-2 nuclear extract. Competition experiments (Comp.) with unlabelled variant oligonucleotides for a specific competitor for Cdx-2 (a) or HNF-4 α (b) indicate binding specifity of these TFs, which is shown further by supershift experiments with antibodies (Anti-) against Cdx-2 (a) and HNF-4 α (b). A non-specific antibody was used as negative control (b). Note that lanes were cut out of the gel pictures but lanes are from the same gels. Bands were classified as non-specific binding when they were only present on some of the gels and were reduced or removed with the use of less extract.

4.5.2 Transcription factor binding at -13779 G>C

As can be seen in Figure 4.4, both variants of the -13779 G>C SNP gave similar strong DNA-protein bands that could be competed with the unlabeled probe containing the ancestral or derived allele in a similar way.

Both variant probes show a slight competition with the Cdx-2 (Figure 4.4a) and only the ancestral allele showed a minor change in band intensity with the Oct-1 and HNF-1α competitors. A complete displacement was in contrast seen with the oligonucleotide containing an HNF-4α binding site and binding of this transcription factor could be confirmed with the supershift of the complex formed with the HNF-4α antibody. The shift was slightly stronger for the derived variant 13779C probe on Figure 4.4b but this could be due to the general stronger probe on that gel.

Further competition experiments with Affymetrix competitor oligonucleodites for the predicted transcription factors LEF-1, GKLF-1 and AML-1 only revealed a slight competition with the GKLF-1 and AML-1 competitors similar for both, the ancestral and derived allele probes (not shown).

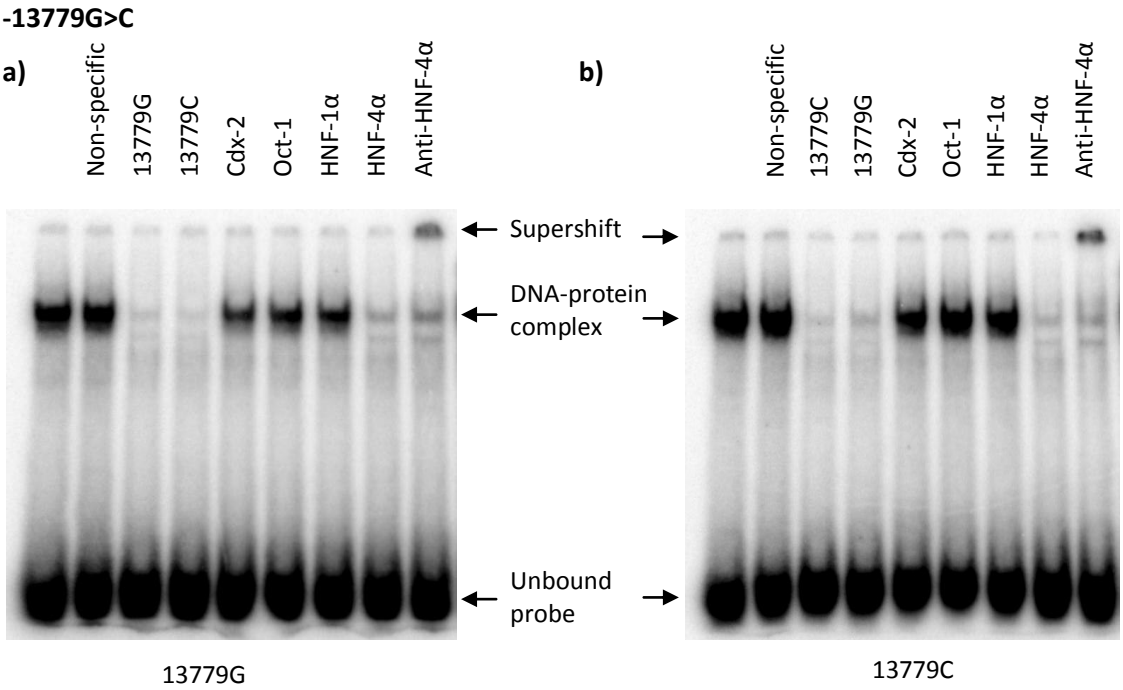


Figure 4.4: EMSA pictures of competition and supershift experiments for the ancestral (a) and derived (b) variant probes of the -13779G>C SNP. Competition experiments with unlabelled variant oligonucleotides, oligonucleotides with a non-specific sequence and binding sequences for different TFs are shown. These, as well as an antibody (Anti-) against the HNF-4α TF for a supershift experiment, were used as indicated above the pictures. The specific DNA probe-protein complex could be clearly supershifted with this antibody (strong band on top of the gel).

4.5.3 Transcription factor binding at the positions -14011, -14010 and -14009

In a collaborative project, lead by Tine Jensen of the Copenhagen lab, we had identified transcription factor binding sites surrounding -14010G>C. The SNP is located in an Oct-1 binding site (-14021 to 14009) and resides close to an HNF-1α binding site (-14007 to -13993). The lactase persistence associated *-14010*C* showed stronger binding to Oct-1 than the ancestral allele. However, this effect could only be shown with oligonucleotides spanning both transcription factors, indicating that it is mediated by HNF-1α (Jensen et al. 2011).

Since *-14011*T* and *-14009*G* are located adjacent to that variant and are part of the same Oct-1 binding site, it seemed likely they would cause a similar effect. The oligonucleotides carrying the two alleles were therefore designed to cover the same length of sequence, including the Oct-1 and HNF-1α binding sites as the 14010C probe (Jensen et al. 2011), and this probe was added to the assays for comparison.

Gel shift analysis indeed revealed similarities of all variant probes in forming protein complexes, visible as several bands on the gels. The upper band was competed in a similar way for the ancestral probe and the three variant probes with unlabelled Cdx-2, HNF-1α and Oct-1 competitors (partly shown in Figure 4.5a). The strongest competition was visible with the HNF-1α oligonucleotide, followed by the one containing an Oct-1 site and only minor competition was shown with the Cdx-2 competitor.

The 14011T probe showed the same strong protein binding as the 14010C probe, a slightly more intense band than the ancestral variant probe (14011C). Additionally, the supershift with an Oct-1 antibody, generated with all variant probes, was slightly stronger for the derived 14011T probe, similar to 14010C (Figure 4.5b). With the Cdx-2 and HNF-1α antibodies a slight inhibition of the formation protein probe complex for all probes and in a similar way (not shown).

A slight difference seen for 14011T, compared to the other probes, was a slightly faster migrating DNA-protein complex visible as the lower band on the gels. Additional binding of GATA-3 and 4 to *-14011*T* was predicted bioinformatically and gelshift assays were performed to test this. Preliminary results show a competition of both 14011 probes with different GATA competitor oligonucleotides, which mainly affected the upper bands on the

gels (not shown). However this has not yet been confirmed with further experiments or in supershifts.

The 14009G probe showed a very interesting pattern that differed from the three other probes tested. In addition to a weak upper band, which could be shifted with the Oct-1 antibody and slightly inhibited with the Cdx2 and HNF-1α antibodies, another DNA-protein complex was formed (shown by blue arrows in Figure 4.5). This suggests that besides the two other transcription factors binding to the 14009G probe, another protein in the nuclear extract is specifically binding to the *-14009*G* containing sequence..

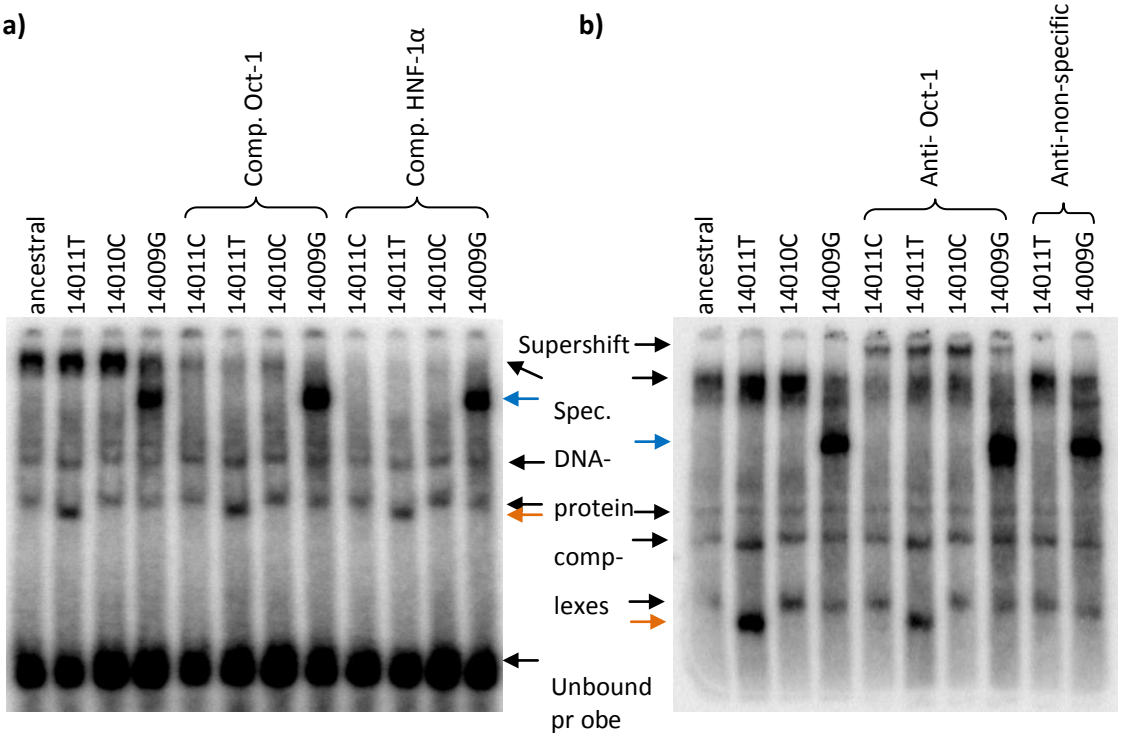


Figure 4.5: Phosphoimaging pictures of gelshift assays of competition (a) and supershift experiments (b) for the 14011T, 14010C and 14009G variant probes compared to the ancestral version. Competitors (Comp.) covering known binding sequences for TFs and antibodies (Anti-) were used as indicated above the gel images, a non-specific antibody was used as negative control. Different specific probe-protein complexes were formed of which the upper ones could be competed with Oct-1 and HNF-1α competitors (a) and supershifted (b) with an Oct-1 antibody (red arrow) for all 4 probes, whereas the DNA-protein complex formed with the 14009G probe (blue arrow) was not shifted. A slightly different binding pattern was seen for the lower band for 14011T (orange arrow).

4.5.4 Binding of an Ets transcription factor to *-14009*G*

Bioinformatic analyses with different software programs (see section 2.3.13) predicted additional possible transcription factors that would bind to the 14009G oligonucleotide sequence, mainly members of the Ets transcription factor family. Further investigation of this sequence with the TRANSFAC database with a decreased matrix match threshold of 0.70 revealed further candidate transcription factors, as shown Appendix E, that potentially bind to the derived but not the ancestral variant probe for the 14009T>G SNP and are expressed in Caco-2 cells. Competitor oligonucleotides were designed taking into account specific transcription factor binding motifs, found in TRANSFAC (Table 4.1, above).

Table 4.1: Design of competitor oligonucleotides for EMSA experiments examining the -14009 T>G SNP. Information about candidate transcription factors, with high differences in MATCH scores between the ancestral and derived variant, were taken from the TRANSFAC database. Letters in capitals represent the core sequence of the binding matrix of a certain TF stored as identifier sequence in TRANSFAC.

Transcription factor	TRANSFAC consensus sequence	Matrix accession number	Identifier	Designed Oligonucleotide sequence	Name
c-Ets-1 (p54)	NNNRCCGGAWRYNNNN	M01078	V$CETS1P54_03	aatcc**TCTACCGGA**t**GTAGGT**cgac	c-Ets-1
	NNACMGGAWRTNN		V$CETS1P54_02		
Ets	ACTTCCTS	M00971	V$ETS_Q6	atcttact**ACTTCCTC**gctgaccgt	Ets/Tel-2
NF-kappaB	NNNNKGGRAANTCCCN	M00774	V$NFKB_Q6_01	agt**GGCGGGGAAA**g**TCCCC**agaatc	NF-kappaB
Pax	CTGGAACTMAC	M00808	V$PAX_Q6	aagaag**TGGAACTCAC**gatcgtgct	Pax
Pax-4	NGNVGTCANGCGTGNNSNNYN	M00373	V$PAX4_01	ac**GGCGTTCATGCGTGAGCGACC**gt	PAx4-8
Pbx	GATTGATKGNNS	M00998	V$PBX_Q3	agatgg**GATTGATGGTAG**ccgtatt	Pbx
Tel-2	YTACTTCCTG	M00678	V$TEL2_Q6	see Ets	

Figure 4.6 shows the EMSA experiments with the first set of oligonucleotide competitors. It can clearly be seen that the upper DNA-protein complex could be competed with the ancestral sequence on both variant probes. The lower complex band with the -14009G probe could not be competed with the ancestral variant competitor, which confirms the specificity of the factor binding to the derived allele. The c-Ets-1 sequence was only one that competed strongly with the lower complex of 14009G. A slight competition of the binding of both variant probes to the upper band was seen for NF-kappaB, Pax, Pax4-8 and Pbx.

From the Affymetrix competitor set two oligonucleotides competed with the 14009G specific band: Aff_c-Ets-1 and Aff_ELK-1. However, the protein binding to the ancestral probe 14011C (shown as band for the specific DNA-protein complex in Figure 4.6a) was also competed with the Aff_ELK-1 competitor and slightly with Aff_ELF and Aff_ETS (1) competitors (not shown).
The reciprocal experiments, using c-Ets-1 and Aff_ELK-1 as probes showed that Aff-ELK-1 was competed with both competitor sequences, 14011C and 14009G, whereas the c-Ets-1 probe could only be completely competed with 14009G and only slightly with the ancestral version (not shown). So far it was not possible to confirm this in supershift experiments when an antibody for Ets1/2 or Ets-1 was added to the assay.

-14009 T>G

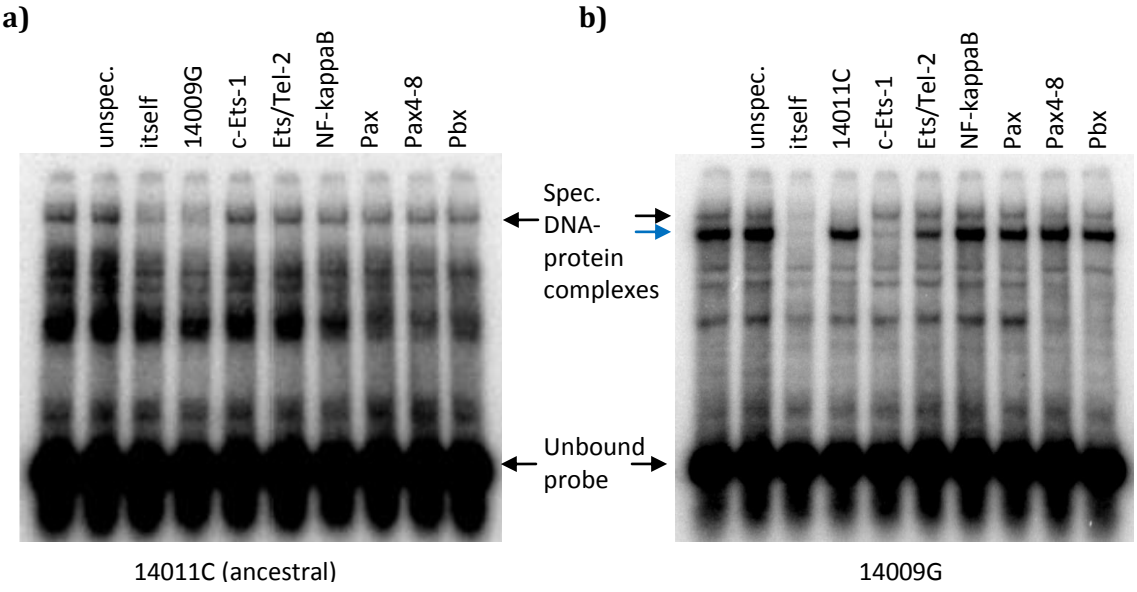


Figure 4.6: Images of gelshift assays of competition experiments for the ancestral (a) and derived (b) variant probes of the -14009 SNP. Competitors were used as indicated above the pictures. The specific DNA-protein complex for 14009G (blue arrow) could not be competed with the ancestral sequence competitor but was with competitor oligonucleotides of the Ets family.

4.6 Experimental strategy for reporter gene assays

Different strategies were pursued for the insertion of the *LCT* enhancer and the promoter variants into the reporter gene plasmids, upstream of the firefly luciferase gene (*luc+*), as described in chapter 2. Results for the experiments with the *LCT* promoter variants are discussed in chapter 6. Enhancer variants were generated via site directed mutagenesis and inserted via enzyme digest into the luciferase reporter vector pGL3 hLPH1085 (Troelsen et al. 2003) as shown in Figure 4.7, about 2.8 kb upstream of the human *LCT* promoter, already present in the construct. Transient transfections of Caco-2 cells were conducted as described in chapter 2 (section 2.2.7).

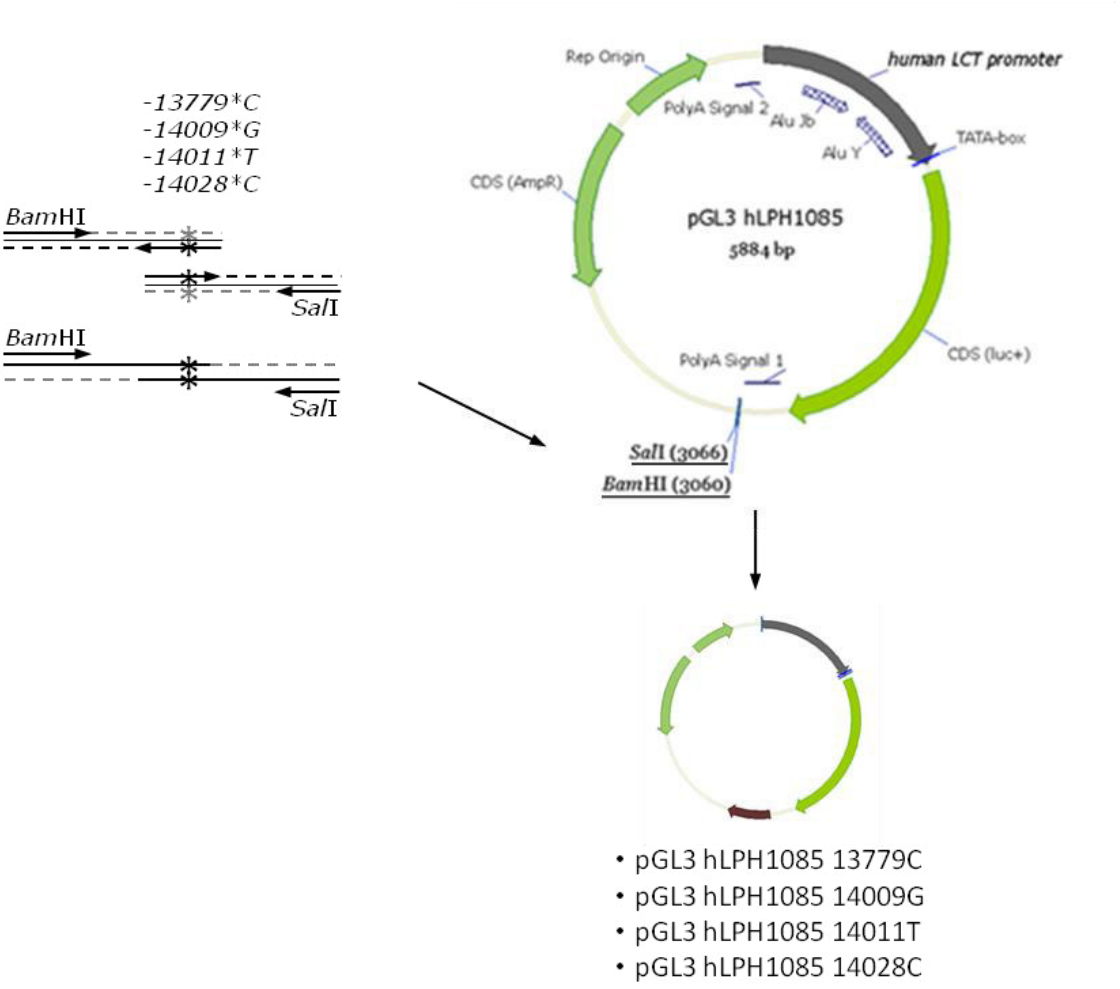


Figure 4.7: Summary of the site directed mutagenesis and insertion of *LCT* enhancer fragments into the luciferase reporter gene plasmids, as described in detail in section 2.2.7 (modification of Figure 2.2 of chapter 2). *LCT* enhancer variants were generated in a two step PCR amplification and inserted via *Bam*HI and *Sal*I into the pGL3vector containing the ancestral promoter region (pGL3 hLPH1085 (Troelsen et al. 2003)).

4.7 The influence of *LCT* enhancer variants on reporter gene expression

Luciferase reporter gene expression of the different enhancer variant constructs was measured in undifferentiated (2 days after transfection) and differentiated Caco-2 cells (9 days after transfection) and results of one experiment are shown in Figure 4.8. The -*13910*T* and -*14010*C* constructs were used as controls since they had previously been reported to influence reporter gene expression, at least in undifferentiated cells (Tishkoff et al. 2007; Troelsen et al. 2003).
The function of the 450 bp region carrying the ancestral alleles (Lewinsky et al. 2005; Troelsen et al. 2003) was confirmed since it enhanced *LCT* promoter function 7 fold in undifferentiated and 13 fold in differentiated cells compared to the promoter only constructs.

Significant differences in luciferase expression compared to the ancestral -*13910*C* constructs were demonstrated for all four newly tested variants in undifferentiated Caco-2 cells (Figure 4.8). The expression was about 1.3 times higher for the -*13910*T*, -*14009*G* and -*14028*C* reporter gene plasmids than the ancestral construct. For -*14010*C* and -*14011*C* this difference was even larger, up to about 1.4 fold expression compared to the ancestral version and the construct with -*13779*C* shows about twice the activity of the ancestral enhancer.

The picture changes slightly with the measurements in the differentiated cells, harvested 9 days after transfection. Only the activities for the control plasmids, -*13910*T* and -*14010*C* and those carrying -*14009*G* and -*14011*T* stayed significant (Figure 4.8). -*13910*T* increased luciferase activity to 1.8 fold compared to the ancestral enhancer and the effect of -*14010*C* was even bigger. It increases enhancer activity up to 2.8 fold in differentiated cells.
The effect of -*14009*G* on reporter gene expression resulted in a 2 fold increase in activity. The significant effects of -*14009*G* and -*14010*C* were also confirmed in another 9 day experiment, which was recently published (Jones et al. 2013).
The enhancer with the -*14011*T* variant is 1.4 (2 days) to 2 fold (9 days) more active than the ancestral version, which was also confirmed in another 9 day experiment (not shown).

These results support those of the gel shifts and suggest that both alleles (-14009 and 14011) alter lactase expression.

Although the constructs with *-14028*C* and *-13779*C* showed a 1.3 and 2 fold increased expression in the 2 day experiments, they did not show a significant increase in activity in the 9 day experiments. Both variants also revealed a high variation in luciferase activity in further experiments (data not shown) so it was not possible to confirm these results in the undifferentiated stage.

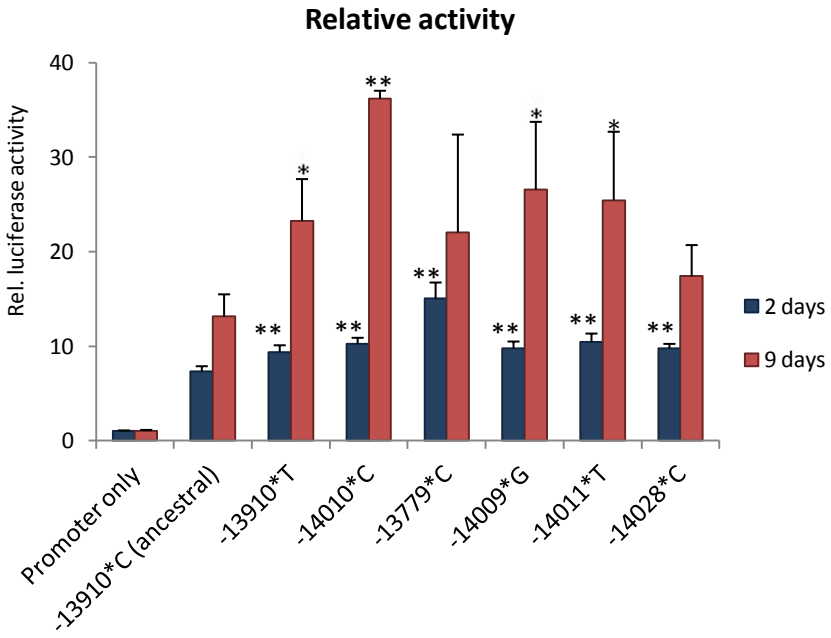


Figure 4.8: Result of luciferase reporter gene assays for *LCT* enhancer variants after 2 and 9 days of transfecting the Caco-2 cells. Luciferase activity of enhancer variant constructs are shown relative to the promoter only construct. Luciferase activities (means±SD, n=4) were corrected for transfection efficiency against ß-galactosidase and normalised to the expression of pGL3 hLPH1085 (promoter only). Significant differences of luciferase activity compared to the ancestral variant were calculated using a student *t* test and are shown by ** ($p<0.01$) and * ($p<0.05$).

4.8 Summary and discussion of the functional studies

Functional evidence for the 4 variants *-14028*C*, *-14011*T*, *- 14009*G* and *-13779*C* in comparison to their ancestral alleles, obtained from EMSA and transfection experiments that were conducted during this thesis work are summarised in Table 4.2. A more detailed summary of the experimental outcome and bioinformatic investigations for each of these variants can be found in Appendix E.

Table 4.2: Summary of the main outcome of the functional studies conducted during this thesis. Abbreviations: +: variant has an effect on reporter gene expression or shows binding to a certain transcription factor, bold crosses: strong effect, - no effect, (+) weak effect respectively, n/a not applicable.

	-14028*T	-14028*C	-14011*C	-14011*T	-14009*G	-13779*G	-13779*C
Transfections							
2 days	control	+	control	+	+	control	+
9 days	control	-	control	+	+	control	-
EMSAs							
Cdx-2	**+**	+	+	+	+	(+)	(+)
Oct-1	+	+	+	**+**	+	(+)	-
HNF-1α	+	+	+	+	+	(+)	-
HNF-4α	-	**+**				**+**	**+**
c-Ets-1	n/a	n/a	-	n/a	**+**	n/a	n/a
ELK-1, ETS	n/a	n/a	+	n/a	**+**	n/a	n/a

It was clearly shown that sequences containing the two alleles, *-14011*T* and*-14009*G* differ functionally from their ancestral forms *in vitro*, as concluded from transfection experiments and evidence from EMSAs. The high activity of the enhancer constructs carrying those alleles, compared to the ancestral variant, could be explained by an altered binding to different transcription factors in each case: Slightly stronger binding of *-14011*T* to Oct-1 and probably also stronger binding to a GATA factor and the binding of *-14009*G* to an additional transcription factor, most probably a member of the Ets family.

Binding of *-14009*G* to the Ets-1 transcription factor could not be confirmed in EMSA experiments with either of the Ets-1/2 and Ets-1 antibodies. However, these commercially available antibodies might not be specific enough or even suitable for EMSA experiments and at the time no other Ets antibody was available that did have confirmed function in binding to Ets-1 in EMSA studies.
Although direct evidence from supershift assays is missing the competition with commercially available and target designed competitors strongly suggest the binding to c-Ets-1 or a member of the same transcription factor family. Ets (E26 transformation-

specific) factors play an important role in the differentiation, survival and proliferation of cells. These factors usually interact with other factors to regulate DNA-binding or transcription (for example Sharrocks 2001) and could well play a role in the regulation of *LCT* expression.

The -14028 variant tested apparently slightly influenced *LCT* promoter activity in luciferase reporter gene assays (but only significantly in undifferentiated Caco2 cells) and the results were not replicated during the course of this work. There was however some evidence for alteration of transcription factor binding in EMSAs. It could be shown that the derived *-14028*C* leads to a loss of binding to Cdx-2 and seems to create a HNF-4α binding site, but it can be concluded that this does not seem to change enhancer activity dramatically *in vitro*, at least under the conditions tested.

In contrast to all other variants tested, a direct functional effect was not supported for -13779 G>C. The probes were shown to contain an HNF-4α binding site, which is probably the cause for the footprint around that position shown in Figure 4.2, but the -13779 G>C polymorphism doesn't seem to have an effect on protein binding. Mutation analysis might be useful to define the exact binding site of this transcription factor along sequence the studied here, but there is no real hint that this particular substitution has any effect. The suggestive evidence for a functional role of *-13779*C* had been not very strong as it mainly came from its relatively high frequencies in milk drinking populations in India, but it had in fact previously been reported in one Somali individual who was diagnosed as a lactose maldigester (Ingram et al. 2009b).

The results show that *-14011*T* and*-14009*G,* at least,-give clear evidence of a difference *in vitro* but it should always kept in mind that this does not necessarily prove-a functional difference *in vivo.* Physiological conditions are different in the living organism and other factors influence DNA-protein binding, such as chromatin structure or other flanking sequences of importance that cannot be mimicked in EMSA experiments or transfection studies. It might also be that *in vivo* function of the variants is even larger as in the transfection experiments as Caco2 is only capable of expressing a small amounts of lactase in comparison to the cells of the small intestine (Rousset 1986). In the case of *-14009*G,* the association with lactase persistence further confirms its likely functional role. Unfortunately *-14011*T* was not frequent enough in any one population to make this approach a possibility.

Techniques based on chromatin-immunoprecipitation such as ChIP-chip (chromatin immunoprecipitation coupled to microarray hybridisation) or ChIP-Seq (chromatin immunoprecipitation coupled to high throughput sequencing) give a better insight to specific transcription factor DNA targets and their networks in the intestine (reviewed Olsen et al. 2012) and could be useful to further examine the influence of the *LCT* enhancer alleles on transcription factor binding.

Further experiments would also be necessary to investigate the interaction of the identified transcription factors with other factors involved in lactase expression. Previous studies revealed interesting results. For example was luciferase activity highest for constructs containing *-13915*G* under simultaneous over-expression of Oct-1 and HNF-1α but less with HNF-1α and Oct-1 alone (Enattah et al. 2008). A similar interaction of these two transcription factors in mediating *LCT* enhancer activity was previously shown for constructs containing the -13910C>T polymorphism (Lewinsky et al. 2005).

The transcription factors predicted bioinformatically only overlap to a certain extent with those successful confirmed experimentally. Bioinformatic tools are useful in searching for candidate transcription factors for functional tests, nevertheless they have their limitations as most of the binding matrixes are collected from different organisms and cell types. However, data about transcription factor binding sites will increase dramatically in the next years with the application of new techniques and large scale investigations of functional elements in a great range of cell types in project such as ENCODE (Hoffmann et al. 2012), which will make predictions more accurate.

During the course of this project the question was posed as to whether there were any other polymorphic changes (reflecting the different haplotype background of the variants) that might influence the level of expression in our *in vitro* experiments. To check this the sequence of the promoter used as part of the construct was reviewed in detail and was shown to derive from a B haplotype chromosome. It should thus be noted that only *-14028*C* is tested in a construct with a promoter sequence reflecting its *in vivo* situation, with a *T** allele at position 958, which is part of the B haplotype background of the variant. All other reporter-gene-constructs that were used to test the effect of certain enhancer variants were done using the same pGL3 hLPH1085 vector (Troelsen et al. 2003). Since it had been reported previously that this SNP alters transcription factor binding in gel shift assays (Hollox et al. 1999) this site was inspected bioinformatically and shown to be located within a predicted Oct-1 site. Experiments were therefore conducted to attempt to

replicate the gel shifts. These showed a gel shift which was stronger for the ancestral then the derived variant and could be slightly competed with Oct-1 competitor oligonucleotides. However, these results are only preliminary and would need further confirmation.
A clear line for future research would be to analyse the possible combinatorial effect of this allele with the various enhancer alleles.

Lactase is only fully expressed in differentiated cells and therefore the activity measurements of the cells harvested 9 days after transfection might better represent the influence of the variants. The differences in expression from 2 to 9 days for all enhancer variants tested show that the mean luciferase activity rises more than 2 fold from the undifferentiated to the differentiated cell state for these variants showing a significantly different expression to the ancestral variant of the enhancer. One can speculated that it is the ability of a variant to up-regulate expression during the differentiation process that is the crucial effect for lactase persistence. Unfortunately, it is not possible to study the maturation of cells and the influence of the lactase persistence associated alleles on expression during that development *in vitro*. Model organisms could be used to study this as recently done for *-13910*T* in mice (Fang et al. 2012). However for cost and logistic reasons the only practical ways to study several novel alleles is the combination of *in vitro* functional studies, as was done here. For any that are more frequent, phenotype association studies, perhaps in combination with mRNA expression, is highly informative but cannot provide direct evidence because of linkage disequilibrium

5 *Evolutionary background of LCT enhancer variants*

5.1 Introduction

In chapter 3, a number of different derived enhancer variants were documented, in a large number of European and Asian populations. This chapter aims to combine information from several sources, about the common enhancer variants and determine their haplotype background to get a better picture about their evolution and possible signatures of selection in the genetic regions examined.

Previous studies by different members of our group, which attempted to determine the core *LCT* haplotype background of the different *LCT* enhancer variants used slightly different combinations of genetic markers, and more limited sample sets (Ingram et al. 2009b; Jones et al. 2013; Poulter et al. 2003). The next step in getting a better and more consistent picture about the haplotype background of enhancer variants would be to add more and missing markers and to combine more information about populations from Africa, Europe, the Middle East and other parts of Asia in the analysis.

It is widely agreed that the lactase persistence associated alleles occur on more than one haplotype, which was part of the evidence that lactase persistence evolved independently several times (Enattah et al. 2008; Ingram 2008; Ingram et al. 2007; Tishkoff et al. 2007), but it is less clear whether the same mutations might have recurred. One publication claimed a different haplotype background for some of the *-13910*T* alleles in groups of Russia and Iran and therefore concluded that this site had mutated more than once (Enattah et al. 2007). It is of interest to see if there is any evidence for that from looking at similar extended haplotypes in a different dataset, but including Iranians, which might or might not support this suggestion.

The B haplotype, which is quite common across the globe (Hollox et al. 2001), showed evidence of a frequent extended haplotype across at least 500 kb in a collaborative study focussing on South Asia (Gallego Romero et al. 2012), see Appendix G. My contribution to this work involved extracting SNP data from families of the 'Human Genome Diversity Panel-Centre d'Etude du Polymorphisme Humain ' (CEPH) (Dausset et al. 1990) from

previous work in our lab and merging it with public data for the CEPH (CEU) samples from the HapMap project (http://hapmap.ncbi.nlm.nih.gov/) in order to infer which were the B haplotypes.
We were curious that this haplotype was so frequent yet and did not carry any of the main derived lactase persistence associated alleles. Only the *-14028*C* variant on a high lactase expressing chromosome was associated with the core B haplotype (Poulter et al. 2003), but as seen in chapter 3 it is not frequent in any of the populations studied. This led to suggestions that the length of chromosome known as the B haplotype might have been selected for by another advantageous mutation located on this sequence.

5.2 Chapter aims

Thus this chapter aims to examine the extended haplotype background of the derived enhancer variants in this combined dataset from various geographic regions, with the target of obtaining further insight into the demographic history and origin of these alleles. I also investigate the distribution of the extended B haplotype to get insights into the selective or demographic explanations for its frequent appearance. I examine patterns of linkage disequilibrium and also test methods to detect evidence for different types of selection.

5.3 Population selection

Of the new dataset described in chapter 3, the Middle Eastern populations showed the most variation across the *LCT* enhancer region. Middle Eastern sample sets (Bedouin groups, Israel and Palestinian Arabs) previously analysed by our group (Ingram et al. 2007) were also included for further analysis as well as randomly chosen samples from several African populations (Afar, Beni Amer, Chagga) that showed high frequencies of lactase persistence associated alleles (Jones 2012). Of interest were also populations (Asante, Chewa, Mambila, Shabo) that do not have a pastoralist lifestyle (see also chapter 2, Table 2.2), in which enhancer variation was mainly absent (Jones 2012).
To get a wider picture of the haplotype distribution, several European and Asian groups, analysed in chapter 3 were included in the sample set.

To look specifically at differences between lactase persistent and non-persistent individuals and compare with previous findings, phenotyped samples for Europe (Northern Europeans and Italians) and Africa (Amhara, Oromo, Jaali) were also part of the dataset.

5.4 Sequencing/Genotyping strategy

Five regions in and around the *LCT* gene were included in the initial haplotype analysis spanning a region of about 80 kb, as shown in Figure 5.1.

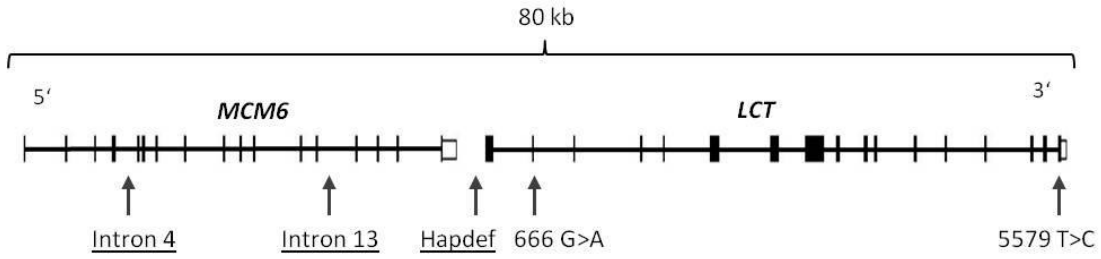


Figure 5.1: Sequence regions and SNPs included in haplotype analysis, spanning a region of 80kb.

To obtain as much as possible of the *LCT* core haplotype information as described by Hollox (2001) the haplotype defining region upstream of *LCT,* called the 'hapdef' region within this thesis as it was in Jones et al. (2013), was sequenced and information about the two haplotype markers in exon 2 and exon 17 of *LCT*, 666 G>A and 5579 T>C mainly obtained via genotyping was also collected as described below. The *MCM6* intron 4 region also described in Jones et al. (2013) was also sequenced to get haplotype information in relation to intron 13 of *MCM6*. The hapdef and intron 4 information was also used to serve as control regions for the enhancer to compare for diversity. Both regions are located near enough to the enhancer to share a similar history, with low recombination between them and little difference in mutation rates (Jones et al. 2013).

All sequencing data used in this chapter that formed part of previous projects of the group were re-analysed and sequencing repeated to get complete datasets, for example to extend the *LCT* enhancer region further 3', or if readable sequences were not found. All Middle Eastern samples analysed by Kate Ingram (Ingram et al. 2007) and those of chapter 3, the European and Asian groups with complete enhancer data were sequenced for the *MCM6* intron 4 and *LCT* hapdef regions specifically for this project. Additional samples of the Saudi Bedouins were included as well and sequenced for all regions.

The hapdef sequence region upstream of the *LCT* promoter spans 701 bp (Chr2:136595212-136595912, GRCh37/hg19). This region was successfully amplified with primers applied previously (Jones et al. 2013) and sequences were readable for all samples from bp position -1100 until -630 from the start of transcription of *LCT*. An InDel

at the 3' end of the sequence and the consequent need for a different internal sequencing primer restricted the length of evaluable sequence region. The control region of *MCM6* intron 4 covers a 683 bp region (chr2:136624515-136625197, GRCh37/hg19) and sequences were readable from bp position -30292 to -29922 upstream of the *LCT* transcription start. The sequences for the *LCT* enhancer region were obtained as described in chapter 3.
The haplotype markers 666 G>A and 5579 T>C were part of the marker set (see below), genotyped by LGC Genomics (KBioscience). The regions of exon 2 and 17 of *LCT* were also subjected to sequencing for a subset of the Middle Eastern samples within an undergraduate project of Lana Couzens (Couzens 2011). These data were used to crosscheck for quality of the genotype data delivered by LGC Genomics and to complement missing data.
Details about all primer sequences can be found in Table 2.3 of chapter 2.

5.4.1 Marker selection for extended haplotype analysis

Markers for extended haplotype analysis were chosen to span a sequence region of about 1.8 Mb around *LCT.* Certain criteria were deployed to get an informative marker selection. All SNPs are part of the Illumina 650k platform typed for most of the HapMap populations but also used for the haplotype analysis in our recently published paper about Indian and European cattle herders (Gallego Romero et al. 2012). With the haplotype information of these data it was possible to exclude close SNPs in complete LD with each other. To further guide the choice of markers LDU maps were constructed (see below, 5.5.8.3), that were particularly useful in the region downstream of *LCT* where there is a long region of very high LD between the genes *R3HDM1*and *RAB3GAP1.*

The first set of markers was chosen to be broadly 50 kb apart, with the distance increasing for the outer SNPs, to have a minor allele frequency of at least about 8% and to be as different as possible in frequency between the CEPH (European), Yoruban and Han-Chinese HapMap populations. Downstream of *LCT* the region was extended up to about position 135.8 Mb (b37, hg 19) on chromosome 2 to include the region where a sharp drop in extended haplotype homozygosity (EHH) was seen in European and Asian populations (Gallego Romero et al. 2012). Later the *LCT* haplotype SNPs 666 G>A (rs3754689) and 5579 C>T (rs2278544) were also included, as well as the lactase persistence associated SNP 22kb upstream of *LCT* (-22018 G>A, rs182549).
The chromosomal location and further information about the 36 SNPs chosen can be found in Figure 5.2 and Table 5.1.

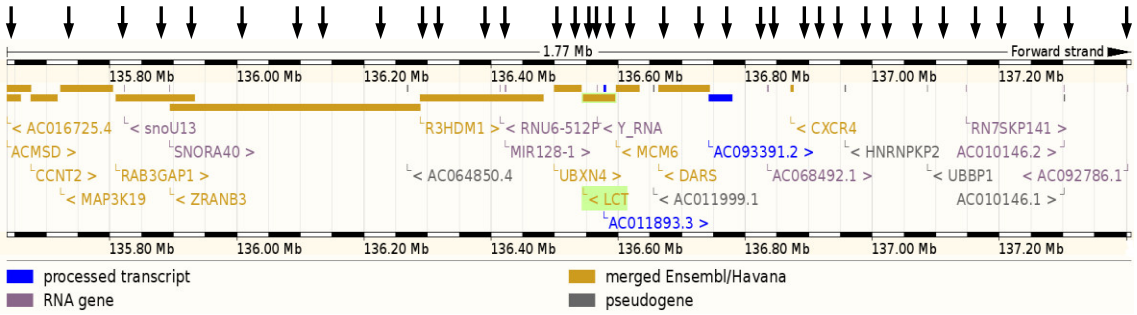


Figure 5.2: Location of the SNPs chosen for extended haplotype analysis. The broad position on chromosome 2 of all SNPs is indicated by arrows. The picture shows all genes documented on Ensembl for the chosen region and their position in chromosomal+ direction (b37), taken and modified from the Ensembl website (http://www.ensembl.org). Note that *LCT* and *MCM6* are transcribed from the chromosomal negative strand accounting for the location of *LCT* (highlighted in green) on the left of *MCM6*.

Table 5.1: Chromosomal location and allelic information about all 36 SNPs chosen for extended haplotype analysis.

No.	SNP	Position on Chromosome b36 (Kb)	Position on Chromosome b37 (kb)	Dist. (kb)	SNP (Chr +)	Anc. allele	MAF 1000 genomes	MAF in HapMap pop. CEU	MAF in HapMap pop. YRI	MAF in HapMap pop. CHB
1	rs1446525	135354.317	135637.847		A/G	A	G = 0.288/629	0.60	0.16	0.07
2	rs4954209	135454.378	135737.908	100	G/T	T	T = 0.348/761	0.18	0.37	0.62
3	rs2874739	135535.377	135818.907	81	C/T	T	T = 0.347/757	0.12	0.75	0.20
4	rs1869829	135594.032	135877.562	59	A/G	G	A = 0.385/840	0.81	0.02	0.38
5	rs2305248	135644.782	135928.312	51	A/G	G	G = 0.343/748	0.12	0.74	0.20
6	rs1900741	135718.970	136002.500	74	C/T	T	C = 0.431/941	0.91	0.25	0.14
7	rs1561277	135808.531	136092.061	90	A/C	A	C = 0.264/577	0.74	1.00	1.00
8	rs9798267	135846.261	136129.791	38	A/G	A	G = 0.277/604	0.08	0.66	0.15
9	rs6709132	135949.042	136232.572	103	A/G	G	G = 0.211/460	0.08	0.31	0.15
10	rs3806502	136004.743	136288.273	56	C/T	T	T = 0.327/714	0.12	0.76	0.15
11	rs4954265	136040.695	136324.225	36	A/G	G	G = 0.270/590	0.06	0.67	0.15
12	rs961360	136110.128	136393.658	69	A/G	A	G = 0.315/687	0.08	0.30	0.43
13	rs4954278	136124.761	136408.291	15	C/T	C	T = 0.181/396	0.08	0.38	0.15
14	rs6430585	136223.397	136506.927	99	A/C	C	A = 0.292/637	0.14	0.30	0.22
15	rs10188066	136255.983	136539.513	33	A/G	G	G = 0.455/993	0.17	0.79	0.34
16	rs2278544	136262.580	136546.110	7	A/G	A	G = 0.492/1074	0.82	0.21	0.40
17	rs2304370	136278.205	136561.735	16	A/G	G	A = 0.254/555	0.11	0.38	0.20
18	rs3754689	136307.216	136590.746	29	C/T	C	T = 0.339/740	0.09	0.49	0.41
19	rs182549	136333.224	136616.754	26	C/T	C	T = 0.234/510	-	-	-
20	rs309152	136373.722	136657.252	40	A/G	G	G = 0.321/702	0.08	0.44	0.42
21	rs6430594	136435.643	136719.173	62	A/G	A	G = 0.198/432	0.09	0.09	0.16
22	rs309137	136482.421	136765.951	47	C/T	C	T = 0.376/821	0.79	0.02	0.35
23	rs2090660	136535.189	136818.719	53	C/T	C	T = 0.269/588	0.20	0.11	0.23
24	rs6430600	136552.835	136836.365	18	A/G	A	A = 0.343/749	0.29	0.62	0.19
25	rs12691874	136596.944	136880.474	44	A/G	G	A = 0.339/740	0.60	0.10	0.19
26	rs953387	136623.640	136907.170	27	A/C	A	T = 0.460/1004	0.71	0.56	0.16
27	rs1016269	136657.132	136940.662	33	A/G	G	A = 0.279/610	0.17	0.24	0.38
28	rs7371043	136693.406	136976.936	36	C/T	T	T = 0.158/344	0.08	0.18	0.30
29	rs4074120	136743.057	137026.587	50	C/T	T	C = 0.371/810	0.20	0.28	0.51
30	rs12465599	136791.320	137074.850	48	A/G	G	G = 0.439/958	0.61	0.55	0.27
31	rs6715450	136838.201	137121.731	47	A/G	A	A = 0.346/756	0.27	0.35	0.40
32	rs543721	136878.027	137161.557	40	G/T	G	T = 0.411/897	0.37	0.26	0.58
33	rs12618749	136921.944	137205.474	44	C/T	C	T = 0.228/497	0.09	0.20	0.41
34	rs16834591	136979.740	137263.270	58	A/G	A	A = 0.260/567	0.09	0.31	0.37
35	rs580879	137030.609	137314.139	51	C/T	T	T = 0.257/562	0.19	0.53	0.14

36	rs6711718	137123.482	137407.012	93	C/T	T	C = 0.451/985	0.50	0.48	0.20

5.5 Results

5.5.1 Distribution of the variation of *LCT* enhancer and flanking regions in a combined population set

Full data for the three sequenced regions (intron 4 and intron 13 of *MCM6* and the haplotype defining region hapdef of the *LCT* promoter) and the additional two haplotype markers 666 G>A and 5579 T>C were obtained from a final dataset of 872 samples. This sample set also contained 7 single individuals from different sample groups with rare enhancer SNPs, which were included in later haplotype analyses. 865 samples belonging to 28 populations were analysed with the Arlequin software and all variants were checked, in each of the 28 populations, for deviation from HWE, mainly to reassure the exclusion of sampling bias and genotyping problems. Only in four populations was a borderline *p*-value observed for single SNPs, which did not remain significant after Bonferroni correction for 28 tests.

Table 5.2 toTable 5.4 summarise the allele frequencies for the SNPs within the three sequenced regions and the two haplotype SNPs for the 28 populations.

Table 5.2: Allele frequencies for the SNPs in the *MCM6* intron 4 region in the combined sample set of 28 populations, N: Number of chromosomes.

Region	Country	Populations	N	rs1435577 -30210 G>C	- -30203 G>Del	- -30196 A>G	rs56263017 -30182 A>G	- -30160 A>T	- -30071/70 TC>AA	rs4988172 -29949 G>C
Africa	Cameroon	Mambila	40	0.950	-	-	0.200	-	-	0.075
	Ethiopia	Afar	124	0.677	-	-	0.242	-	-	-
		Amhara	80	0.588	0.025	-	0.250	-	-	-
		Oromo	124	0.500	-	-	0.210	-	-	0.024
		Shabo	44	0.841	-	0.023	0.318	0.091	-	0.159
	Ghana	Asante	40	0.825	-	-	0.200	-	-	0.175
	Malawi	Chewa	40	0.875	-	-	0.350	-	0.025	0.150
	Sudan	Beni Amer	130	0.800	-	-	0.346	-	0.008	0.008
		Jaali	76	0.750	-	-	0.329	-	-	0.026
	Tanzania	Chagga	82	0.646	-	-	0.159	-	-	0.110
Central Asia	Mongolia	Khalka	38	0.789	-	-	0.132	-	-	0.053
	Nepal	Tharu	40	0.750	-	-	0.125	0.025	-	0.050
Europe		Northern Europeans	40	0.900	-	-	0.125	-	-	0.050
	Italy	Italians	32	0.656	-	-	0.219	0.063	-	0.063
	Norway	Norwegians	40	0.975	-	-	0.100	-	-	-
	Romania	Romanians	40	0.675	-	-	0.275	0.025	-	0.025
	Ukraine	Ukrainians	40	0.675	-	-	0.075	-	-	0.125
Middle East	Iran	Iranians	78	0.579	-	-	0.355	0.026	-	0.026
	Israel	Israeli Arabs	40	0.575	-	-	0.200	0.075	0.025	0.050
		Israeli Bedouin	32	0.719	-	-	0.375	-	-	0.063
		Palestinians	38	0.632	-	0.026	0.316	-	0.026	0.105
	Jordan	Jordanian Bedouin	44	0.682	-	-	0.409	-	-	-
	Kuwait	Kuwaiti	62	0.714	-	-	0.519	-	-	0.018
	Saudi Arabia	Saudi Bedouin	40	0.725	-	-	0.500	-	-	-
	Syria	Syrians	82	0.512	-	-	0.232	-	-	0.037
	Turkey	Anatolian-Turks	40	0.500	-	-	0.200	0.050	-	0.125
	Yemen	Yemeni Hadramaut	156	0.756	0.013	-	0.467	0.006	-	0.026
		Yemeni Sena	68	0.853	-	-	0.382	0.074	-	0.060

Table 5.3: Allele frequencies for the SNPs in the *LCT* 'hapdef' region and two haplotype markers in the combined sample set of 28 populations, N: Number of chromosomes.

Region	Country	Populations	N	rs56064699 -958 C>T	-943 C>G	rs148142676 -943/42 TC>Del	-931 T>C	rs78205226 -875 G>A	-815 A>G	-811 A>G	rs56211644 -678 A>G	rs3754689 666 G>A	rs2278544 5579 T>C
Africa	Cameroon	Mambila	40	-	-	0.525	-	-	-	-	0.200	0.575	0.325
	Ethiopia	Afar	124	0.290	-	0.040	-	-	-	-	0.242	0.363	0.351
		Amhara	80	0.388	-	0.063	-	-	-	-	0.238	0.500	0.171
		Oromo	124	0.452	-	0.113	-	0.008	0.008	-	0.205	0.607	0.237
		Shabo	44	0.024	-	0.167	-	-	-	-	0.333	0.295	0.432
	Ghana	Asante	40	0.100	-	0.350	-	-	-	-	0.200	0.500	0.325
	Malawi	Chewa	40	0.025	-	0.300	-	-	-	-	0.375	0.425	0.300
	Sudan	Beni Amer	130	0.138	-	0.146	-	-	-	0.015	0.362	0.346	0.349
		Jaali	76	0.211	-	0.132	-	0.026	-	-	0.329	0.382	0.276
	Tanzania	Chagga	82	0.098	-	0.341	-	-	-	-	0.159	0.695	0.200
Central Asia	Mongolia	Khalka	38	0.211	-	0.211	-	-	-	-	0.132	0.375	0.500
	Nepal	Tharu	40	0.289	0.026	0.289	-	-	-	-	0.118	0.600	0.300
Europe		Northern Europeans	40	0.100	-	-	-	-	-	-	0.100	0.111	0.789
	Italy	Italians	32	0.313	-	-	-	0.063	-	-	0.250	0.333	0.467
	Norway	Norwegians	40	0.025	-	-	-	-	-	-	0.100	0.025	0.875
	Romania	Romanians	40	0.325	-	-	-	0.100	-	-	0.225	0.325	0.500
	Ukraine	Ukrainians	40	0.325	-	-	-	0.025	-	-	0.075	0.316	0.600
Middle East	Iran	Iranians	78	0.410	-	-	-	0.013	-	-	0.308	0.410	0.333
	Israel	Israeli Arabs	40	0.400	-	0.025	-	-	-	-	0.200	0.450	0.400
		Israeli Bedouin	32	0.250	-	0.031	-	-	-	-	0.375	0.313	0.313
		Palestinians	38	0.342	-	0.053	-	0.026	-	-	0.237	0.421	0.306
	Jordan	Jordanian Bedouin	44	0.318	-	-	-	0.068	-	-	0.409	0.341	0.250
	Kuwait	Kuwaiti	62	0.293	-	0.067	-	0.019	-	-	0.481	0.397	0.172
	Saudi Arabia	Saudi Bedouin	40	0.250	-	0.050	-	-	-	-	0.475	0.316	0.200
	Syria	Syrians	82	0.488	0.012	-	-	0.037	-	-	0.220	0.488	0.408
	Turkey	Anatolian-Turks	40	0.450	-	-	-	0.025	-	-	0.225	0.450	0.425
	Yemen	Yemeni Hadramaut	156	0.237	-	0.064	0.013	0.038	-	-	0.436	0.314	0.264
		Yemeni Sena	68	0.152	-	0.015	-	-	-	-	0.333	0.152	0.438

Table 5.4: Allele frequencies for the SNPs in the *LCT* enhancer region in the combined sample set of 28 populations, N: Number of chromosomes.

Populations	N	- -14028 T>C	rs4988233 -14011 C>T	rs145946881 -14010 G>C	ss820486563 -14009 T>G	ss820496524 -13957 A>G	rs41380347 -13915 T>G	rs41456145 -13913 T>C	rs4988235 -13910 C>T	rs41525747 -13907 C>G	ss820496565 -13806 A>G	rs144412793 -13800 G>T	- -13779 G>C	- -13744 C>G	rs4954492 -13730 T>G	rs56348046 -13603 C>T	rs4954490 -13495 C>T
Mambila	40	-	-	-	-	-	-	-	-	-	-	-	-	-	0.075	0.050	0.125
Afar	124	-	-	-	0.008	0.008	0.185	0.008	-	0.274	0.032	-	-	-	0.040	-	0.360
Amhara	80	-	-	-	0.038	-	0.050	0.013	-	0.025	0.025	-	-	-	0.050	-	0.241
Oromo	124	-	-	0.008	0.024	-	0.145	0.016	-	0.048	0.032	-	-	-	0.024	0.025	0.162
Shabo	44	-	-	-	-	0.023	-	-	-	-	-	0.023	-	-	0.023	-	0.114
Asante	40	-	-	-	-	-	-	-	-	-	-	-	-	-	0.075	-	0.075
Chewa	40	-	-	-	-	-	-	-	-	-	-	-	-	-	0.025	0.025	0.125
Beni Amer	130	-	-	-	0.138	-	0.262	-	-	0.169	-	-	-	-	0.047	-	0.230
Jaali	76	-	-	-	0.092	-	0.197	0.013	0.013	0.013	-	-	-	-	0.079	0.039	0.172
Chagga	82	-	-	0.146	-	-	-	-	-	-	-	0.012	-	-	0.061	-	0.063
Khalka	38	-	0.026	-	-	-	-	-	0.026	-	-	-	-	-	-	-	0.421
Tharu	40	-	-	-	-	-	-	-	0.075	-	-	-	-	-	-	-	0.211
Northern Europeans	40	0.025	-	-	-	-	-	-	0.650	-	-	-	-	-	-	-	0.737
Italians	32	-	0.031	-	-	-	-	-	0.031	-	-	-	-	-	-	-	0.344
Norwegians	40	-	-	-	-	-	-	-	0.850	-	-	-	-	-	-	-	0.875
Romanians	40	-	-	-	-	-	-	-	0.150	-	-	-	-	-	-	-	0.200
Ukrainians	40	-	0.025	-	-	-	-	-	0.300	-	-	-	-	-	-	-	0.444
Iranians	78	-	0.013	-	-	-	-	-	0.026	-	-	-	-	-	-	-	0.231
Israeli Arabs	40	-	-	-	0.025	-	0.025	0.050	0.025	-	0.025	-	-	-	0.025	-	0.200
Israeli Bedouin	32	-	-	-	-	-	0.188	-	0.031	-	-	-	-	-	0.031	0.031	0.219
Palestinians	38	-	-	-	-	-	0.026	-	0.026	-	-	-	-	-	-	-	0.237
Jordanian Bedouin	44	-	-	-	-	-	0.386	-	0.068	-	-	-	-	-	-	0.114	0.273
Kuwaiti	62	-	-	-	0.016	-	0.267	-	0.032	-	-	-	-	-	0.016	0.016	0.133
Saudi Bedouin	40	-	0.025	-	-	-	0.450	0.025	-	-	-	-	-	-	0.025	-	0.150
Syrians	82	-	-	-	-	-	0.037	-	-	-	-	-	-	-	-	-	0.244
Anatolian-Turks	40	-	-	-	-	-	-	-	0.075	-	-	-	-	-	-	-	0.125
Yemeni Hadramaut	156	-	-	0.006	-	-	0.237	-	0.019	0.013	0.006	-	0.032	0.006	0.006	-	0.221
Yemeni Sena	68	-	-	-	-	-	0.294	-	0.147	-	-	-	-	0.059	0.029	0.059	0.294

5.5.2 Population differentiation

Pairwise F_{ST} values were calculated in Arlequin with genotype data for the three sequenced regions and the 2 additional haplotype SNPs. Genotype data for the variant at -13495 were excluded because of too much missing data. Genetic distances between populations are visualised in the PCO plot of Figure 5.3, which shows the first two coordinates as calculated in R.

The Northern European populations are displaced to the right, whereas all other populations are clustered closer together to the left. The Middle Eastern populations are seen together with the Ethiopian and Sudanese samples in the upper part of the cluster. The Asante, Chewa, Chagga and Mambila, which are geographically more distant from Ethiopia and Sudan are-located away from the other African populations in the lower part of the plot.

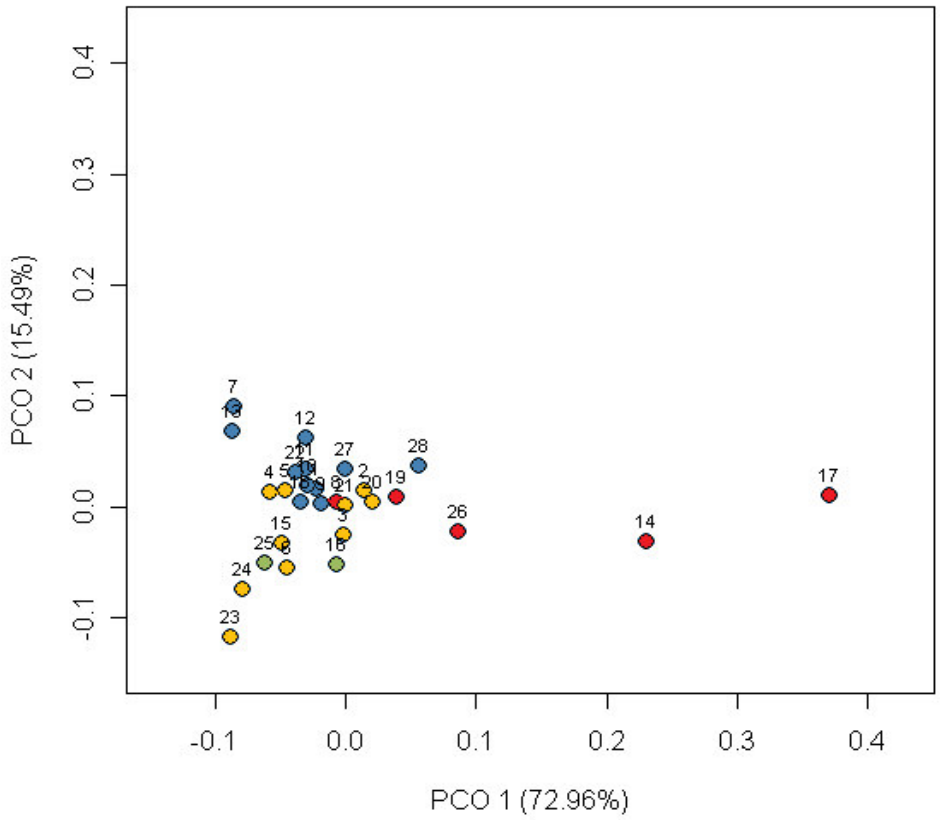


Figure 5.3: Principal co-ordinates plot of genetic distances between populations from pairwise F_{ST} values calculated from genotype data of *LCT* enhancer and the flanking regions. Populations included: 1: Anatolian-Turks, 2: Afar, 3: Shabo, 4: Amhara, 5: Oromo, 6: Asante, 7: Saudi Bedouin, 8: Italians, 9: Israeli Arabs, 10: Israeli Bedouin, 11:Iranians, 12: Jordanian Bedouin, 13: Kuwaiti, 14: Northern Europeans, 15: Chewa, 16: Khalka, 17: Norwegians, 18: Palestinians, 19: Romanians, 20: Beni Amer, 21: Jaali, 22: Syrians, 23: Mambila, 24: Chagga, 25: Tharu, 26: Ukrainians, 27: Yemeni Hadramaut, 28:Yemeni Sena. Colour code: red: European, green: Central Asian, blue: Middle Eastern, yellow: African populations.

5.5.3 Molecular diversity and neutrality tests

Diversity and neutrality measures were calculated using DnaSP with haplotype information of 855 individuals of 28 populations inferred by PHASE (see 5.5.4 below). A combined sequence of the readable sequence parts of 3 regions was compared: 371 bp of *MCM6* intron 4, 559 bp of the *LCT* enhancer region in intron 13 of *MCM6* and 471 bp of the haplotype defining region (hapdef) upstream of *LCT*.

Table 5.5: Haplotype diversity measures for 28 populations across the two control regions and the *LCT* enhancer. N: number of chromosomes, Comp.: Graphic comparison of the regions shows the values for haplotype diversity (Neis'*H*) for all populations examined. For this analysis -13945 C>T was included as it was possible to infer missing data at that position with PHASE. As previously shown in Jones (2012) but with a shorter enhancer sequence which notably excludes -13495, the non milk drinking African groups (Asante, Chewa, Mambila and Shabo) have a lower heterozygosity level across the enhancer compared to the two flanking regions. However lower heterozygosity of the enhancer can be also seen in some of the European and Asian and Middle Eastern groups as Italians, Romanians, Tharu, Anatolian Turks, Israeli Arabs, Palestinians and Syrians.

A full table with all diversity measures and outcomes of neutrality tests can be found in Appendix C1. Significant values for deviation from neutrality were only obtained for the Israeli Arabs for the enhancer region and the Jordanian Bedouins for the whole sequence for Fu and Li's *F* statistic and the Yemen Sena for Fu and Li's *D* for the whole sequence. It is noticeable that values for Tajima's *D* are usually much lower (although not significantly) in the enhancer than the two flanking regions, except in some European populations such as the Norwegians, Northern Europeans, Ukrainians and Romanians, and also the Beni Amer of Africa.

Table 5.5: Haplotype diversity measures for 28 populations across the two control regions and the *LCT* enhancer. N: number of chromosomes, Comp.: Graphic comparison of the regions.

			Haplotype diversity (Neis'*H*) (± SD)			
Region	**Population**	**N**	**Intron 4**	**Enhancer**	**Hapdef**	**Comp.**
Africa	Afar	124	0.653 (0.000)	0.754 (0.000)	0.644 (0.000)	
	Amhara	80	0.662 (0.000)	0.600 (0.003)	0.661 (0.003)	
	Asante	40	0.714 (0.002)	0.273 (0.008)	0.472 (0.006)	
	Beni Amer	128	0.654 (0.000)	0.786 (0.000)	0.612 (0.001)	
	Chagga	82	0.703 (0.001)	0.459 (0.004)	0.417 (0.004)	
	Chewa	40	0.733 (0.001)	0.273 (0.008)	0.512 (0.002)	
	Jaali	76	0.682 (0.000)	0.752 (0.001)	0.654 (0.001)	
	Mambila	40	0.509 (0.006)	0.355 (0.008)	0.328 (0.006)	
	Oromo	124	0.643 (0.001)	0.598 (0.002)	0.655 (0.000)	
	Shabo	42	0.791 (0.001)	0.297 (0.007)	0.487 (0.003)	
Central Asia	Khalka	38	0.585 (0.005)	0.542 (0.003)	0.519 (0.006)	
	Tharu	36	0.668 (0.003)	0.341 (0.009)	0.595 (0.003)	
Europe	Italians	32	0.752 (0.001)	0.504 (0.006)	0.702 (0.002)	
	N. Europeans	40	0.458 (0.008)	0.522 (0.005)	0.349 (0.008)	
	Norwegians	40	0.229 (0.007)	0.268 (0.007)	0.229 (0.007)	
	Romanians	40	0.713 (0.001)	0.344 (0.007)	0.704 (0.002)	
	Ukrainians	40	0.664 (0.002)	0.627 (0.003)	0.558 (0.004)	
Middle East	Anatolian Turks	40	0.694 (0.003)	0.232 (0.007)	0.679 (0.001)	
	Iranians	76	0.675 (0.001)	0.362 (0.003)	0.674 (0.000)	
	Israeli Arabs	40	0.736 (0.002)	0.555 (0.007)	0.656 (0.001)	
	Israeli Bedouin	32	0.720 (0.001)	0.641 (0.006)	0.677 (0.001)	
	Jordanian Bedouin	44	0.672 (0.001)	0.725 (0.002)	0.707 (0.001)	
	Kuwaiti	56	0.647 (0.002)	0.639 (0.003)	0.662 (0.001)	
	Palestinians	38	0.758 (0.001)	0.422 (0.007)	0.684 (0.001)	
	Saudi Bedouin	40	0.640 (0.002)	0.674 (0.002)	0.653 (0.001)	
	Syrians	82	0.656 (0.001)	0.427 (0.002)	0.680 (0.001)	
	Yemeni Hadramaut	154	0.660 (0.000)	0.675 (0.001)	0.680 (0.000)	
	Yemeni Sena	66	0.716 (0.001)	0.770 (0.001)	0.610 (0.001)	

The comparison of haplotype diversity between lactose digesters and non-digesters of European and African populations shows a clear difference between these populations as well as between the three sequence regions in phenotyped Ethiopians that we have published so far (Jones et al. 2013). Whereas heterozygosity is reduced across the *LCT* enhancer compared to the flanking regions in the non-digester Ethiopian groups, as previously reported, it is *not* different for the Europeans. Furthermore, compared to Africans groups, which show similar diversity in the digesters and non digesters, the Europeans show a higher diversity of the flanking regions in the non-digesters than in the digesters (see Appendix C1). This difference in pattern seems to reflect the two different types of selection acting on that region.

The greater haplotype diversity of the enhancer in the digester Africans than in the non-digesters, is likely to be caused by the co-occurrence of several functional alleles in the digesters due probably to their co-selection (so-called soft selective sweep) but could possibly be due to a reduction in purifying selection. The lower haplotype diversity in the regions flanking the enhancer in the European lactose digesters as compared with the non-digesters instead reflects the much talked about 'hard' selective sweep, caused by increased homozygosity of the extended haplotype carrying the single selected mutation. A full table with results of neutrality tests and diversity measures compared between digesters and non-digesters can be found in Appendix C2.

5.5.4 Haplotype analysis of the *LCT* enhancer and flanking regions

Haplotypes were reconstructed with PHASE from the combined genotype data for the tree sequencing regions (intron 4 and intron 13 of *MCM6* and the hapdef region of the *LCT* promoter) and the two additional haplotype markers 666 G>A and 5579 T>C, spanning a region of about 80 kb, see Figure 5.4. Positions with alleles occurring only once, as well as samples with missing data at more than 4 of the remaining 31 positions were excluded from the PHASE analysis. This left 862 individuals for haplotype analysis, including 7 individuals carrying rare alleles *(-14028*C, -14011*T* and-*13914*A*) who were selected for inclusion and did not strictly belong to any of the 28 population samples.

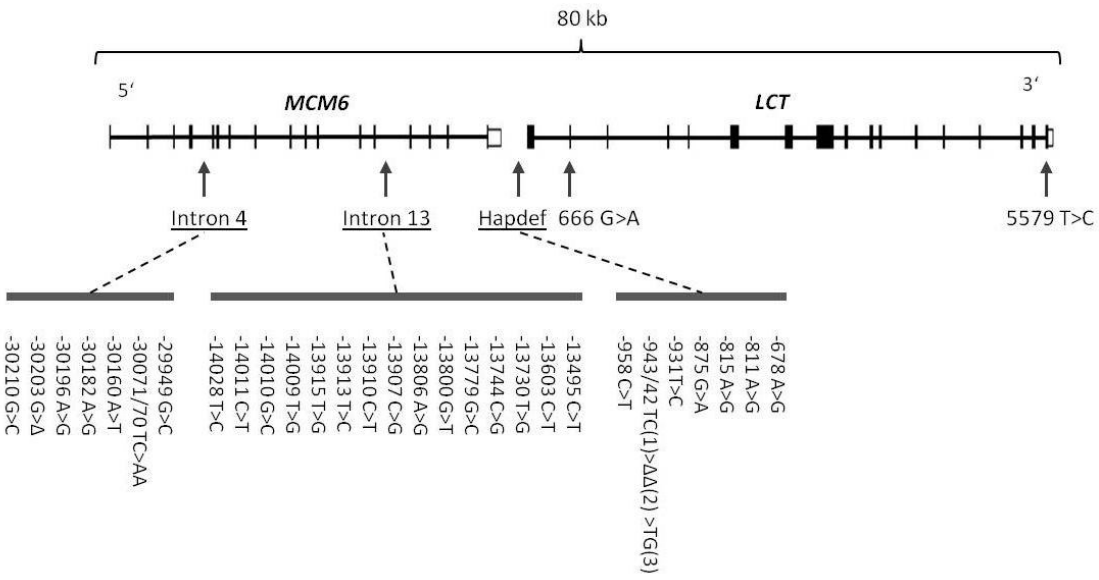


Figure 5.4: Variants included in haplotype analysis of the 80 kb region and their position in relation to the sequenced regions.

Eight of the 11 core haplotype markers, previously defined by Hollox et al. (2001), were used to define the *LCT* haplotype background. One of the set of markers was also the SNP at -946 A>G, but as all samples were homozygous A at that position this was not included in PHASE analysis. The haplotypes inferred by PHASE, especially the rarer ones, were inspected by eye for anomalies and obvious wrong assignments, which left 76 haplotypes (HT) for further analysis. Common haplotypes, linked to the *LCT* core haplotypes, are shown in Table 5.6 and the appearance of the most frequent haplotypes in the populations tested is shown in Table 5.7.

The most frequent haplotype found in this 'global' population is that previously described as B (HT 71 containing *-30210*C*, which does not carry any derived enhancer allele). The B haplotype is relatively frequent across all four geographic regions but most frequent in the Middle East and Asia and the Afro-Asiatic speaking African groups tested (see Figure 5.5 and Table 5.7). The C haplotype containing the derived allele *-13915*G* (HT 56, *-13915*G*, *-30182*G*) is most common in the Saudi and Jordanian Bedouin and only found in the Middle East, Ethiopia and Sudan whereas the ancestral C haplotype (HT 43, *-30182*G*) is present in all populations with highest frequencies in the Shabo and Chewa. The ancestral A haplotype (HT 10) carrying the *-13495*T* allele can be also found across all populations and is most frequent in the Khalka of Mongolia. As expected the *-13910*T* carrying A haplotype (HT 24, *-13910*T*, *-13495*T*) is most prevalent in the Northern Europeans but also present in other European, Asian and Middle Eastern populations, except the Saudi Bedouin and Syrians. In the Africans in this dataset it can be only found in the Jaali, whereas the *-13907*G* carrying A haplotype (HT 21, *-13907*G*, *-13495*T*) is restricted to the Afro-Asiatic speaking African populations and the Yemeni from Hadramaut.

Table 5.6: Inferred haplotypes by PHASE. Low frequency haplotypes (<5) are not shown. Haplotypes in bold are the most probable equivalents (Hollox et al. 2001). N: number of chromosomes. The colour scheme is consistent with the pie charts and network diagram below.

Haplotype ID	-30210 G>C	-30203 G>Del	-30196 A>G	-30182 A>G	-30160 A>T	-30071/70 TC>AA	-29949 G>C	-14028 T>C	-14011 C>T	-14010 G>C	-14009 T>G	-13915 T>G	-13913 T>C	-13910 C>T	-13907 C>G	-13806 A>G	-13800 G>T	-13779 G>C	-13744 C>G	-13730 T>G	-13603 C>T	-13495 C>T	-958 C>T	-943/42 TC>Del>TG (1 TC, 2 DEL)	-931T>C	-875 G>A	-815 A>G	-811 A>G	-678 A>G	666 G>A	5579 T>C	N	Haplotypes
71	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	A	.	388	**B**, G
56	.	.	.	G	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	204	**C**, M, j, k
43	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	200	**C**, M, j, k
10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	173	**A**, J
24	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	110	**A**, J
7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	A	.	83	O, S, **U**, f, g, p
21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	63	**A**, J
31	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	56	**A**, J
66	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	48	**P**, W, Y
17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	41	H, K, Q, Z
26	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	A	C	33	**X**, m, n
6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	A	C	30	**X**, m, n
42	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	C	27	E
75	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	A	.	.	.	A	.	26	D
15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	.	.	C	19	**A**, J
37	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	19	**A**, J
11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	17	H, K, Q, Z
82	C	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	14	**P**, W, Y
49	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	14	**A**, J
33	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	A	C	14	**X**, m, n
70	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	A	C	13	F, **I**
53	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	12	**C**, M, j, k
2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	10	**A**, J
80	C	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	A	.	8	**B**, G
12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	6	**P**, W, Y
30	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	6	**A**, J
52	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	5	**C**, M, j, k
34	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	A	.	5	O, S, **U**, f, g, p
41	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	H, K, Q, Z

Table 5.7: Distribution of the most common haplotypes across 28 populations studied. Abbreviations for language families: NC: Niger-Congo, AA: Afro-Asiatic, NS: Nilo-Saharan, A: Altaic, IE: Indo-European. N: Number of chromosomes.

Region	Country	Populations	Lang. family	N	71 (**B**,G)	56 (**C**, M, j, k) -30182*G, -13915*G	43 (**C**, M, j, k) -30182*G	10 (**A**, J) -13495*T	24 (**A**,J) -13910*T, -13495*T	7 (O, S, **U**, f g, p)	21 (**A**,J) -13907*G, -13495*T	31 (**A**,J) -29949*C	66 (**P**, W, Y) -30210*C	17 (**H**, K, Q, Z) -13730*G	26 (**X**, m, n) -14009*G	6 (**X**, m, n)
Africa	Cameroon	Mambila	NC	40	-	-	0.200	0.075	-	0.325	-	0.025	0.050	0.075	-	0.150
	Ethiopia	Afar	AA	124	0.274	0.185	0.024	0.065	-	0.024	0.242	-	0.032	0.032	0.008	0.008
		Amhara	AA	80	0.375	0.050	0.138	0.125	-	0.013	0.025	-	0.025	0.050	0.038	0.013
		Oromo	AA	124	0.403	0.145	0.024	0.065	-	0.032	0.048	0.024	0.032	0.024	0.024	0.032
		Shabo	NS	42	0.024	-	0.310	0.071	-	0.071	-	0.143	0.071	0.024	-	0.071
	Ghana	Asante	NC	40	0.100	-	0.175	0.075	-	0.175	-	0.075	0.075	0.075	-	0.050
	Malawi	Chewa	NC	40	0.025	-	0.325	0.100	-	0.125	-	0.050	0.100	0.025	-	0.075
	Sudan	BeniAmer	AA	128	0.125	0.266	0.078	0.047	-	-	0.172	0.008	0.063	0.039	0.133	0.008
		Jaali	AA	76	0.171	0.197	0.118	0.053	0.013	0.013	0.013	0.026	0.039	0.079	0.092	0.026
	Tanzania	Chagga	NC	82	0.085	-	0.146	0.049	-	0.195	-	0.037	0.110	0.049	-	0.049
Central	Mongolia	Khalka	A	38	0.211	-	0.105	0.342	0.026	0.211	-	0.026	-	-	-	-
Asia	Nepal	Tharu	IE	36	0.250	-	0.111	0.111	0.083	0.278	-	0.056	-	-	-	-
Europe		N. Europeans	IE	40	0.075	-	0.075	0.050	0.650	-	-	0.025	-	-	-	-
	Italy	Italians	IE	32	0.219	-	0.156	0.219	0.031	-	-	0.063	-	-	-	-
	Norway	Norwegians	IE	40	0.025	-	0.100	0.025	0.850	-	-	-	-	-	-	-
	Romania	Romanians	IE	40	0.225	-	0.150	0.050	0.150	-	-	0.025	-	-	-	-
	Ukraine	Ukrainians	IE	40	0.300	-	0.075	0.150	0.300	-	-	0.125	-	-	-	-
Middle	Iran	Iranians	IE	76	0.395	-	0.263	0.145	0.026	-	-	0.026	-	-	-	-
East	Israel	Israeli Arabs	AA	40	0.325	0.025	0.100	0.175	0.025	-	-	0.050	0.025	0.025	0.025	-
		Israeli Bedouin	AA	32	0.250	0.188	0.125	0.125	0.031	0.031	-	0.063	0.031	0.031	-	-
		Palestinians	AA	38	0.316	0.026	0.105	0.158	0.026	0.026	-	0.105	0.026	-	-	-
	Jordan	Jordanian Bedouin	AA	44	0.250	0.386	0.023	0.068	0.068	-	-	-	-	-	-	-
	Kuwait	Kuwaiti	AA	56	0.268	0.268	0.161	0.054	0.036	0.036	-	-	-	0.018	0.018	-
	Saudi Arabia	Saudi Bedouin	AA	40	0.150	0.425	0.050	0.100	-	0.025	-	-	0.050	0.025	-	0.025
	Syria	Syrians	AA	82	0.390	0.037	0.122	0.232	-	-	-	0.024	-	-	-	-
	Turkey	AnatolianTurks	A	40	0.375	-	0.150	0.050	0.075	-	-	0.125	-	-	-	-
	Yemen	Yemeni Hadramaut	AA	154	0.182	0.234	0.117	0.136	0.019	0.045	0.013	0.013	0.006	0.006	-	0.006
		Yemeni Sena	AA	66	0.152	0.212	0.061	0.076	0.136	-	-	0.061	-	0.030	-	0.015

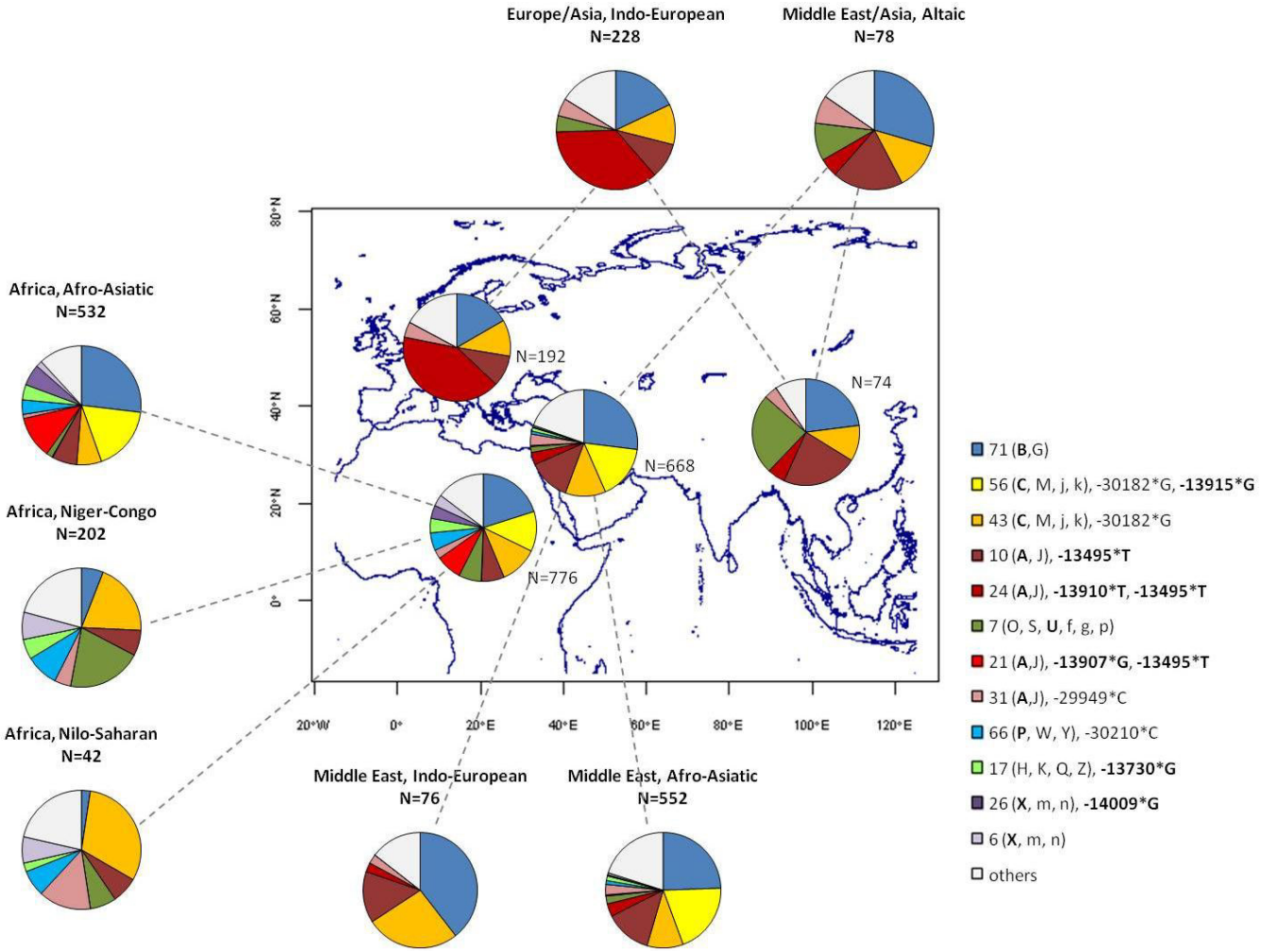


Figure 5.5: Distribution of haplotypes in Africa, Europe, Middle East and Central Asia (within the map) as well as haplotype distribution, classified by language groups. N: Number of chromosomes. Derived enhancer alleles and most probable haplotypes are shown in bold. The colour scheme broadly corresponds to that used in Table 5.6.

5.5.5 Haplotype association of derived *LCT* enhancer alleles

Table 5.6 and Table 5.8 show the relationship of the derived *LCT* enhancer alleles to their haplotype background and to the most common 80 kb haplotypes investigated. Consistent with previous studies (Bersaglieri et al. 2004; Coelho et al. 2005; Mulcare 2006; Poulter et al. 2003) nearly all *-13910*T* alleles were on the same 80 kb A haplotype (HT 24, *-13910*T*, *-13495*T*). Only one **T* allele resided on a P haplotype (or W, Y), which is most likely caused by a recombination event between the hapdef region (-678 A>G) and exon 2 (666 G>A) of *LCT,* as the haplotype is the same as HT 24 (A, *-13910*T*, *-13495*T*) up to position -678, as indicated in Table 5.8. Myles et al. (2005) found 8 similar cases in Moroccan and Algerian Berber populations.

With the combination of markers used in this thesis it was possible to distinguish between B and P haplotypes and to re-evaluate the occurrence of *-14010*C* in the previously investigated African samples (Ingram et al. 2009b), as also shown in Jones et al. (2013), and include a further 11 cases. *-14010*C* exclusively lies on a P haplotype (HT 82) background.

The vast majority of *-13915*G* alleles have the C haplotype as background (HT 56), which also seems to provide the background for the newly discovered rarer variant *-13944*G* (HT 59). However in 3 cases *-13915*G* was found on other haplotypes (E, H and L, as shown in Table 5.8).

Except for three alleles on the ancestral haplotype H, *-13907*G* is located on an A haplotype background (HT 21), as is*-13603*T* (HT 15) and most of the *-14011*T* variants. However two *-14011*T*s were found on different B haplotype backgrounds and if the assignments are correct that might suggest this mutation happened several times independently, probably in geographically quite distinct places.

*-14009*G* mostly occurred on an X haplotype (HT 26) and one on the ancestral H, which is also the background of the majority of the *-13730*G* variants, though 3 cases were found on A and C haplotypes. The variants *-13806*G* and *-13779*C* exclusively occur on the C haplotype background (HT 53 and 52 respectively). *-13913*C* resides on the B haplotype (HT 80).

The location of *-14028*C,* the variant described in functional studies in the previous chapter, on a B haplotype, was confirmed for both cases. Another rarer variant, *-13800*T*, has U as background haplotype.

To summarise, the derived enhancer alleles in most cases lie on a characteristic haplotype:*-13910*T* on A haplotype as well as *-13907*G* and *-13603*T, -13913*C* on a B haplotype,*-13915*G* and *-13806*G* on a C haplotype, *-14010*C* on a P haplotype, *-13730*G* on an H haplotype and*-14009*G* on an X haplotype (Table 5.6 and Table 5.8). However *-13495*T*, although mostly on A, can in contrast be found on several different haplotype backgrounds, and seems to be evolutionarily much older, and there has clearly been some recombination between it and markers within *LCT* (see also network and LD analyses below).

These results broadly confirm previous findings (Ingram 2008; Jones 2012) but extends them by providing much more information on *-13495*T* and the rare haplotypes, as well as more definitive assignment of *-14009*G* and *-14010*C*.

Table 5.8: Haplotype backgrounds of the lactase persistence associated enhancer alleles. Lactase persistence associated alleles are shown in bold, boxes represent possible recombination events. N: Number of chromosomes. Note that haplotype 25 is like haplotype A for much of the lenght.

Haplotype ID	-30210 G>C	-30203 G>Del	-30196 A>G	-30182 A>G	-30160 A>T	-30071/70 TC>AA	-29949 G>C	-14028 T>C	-14011 C>T	-14010 G>C	-14009 T>G	-13915 T>G	-13913 T>C	-13910 C>T	-13907 C>G	-13806 A>G	-13800 G>T	-13779 G>C	-13744 C>G	-13730 T>G	-13603 C>T	-13495 C>T	-958 C>T	-943/42 TC>Del>TG (1 TC, 2 DEL)	-931T>C	-875 G>A	-815 A>G	-811 A>G	-678 A>G	666 G>A	5579 T>C	N	Haplotypes
	-14010*C																																
82	C	1	A	A	A	TC	G	T	C	**C**	T	T	T	C	C	A	G	G	C	T	C	C	C	1	T	G	A	A	A	A	T	14	P, W, Y
	-14009*G																																
26	G	1	A	A	A	TC	G	T	C	G	**G**	T	T	C	C	A	G	G	C	T	C	C	C	2	T	G	A	A	A	A	C	33	X, m, n
27	G	1	A	A	A	TC	G	T	C	G	**G**	T	T	C	C	A	G	G	C	T	C	C	C	2	T	G	A	A	A	A	T	1	O, S, U, f, g, p
	-13915*G																																
56	G	1	A	G	A	TC	G	T	C	G	T	**G**	T	C	C	A	G	G	C	T	C	C	C	1	T	G	A	A	G	G	T	204	C, M, j, k
59	G	1	A	G	A	TC	G	T	C	G	T	**G**	T	C	C	A	G	G	G	T	C	C	C	1	T	G	A	A	G	G	T	4	C, M, j, k
55	G	1	A	G	A	TC	G	T	C	G	T	**G**	T	C	C	A	G	G	C	T	C	C	C	1	T	G	A	A	G	G	C	1	E
58	G	1	A	G	A	TC	G	T	C	G	T	**G**	T	C	C	A	G	G	G	T	C	C	C	1	T	G	A	A	A	G	T	1	H, K, Q, Z
57	G	1	A	G	A	TC	G	T	C	G	T	**G**	T	C	C	A	G	G	C	T	C	C	T	1	T	G	A	A	A	G	T	1	L, i
	-13910*T																																
24	G	1	A	A	A	TC	G	T	C	G	T	T	T	**T**	C	A	G	G	C	T	C	T	C	1	T	G	A	A	A	G	C	110	A, J
25	G	1	A	A	A	TC	G	T	C	G	T	T	T	**T**	C	A	G	G	C	T	C	T	C	1	T	G	A	A	A	A	T	1	P, W, Y
	-13907*G																																
21	G	1	A	A	A	TC	G	T	C	G	T	T	T	C	**G**	A	G	G	C	T	C	T	C	1	T	G	A	A	A	G	C	63	A, J
20	G	1	A	A	A	TC	G	T	C	G	T	T	T	C	**G**	A	G	G	C	T	C	C	C	1	T	G	A	A	A	G	C	1	A, J
22	G	1	A	A	A	TC	G	T	C	G	T	T	T	C	**G**	A	G	G	C	T	C	T	C	1	T	G	A	A	A	G	T	3	H, K, Q, Z

5.5.6 Haplotype networks

A network was constructed using the Network software and the haplotype information as shown in Table 5.6. The Network is made by assuming single stepwise mutational changes. Figure 5.6 illustrate the relationships of the different haplotypes to each other and locates the derived enhancer and intron 4 alleles of *MCM6*. With more than one appearance of a mutation, recombination can be inferred, as indicated by the ovals. The two exonic SNPs of *LCT* play a large role in haplotype distinction. The network indicates recombination events that could have occurred, especially between the enhancer region and the 5579 SNP in exon 17 of *LCT*, leading to less frequent haplotypes. Most of these events are located between the *LCT* SNPs and position -13495.

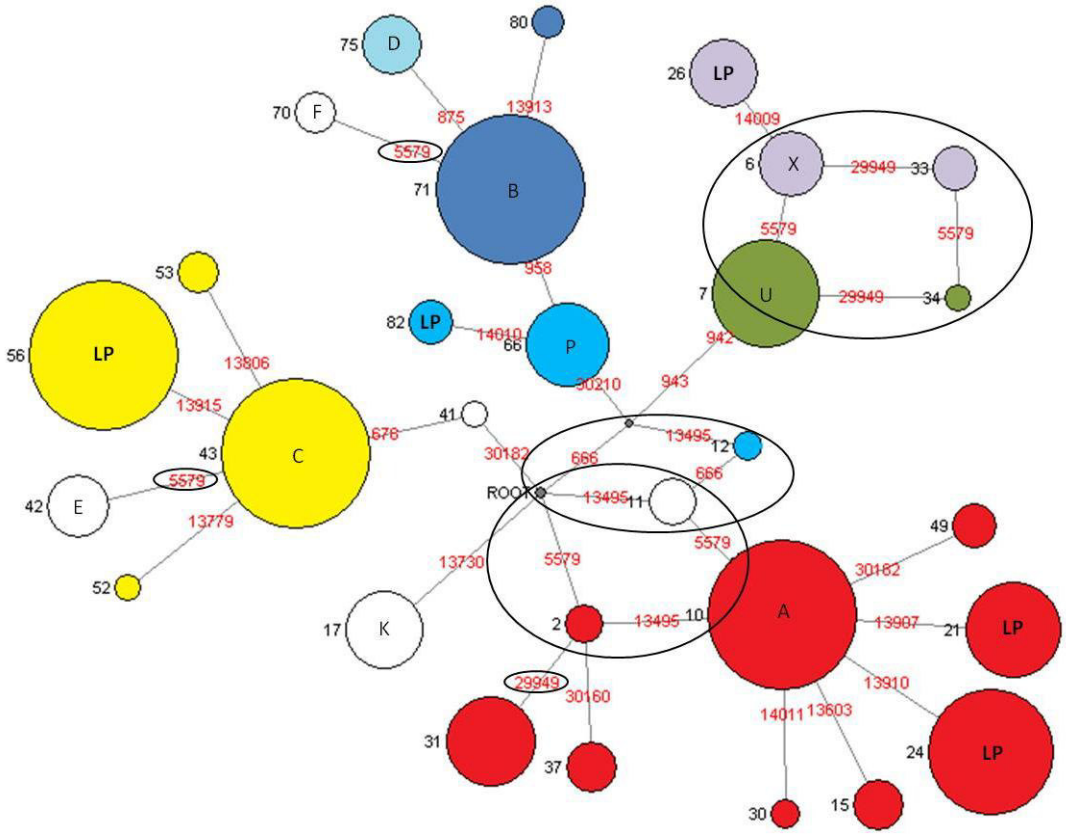


Figure 5.6: Maximum parsimony neighbour joining network of the haplotypes shown in Table 5.6. Haplotype nomenclature correspondents to the haplotype IDs of Table 5.6 and nodes are proportional to haplotype frequency. Names of the most probable *LCT* core haplotypes and those associated with lactase persistence (LP) are indicated, colours represent similar *LCT* core haplotypes and circles indicate recombination events. Note that branch lengths do not represent evolutionary scales.

5.5.7 Linkage disequilibrium

Pairwise linkage disequilibrium across the 80 kb haplotype region was examined using PowerMarker software and values were also checked in DnaSP. The *D'* measure was calculated with phased haplotype information of 855 individuals. These were grouped in various ways and association patterns differed slightly between the groups examined, partly reflecting the alleles present. An example of the pairwise calculations is shown in Table 5.9, all others in Appendix F.

Linkage disequilibrium was consistently high (*D'* of 1) between the *MCM6* intron 13 markers, with the exception of the association of -13945 and the two close SNPs at -13907 and -13915 in Afro-Asiatic speaking groups. The lack of statistical significance for the association of some pairs of markers with *D*' values of 1 (no evidence of recombination) results from low allele frequencies.
A noticeable drop in LD between the 5579 SNP and the markers of the hapdef region is seen in most groups, and also some markers of *MCM6* intron 4. Only in the African groups were there reduced *D'* values between 5579 and some enhancer variants, probably due to the higher variability of the enhancer in these groups. *D'* values were lower between -13495 and at least one of the *MCM6* intron 4 SNPS in all language groups.
The Nilo-Saharan population additionally showed reduced *D*' values between of both exonic markers 5579 and 666 and the -13495 enhancer variant, undoubtedly because the -*13495 *T* is found on other core haplotypes than A. Both African-Asiatic groups showed a drop in LD between the haplotype markers -958 and -943/942 as well as the Niger-Congo populations between 666 and -943/942.

A more useful graphical way of presenting a picture of historic recombinations is the construction of Linkage Disequilibrium Unit (LDU) maps as was done for extended haplotype analysis as shown below (section 5.5.8.3).

Table 5.9: Pairwise linkage disequilibrium *D'* across the 80kb haplotype region. Monomorphic markers were excluded from analysis. Statistically significant values after Bonferroni correction for multiple testing for the respective amount of populations per group are shaded in green.

Africa Afro-Asiatic

	-30210	-30203	-30182	-30071/70	-29949	-14010	-14009	-13915	-13913	-13910	-13907	-13806	-13730	-13603	-13495	-958	-943/42	-875	-815	-811	-678	666
-30203	1																					
-30182	1	0.54																				
-30071/70	1	1	1																			
-29949	1	1	1	1																		
-14010	1	1	1	1	1																	
-14009	1	1	1	1	1	1																
-13915	1	1	1	1	1	1	1															
-13913	1	1	1	1	1	1	1	1														
-13910	1	1	1	1	1	1	1	1	1													
-13907	1	1	1	1	1	1	1	1	1	1												
-13806	1	1	1	1	1	1	1	1	1	1	1											
-13730	1	1	1	1	1	1	1	1	1	1	1	1										
-13603	1	1	1	1	1	1	1	1	1	1	1	1	1									
-13495	1	0.14	1	1	1	1	1	1	1	1	0.98	1	1	1								
-958	0.99	1	1	1	1	1	1	1	1	1	1	1	1	1	1							
-943/42	0.89	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.94						
-875	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
-815	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
-811	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
-678	1	0.54	0.99	1	1	1	1	1	1	1	1	1	0.85	1	0.97	1	1	1	1	1		
666	1	1.0	1	1	0.68	1	1	1	1	1	1	1	1	1	0.93	1	1	1	1	1	1	
5579	0.96	0.06	0.95	1	0.80	1	0.96	1	1	1	0.93	1	0.86	1	0.78	0.96	0.65	1	1	1	0.95	0.40

5.5.8 Extended haplotype analyses

To investigate how the *LCT* haplotypes carrying the lactase persistence associated enhancer alleles differ from their ancestral forms and how far they expand beyond *LCT* in the combined sample set, additional markers were selected (see 5.4.1). They extend the haplotype analysis across a region of about 1.8 Mb. For 880 individuals 36 markers were genotyped by LGC Genomics. After quality control the genotype data of this set were merged with the data for *MCM6* intron 4 and 13 and the *LCT* hapdef region of the corresponding individuals. All individuals with missing data at more than 6 positions across the whole set or more than 3 positions in the hapdef region, as well as SNPs occurring twice or less were excluded from further analysis. Haplotypes were inferred using PHASE for a final set of 60 SNPs for 837 individuals.

5.5.8.1 Haplotype background the derived enhancer alleles

For comparison the haplotype lengths between chromosomes carrying the ancestral and derived variants on the same haplotype core background A, B, C, P and X (as defined by Hollox et al. 2001) were calculated in bp and plotted.

Figure 5.7 shows all A haplotypes carrying the derived *-13910*T* variant (N=104). The region of about 310 kb is identical for all **T* carrying chromosomes (except one with the ancestral G allele at -22 kb) and spans from SNP rs6430594 located in the *DARS* gene upstream of *LCT* up to the polymorphism rs4954278 in the gene *R3HDM1*.

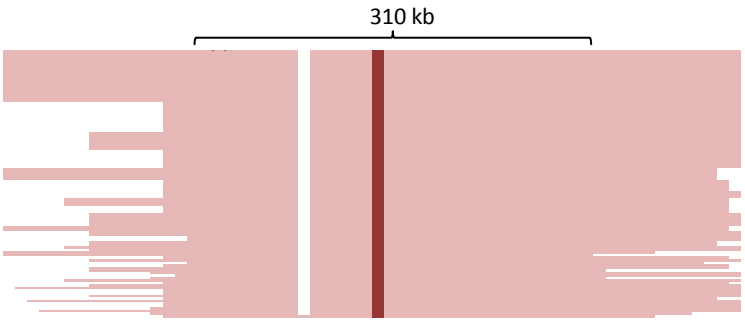


Figure 5.7: Graphical representation of *-13910*T* (red) carrying haplotypes. *-22018*A* (left white bar) is on the same haplotype background. All sequences are identical for at least 310 kb.

As a further example, Figure 5.8 shows the graphical representation of all X haplotypes with ancestral and derived variants at position -14009. The length differences between the derived and ancestral forms can clearly be seen.

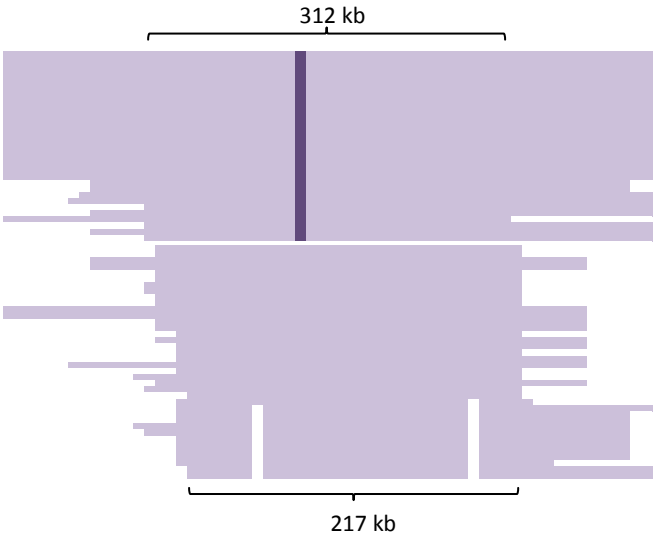


Figure 5.8: Plot of X haplotypes carrying *-14009*G* (above, with lilac bars) and the ancestral version (below). Kb values represent the minimum haplotype length. The white left bar in the ancestral haplotypes represent *-29949*C* and the right bar rs2304370, which is located between the *LCT* haplotype defining SNPs at position 666 and 5579. Gene conversion and/or recombination can be inferred for these haplotypes particularly from -29949 upstream which would shorten the minimum length of the ancestral haplotype to 55 kb and the mean to 655 kb.

The mean lengths were calculated for all haplotypes carrying the derived allele versus all others in calculating the haplotype length of each chromosome as sequence region shared with the most frequent haplotype (for example shown as broad upper bar in Figure 5.7), and results are shown in Table 5.10. For comparison of each of the two groups, a two tailed Mann-Whitney U test was performed. The mean lengths of the haplotypes carrying the derived lactase persistence associated alleles are significantly longer (p<0.001) than those carrying the ancestral alleles on the SNP positions.
Interestingly, the ancestral B haplotype is also relatively long and the *-13913*C* variant, which is not associated with lactose digester status, also occurs on a slightly longer B haplotype, though this difference is not statistically significant (p=0.68). This smaller difference presumably indicates that this non-functional mutation is much older than the others.

Table 5.10: Mean haplotype length of core haplotypes carrying derived and ancestral alleles.

Haplotype	variant	Haplotype length (kb) derived	ancestral
A	*-13910*T*	1342	658
	*-13007*G*	1580	
C	*-13915*G*	1473	549
X	*-14009*G*	1578	675
P	*-14010*C*	1417	849
B	*-13913*C*	1076	1004

5.5.8.2 Extended haplotype homozygosity (EHH)

Previous studies of long range haplotypes across the *LCT* region were mainly done considering the -13910 C>T to be functional for lactase persistence (Bersaglieri et al. 2004; Sabeti et al. 2007). These and other lactase persistence associated alleles were compared to all haplotypes carrying the ancestral variant at that position (Enattah et al. 2008; Tishkoff et al. 2007), i.e. all other *LCT* core haplotypes combined together. In a more recent study we conducted a long range haplotype test on the A carrying *13910*T* haplotype in comparison with ancestral A, B and C (Gallego Romero et al. 2012). Remarkably, the B haplotype carrying none of the associated alleles showed a long stretch (500 kb) of relatively high extended haplotype homozygosity (EHH).

To investigate selection patterns in the region around *LCT* in the combined dataset, EHH tests were first conducted using the Sweep software for the different lactase persistence associated alleles as reported previously by others, using the single derived allele as core and the EHH values plotted in R (see also section 2.3.6 of chapter 2). Figure 5.9 toFigure 5.11 show the plots of for the decay of EHH against genetic distance from the core SNPs. It should be noted that all plots illustrate are shown in the chromosomal + direction, as also in Figure 5.2, which is the opposite way round to the previous graphs and tables of this chapter, which are shown in the direction of orientation of the *LCT* and *MCM6* genes.

Haplotypes carrying the derived lactase persistence associated alleles were analysed against all other haplotypes of the whole sample set, except for *-14009*G* and *-14010*C* where analyses were done on a subset of samples because of their low frequency in the overall sample set, on the African Afro-Asiatic speaking (522 chromosomes) and Niger-Congo speaking (202 chromosomes) groups respectively. To get more information about the B haplotype the same was done for the -958 SNP as core (Figure 5.10).

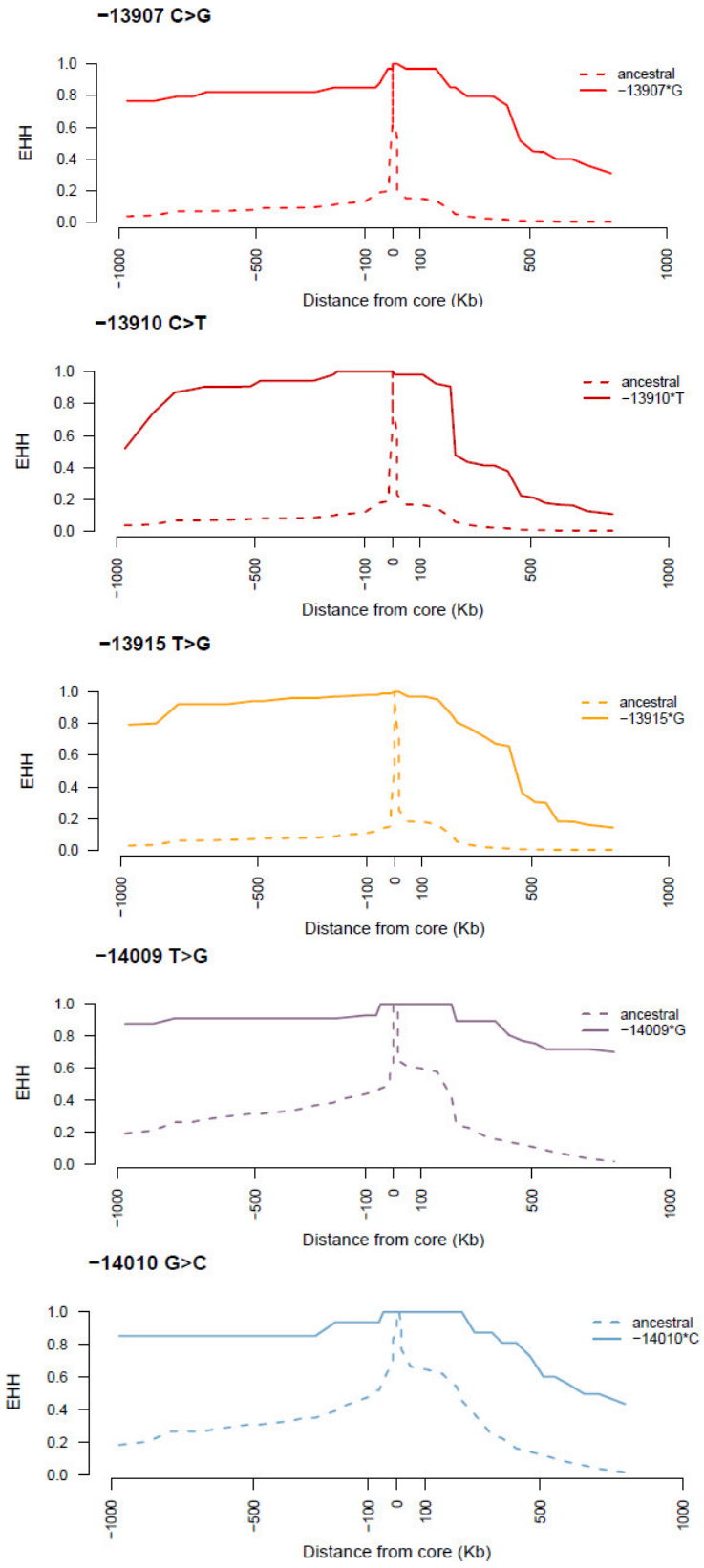


Figure 5.9: Decay of extended haplotype homozygosity for the haplotypes carrying derived LP associated alleles of the *LCT* enhancer compared to those carrying ancestral alleles. For varians at-13907, -13910 and -13915 these were compared to all other haplotypes of the set of 1674 chomosomes and for -14009 and -14010 for a subset of 522 and 202 chromosomes respectively.

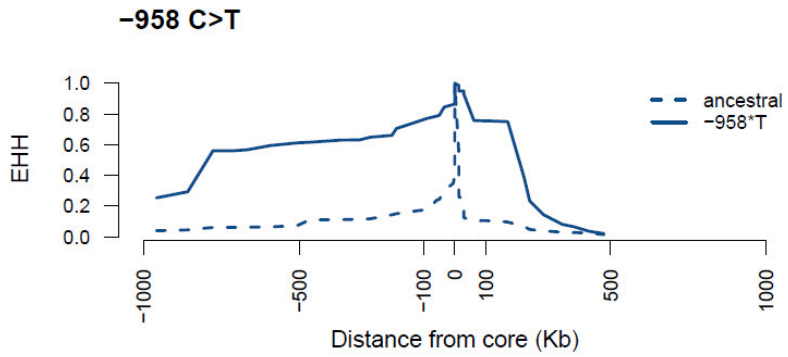


Figure 5.10: EHH decay of the haplotypes carrying *-958*T* (as key marker for the B haplotype) against all other chromosomes of the whole sampleset (n=1674).

All haplotypes carrying the derived enhancer alleles show greater EHH on both sides of the core SNPs in comparison to all other haplotypes (Figure 5.9 and Figure 5.10). This shows evidence for recent origin of these alleles and is also consistent with selection of these variants, and is consistent with previous findings for some of the variants (Bersaglieri et al. 2004; Gallego Romero et al. 2012; Poulter et al. 2003; Sabeti et al. 2007; Tishkoff et al. 2007; Voight et al. 2006). It should be noted that the EHH decays in a non-symmetrical manner. It drops more quickly on the right side of core SNP on the chromosome (i.e. upstream of the enhancer in *LCT* gene direction, chromosome -), which is particularly apparent for the *-13910*T* carrying haplotype. It is also of interest that this same extended haplotype homozygosity is detectable for *-13907*G*, , which is located on the same haplotype background, even though there are more *13910*T* alleles in the dataset.

To be able to compare between the *LCT* core haplotypes with the derived or ancestral alleles of the lactase persistence associated enhancer SNPs, the A, C, P and X haplotype chromosomes were selected from the data and EHH tests done for each of them separately but the results combined as shown in Figure 5.11. In addition to that the B haplotype chromosomes were selected from the overall dataset and setting -958 as core SNP against all other preselected chromosomes. In this comparison *-958*T* alleles carried on derived and recombinant chromosomes D, F, I, R, N, a, L, i and B *-13913*C* are excluded.

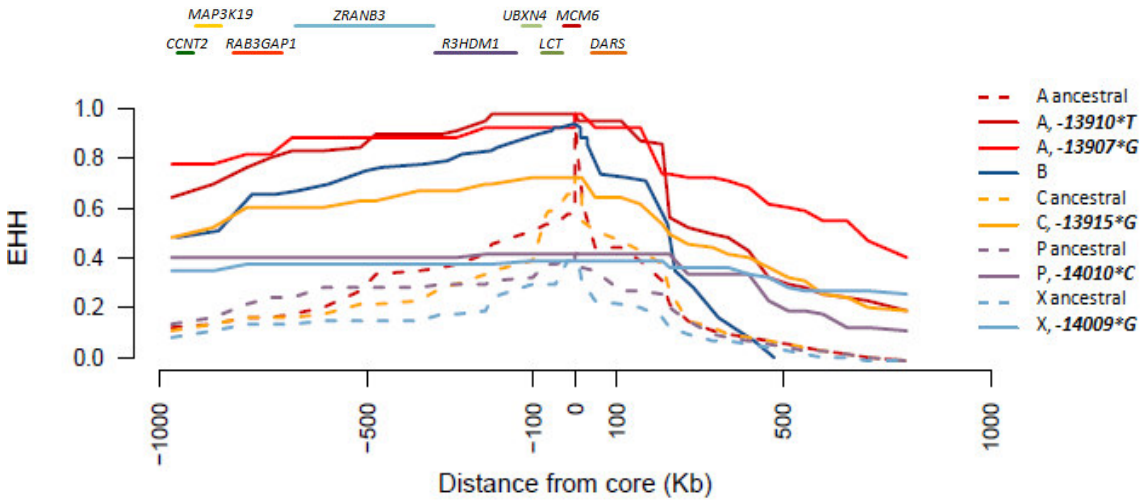


Figure 5.11: EHH decay over physical distance for the haplotypes carrying derived LP associated alleles compared to their ancestral haplotypes and all B haplotype carrying chromosomes. Note that the derived haplotypes for P and X represent less frequent chromosomes of n=13 and n= 31 respectively. The broad location of the genes on the chromosome is indicated above the plot.

A large step in EHH decay is located about 230 kb chromosomal upstream of the core SNPs (i.e. to the right hand-side on the diagram on Figure 5.11), which is similar for all variant haplotypes. To the other side, the haplotypes show a long stretch of EHH at relatively high levels, with a small step of decay for the *-13910*T* and *-13907*G* carrying haplotypes far beyond *LCT* in gene *R3HDM1* and further away about 800 kb in gene *RAB3GAP1*.

Interestingly this long stretch of EHH can also be seen in the B haplotypes that do not carry any of the functional derived alleles. Even in Figure 5.10, where only the SNP at -958 distinguishes between the haplotypes, a stretch of about 500 kb can be seen in about 60% of the chromosomes. Note that in this case of the *-958*T* carrying haplotypes that the small sharp drop in EHH close to the core SNP is probably mainly due to the recombination event, which leads to the F haplotype (see Figure 5.6).
The critical question is whether this pattern is likely to be caused by selection of another allele outside of the enhancer but within the 500 kb region, or other factors. There is clear evidence that this haplotype is old since it occurs on many continents.

From the Genome Browsers the recombination rates in the tested genomic region are shown to be relatively low. The estimated recombination rates between the regions 135-136 and 136-137 Mb of chromosome 2 show only a minor difference of 0.5 cM/Mb and 0.4 cM/Mb respectively (deCODE sex averged rates, 0.4 and 1.1 Marshfield and 0.8 both Genethon sex averaged rates respectively, UCSC genome browser) but increase to 0.7 (2.0

and 1.7 respectively) from 137 Mb upstream. Linkage disequilibrium maps were therefore used to get further information about recombination patterns of the examined 1.8 Mb region in different HapMap populations.

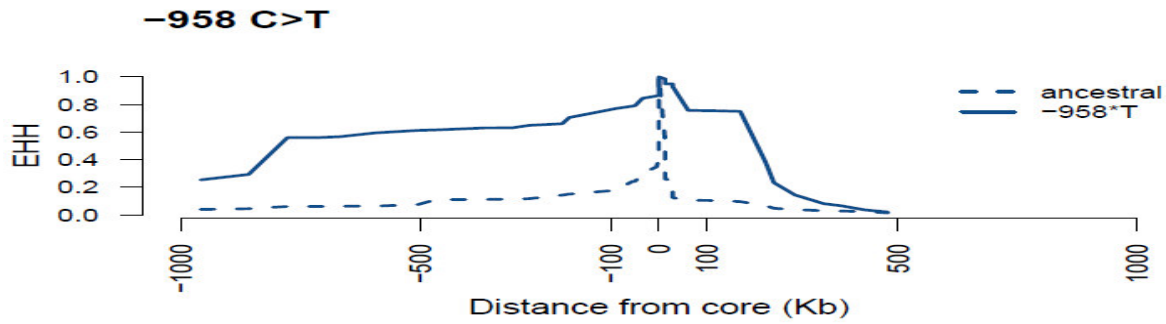
5.5.8.3 Linkage Disequilibrium Unit (LDU) maps

The construction of LDU maps is based on the idea that LD decays with distance and pairwise LDU for each SNP are calculated using $\varepsilon_i d_i$, the product of the exponential decline ε of association between SNPs and their distance d in kb (Maniatis et al. 2002, see also chapter 2, 2.3.5). The information is combined over many adjacent SNPs sequentially and converted to LD units. These units are plotted against distance resulting in a graph that shows a series of steps, where the steps in the line represent the breakdown in LD due to recombination and the plateaus high LD. LD units can also be plotted as LDU/kb against distance to get a series of peaks, which also represent recombination hotspots. With the help of Winston Lau this has been done on publically available Hap Map Phase III data of release 28.

From Figure 5.12 it can be seen that the patterns of LD and areas of a breakdown of linkage are similar in all populations. Where the data are plotted to show LDU/kb against kb and it can be seen more clearly that the peaks that represent recombination hotspots are also similar. A long stretch of no recombination across *ZRANB3* can clearly be seen in all populations in these plots, irrespective of whether they are known to have frequent LP alleles. This extends in the CEU population until *MAP3K19.* In the other two populations there is a small amount of recombination between *ZRANB3* and *MAP3K19.* The region upstream of *LCT* including *MCM6* and *DARS* also shows relatively low recombination, with only a small peak between *MCM6* and *DARS* in the African in European populations. Interestingly, there is a high peak of recombination within *LCT*, at the 3' end of the gene, in the Chinese, which is lower for the other populations and which could reflect the region around the 5579 SNP.

Even more noteworthy is the observation that the LDU maps show considerable correspondence to the EHH plots for the derived alleles and for the B haplotype (Figure 5.12). These show a large region of haplotype homozygosity that stretches along *ZRANB3* partly including *RAB3GAP1* and *R3HDM1*. On the other side, beyond *DARS*, haplotypes seem to break down, which is similar to the LDU plots. Figure 5.13 shows the same plot more focussed on the genotyped region.

a)



b)

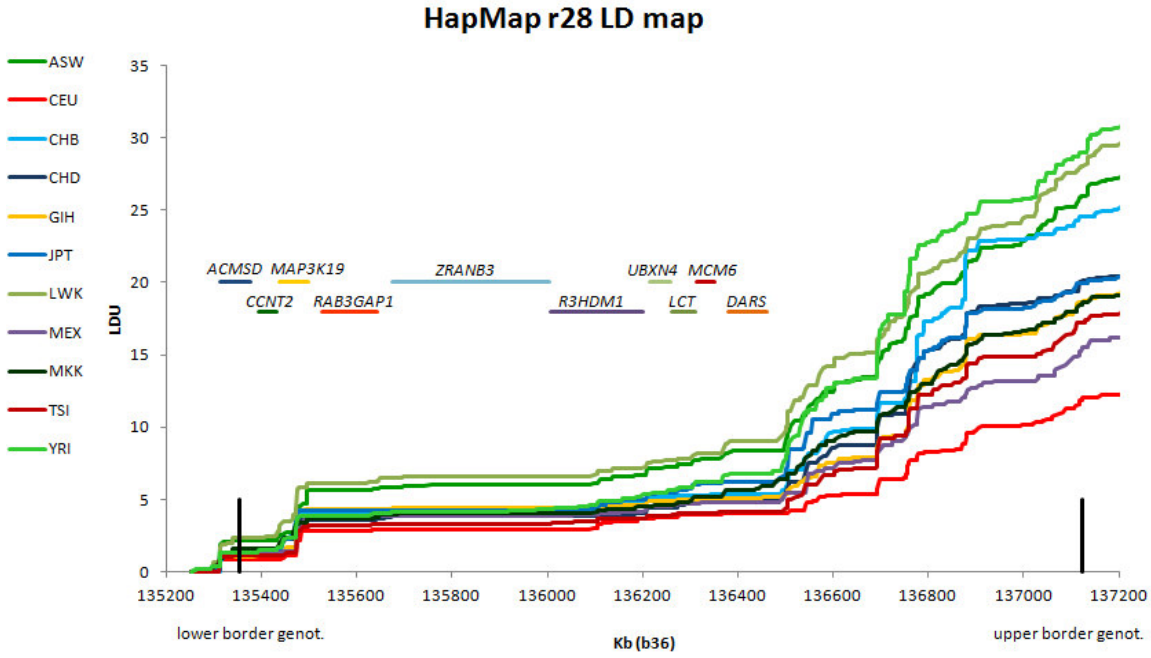


Figure 5.12: Linkage disequilibrium map showing LDU against distance (b) in comparison to the EHH plot of -958 C>T (a). The borders on the x-axis of the LD map represent the region genotyped for the samples in this chapter. The colour codes represent different HapMap populations: ASW: South western Americans with African ancestry, CEU: CEPH, Utah residents with Northern and Western European ancestry, CHB: Han-Chinese from Bejing, CHD: Chinese from Denver , GIH: Gujarati Indians from Houston, JPT: Japanese, LWK: Luhya from Webuye, Kenya, MEX: Mexican ancestry from Los Angeles, MKK: Maasai from Kinyawa, Kenya, TSI: Toscani from Italia, YRI: Yoruba from Ibadan, Nigeria.

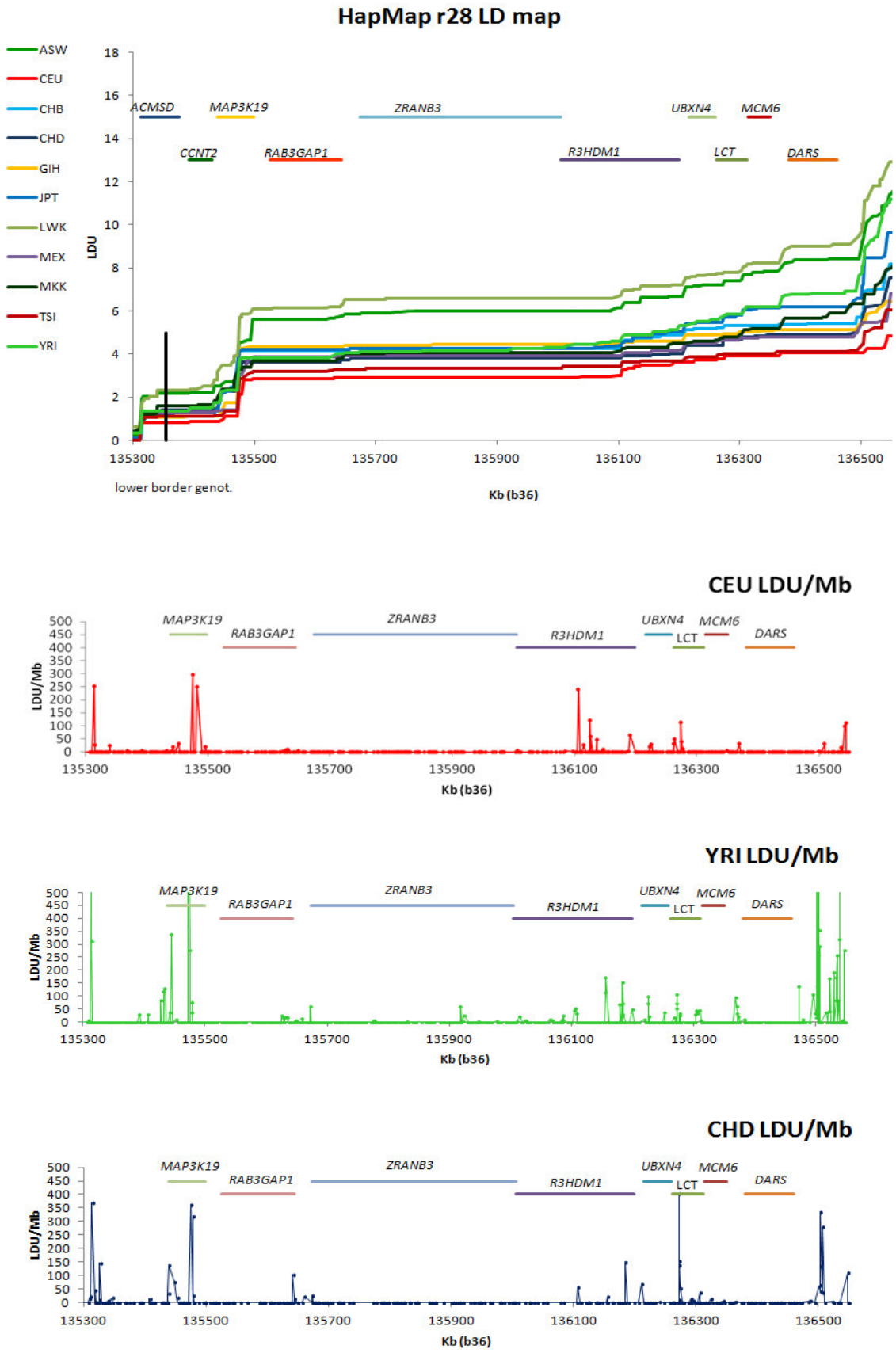


Figure 5.13: LDU against distance plot (upper plot) for the genetic region included in extended haplotype analysis and corresponding LDU/Mb against distance plot for several HapMap populations (abbreviations in legend Figure 5.12) showing peaks of recombination hotspots.

5.6 Discussion

In this chapter genotype and haplotype data from a wide set of samples from different continents were collected and interpreted to get further insights into the evolution of the *LCT* enhancer alleles and detect possible signatures of selection. Different approaches were used to examine the *LCT* haplotypes.

The analysis of the 80 kb haplotype around *LCT* allowed the construction of a Network and confirmed a tight association of the lactase persistence associated enhancer variants with particular haplotypes as described previously (Enattah et al. 2008; Ingram 2008; Ingram et al. 2007; Tishkoff et al. 2007). The patterns of LD and haplotype diversity show differences in the various populations, with the least diversity in Northern Europe. Comparing the *LCT* enhancer region with the two regions upstream and downstream, shows a reduction in diversity of the flanking regions in lactose digesters in comparison with non-digesters, unlike the situation in Africans where the digesters show equivalent or more diversity in the enhancer. This pattern reflects the different consequences of positive selection for one allele on a single haplotype background, and several alleles on different haplotypes (e.g. Bersaglieri et al. 2004; Ingram et al. 2009b; Jones et al. 2013).

The high frequency and wide distribution of the B haplotype, also shown in other studies (Gallego Romero et al. 2012; Hollox et al. 2001; Ingram et al. 2007; Jones et al. 2013) was confirmed in the combined dataset, and its appearance across the continents suggests that this is an old haplotype.

With the extension of the haplotype analysis to about 1.8 Mb it was possible to look further for possible signatures of selection for the lactase persistence associated alleles in the dataset. Different approaches were used, which came to similar conclusions. It was shown that the haplotypes carrying the derived lactase persistence associated alleles are much longer than their ancestral haplotypes, as calculated from their mean haplotype length. This pattern could also be shown with EHH analysis. The extended haplotype homozygosity of the lactase persistence haplotypes spans longer regions compared to all other haplotypes of the datasets as well as the haplotypes on which they arose. These differences support the recent dating of these variants with lack of time and therefore recombination events to occur and break them up (Sabeti et al. 2006). Proper examination of their frequency in less heterogeneous datasets would be likely to confirm a signature of selection by methods used by others (Sabeti et al. 2006; Voight et al. 2006). However,

aside from possible sampling issues, this chapter highlights one problem with the analyses reported so far. That is that the comparison of a lactase persistence associated allele carrying haplotype against all other haplotypes somewhat exaggerates this effect, because the non-carrier chromosomes include a mixture of chromosomes of different the *LCT* core haplotype.

The analysis of derived and ancestral variant haplotypes of the same *LCT* core haplotype (as first shown by Gallego Romero et al. 2012) seemed to be a better way of representing the effect of a certain allele. The haplotype lengths of all chromosomes with an A, B, C, P and X haplotype backgrounds were calculated from all phased chromosomes clustered using the haplotype definition of Hollox et al. (2001) and were shown to be significantly different for the functional enhancer alleles. Results from EHH analysis revealed longer stretches of haplotype homozygosity for the lactase persistence variant alleles than their ancestral alleles (only shown visually). However, extended haplotype homozygosity was also observed for the B haplotype, which does not carry any functionally important enhancer alleles. This is a pattern that had been noticed before (Gallego Romero et al. 2012).

Regions of LD as seen in the EHH plots interestingly overlap with those from LDU plots for different populations and are similar for all haplotypes. LD decays rapidly upstream of *LCT* (downstream in chromosomal direction), after *DARS (*aspartyl-tRNA synthetase). However, the region of LD to the other side of *LCT* is much longer and spans across *ZRANB3* (zinc finger, RNA-binding domain containing 3), up to *MAP3K19 (*Mitogen-activated protein kinase kinase 19), where a decay of LD can be seen.

The LDU plots show that *ZRANB3* region has not been disrupted by recombination. This might somehow have been an effect of selection for reduced recombination due to its functional importance. *ZRANB3* was shown to be a DNA annealing helicase and endonuclease with function for genome stability (UniProtKB, http://www.uniprot.org/). It is important for replication stress response and was for example shown to be involved in the repair of DNA lesions that block the replication process (Weston et al. 2012). Whether or not as a consequence of selection affecting this or another gene, it may be that there is a lack of recombination motifs (Myers et al. 2010) or even a structural rearrangement of the chromosome in this region. Such repression of recombination will have distorted inferences about hard selective sweeps.

From the combined results of this chapter it can be concluded that all *-13910*T alleles* occur on one haplotype which extends at least 310 kb and in many cases much longer (mean length 1.3 Mb). This supports the conclusion of previous findings which suggest one single origin of *-13910*T* and its presence on one selected haplotype (for example Bersaglieri et al. 2004; Poulter et al. 2003). Our failure to find *-13910*T* on any different background even in Iran does not provide support for the existence of a convergent evolution of the allele as claimed previously (Enattah et al. 2007). Inspection of the haplotype networks in that paper make errors in haplotype inference or gene conversion more likely, as the 22kb SNP occurred on two different branches of the network.

The same pattern of extended haplotypes is also seen for all other variants that have been shown to be functional, apart from-14011 C>T, for which there is some break-up of haplotype background. It is of interest to note that in the case of *-13913**C, for which there is strong evidence that it does not cause lactase persistence, the derived allele is also located on a long haplotype in the 7 chromosomes that we had to analyse, but this is not significantly longer than that of the ancestral B variant on which it arose, even though its low frequency would suggest a very recent origin. So in that case one might speculate that the allele was spread by drift.

The observations on the extended region of high LD in all populations, including many without LP and of the extended B haplotype, show that new methods ought to be developed to examine selection to take demography and recombination into greater account

6 *LCT immediate promoter and intron 2 - a search for yet undiscovered functional variants*

6.1 Introduction

The genotype-phenotype comparison maps in chapter 3 show areas where lactase persistence frequencies of all 5 variants thought to be causal of the trait, combined, fails to match the observed phenotype data. Besides sampling issues and incomplete datasets of genotype- or phenotype data these interpolated maps show geographic candidate regions where other causes of lactase persistence might yet be uncovered. One example for missing lactase persistence alleles, as already mentioned in chapter 1, comes from the unexpected lack of enhancer alleles in the Wolof (Ingram et al. 2009b; Jones et al. 2013; Mulcare et al. 2004) with half the sample tested being judged lactase persistent by lactose tolerance tests (Arnold et al. 1980).

With the discoveries of the last years, revealing more lactase persistence variation in the enhancer region of *LCT*, the doubts about this region being of critical functional importance have subsided, but speculations remain that other causes for lactase persistence exist and it cannot be excluded, and it is indeed likely, that other sequence regions influence lactase expression.

A recent reinvestigation of the exonic region around the active sites of lactase in Africans did not reveal any variation (Jones 2012) which might have altered lactase expression and all lines of evidence pointed to regulatory variation being more likely to be responsible for inter individual difference in adult lactase expression. In an effort to find other variation of possible function, a study at the outset of this thesis work showed variation in the immediate promoter of the lactase gene in several African individuals, which was slightly surprising since this region had been shown to be a quite conserved part of the gene (Troelsen 2005). Two alleles, *-17*T* and *-36*T* (rs111985625), were of interest as they were only found in individuals from milk drinking populations (Jones 2012).

6.2 Chapter aims

These new findings inspired the examination of the immediate promoter region in further samples of European and Middle Eastern origin. A further aim of this chapter was to use bioinformatic tools and publicly available databases to determine new genomic regions in and around the *LCT* gene with a possible functional influence on lactase persistence. One such region would be selected and examined. Depending on the results *in vitro* experiments would be designed to examine the effect of nucleotide changes in one or both regions.

6.3 Strategy

For the promoter region of *LCT* previously tested primers were used (Jones 2012) to amplify a 675 bp sequence region (Chr2:136594559-136595233, GRCh37/hg19) of which a sequence of bp -360 until +82 from the *LCT* transcription start was readable for all samples. Details of primer sequences are shown in Table 2.3, chapter 2.

To identify candidate regions in and around *LCT* with possible function the UCSC genome browser (http://genome.ucsc.edu) was scanned in particular for bioinformatically identified transcription factor binding sites and conserved regions between species. Especially useful was information from two publications that used a combination of chromatin immunoprecipitation and next generation sequencing (ChiP-Seq) to identify potential Cdx-2 and HNF-4α binding sites in Caco-2 cells (Boyd et al. 2009; Boyd et al. 2010). As mentioned in chapter 4, both transcription factors play an important role in the intestine. One region in intron 2 of *LCT* showed high binding affinity to both transcription factors in these assays (as shown in Figure 6.1) and was chosen for further investigation via sequencing.

Primers were designed to include the most of the sequence region of interest. With these primers a region of 839 bp was amplified (Chr2:136588631-136589469, GRCh37/hg19) of which bp positions 5345-6052 of *LCT* were readable.

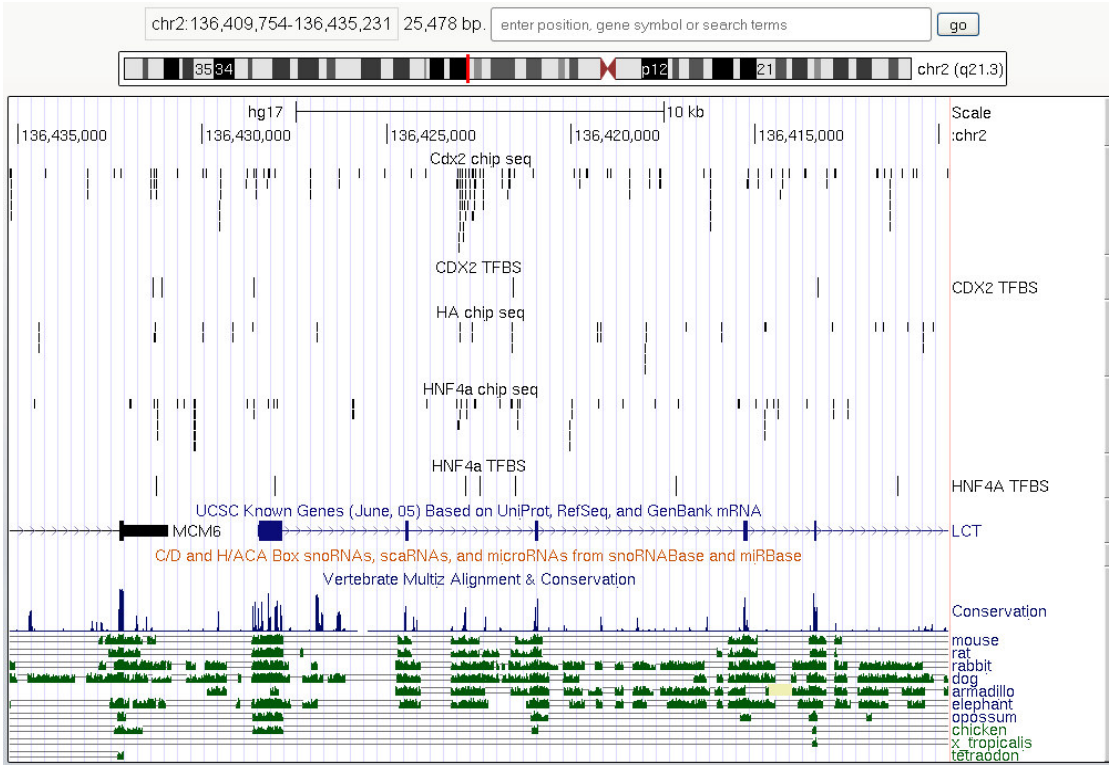


Figure 6.1: Output of the UCSC Genome Browser of the human sequence (http://genome.ucsc.edu, NCBI35/hg17) showing the positions of Cdx-2 and HNF-4α binding in intron 2 of *LCT* in ChIP-Seq assays (Boyd et al. 2009; Boyd et al. 2010).

6.4 Variation of the immediate *LCT* promoter

The immediate promoter was successfully sequenced for 1018 individuals from 45 populations. Table 6.1 shows all variants found in the *LCT* promoter region. 6 new variants were identified (-333 C>T, -294 C>A, -237 G>A, -211 C>T, -204 C>T, 41 T>C) most of which are rare. Only *-294*A*, *-211*T* and *-17*T* occur in more than one population. Of the more frequent African alleles (Jones 2012) *-173*A* was found in two Yemeni samples from Hadramaut and *-17*T* in 6 populations of which most are also from the Middle East.

Table 6.1: *LCT* promoter variation (allele counts) in the samples sequenced from 45 European and Asian population groups, N: number of chromosomes.

Region	Country	Population	N	-333 C>T	-294 C>A	-237 G>A	-211 C>T	-204 C>T	-173 G>A	-17 C>T	41 T>C
Northwest/ Central Europe	Germany	Germans	60	-	-	-	-	-	-	-	1
		Sorbs	64	-	-	-	-	-	-	-	-
	Ireland	Irish	68	-	-	-	-	-	-	-	-
	Netherlands	Frisians	54	-	-	-	-	-	-	-	-
	Norway	Norwegians	54	-	1	-	-	-	-	-	-
	Slovakia	Roma	64	-	-	-	-	-	-	-	-
	Sweden	Sami	10	-	-	-	-	-	-	-	-
		Swedes	14	-	-	-	-	-	-	-	-
	UK	Ashkenazi-Jews	4	-	-	-	-	-	-	-	-
		English	52	-	-	-	-	-	-	-	-
	Other	Other C/NW Europeans	44	-	-	-	-	-	-	-	-
South Europe	Greece	Greeks	40	-	-	-	-	-	-	-	-
	Italy	Tyroleans Bozen	62	-	-	-	-	-	-	-	-
		Tyroleans Gadertal	58	-	-	-	-	-	-	-	-
		Tyroleans Vinschgau	60	-	-	-	-	-	-	-	-
	Portugal	Portuguese	38	-	-	-	-	-	-	-	-
	Spain	Catalans	60	-	-	-	-	-	-	-	-
		Spanish	62	-	-	-	-	-	-	1	-
East/ Southeast Europe	Belarus	Belarusians	44	-	-	-	-	-	-	-	-
	Macedonia	Macedonians	64	-	-	-	-	-	-	-	-
	Romania	Romanians	64	-	-	-	-	-	-	-	-
	Russia	Erzja	2	-	-	-	-	-	-	-	-
		Russians Perm	34	-	-	-	-	-	-	-	-
	Ukraine	Ukrainians	28	-	-	-	-	-	-	-	-
	Other	Other E/SE Europeans	12	-	-	-	-	-	-	-	-
Middle East	Cyprus	Greek Cypriots	50	1	-	-	-	-	-	-	-
	Iran	Iranians	90	-	1	-	-	-	-	-	-
	Kuwait	Kuwaiti	50	-	-	-	3	-	-	-	-
	Syria	Syrians	114	-	-	-	-	1	-	2	-
	Turkey	Anatolian-Turks	40	-	-	-	-	-	-	-	-
	Yemen	Yemeni Hadramaut	134	-	1	-	-	-	2	2	-
		Yemeni Sena	28	-	4	-	-	-	-	1	-
	Other	Other Middle Easterns	48	-	-	-	-	-	-	1	-
West Asia	Armenia	Armenians	50	-	1	-	-	-	-	-	-
	Azerbaijan	Azeri	50	-	-	-	1	-	-	-	-
	Georgia	Georgians	38	-	-	-	-	-	-	1	-
Central/ South Asia	Afghanistan	Pashtuns/Afghans	32	-	-	-	-	-	-	-	-
		Tadjiks	12	-	-	-	-	-	-	-	-
		Uzbeks	12	-	-	-	-	-	-	-	-
	Nepal	Nepalese	10	-	-	-	-	-	-	-	-
		Tharu	42	-	-	-	-	-	-	-	-
	Uzbekistan	Uzbeks	42	-	-	1	-	-	-	-	-
	Other	Other C/South Asians	14	-	-	-	-	-	-	-	-
Central/East/ Southeast Asia	Mongolia	Khalka	52	-	-	-	-	-	-	-	-
		Mongols	12	-	-	-	-	-	-	-	-

6.5 Reporter gene assays on *LCT* promoter variants

Functional studies were conducted to examine the two most interesting *LCT* promoter variants. The variants *-17*T* and *-36*T* were chosen for analysis because, as described above, *-17*T* was found in many African and some other populations and *-36*T* was more common in pastoralists groups from Africa (Jones 2012). Their locations in a region of binding for several transcription factors (Figure 6.2) also make them interesting for study. With the comparison of the promoter region with other primate species (Appendix A, or also seen in Figure 6.2) it was recognised that *-36*T* is actually the ancestral variant for this SNP. However, *-36*C* is the common variant in all populations so far and for comparison reasons with previous work (Jones 2012) I will further refer to *-36*T* as the variant.

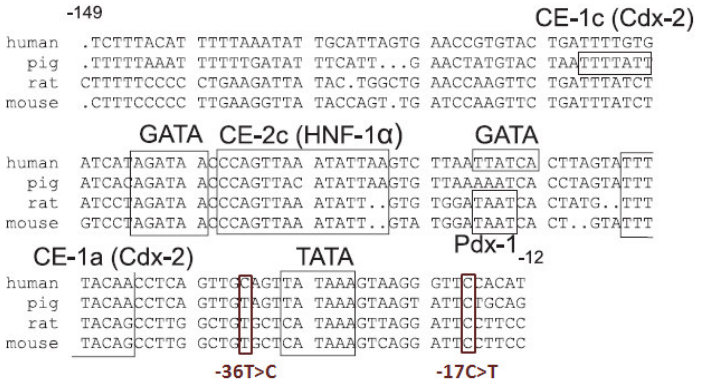


Figure 6.2: Alignment of 150 bp of the proximal *LCT* promoter showing the location of the two variants chosen for transfection experiments (red boxes). *Cis*-elements are indicated by black boxes and transcription factors binding are shown. Picture modified from Troelsen (2005).

The two *LCT* promoter variants were introduced into the reporter gene plasmids containing the ancestral enhancer sequence, upstream of the firefly luciferase gene (*luc+*) (for details see section 2.2.7.2) and transfected into Caco2 cells as described in chapter 2 (section 2.2.7). Gel shift assays were not conducted since these are inappropriate for promoter regions where multiple proteins are binding.

Transfection studies revealed no significant difference in reporter gene expression between the control plasmids and the constructs containing the *-36*T* and *-17*T.* Figure 6.3 shows the relative luciferase activity after 2 and 9 days of transfection compared to the ancestral promoter only plasmid and also that containing the ancestral enhancer sequence, with promoter enhancer construct carrying *-17*C* and *-36*C*.

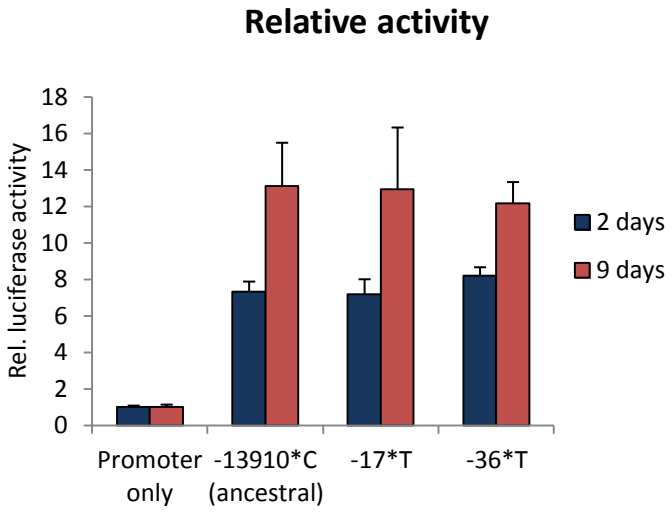


Figure 6.3: Result of luciferase reporter gene assay for the *LCT* promoter variants in constructs including the ancestral enhancer sequence, after 2 and 9 days of transfecting the Caco-2 cells. Luciferase activity of promoter variant constructs are shown in comparison with the promoter only construct. Luciferase activities (means ± SD, n=4) were corrected for transfection efficiency and normalised to the expression of pGL3 hLPH1085 (promoter only). Neither variant allele showed significantly different expression from the ancestral allele.

6.6 Variation in intron 2 of *LCT*

Data for *LCT* intron 2 were collected for a selection of European and Middle Eastern populations from chapter 1 and phenotyped samples from Europe, Asia and Africa. The *LCT* intron 2 region was successfully sequenced for 308 samples and Table 6.2 shows the allele frequencies in 11 populations tested. 13 variants were detected, of which 5372 C>T, 5473 C>T and a 4 bp deletion, around position 5788 from the start of transcription of *LCT*, were present in all populations studied.

To explore a possible association with the lactase persistence phenotype previously phenotyped populations (see also chapter 2) from Northern Europe (Harvey et al. 1995a) and Italy (Harvey et al. 1998), the Siberian Yakuts (Hollox 2000) and the African Jaali and Somali (Ingram et al. 2007; Ingram et al. 2009b) and an additional set of 26 phenotyped Finnish individuals (Harvey et al. 1998; Hollox 2000) were further analysed. After exclusion of samples with missing data at more than 4 positions in the *LCT* intron 2 or enhancer sequence regions 172 samples were examined.

Allele counts for the entire group and for the group divided into European + Asian and African samples are shown in Table 6.3. For simplicity lactase persistent samples as tested

by biopsy and lactose digesters tested by breath hydrogen tests were both grouped as lactase persistent (and vice versa for non-persistent individuals). Phenotype counts were then used to build a 2x2 table to examine association with Fishers exact test.

The two alleles *5473*T* and *5586*A* show a significant association with lactase persistence in the overall sample set, whereas *5372*T* and the 4 bp deletion at position 5788 are instead associated with lactase non-persistence. Since it seemed to be very likely that those associations reflect demographic patterns, samples were split according to their origin into a European/Asian and African sample set.
*5473*T* was found in both, the European/Asian and African sample sets and is not significant associated with persistence when both sets were analysed separately (Table 6.3). As shown in Table 6.2, *5586 *A* was only found in Europeans and 18/21 of the chromosomes were in persistent people (Table 6.3). This suggested that it might be associated with the *-13910*T* carrying chromosomes. However, the three *5586*A* variants occurred in non-persistent individuals who do not carry the **T* allele but all of whom were previously shown by former members of the lab to carry at least one A haplotype which would support the suggestion that *5586*A* occurs on an A haplotype background and preliminary PHASE analysis points to this.

The 4 bp deletion at position 5788 remained significantly associated with lactase non-persistence in the African sample subset. This variant was mainly found on a B haplotype background (*-958*T, 666*A*) but was also found on the core A, C and other haplotypes as assessed by visual inspection of the data and preliminary PHASE analysis. This result would need further confirmation but suggests that some gene conversion has occurred.

Critically, no alleles were found exclusively in persistent people, including those with no functional enhancer alleles.

Table 6.2: Allele frequencies of *LCT* all intron 2 variants observed in a subset of European and Middle Eastern populations from chapter 3 and additional phenotyped samples, N: Number of chromosomes.

Population	N	rs11886852 5372 C>T	5409 C>T	rs9636213 5473 C>T	5586 G>A	5623 G>A	rs10928551 5669 C>A	5698 C>T	5742 G>A	rs10928550 5783 G>C	5788 4 bp Indel (1=del)	5807 C>T	5810 C>T	6013 C>T
Norwegians	52	0.058	-	0.942	0.019	0.019	-	-	-	-	0.058	-	-	-
Greek-Cypriots	56	0.446	-	0.536	0.036	-	-	-	0.107	-	0.464	-	-	-
Kuwaiti	44	0.341	-	0.591	-	-	0.023	-	0.068	0.023	0.341	-	-	0.023
Syrians	32	0.406	-	0.594	0.063	-	-	-	0.031	-	0.375	-	-	-
Yemeni Hadramaut	60	0.350	0.017	0.567	0.017	-	0.017	0.033	0.033	0.017	0.317	-	-	0.050
Other_ME	12	0.250	-	0.750	-	-	-	-	0.167	-	0.250	-	-	-
Phenotyped samples														
Northern Europeans	40	0.100	-	0.900	0.425	-	-	-	-	-	0.100	-	-	-
Italians	32	0.281	-	0.719	0.031	-	-	-	0.063	-	0.313	-	-	-
Yakuts	56	0.143	-	0.821	-	-	-	-	-	-	0.179	0.018	0.018	-
Jaali	110	0.382	-	0.564	-	-	0.082	0.009	0.036	0.091	0.273	-	-	0.018
Somali	122	0.467	-	0.352	-	-	0.098	0.008	-	0.057	0.484	-	-	0.008

Table 6.3: Allele counts and association of the *LCT* intron 2 variants. Statistically significant *p*-values after Bonferroni correction for 11 tests (0.0045) are shaded. P: lactase persistent (or lactose digesters), NP: lactase non-persistent (or lactose non-digesters), N: Number of chromosomes.

Group		N	5372 C>T	5473 C>T	5586 G>A	5669 C>A	5698 C>T	5742 G>A	5783 G>C	5788 4 bp Indel	5807 C>T	5810 C>T	6013 C>T
All	P	162	35	115	18	6	0	2	6	26	1	1	0
	NP	182	73	94	3	9	1	3	6	72	0	0	1
p-value			<0.001	<0.001	<0.001	0.609	1	1	1	<0.001	0.471	0.471	1
Europeans/	P	82	10	71	18	0	0	1	0	11	1	1	0
Asians	NP	90	24	62	3	0	0	2	0	26	0	0	0
p-value			0.021	0.015	<0.001`	-	-	1	-	0.015	0.482	0.482	-
Africans	P	80	25	44	0	6	0	1	6	15	0	0	0
	NP	94	49	32	0	9	1	1	6	46	0	0	1
p-value			0.006	0.006	-	0.788	1	1	0.774	<0.001	-	-	1

6.7 Discussion

During this study the *LCT* promoter near to the start of transcription was surveyed in many European and Middle Eastern samples. As previously suggested for Europeans, this region is quite conserved and only a few variants were found. This same region was also studied in a parallel project in Africans (Jones 2012). Of the frequent African variants, *-173*A* and *-17*T* were also present in the sample set of this study, mainly in the Middle East, although at low frequencies.
However, the evidence from the African groups of a possible connection between a pastoralist lifestyle and the occurrence of *-36*T* and *-17*T* motivated the inclusion of these two variants in functional studies. Transfection studies were conducted to test the effect of the varying promoter sequences at these positions. No influence on reporter gene expression could be detected. This negative finding however agrees with the further developments of the parallel African studies, namely association studies with lactose digester status in Ethiopians, which failed to show any evidence of association with persistence making a functional role unlikely (Jones 2012).

In a search for possible new regions with variation that could be causal for lactase persistence, results of ChiP-Seq analysis conducted by our colleagues in the laboratory in Copenhagen (Boyd et al. 2009; Boyd et al. 2010) were used to identify a region in intron 2 of *LCT* which had been shown to be target for Cdx-2 and HNF-4α binding.

In contrast to the promoter, the other region of interest, intron 2 of *LCT*, revealed relatively more variations. Some of the alleles were rather frequent and present in both persistent and non-persistent people discounting a direct functional role. Some were more frequent in lactase non-persistent individuals and others were more frequent in the persistent individuals.
The only possible candidate SNP for function is 5586 G>A since only three individuals classed as non-persistent carried the allele. However all 18 derived alleles that occurred in persistent individuals also carried at least one *-13910*T* allele. Clearly not all *-13910*T* allele carrying persistent individuals carry *5586*A* – since the allele was at low frequency in some places where the **T* allele is common, such as Norway. This suggests that it is a derived allele on the background of A, *-13910*T* chromosomes, and thus non-essential for lactase persistence. Likewise the three individuals who carry this allele and not *-13910*T* are non-persistent showing that it cannot on its own cause persistence, and indicating some recombination/gene conversion has occurred.

Nevertheless, it can be speculated that binding of HNF-4α and Cdx-2 may differ for the ancestral and derived allele at this position and this might interplay with the Oct-1 binding of the -13910 position and the transcription factors binding to the promoter to modulate *LCT* transcription, possibly causing inter-individual differences in lactase expression and persistence. Further functional studies could be conducted to test this *in vitro*, to search for evidence of that SNP altering the binding of specific transcription factors. The case for doing that is however not supported by the bioinformatic examination of the effect of the SNP on the probability of the predicted Cdx-2 binding which did not differ between the derived and ancestral allele.

In summary, no new highly putative causal alleles were found. Disappointingly it was not possible to include the Wolof in these DNA analyses since the quality of the remaining DNA samples in the lab was poor. The variants in intron 2 appear to disrupt the haplotypes described in Chapter 5 in a way that was unexpected, and suggesting gene conversion events and this might have merited further examination, but the datasets were not completely overlapping and further testing was not warranted by the lack of good candidates in this region.

7 General Discussion

One of the main aims of this thesis was to examine the genetic regions in and around *LCT* for variation and collect data about known lactase persistence associated alleles in geographic regions where there was less dense information available, mainly Eastern European and West Asian populations. The vast amount of enhancer variation data assembled in chapter 3 should significantly contribute to the knowledge about the distribution of *LCT* enhancer alleles. The Middle East, as contact zone between Europe, Africa and Asia revealed variation patterns reflecting influences from all these continents in both, the enhancer and the *LCT* surrounding markers as seen in the haplotype variation.

In chapter 4 it could be shown that two further *LCT* enhancer alleles alter function *in vitro*. This increases the number of functional variants likely to be responsible for the lactase persistence trait to six: *-14011*T, -14010*C, -14009*G, -13915*G,-13910*T* and *-13907*G*. Our recently published study (Jones et al. 2013) revealed the association of*-14009*G* with lactase persistence in a 350 individuals Ethiopian cohort and confirmed the association of -1*3915*G* and *-13907*G.*
The functional studies described here show that *-14009*G* and also *-14011*T* have a clear effect on transcription factor binding *in vitro.* That in the case of *-14009*G* is quite different from *-14010*C* and *-14011*T*, showing that even when a small DNA binding region is involved more than one different route can lead to alteration of function. *-14011*T* is not frequent enough to test for association. It was found scattered as a rare variant in several populations in Europe and the Middle East in this thesis and in further studies in Estonia, Ethiopia and even in indigenous populations from Brazil (Friedrich et al. 2012; Jones et al. 2013; Lember et al. 2006). This together with the finding of the haplotype analyses of chapter 5, that showed both A and B haplotypes as background, leads to the suggestion that this allele might have arisen by mutation independently several times. The lack of evidence of a single extended haplotype is slightly surprising and is not consistent with recent selection. A further allele *-14028*C* is almost certainly functional as well since it was associated with high lactase RNA expression (Poulter et al. 2003).

Further downstream, variants *-13806*G* and *-13730*G* were shown not be lactase persistence associated (Jones et al. 2013) although located in the same region with enhancer function (Troelsen et al. 2003). These and other new variants found in the

investigations reported in chapter 3 and 5 are single variants or not frequent enough to be worth examining their influence on *LCT* expression without other reason.

One of the relatively strong candidate functional alleles *-13779*C*, frequent in Indians tested in this thesis and elsewhere (Gallego Romero et al. 2012) showed no difference in protein binding in the EMSA studies and no significant difference in reporter gene activity in the experiments conducted during this thesis. From this evidence it can be concluded that it does not alter function. However, the -13779 position or the surrounding 12 bp sequence contains a binding site for HNF-4α which might not have direct influence on the lactase persistence trait but is possibly part of the transcription factor network involved in regulation.

It is striking that the functional variants cluster in two regions indicating that these are the 'functional hotspots' of the enhancer region. However this does not *per se* mean that mutations there always influence function, as shown for *-13913*C* (Enattah et al. 2008; Ingram et al. 2007) which is not associated with lactase persistence (Jones et al., 2013). This is another example of how very closely located mutations can have a different effect of transcription factor binding.

The complex interactions of different transcription factors binding to the enhancer and promoter of the *LCT* gene were reported previously (reviewed in Montgomery et al. 2007) and further experiments would be needed to investigate the role of the two SNPs -14011 C>T and 14009 T>G in this network of *cis* and *trans* regulatory elements. This could be done for example in further reporter gene essays under the co-transfection of other transcription factor proteins, known to be involved in intestinal regulation.

One of the important findings of chapter 4 was the binding of an additional protein to *-14009*G,* which is suggested to be an Ets factor, as evidenced from EMSA experiments with tailored competitor oligonucleotide sequences. It is quite likely that a factor of this protein family is involved *in vivo* as they play a role in cell differentiation and maintenance (Sharrocks 2001). EMSA experiments with further antibodies against c-Ets-1, ELK-1 or other Ets factors might be able to confirm an interaction with the variant at -14009. Another way of identifying the protein would of course be its isolation from gels, purification and protein sequencing, or with the fractionation of the Caco2 nuclear extract as done by Lewinsky and colleagues for the *-13910*T* binding Oct-1 and GAPDH proteins in Caco2 and HeLa cells (Lewinsky et al. 2005). However, this requires an enormous

amount of time and material and the antibody approach might therefore be the first choice.

Of course *in vitro* studies do have their limitations, as conditions in the living organism differ and the complex biochemical settings in a cell of a living organism might differ from the *in vitro* situation, though ideally, the effects of variants are better studied *in vivo* as recently done for *-13910*T*. A 218 bp enhancer fragment carrying the variant, fused to a 2 kb rat promoter was sufficient to prevent the down-regulations of lactase in mice (Fang et al. 2012).
However, it is difficult and costly to establish transgenic mouse lines and many aspects of *LCT* regulation are different in the mouse. It might be more appropriate to study lactase persistence directly in measuring mRNA levels and chromatin alterations in the human proximal jejunum. In projects that collaborate with hospitals outside Europe and collect intestinal biopsy material for other reasons, the association of allelic expression levels with other enhancer variants could be examined, together with chromatin technologies to examine TF/chromatin complexes in the context of development. This way changes in transcription factor binding as it occurs *in vivo* can be confirmed.

Another point was recognised during the experimental phase for chapter 4. As already mentioned in the discussion of that chapter, the enhancer variants tested in all the transfection studies in the Danish lab were tested in a construct with *-958*T* in the promoter sequence, which does not reflect the *in vivo* situation for variants other than *-14028*C*.
To exclude possible confounding influence of *-958*T* and other variable promoter positions, experiments should be repeated with a promoter carrying *-958*C.* This would be especially important as the *T** allele at that position was shown in previous EMSA experiments to disrupt binding of a protein (Hollox et al. 1999), which is probably Oct-1 as suggested from experiments of this thesis, and this variant also decreases reporter gene expression (Chitkara et al. 2001). For the -13915 T>C variant mutagenesis of the *-678*A* to *-678*G* should also be considered to mirror the haplotype background of this allele. A further mutation of the construct for *-14009*G* at the InDel position -943/942 is probably less important as this was shown to not affect protein binding (Hollox et al. 1999).

However, functional studies of both kinds might also lack genetic regions in the region of *LCT* that could have an influence on expression. The updated GenoPheno correlation map of chapter 3 revealed that there are still geographic regions where there is a difference

between the lactase persistence genotype and phenotype frequency, which implies the possible existence of other functional elements responsible for lactase persistence.

Scanning the genome for transcription factor binding sites is possible with immunoprecipitation methods, which can also identify other regulatory influences, such as DNA methylation and the binding of RNA polymerase II, which could be helpful in identifying possible new target regions (reviewed in Mikkelsen et al. 2007; Olsen et al. 2012). Data for proteins with a restricted expression in the intestine and therefore relevant to lactase can be collected from Caco2 cells as it has been done with HNF-4 and Cdx-2 by Boyd and colleagues (2009; 2010). Preliminary experiments failed so far to establish a similar assay for Oct-1 (Troelsen, personal communication).

The next generation sequencing technologies allow large scale genome scans for elements regulating gene expression, also including the detection of small interfering RNA molecules and allelic transcript expression by RNA sequencing (RNA-Seq, for example Chia et al. 2010). Further techniques can scan for open chromatin structures such as DNase-Seq (DNase I hypersentitivity sequencing) or FAIRE (Formaldehyde-Assisted Isolation of Regulatory elements)-Seq (Giresi et al. 2007), which was recently modified to work with gene chips (FAIRE-gen) to be suitable for GWAS (Smith et al. 2012). Projects like ENCODE (Bernstein et al. 2012, https://genome.ucsc.edu/encode/) have the goal to combine and annotate data collected with these techniques and accumulate knowledge about regulatory elements and gene expression.

Research in the past focussed on protein structure or exon sequencing to identify genetic causes for monogenic or polygenic traits. That this still is case shows the data of one of the most popular research projects, the 1000 genomes project. It provides high coverage data for exonic regions (at least 20 x) but only low (1-2 times on average) for non-exonic regions (Abecasis et al. 2010, http://www.1000genomes.org/), though the coverage is higher for the first 100 -200 bp of non-coding regions. It was for example not possible to use 1000 genomes data for this thesis as the enhancer region was only covered 2x on average for each individual which of course cannot reliably distinguish heterozygotes, even an expansion of the data to a data coverage of 4-8x (Tyler-Smith, personal communication 2012) would not have been reliable enough.

The haplotype analyses in chapter 5 gave further insights into the occurrence and constitution of the haplotypes of the region surrounding *LCT*. Remarkable was the

frequent occurrence of a frequent B haplotype, which was of similar length to those haplotypes carrying functional variants and suggested to be the result of selection events. Rather surprising was the paucity of derived functional variants on the B haplotype. This haplotype forms the background of only two relatively rare derived enhancer alleles: -*13913*C*, which does not seem to cause LP, and -*14028*C*. The latter has been found in two European individuals so far (in one sample examined in this thesis) and was shown to reside on a high lactase expressing chromosome (Poulter et al. 2003). It was confirmed that the -14028 position is part of a Cdx-2 binding site as previously shown (Lewinsky et al. 2005). The derived allele leads to a visible loss of binding to that transcription factor in the EMSA studies conducted but also gains binding to another transcription factor shown to be highly involved in lactase expression, HNF-4α. However, a functional effect could not be shown with the transfection studies conducted in this thesis (chapter 4).
It should also be noted that the British person carrying -*14028*C* was only 21 years of age, which could mean that the decline of lactase expression had not fully happened, as shown for some Finnish individuals at the same age in a different study (Sahi et al. 1983). In this context -*14028*C* may have a function by simply delaying this process.

In chapter 5, the results from EHH analysis showed long stretches of homozygosity in the same regions where the linkage disequilibrium maps showed clear evidence of suppression of recombination, which exists in populations with and without lactase persistence. This leads to the question to what degree methods based on extended haplotype homozygosity to detect recent selection events are influenced by historical recombination patterns reflected for example in recombination hot or cold spots of the genome. It is of noteworthy that *LCT* is surrounded by genes that are important for cell function, translation and replication stress response, which might possibly favour reduced diversity.
Allele specific suppression of recombination in the region around *LCT* has previously been discussed as a possible factor that might create a similar pattern of an extended haplotype to that of a recent selection event (Hollox 2005), in the context of the -*13910*T* allele. A different strategy involving examination of microsatellite diversity around *LCT* did in that case however support the evidence for selection (Coelho et al. 2005).

Although *MCM6* and *LCT* are expressed at different developmental stages of an intestinal cell, (*MCM6* is highest in foetuses and high in intestinal crypts but down regulated when lactase activity is highest in the differentiated state), it cannot be excluded that they do not influence each other. It was suggested that *MCM6* expression in less differentiated cells has

an impact as precursor on chromatin packaging that also allows easier accessibility of transcription factors to the *LCT* promoter (Troelsen 2005). Maybe other neighbouring genes have an influence on the chromatin state of that region.

LDU maps take current and historical recombination patterns into account when comparing linkage disequilibrium between alleles in populations. Regions of similar LDU therefore reflect population specific evolutionary influences such as mutation, drift and most importantly, selection. It would therefore be desirable to examine the populations for the markers used for extended haplotype analyses in chapter 5 with both the LDU approach and EHH tests in each of the populations to be able to compare both methods and get quantitative calculable evidence for selection (Tapper 2007). The SNP data generated here were unfortunately of much lower density than that used for the LDU maps.

The main questions that remain unanswered and could be the motivation for future work are: Is there something about the B haplotype that impedes adult *LCT* expression when mutations do occur? Does the long region of LD included within the B haplotype give us any clues? What is suppressing recombination in this region and could this be allele specific? Can we exclude that there is not a large inversion within this long sequence region, perhaps on B haplotype chromosomes?
However, the occurrence of the lactase persistence trait in various parts of the world, enabled by different recent variants in the enhancer element of the *LCT* gene and each on long haplotypes, is strongly supportive evidence, of selection triggered by the adaptation to a new diet. The question is how this new diet influenced the life of the early Neolithic farmers and how milk consumption was so beneficial.

The shift to agriculture was associated with a reduction in diversity of the nutritional intake but also an excess of caloric availability, mainly in form of starch (Luca et al. 2010). For early Neolithic farmers the transition from a hunter-gatherer lifestyle to farming was most probably accompanied by periods of crop failure and domesticated animals built an additional food supply during these times. Even when taking the lower productivity of early domesticated cows into account and subtracting the milk needed for raising calves, a remarkable surplus of milk would have been produced (Gerbault et al. 2011).

Even if evidence for milk use exists in early Neolithic societies from Anatolia and the Balkans (Craig et al. 2005; Evershed et al. 2008) it is unlikely that people were able to

consume large quantities of fresh milk, as ancient DNA data suggest (Figure 7.1). The earliest evidence for the lactase persistence associated allele was found in samples from Scandinavian late hunter-gatherers from about 5400-3400 BP with a frequency of 5% (Malmstrom et al. 2010) and in Spanish early Neolithic farmers, dated to about 5000-4500 BP with a *T** allele frequency of 11 and 26% (Plantinga et al. 2012). Data from Mesolithic and early Neolithic samples of other regions of Europe suggest the absence of the lactase persistence trait in these populations (Burger et al. 2007; Lacan et al. 2011a; Lacan et al. 2011b).

It is very likely that early farmers would have gained nutritional benefits from dairying in form of processed milk products such as yoghurt or cheese. This would have also been more feasible as these are more durable forms of milk which are easier to transport, probably important for early farmers that had rather seasonal settlements (reviewed in Leonardi et al. 2012). A recent study on pot shards from Poland dated to about 7400-6800 BP, which show the shape of sieves used for cheese production and contain ruminant fat residues, directly supports this argument (Salque et al. 2013). For the milk residues found earlier, 8500 BP in Anatolia and the Balkans from around 7900-7500 BP it was also suggested that they could only last this long time if fermented milk was used in these vessels (Craig et al. 2005).

The fermentation process would have reduced the amount of lactose, which would have also allowed non-persistent farmers to profit from the nutritional value of milk. Therefore Burger and Thomas (2011) suggest the rise of *-13910*T* would have started after 5500 when dairying practises had been established and with a better supply of fresh milk the ability to drink it became more advantageous.

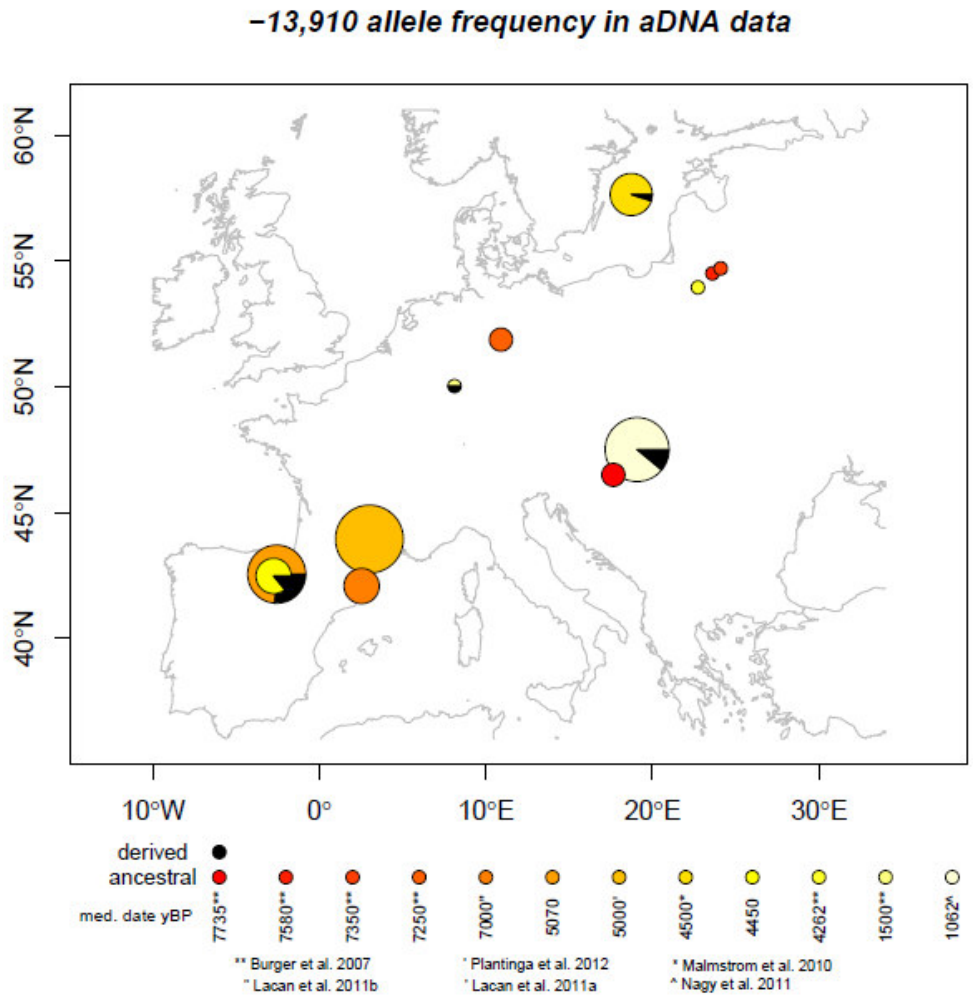


Figure 7.1: Frequency of *-13910*T* as revealed by ancient DNA analysis and location of the sample origins. Circle diameter correspond to sample sizes (map kindly provided by P. Gerbault)**.**

The characteristic molecular signatures of purifying and positive selection are a decrease in diversity in the adjacent genetic regions of the selected mutation and an increase of this diversity for balancing selection and soft selective sweeps. Both are seen for the *LCT* enhancer. However, the ways selection operates to increase, decrease or maintain genetic diversity are complex i.e. allele frequencies can be variable and processes involved difficult to retrace through space and time. In general, genetic drift and demography affect the whole genome whereas selection tends to act on the selected genes or regions only. As mentioned earlier, it can be difficult to distinguish selection from demographic events and statistical tests can help to build up the evidence for a certain scenario. The combination of different tests and simulation models that compare simulated variables with actual data from various scientific disciplines such as archaeology and anthropology as described above (chapter 1, section 1.2.15) seems to be a good way forward.

The most recent study that builds upon that approach was done by Pascale Gerbault (2013) also using a coalescent simulation model paired with approximate Bayesian computation (ABC) to model the spread and origins of lactase persistence across Europe but taking further and weighted data into account. Statistics that fit best with data from archaeological records including dairy fat residues in potsherds, and ancient and modern DNA data (partly from this thesis) for *-13910*T* were evaluated. Exchanges between hunter-gatherers and farmers were allowed in form of cultural diffusion, gene flow and density-dependent competition between hunter-gatherers and farmers. In comparison to Itan et al. (2009) this model suggests a more recent time of about 5000 years to be the starting point of selection for lactase persistence, placing it in Northern rather than Eastern Europe (Gerbault 2013). This shows how sensitive this kind of study is to the correct selection of realistic parameters.

Nevertheless, with the further development of the simulation models it might also be possible in the future to take other *LCT* enhancer alleles into account and model the spread of lactase persistence outside of Europe, which might reveal further insights into the evolutionary past of this trait and provide answers in relation to the different sweeps involved.

References

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, and McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-1073.

Akey JM, Zhang G, Zhang K, Jin L, and Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12(12):1805-1814.

Al-Abri AR, Al-Rawas O, Al-Yahyaee S, Al-Habori M, Al-Zubairi AS, and Bayoumi R. 2012. Distribution of the lactase persistence-associated variant alleles -13910* T and -13915* G among the people of Oman and Yemen. Hum Biol 84(3):271-286.

Almon R, Alvarez-Leon EE, Engfeldt P, Serra-Majem L, Magnuson A, and Nilsson TK. 2010. Associations between lactase persistence and the metabolic syndrome in a cross-sectional study in the Canary Islands. Eur J Nutr 49(3):141-146.

Ammerman AJ, and Cavalli-Sforza LLL. 1984. The Neolithic Transition and the Genetics of Populations in Europe: Princeton University Press.

Anagnostou P, Battaggia C, Coia V, Capelli C, Fabbri C, Pettener D, Destro-Bisol G, and Luiselli D. 2009. Tracing the distribution and evolution of lactase persistence in Southern Europe through the study of the T(-13910) variant. Am J Hum Biol 21(2):217-219.

Anderson B, and Vullo C. 1994. Did malaria select for primary adult lactase deficiency? Gut 35(10):1487-1489.

Aoki K. 1986. A stochastic model of gene-culture coevolution suggested by the "culture historical hypothesis" for the evolution of adult lactose absorption in humans. Proc Natl Acad Sci U S A 83(9):2929-2933.

Aoki K. 2001. Theoretical and empirical aspects of gene-culture coevolution. Theor Popul Biol 59(4):253-261.

Arnold J, Diop M, Kodjovi M, and Rozier J. 1980. Lactose intolerance in adults in Senegal. C R Seances Soc Biol Fil 174(6):983-992.

Arribas JC, Herrero AG, Martin-Lomas M, Canada FJ, He S, and Withers SG. 2000. Differential mechanism-based labeling and unequivocal activity assignment of the two active sites of intestinal lactase/phlorizin hydrolase. Eur J Biochem 267(24):6996-7005.

Auricchio S. 1998. Lactase deficiency phenotype has not been selected by malaria. Ital J Gastroenterol Hepatol 30(5):494-495.

Auricchio S, Rubino A, Landolt M, Semenza G, and Prader A. 1963. Isolated Intestinal Lactase Deficiency in the Adult. Lancet 2(7303):324-326.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, and Struhl K. 2002. Current Protocols in Molecular Biology. New York: Wiley.

Bandelt HJ, Forster P, and Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1):37-48.

Bayless TM, and Rosensweig NS. 1966. A racial difference in incidence of lactase deficiency. A survey of milk intolerance and lactase deficiency in healthy adult males. JAMA 197(12):968-972.

Bayoumi RA, Flatz SD, 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(2):173-178.

Bayoumi RA, Saha N, Salih AS, Bakkar AE, and Flatz G. 1981. Distribution of the lactase phenotypes in the population of the Democratic Republic of the Sudan. Hum Genet 57(3):279-281.

Behrendt M, Polaina J, and Naim HY. 2010. Structural hierarchy of regulatory elements in the folding and transport of an intestinal multidomain protein. J Biol Chem 285(6):4143-4152.
Bellwood P. 2008. The first farmers: origins of agricultural societies. Oxford: Blackwell Publishing.
Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, and Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.
Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, and Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6):1111-1120.
Biswas S, and Akey JM. 2006. Genomic insights into positive selection. Trends Genet 22(8):437-446.
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(5):889-902.
Bollongino R, Elsner J, Vigne JD, and Burger J. 2008. Y-SNPs do not indicate hybridisation between European aurochs and domestic cattle. PLoS One 3(10):e3418.
Bollongino R, Nehlich O, Richards MP, Orschiedt J, Thomas MG, Sell C, Fajkosova Z, Powell A, and Burger J. 2013. 2000 years of parallel societies in Stone Age Central Europe. Science 342(6157):479-481.
Bosse T, Piaseckyj CM, Burghard E, Fialkovich JJ, Rajagopal S, Pu WT, and Krasinski SD. 2006a. Gata4 is essential for the maintenance of jejunal-ileal identities in the adult mouse small intestine. Mol Cell Biol 26(23):9060-9070.
Bosse T, van Wering HM, Gielen M, Dowling LN, Fialkovich JJ, Piaseckyj CM, Gonzalez FJ, Akiyama TE, Montgomery RK, Grand RJ et al. 2006b. 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(5):G1016-1024.
Boudreau F, Rings EH, van Wering HM, Kim RK, Swain GP, Krasinski SD, Moffett J, Grand RJ, Suh ER, and Traber PG. 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 sucrase-isomaltase gene. J Biol Chem 277(35):31909-31917.
Boyd M, Bressendorff S, Moller J, Olsen J, and Troelsen JT. 2009. Mapping of HNF4alpha target genes in intestinal epithelial cells. BMC Gastroenterol 9:68.
Boyd M, Hansen M, Jensen TG, Perearnau A, Olsen AK, Bram LL, Bak M, Tommerup N, Olsen J, and Troelsen JT. 2010. Genome-wide analysis of CDX2 binding in intestinal epithelial cells (Caco-2). J Biol Chem 285(33):25115-25125.
Bradford MM. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248-254.
Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P, Tambets K, Antanaitis-Jacobs I, Haidle MN, Jankauskas R, Kind CJ et al. 2009. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science 326(5949):137-140.
Braverman JM, Hudson RR, Kaplan NL, Langley CH, and Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140(2):783-796.
Breasted JH. 1916. Ancient Times, a History of the Early World: An Introduction to the Study of Ancient History and the Career of Early Man: Ginn.
Breasted JH. 1938. The Conquest of Civilization.
Briet F, Pochart P, Marteau P, Flourie B, Arrigoni E, and Rambaud JC. 1997. Improved clinical tolerance to chronic lactose ingestion in subjects with lactose intolerance: a placebo effect? Gut 41(5):632-635.

Bronstein I, Martin CS, Oelsen CEM, and Voyta JC. 1997. Combined luminescent assays for multiple enzymes. In: Hastings JW, Kricka LJ, and P.E. S, editors. Bioluminescence and Chemiluminescence: Molecular Reporting with photons. Chichester: John Wiley. p 451-457.
Burger J, Kirchner M, Bramanti B, Haak W, and Thomas MG. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc Natl Acad Sci U S A 104(10):3736-3741.
Burger J, and Thomas MG. 2011. The Palaeopopulationgenetics of Humans, Cattle and Dairying in Neolithic Europe. Human Bioarchaeology of the Transition to Agriculture: John Wiley & Sons, Ltd. p 369-384.
Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, and Werner T. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21(13):2933-2942.
Cavalli-Sforza LL, Menozzi P, and Piazza A. 1994. The History and Geography of Human Genes: Princeton University Press.
Chantret I, Barbat A, Dussaulx E, Brattain MG, and Zweibaum A. 1988. Epithelial polarity, villin expression, and enterocytic differentiation of cultured human colon carcinoma cells: a survey of twenty cell lines. Cancer Res 48(7):1936-1942.
Chia NY, Chan YS, Feng B, Lu X, Orlov YL, Moreau D, Kumar P, Yang L, Jiang J, Lau MS et al. 2010. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468(7321):316-320.
Childe VG. 1936. Man Makes Himself.
Chitkara DK, Chumpitazi B, Krasinski SD, Grand RJ, and Montgomery RK. 2001. Regulation of human lactase-phlorizin hydrolase (LPH) gene by proteins binding to sites 5′ to the alu sequence. Gastroenterology 120(5):A304.
Coelho M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, Destro-Bisol G, and Rocha J. 2005. Microsatellite variation and evolution of human lactase persistence. Hum Genet 117(4):329-339.
Cook GC, and al-Torki MT. 1975. High intestinal lactase concentrations in adult Arbs in Saudi Arabia. Br Med J 3(5976):135-136.
Copley MS, Berstan R, Dudd SN, Docherty G, Mukherjee AJ, Straker V, Payne S, and Evershed RP. 2003. Direct chemical evidence for widespread dairying in prehistoric Britain. Proc Natl Acad Sci U S A 100(4):1524-1529.
Cordain L, Hickey MS, and K. K. 2012. Malaria and rickets represent selective forces for the convergent evolution of adult lactase persistence. In: Geptis P, Famula, T.R, Bettinger R. L.,, editor. Biodiversity in agriculture: domestication, evolution, and sustainablility. Cambridge: Cambridge University Press.
Corella D, Arregui M, Coltell O, Portoles O, Guillem-Saiz P, Carrasco P, Sorli JV, Ortega-Azorin C, Gonzalez JI, and Ordovas JM. 2011. Association of the LCT-13910C>T polymorphism with obesity and its modulation by dairy products in a Mediterranean population. Obesity (Silver Spring) 19(8):1707-1714.
Couzens L-M. 2011. [BSc Thesis]. London: University College London.
Craig O, Chapman J, Heron C, Willis L, Bertosiewics L, Taylor G, Whittle A, and Collins M. 2005. Did the first farmers of central and eastern Europe produce dairy foods? Antiquity 79:882-894.
Craig OE, Steele VJ, Fischer A, Hartz S, Andersen SH, Donohoe P, Glykou A, Saul H, Jones DM, Koch E et al. 2011. Ancient lipids reveal continuity in culinary practices across the transition to agriculture in Northern Europe. Proc Natl Acad Sci U S A 108(44):17910-17915.
Dahlqvist A, Hammond JB, Crane RK, Dunphy JV, and Littman A. 1963. Intestinal Lactase Deficiency and Lactose Intolerance in Adults. Preliminary Report. Gastroenterology 45:488-491.
Dahlqvist A, and Lindquist B. 1971. Lactose intolerance and protein malnutrition. Acta Paediatr Scand 60(4):488-494.

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, and White R. 1990. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6(3):575-577.
Day AJ, Canada FJ, Diaz JC, Kroon PA, McLauchlan R, Faulds CB, Plumb GW, Morgan MR, and Williamson G. 2000. Dietary flavonoid and isoflavone glycosides are hydrolysed by the lactase site of lactase phlorizin hydrolase. FEBS Lett 468(2-3):166-170.
Diamond J, and Bellwood P. 2003. Farmers and their languages: the first expansions. Science 300(5619):597-603.
Dudd SN, and Evershed RP. 1998. Direct demonstration of milk as an element of archaeological economies. Science 282(5393):1478-1481.
Dunne J, Evershed RP, Salque M, Cramp L, Bruni S, Ryan K, Biagetti S, and di Lernia S. 2012. First dairying in green Saharan Africa in the fifth millennium BC. Nature 486(7403):390-394.
Edwards CJ, Bollongino R, Scheu A, Chamberlain A, Tresset A, Vigne JD, Baird JF, Larson G, Ho SY, Heupink TH et al. 2007. Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. Proc Biol Sci 274(1616):1377-1385.
Ehrenkranz JR, Lewis NG, Kahn CR, and Roth J. 2005. Phlorizin: a review. Diabetes Metab Res Rev 21(1):31-38.
Enattah NS, Forsblom C, Rasinpera H, Tuomi T, Groop PH, and Jarvela I. 2004. The genetic variant of lactase persistence C (-13910) T as a risk factor for type I and II diabetes in the Finnish population. Eur J Clin Nutr 58(9):1319-1322.
Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF 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(1):57-72.
Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, and Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30(2):233-237.
Enattah NS, Trudeau A, Pimenoff V, Maiuri L, Auricchio S, Greco L, Rossi M, Lentze M, Seo JK, 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(3):615-625.
Escher JC, de Koning ND, van Engen CG, Arora S, Buller HA, Montgomery RK, and Grand RJ. 1992. Molecular basis of lactase levels in adult humans. J Clin Invest 89(2):480-483.
Evershed RP, Payne S, Sherratt AG, Copley MS, Coolidge J, Urem-Kotsu D, Kotsakis K, Ozdogan M, Ozdogan AE, Nieuwenhuyse O et al. 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. Nature 455(7212):528-531.
Excoffier L, and Lischer HE. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10(3):564-567.
Fajardo O, Naim HY, and Lacey SW. 1994. The polymorphic expression of lactase in adults is regulated at the messenger RNA level. Gastroenterology 106(5):1233-1241.
Fang L, Ahn JK, Wodziak D, and Sibley E. 2012. The human lactase persistence-associated SNP -13910*T enables in vivo functional persistence of lactase promoter-reporter transgene expression. Hum Genet 131(7):1153-1159.
Fang R, Olds LC, Santiago NA, and Sibley E. 2001. GATA family transcription factors activate lactase gene promoter in intestinal Caco-2 cells. Am J Physiol Gastrointest Liver Physiol 280(1):G58-67.
Fay JC, and Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155(3):1405-1413.
Fitzgerald K, Bazar L, and Avigan MI. 1998. GATA-6 stimulates a cell line-specific activation element in the human lactase promoter. Am J Physiol 274(2 Pt 1):G314-324.

Flatz G. 1984. Gene-dosage effect on intestinal lactase activity demonstrated in vivo. Am J Hum Genet 36(2):306-310.
Flatz G. 1987. Genetics of lactose digestion in humans. Adv Hum Genet 16:1-77.
Flatz G, and Rotthauwe HW. 1973. Lactose nutrition and natural selection. Lancet 2(7820):76-77.
Food Standards Agency. 2002. McCance and Widdowson's The composition of foods. Cambridge.
Freeman B, Smith N, Curtis C, Huckett L, Mill J, and Craig IW. 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(1):67-72.
Freund JN, Duluc I, Foltzer-Jourdainne C, Gosse F, and Raul F. 1990. Specific expression of lactase in the jejunum and colon during postnatal development and hormone treatments in the rat. Biochem J 268(1):99-103.
Freund JN, Jost B, Duluc I, and Morel G. 1995. Ultrastructural study of intestinal lactase gene expression. Biol Cell 83(2-3):211-217.
Friedrich DC, Santos SE, Ribeiro-dos-Santos AK, and Hutz MH. 2012. Several different lactase persistence associated alleles and high diversity of the lactase gene in the admixed Brazilian population. PLoS One 7(9):e46520.
Fu YX, and Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133(3):693-709.
Gallego Romero I, Basu Mallick C, Liebert A, Crivellaro F, Chaubey G, Itan Y, Metspalu M, Eaaswarkhanth M, Pitchappan R, Villems R et al. 2012. Herders of Indian and European cattle share their predominant allele for lactase persistence. Mol Biol Evol 29(1):249-260.
Gerbault P. 2013. Modeling demographic and evolutionary history: Integrating genetic and archaeological data. [PhD thesis]. London: University College London.
Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, and Thomas MG. 2011. Evolution of lactase persistence: an example of human niche construction. Philos Trans R Soc Lond B Biol Sci 366(1566):863-877.
Gerbault P, Moret C, Currat M, and Sanchez-Mazas A. 2009. Impact of selection and demography on the diffusion of lactase persistence. PLoS One 4(7):e6369.
Ghosh A. 2010. The genetics of lactase persistence in South and East Asia [BSc Thesis]. London: University College London.
Gillis RE. 2003. A Study of the Development of Lactose Tolerance, as an Example of Nutritional Genetic Adaptation to Domesticated Animal Products in the Old World Populations, Through the Analysis of Archaeological and Biomolecular Evidence: University of Sheffield, Department of Archaeology.
Giresi PG, Kim J, McDaniell RM, Iyer VR, and Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17(6):877-885.
Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-328.
Gronenborn D. 2007. Beyond the models: 'Neolithisation' in Central Europe. In: AlasdairWhittleDistinguished Research Professor SoH, Archaeology CU, and Fellow of the British Academy VickiCummingsLecturer in Archaeology UoCL, editors. Going Over: The Mesolithic-Neolithic Transition in North-West Europe: 'British Academy'.
Grunberg J, and Sterchi EE. 1995. Human lactase-phlorizin hydrolase: evidence of dimerization in the endoplasmic reticulum. Arch Biochem Biophys 323(2):367-372.
Guo SW, and Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48(2):361-372.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95-98.
Hammer HF, and Hammer J. 2012. Diarrhea caused by carbohydrate malabsorption. Gastroenterol Clin North Am 41(3):611-627.
Hammer HF, Petritsch W, Pristautz H, and Krejs GJ. 1996. Evaluation of the pathogenesis of flatulence and abdominal cramps in patients with lactose malabsorption. Wien Klin Wochenschr 108(6):175-179.
Harvey CB, Fox MF, Jeggo PA, Mantei N, Povey S, and Swallow DM. 1993. Regional localization of the lactase-phlorizin hydrolase gene, LCT, to chromosome 2q21. Ann Hum Genet 57(Pt 3):179-185.
Harvey CB, Hollox EJ, Poulter M, Wang Y, Rossi M, Auricchio S, Iqbal TH, Cooper BT, Barton R, Sarner M et al. 1998. Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. Ann Hum Genet 62(Pt 3):215-223.
Harvey CB, Pratt WS, Islam I, Whitehouse DB, and Swallow DM. 1995a. DNA polymorphisms in the lactase gene. Linkage disequilibrium across the 70-kb region. Eur J Hum Genet 3(1):27-41.
Harvey CB, Wang Y, Darmoul D, Phillips A, Mantei N, and Swallow DM. 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(2-3):135-140.
Harvey CB, Wang Y, Hughes LA, Swallow DM, Thurrell WP, Sams VR, Barton R, Lanzon-Miller S, and Sarner M. 1995b. Studies on the expression of intestinal lactase in different individuals. Gut 36(1):28-33.
Hauri HP, Sterchi EE, Bienz D, Fransen JA, and Marxer A. 1985. Expression and intracellular transport of microvillus membrane hydrolases in human intestinal epithelial cells. J Cell Biol 101(3):838-851.
Heitlinger LA, Rossi TM, Lee PC, and Lebenthal E. 1991. Human intestinal disaccharidase activities: correlations with age, biopsy technique, and degree of villus atrophy. J Pediatr Gastroenterol Nutr 12(2):204-208.
Hermisson J, and Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169(4):2335-2352.
Hertzler SR, and Savaiano DA. 1996. Colonic adaptation to daily lactose feeding in lactose maldigesters reduces lactose intolerance. Am J Clin Nutr 64(2):232-236.
Heyer E, Sibert A, and Austerlitz F. 2005. Cultural transmission of fitness: genes take the fast lane. Trends Genet 21(4):234-239.
Hillier LW, Graves TA, Fulton RS, Fulton LA, Pepin KH, Minx P, Wagner-McPherson C, Layman D, Wylie K, Sekhon M et al. 2005. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature 434(7034):724-731.
Ho MW, 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(4):650-657.
Hofer T, Foll M, and Excoffier L. 2012. Evolutionary forces shaping genomic islands of population differentiation in humans. BMC Genomics 13:107.
Hoffmann S, Tomasik G, and Polanski Z. 2012. DNA methylation, histone modifications and behaviour of AKAP95 during mouse oocyte growth and upon nuclear transfer of foreign chromatin into fully grown prophase oocytes. Folia Biol (Krakow) 60(3-4):163-170.
Holden C, and Mace R. 1997. Phylogenetic analysis of the evolution of lactose digestion in adults. Hum Biol 69(5):605-628.
Hollox E. 2005. Evolutionary genetics: genetics of lactase persistence--fresh lessons in the history of milk drinking. Eur J Hum Genet 13(3):267-269.
Hollox EJ. 2000. Molecular and population genetic analyses of variation within and surrounding the human lactase gene [PhD Thesis]. London: University of London.

Hollox EJ, Poulter M, Wang Y, Krause A, and Swallow DM. 1999. Common polymorphism in a highly variable region upstream of the human lactase gene affects DNA-protein interactions. Eur J Hum Genet 7(7):791-800.
Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, and Swallow DM. 2001. Lactase haplotype diversity in the Old World. Am J Hum Genet 68(1):160-172.
Hollox EJ, and Swallow DM. 2002. Lactase Deficiency: Biological and Medical Aspects of the Adult Human Lactase Polymorphism. In: Richard A. King JIR, Arno G. Motulsky, editor. The Genetic Basis of Common Diseases. 2 ed. New York: Oxford University Press. p 250-265.
Hudson RR, Kreitman M, and Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116(1):153-159.
Ihara Y. 2011. Evolution of culture-dependent discriminate sociality: a gene-culture coevolutionary model. Philos Trans R Soc Lond B Biol Sci 366(1566):889-900.
Imtiaz F, Savilahti E, Sarnesto A, Trabzuni D, Al-Kahtani K, Kagevi I, Rashed MS, Meyer BF, 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(10):e89.
Ingram CJ. 2008. The Evolutionary Genetics of Lactase Persistence in Africa and the Middle East [PhD Thesis]: University College London.
Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N et al.. 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? Hum Genet 120(6):779-788.
Ingram CJ, Mulcare CA, Itan Y, Thomas MG, and Swallow DM. 2009a. Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet 124(6):579-591.
Ingram CJ, Raga TO, Tarekegn A, Browning SL, Elamin MF, Bekele E, Thomas MG, Weale ME, Bradman N, and Swallow DM. 2009b. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. J Mol Evol 69(6):579-588.
Ingram CJ, and Swallow DM. 2009a. Lactose Malabsorption. In: McSweeney PLH, and Fox PF, editors. Advanced Dairy Chemistry 3ed. New York: Springer. p 203-229.
Ingram CJE, and Swallow DM. 2009b. Lactose Malabsorption. In: McSweeney PLH, and Fox PF, editors. Advanced Dairy Chemistry Volume 3: Lactose, Water, Salts and Minor Constituents. 3 ed. New York: Springer. p 203-229.
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-945.
Itan Y, Jones BL, Ingram CJ, Swallow DM, and Thomas MG. 2010. A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evol Biol 10:36.
Itan Y, Powell A, Beaumont MA, Burger J, and Thomas MG. 2009. The origins of lactase persistence in Europe. PLoS Comput Biol 5(8):e1000491.
Jacob R, Peters K, and Naim HY. 2002. The prosequence of human lactase-phlorizin hydrolase modulates the folding of the mature enzyme. J Biol Chem 277(10):8217-8225.
Jacob R, Radebach I, Wuthrich M, Grunberg J, Sterchi EE, and Naim HY. 1996. Maturation of human intestinal lactase-phlorizin hydrolase: generation of the brush border form of the enzyme involves at least two proteolytic cleavage steps. Eur J Biochem 236(3):789-795.
Jarvela I, Torniainen S, and Kolho KL. 2009. Molecular genetics of human lactase deficiencies. Ann Med:1-8.
Jenness R, Regehr EA, and Sloan RE. 1964. Comparative Biochemical Studies of Milk. Ii. Dialyzable Carbohydrates. Comp Biochem Physiol 13:339-352.
Jensen TG, Liebert A, Lewinsky R, Swallow DM, Olsen J, and Troelsen JT. 2011. The -14010*C variant associated with lactase persistence is located between an Oct-1

and HNF1alpha binding site and increases lactase promoter activity. Hum Genet 130(4):483-493.
Jobling MA, Hollox E, Hurles M, Tyler-Smith C, and Kivisild T. 2013. Human Evolutionary Genetics: Taylor & Francis Limited.
Jonas MM, Montgomery RK, and Grand RJ. 1985. Intestinal lactase synthesis during postnatal development in the rat. Pediatr Res 19(9):956-962.
Jones BL. 2012. Lactase enhancer diversity and adaptation for the lactase persistence trait in East African pastoralists [PhD]. London: University College London.
Jones BL, Raga TO, Liebert A, Zmarz P, Bekele E, Danielsen ET, Olsen AK, Bradman N, Troelsen JT, and Swallow DM. 2013. Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. Am J Hum Genet 93(3):538-544.
Jones BL, and Swallow DM. 2011. The impact of cis-acting polymorphisms on the human phenotype. Hugo J 5(1-4):13-23.
Kalendar R, Lee D, and Schulman AH. 2011. Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. Genomics 98(2):137-144.
Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, and Wingender E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31(13):3576-3579.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The human genome browser at UCSC. Genome Res 12(6):996-1006.
Keusch GT, Troncale FJ, Miller LH, Promadhat V, and Anderson PR. 1969. Acquired lactose malabsorption in Thai children. Pediatrics 43(4):540-545.
Khabarova Y, Torniainen S, Savilahti E, Isokoski M, Mattila K, and Jarvela I. 2010. The -13914G>A variant upstream of the lactase gene (LCT) is associated with lactase persistence/non-persistence. Scand J Clin Lab Invest 70(5):354-357.
Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217(5129):624-626.
Kimura M. 1983. The Neutral Theory of Molecular Evolution: Cambridge University Press.
Klein R. 1989. Specific expression of lactase in the jejunum and colon during. In: Lebenthal E, editor. Human Gastrointestinal Development. New York: Raven Press. p 367-392.
Krasinski SD, Upchurch BH, Irons SJ, June RM, Mishra K, Grand RJ, and Verhave M. 1997. Rat lactase-phlorizin hydrolase/human growth hormone transgene is expressed on small intestinal villi in transgenic mice. Gastroenterology 113(3):844-855.
Krasinski SD, Van Wering HM, Tannemaat MR, and Grand RJ. 2001. Differential activation of intestinal gene promoters: functional interactions between GATA-5 and HNF-1 alpha. Am J Physiol Gastrointest Liver Physiol 281(1):G69-84.
Kretchmer N. 1971. Lactose and lactase--a historical perspective. Gastroenterology 61(6):805-813.
Kruse TA, Bolund L, Grzeschik KH, Ropers HH, Sjostrom H, Noren O, Mantei N, and Semenza G. 1988. The human lactase-phlorizin hydrolase gene is located on chromosome 2. FEBS Lett 240(1-2):123-126.
Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, and Jarvela I. 2003. Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut 52(5):647-652.
Kuokkanen M, Kokkonen J, Enattah NS, 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(2):339-344.
Lacan M, Keyser C, Ricaut FX, Brucato N, Duranthon F, Guilaine J, Crubezy E, and Ludes B. 2011a. Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci U S A 108(24):9788-9791.
Lacan M, Keyser C, Ricaut FX, Brucato N, Tarrus J, Bosch A, Guilaine J, Crubezy E, and Ludes B. 2011b. Ancient DNA suggests the leading role played by men in the Neolithic dissemination. Proc Natl Acad Sci U S A 108(45):18255-18259.

Larson G, Albarella U, Dobney K, Rowley-Conwy P, Schibler J, Tresset A, Vigne JD, Edwards CJ, Schlumbaum A, Dinu A et al. 2007. Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. Proc Natl Acad Sci U S A 104(39):15276-15281.
Larsson SC, Orsini N, and Wolk A. 2006. Milk, milk products and lactose intake and ovarian cancer risk: a meta-analysis of epidemiological studies. Int J Cancer 118(2):431-441.
Lau W, Kuo TY, Tapper W, Cox S, and Collins A. 2007. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics 23(4):517-519.
Lebenthal E, Tsuboi K, and Kretchmer N. 1974. Characterization of human intestinal lactase and hetero-beta-galactosidases of infants and adults. Gastroenterology 67(6):1107-1113.
Lee MF, and Krasinski SD. 1998. Human adult-onset lactase decline: an update. Nutr Rev 56(1 Pt 1):1-8.
Lee SY, Wang Z, Lin CK, Contag CH, Olds LC, Cooper AD, and Sibley E. 2002. Regulation of intestine-specific spatiotemporal expression by the rat lactase promoter. J Biol Chem 277(15):13099-13105.
Leese HJ, and Semenza G. 1973. On the identity between the small intestinal enzymes phlorizin hydrolase and glycosylceramidase. J Biol Chem 248(23):8170-8173.
Lember M, Torniainen S, Kull M, Kallikorm R, Saadla P, Rajasalu T, Komu H, and Jarvela I. 2006. Lactase non-persistence and milk consumption in Estonia. World J Gastroenterol 12(45):7329-7331.
Leonard WR, and Crawford MH. 2002. The Human Biology of Pastoral Populations: Cambridge University Press.
Leonardi M, Gerbault P, Thomas MG, and Burger J. 2012. The evolution of lactase persistence in Europe. A synthesis of archaeological and genetic evidence. International Dairy Journal 22(2):88-97.
Lewinsky RH, Jensen TG, Moller J, Stensballe A, Olsen J, and Troelsen JT. 2005. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Hum Mol Genet 14(24):3945-3953.
Lewis MP, Simons GF, and Fenning CD. 2013. Ethnologue: Languages of the World. 17th Edition: SIL International.
Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49(1):49-67.
Librado P, and Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11):1451-1452.
Liu K, and Muse SV. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21(9):2128-2129.
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(2):524-529.
Lomer MC, Parkes GC, and Sanderson JD. 2008. Review article: lactose intolerance in clinical practice--myths and realities. Aliment Pharmacol Ther 27(2):93-103.
Luca F, Perry GH, and Di Rienzo A. 2010. Evolutionary adaptations to dietary changes. Annu Rev Nutr 30:291-314.
Maiuri L, Raia V, Potter J, Swallow D, Ho MW, Fiocca R, Finzi G, Cornaggia M, Capella C, Quaroni A et al. 1991. Mosaic pattern of lactase expression by villous enterocytes in human adult-type hypolactasia. Gastroenterology 100(2):359-369.
Malmstrom H, Linderholm A, Liden K, Stora J, Molnar P, Holmlund G, Jakobsson M, and Gotherstrom A. 2010. High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. BMC Evol Biol 10:89.

Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, and Morton NE. 2002. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci U S A 99(4):2228-2233.
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(9):2705-2713.
Martin CS, Wight PA, Dobretsova A, and Bronstein I. 1996. Dual luminescence-based reporter gene assay for luciferase and beta-galactosidase. Biotechniques 21(3):520-524.
Mattar R, de Campos Mazo DF, and Carrilho FJ. 2012. Lactose intolerance: diagnosis, genetic, and clinical factors. Clin Exp Gastroenterol 5:113-121.
Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34(Database issue):D108-110.
McCracken RD. 1971a. Lactase Deficiency: An Example of Dietary Evolution Current Anthropology 12:479-200.
McCracken RD. 1971b. Origins and implications of the distribution of adult lactase deficiency in human populations. J Trop Pediatr Environ Child Health 17(1):7-10.
McDonald JH, and Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328):652-654.
Meloni GF, Colombo C, La Vecchia C, Pacifico A, Tomasi P, Ogana A, Marinaro AM, and Meloni T. 2001. High prevalence of lactose absorbers in Northern Sardinian patients with type 1 and type 2 diabetes mellitus. Am J Clin Nutr 73(3):582-585.
Meloni GF, Colombo C, La Vecchia C, Ruggiu G, Mannazzu MC, Ambrosini G, and Cherchi PL. 1999. Lactose absorption in patients with ovarian cancer. Am J Epidemiol 150(2):183-186.
Meloni T, Colombo C, Ogana A, Mannazzu MC, and Meloni GF. 1996. Lactose absorption in patients with glucose 6-phosphate dehydrogenase deficiency with and without favism. Gut 39(2):210-213.
Meloni T, Colombo C, Ruggiu G, Dessena M, and Meloni GF. 1998. Primary lactase deficiency and past malarial endemicity in Sardinia. Ital J Gastroenterol Hepatol 30(5):490-493.
Metneki J, Czeizel A, Flatz SD, and Flatz G. 1984. A study of lactose absorption capacity in twins. Hum Genet 67(3):296-300.
Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153):553-560.
Mitchelmore C, Troelsen JT, 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 Pt 2:529-535.
Montgomery RK, Krasinski SD, Hirschhorn JN, and Grand RJ. 2007. Lactose and lactase--who is lactose intolerant and why? J Pediatr Gastroenterol Nutr 45 Suppl 2:S131-137.
Morton N, Maniatis N, Zhang W, Ennis S, and Collins A. 2007. Genome scanning by composite likelihood. Am J Hum Genet 80(1):19-28.
Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, and Collins A. 2001. The optimal measure of allelic association. Proc Natl Acad Sci U S A 98(9):5217-5221.
Mulcare CA. 2006. The Evolution of the Lactase Persistence Phenotype [PhD Thesis]. London: University of London.
Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, and Thomas MG. 2004. The T allele of a single-nucleotide polymorphism 13.9 kb

upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. Am J Hum Genet 74(6):1102-1110.
Murdock GP. 1967. Ethnographic Atlas: A Summary. Ethnology 6(2):109-236.
Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, and Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327(5967):876-879.
Myles S, Bouzekri N, Haverfield E, Cherkaoui M, Dugoujon JM, and Ward R. 2005. Genetic evidence in support of a shared Eurasian-North African dairying origin. Hum Genet 117(1):34-42.
Naim HY, Jacob R, Naim H, Sambrook JF, and Gething MJ. 1994. The pro region of human intestinal lactase-phlorizin hydrolase. J Biol Chem 269(43):26933-26943.
Naim HY, and Lentze MJ. 1992. Impact of O-glycosylation on the function of human intestinal lactase-phlorizin hydrolase. Characterization of glycoforms varying in enzyme activity and localization of O-glycoside addition. J Biol Chem 267(35):25494-25504.
Naim HY, 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(3):198-208.
Naim HY, Sterchi EE, and Lentze MJ. 1987. Biosynthesis and maturation of lactase-phlorizin hydrolase in the human small intestinal epithelial cells. Biochem J 241(2):427-434.
Nei M. 1987. Molecular Evolutionary Genetics: Columbia University Press.
Nei M, and Saitou N. 1986. Genetic relationship of human populations and ethnic differences in reaction to drugs and food. Prog Clin Biol Res 214:21-37.
Nemeth K, Plumb GW, Berrin JG, Juge N, Jacob R, Naim HY, Williamson G, Swallow DM, and Kroon PA. 2003. Deglycosylation by small intestinal epithelial cell beta-glucosidases is a critical step in the absorption and metabolism of dietary flavonoid glycosides in humans. Eur J Nutr 42(1):29-42.
Newcomer AD, and McGill DB. 1966. Distribution of disaccharidase activity in the small bowel of normal and lactase-deficient subjects. Gastroenterology 51(4):481-488.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet 39:197-218.
Olds LC, Ahn JK, and Sibley E. 2011. 13915*G DNA polymorphism associated with lactase persistence in Africa interacts with Oct-1. Hum Genet 129(1):111-113.
Olds LC, 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(18):2333-2340.
Olsen AK, Boyd M, Danielsen ET, and Troelsen JT. 2012. Current and emerging approaches to define intestinal epithelium-specific transcriptional networks. Am J Physiol Gastrointest Liver Physiol 302(3):G277-286.
Panzer P, Preuss U, Joberty G, and Naim HY. 1998. Protein domains implicated in intracellular transport and sorting of lactase-phlorizin hydrolase. J Biol Chem 273(22):13861-13869.
Parteli O. 1988. Die Zeit von 1918 bis 1970. In: Fontana J, editor. Geschichte des Landes Tirol. Vienna: Athesia Tyrolia. p 3-1499.
Payne S. 1973. Kill-off pattern in sheep and goats: the mandibles of Asvan kale. Anatolian Studies 23:139-147.
Pennings PS, and Hermisson J. 2006a. Soft sweeps II--molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol 23(5):1076-1084.
Pennings PS, and Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2(12):e186.
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(3):217-224.
Peuhkuri K, Vapaatalo H, Korpela R, and Teuri U. 2000. Lactose intolerance-a confusing clinical diagnosis. Am J Clin Nutr 71(2):600-602.
Pichler I, Mueller JC, Stefanov SA, De Grandi A, Volpato CB, Pinggera GK, Mayr A, Ogriseg M, Ploner F, Meitinger T et al. 2006. Genetic structure in contemporary south Tyrolean isolated populations revealed by analysis of Y-chromosome, mtDNA, and Alu polymorphisms. Hum Biol 78(4):441-464.
Pilson ME, and Kelly AL. 1962. Composition of the Milk from Zalophus californianus, the California Sea Lion. Science 135(3498):104-105.
Pinhasi R, Thomas MG, Hofreiter M, Currat M, and Burger J. 2012. The genetic history of Europeans. Trends Genet 28(10):496-505.
Pinto M, Robine-Leon S, Appay MD, Kedinger M, Triadou N, Dussaulx E, Lacroix B, Simon-Assmann P, Haffen K, Fogh J et al. 1983a. Enterocyte-like differentiation and polarization of the human colon carcinoma cell line Caco-2 in culture. Biol Cell(47):323-330.
Plantinga TS, Alonso S, Izagirre N, Hervella M, Fregel R, van der Meer JW, Netea MG, and de la Rua C. 2012. Low prevalence of lactase persistence in Neolithic South-West Europe. Eur J Hum Genet 20(7):778-782.
Potter J, Ho MW, Bolton H, Furth AJ, Swallow DM, and Griffiths B. 1985. Human lactase and the molecular basis of lactase persistence. Biochem Genet 23(5-6):423-439.
Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, and Swallow DM. 2003. The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet 67(Pt 4):298-311.
Richards MP, Schulting RJ, and Hedges RE. 2003. Archaeology: sharp shift in diet at onset of Neolithic. Nature 425(6956):366.
Rosensweig NS, Huang SS, and Bayless TM. 1967. Transmission of lactose intolerance. Lancet 2:777.
Rousset M. 1986. The human colon carcinoma cell lines HT-29 and Caco-2: two in vitro models for the study of intestinal differentiation. Biochimie 68(9):1035-1040.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-837.
Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, and Lander ES. 2006. Positive natural selection in the human lineage. Science 312(5780):1614-1620.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913-918.
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.
Sahi T, Launiala K, and Laitinen H. 1983. Hypolactasia in a fixed cohort of young Finnish adults. A follow-up study. Scand J Gastroenterol 18(7):865-870.
Salque M, Bogucki PI, Pyzel J, Sobkowiak-Tabaka I, Grygiel R, Szmyt M, and Evershed RP. 2013. Earliest evidence for cheese making in the sixth millennium BC in northern Europe. Nature 493(7433):522-525.
Sanger F, Nicklen S, and Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463-5467.
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(4):489-497.

Sharrocks AD. 2001. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol 2(11):827-837.
Shrier I, Szilagyi A, and Correa JA. 2008. Impact of lactose containing foods and the genetics of lactase on diseases: an analytical review of population data. Nutr Cancer 60(3):292-300.
Silberg DG, Swain GP, Suh ER, and Traber PG. 2000. Cdx1 and cdx2 expression during intestinal development. Gastroenterology 119(4):961-971.
Simoons FJ. 1970a. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis. Am J Dig Dis 15(8):695-710.
Simoons FJ. 1970b. The Traditional Limits of Milking and Milk Use in Southern Asia. Anthropos 65(3/4):547-593.
Simoons FJ. 1978. The geographic hypothesis and lactose malabsorption. A weighing of the evidence. Am J Dig Dis 23(11):963-980.
Skovbjerg H, Noren O, and Sjostrom H. 1978. Immunoelectrophoretic studies on human small intestinal brush border proteins. A qualitative study of the protein composition. Scand J Clin Lab Invest 38(8):723-729.
Skovbjerg H, Noren O, Sjostrom H, Danielsen EM, and Enevoldsen BS. 1982. Further characterization of intestinal lactase/phlorizin hydrolase. Biochim Biophys Acta 707(1):89-97.
Skovbjerg H, Sjostrom H, and Noren O. 1981. Purification and characterisation of amphiphilic lactase/phlorizin hydrolase from human small intestine. Eur J Biochem 114(3):653-661.
Smith AJ, Howard P, Shah S, Eriksson P, Stender S, Giambartolomei C, Folkersen L, Tybjaerg-Hansen A, Kumari M, Palmen J et al. 2012. Use of allele-specific FAIRE to determine functional regulatory polymorphism using large-scale genotyping arrays. PLoS Genet 8(8):e1002908.
Stephens M, and Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73(5):1162-1169.
Stephens M, and Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76(3):449-462.
Stephens M, Smith NJ, and Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978-989.
Sterchi EE, Mills PR, Fransen JA, Hauri HP, Lentze MJ, Naim HY, 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(4):1329-1337.
Stoneking M, and Krause J. 2011. Learning about human population history from ancient and modern genomes. Nat Rev Genet 12(9):603-614.
Swallow DM. 2003. Genetics of lactase persistence and lactose intolerance. Annu Rev Genet 37:197-219.
Swallow DM, Harvey C. B. 1993. Genetics of Adult-Type Hypolactasia. In: Aurichhio S, Semenza, G., editor. Common Food Intolerances 2: Milk in Human Nutrition and Adult-Type Hypolactasia. Basel: Dynamic Nutrition Research, Karger. p 85-92.
Swallow DM, and Hollox EJ. 2000. Genetic Polymorphism of Intestinal Lactase Activity in Adult Humans. In: Scriver CR, Beaudet AL, Sly WS, and Valle D, editors. The Metabolic and Molecular Basis of Inherited Disease 8ed. New York: McGraw-Hill. p 1651-1663.
Szilagyi A, Cohen A, Vinokuroff C, Ahmad D, Nathwani U, and Yesovitch S. 2004. Deadaption and readaptation with lactose, but no cross-adaptation to lactulose: a case of occult colonic bacterial adaptation. Can J Gastroenterol 18(11):677-680.
Tag CG, Oberkanins C, Kriegshauser G, Ingram CJ, Swallow DM, Gressner AM, Ledochowski M, and Weiskirchen R. 2008. Evaluation of a novel reverse-hybridization

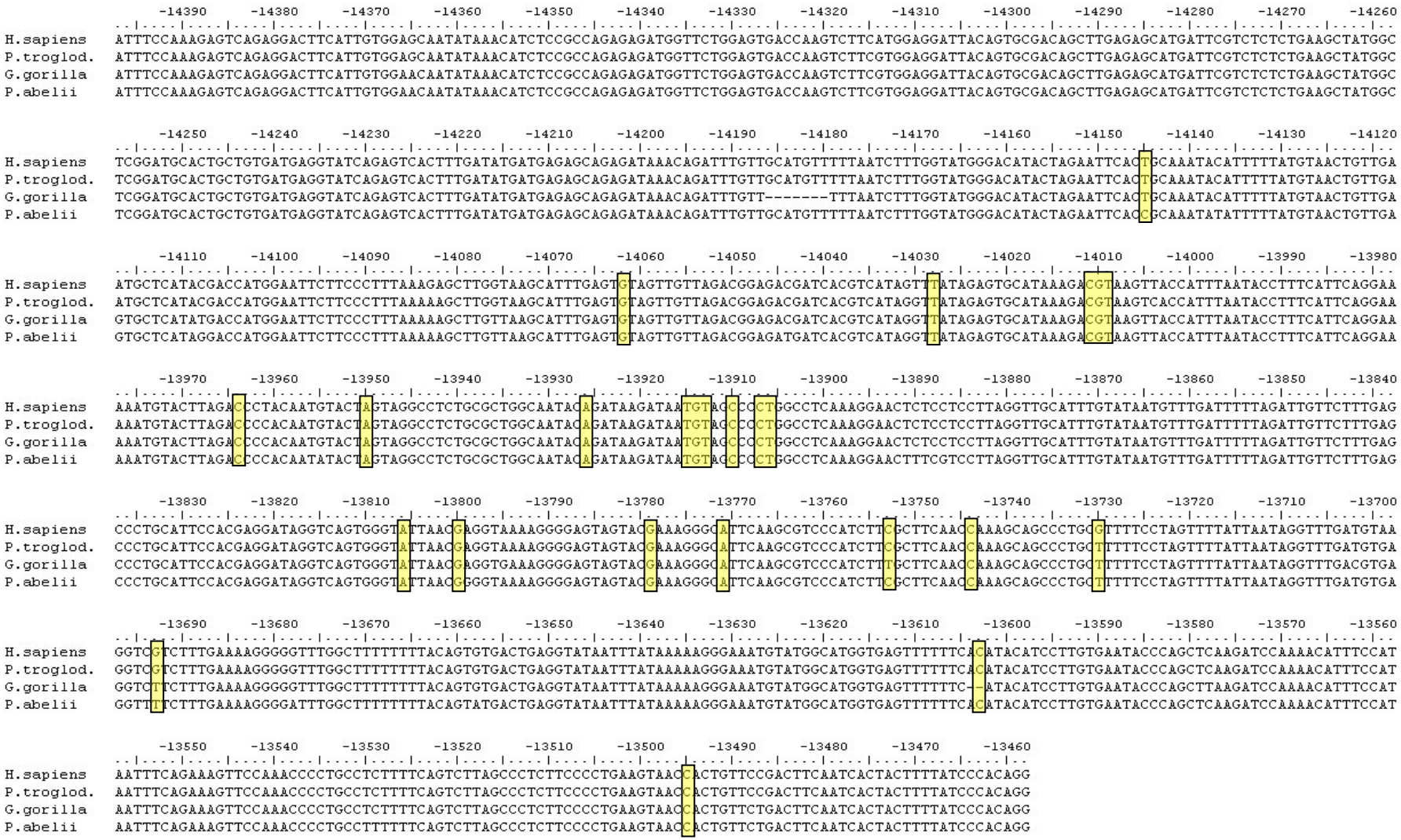
StripAssay for typing DNA variants useful in diagnosis of adult-type hypolactasia. Clin Chim Acta 392(1-2):58-62.
Tag CG, Schifflers MC, Mohnen M, Gressner AM, and Weiskirchen R. 2007. A novel proximal -13914G>A base replacement in the vicinity of the common-13910T/C lactase gene variation results in an atypical LightCycler melting curve in testing with the MutaREAL Lactase test. Clin Chem 53(1):146-148.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585-595.
Tapper W. 2007. Linkage disequilibrium maps and location databases. Methods Mol Biol 376:23-45.
Thompson JD, Higgins DG, and Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673-4680.
Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39(1):31-40.
Torniainen S, Parker MI, Holmberg V, Lahtela E, Dandara C, and Jarvela I. 2009. Screening of variants for lactase persistence/non-persistence in populations from South Africa and Ghana. BMC Genet 10:31.
Tresset A, and Vigne J-D. 2007. Substitution of species, techniques and symbold at the Mesolithic-Neolithic transition in Western Europe. In: Whittle A, and Cummings V, editors. Going Over: The Mesolithis-Neolithic Transition in North West Europe: The Mesolithic-Neolithic Transition in North-West Europe (Proceedings of the British Academy). Oxford: Oxford University Press.
Troelsen JT. 2005. Adult-type hypolactasia and regulation of lactase expression. Biochim Biophys Acta 1723(1-3):19-32.
Troelsen JT, Mehlum A, Olsen J, Spodsberg N, Hansen GH, Prydz H, Noren O, and Sjostrom H. 1994. 1 kb of the lactase-phlorizin hydrolase promoter directs post-weaning decline and small intestinal-specific expression in transgenic mice. FEBS Lett 342(3):291-296.
Troelsen JT, Mitchelmore C, Spodsberg N, Jensen AM, Noren O, and Sjostrom H. 1997. Regulation of lactase-phlorizin hydrolase gene expression by the caudal-related homoeodomain protein Cdx-2. Biochem J 322 (Pt 3):833-838.
Troelsen JT, Olsen J, Moller J, and Sjostrom H. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology 125(6):1686-1694.
Troelsen JT, Olsen J, Noren O, and Sjostrom H. 1992. A novel intestinal trans-factor (NF-LPH1) interacts with the lactase-phlorizin hydrolase promoter and co-varies with the enzymatic activity. J Biol Chem 267(28):20407-20411.
Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, and Bradley DG. 2001. Genetic evidence for Near-Eastern origins of European cattle. Nature 410(6832):1088-1091.
van Wering HM, Bosse T, Musters A, de Jong E, de Jong N, Hogen Esch CE, Boudreau F, Swain GP, Dowling LN, Montgomery RK et al. 2004. Complex regulation of the lactase-phlorizin hydrolase promoter by GATA-4. Am J Physiol Gastrointest Liver Physiol 287(4):G899-909.
van Wering HM, Huibregtse IL, van der Zwan SM, de Bie MS, Dowling LN, Boudreau F, Rings EH, Grand RJ, and Krasinski SD. 2002. Physical interaction between GATA-5 and hepatocyte nuclear factor-1alpha results in synergistic activation of the human lactase-phlorizin hydrolase promoter. J Biol Chem 277(31):27659-27667.
Vigne J-D, and Helmer D. 2007. Was milk a 'secondary product' in the Old World Neolithisation process? Its role in the domestication of cattle, sheep and goats. Anthropozoologica 42(3):9-40.

Villako K, and Maaroos H. 1994. Clinical picture of hypolactasia and lactose intolerance. Scand J Gastroenterol Suppl 202:36-54.
Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol 4(3):e72.
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 with other glycosidases, the membrane anchor of lactase-phlorizin hydrolase. J Biol Chem 267(26):18744-18752.
Wang Y, Harvey C, Rousset M, and Swallow DM. 1994. Expression of human intestinal mRNA transcripts during development: analysis by a semiquantitative RNA polymerase chain reaction method. Pediatr Res 36(4):514-521.
Wang Y, Harvey CB, Hollox EJ, Phillips AD, Poulter M, Clay P, Walker-Smith JA, and Swallow DM. 1998. The genetically programmed down-regulation of lactase in children. Gastroenterology 114(6):1230-1236.
Wang Y, Harvey CB, Pratt WS, Sams VR, Sarner M, Rossi M, Auricchio S, and Swallow DM. 1995. The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. Hum Mol Genet 4(4):657-662.
Weston R, Peeters H, and Ahel D. 2012. ZRANB3 is a structure-specific ATP-dependent endonuclease involved in replication stress response. Genes Dev 26(14):1558-1572.
Witte J, Lloyd M, Lorenzsonn V, Korsmo H, and Olsen W. 1990. The biosynthetic basis of adult lactase deficiency. J Clin Invest 86(4):1338-1342.
Wixman R. 1984. The Peoples of the USSR: An Ethnographic Handbook: M.E. Sharpe.
Wright S. 1931. Evolution in Mendelian Populations. Genetics 16(2):97-159.
Wright S. 1950. Genetical structure of populations. Nature 166(4215):247-249.
Wright SI, and Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168(2):1071-1076.
Wuthrich M, Grunberg J, Hahn D, Jacob R, Radebach I, Naim HY, and Sterchi EE. 1996. Proteolytic processing of human lactase-phlorizin hydrolase is a two-step event: identification of the cleavage sites. Arch Biochem Biophys 336(1):27-34.
Zecca L, Mesonero JE, Stutz A, Poiree JC, Giudicelli J, Cursio R, Gloor SM, and Semenza G. 1998. Intestinal lactase-phlorizin hydrolase (LPH): the two catalytic sites; the role of the pancreas in pro-LPH maturation. FEBS Lett 435(2-3):225-228.
Zeder MA. 2006. Documenting Domestication: New Genetic and Archaeological Paradigms: University of California Press.
Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci U S A 105(33):11597-11604.
Zmarz P. 2010. Evolutionary Genetics of Lactase Persistence [MSc Thesis]. London: University College London.

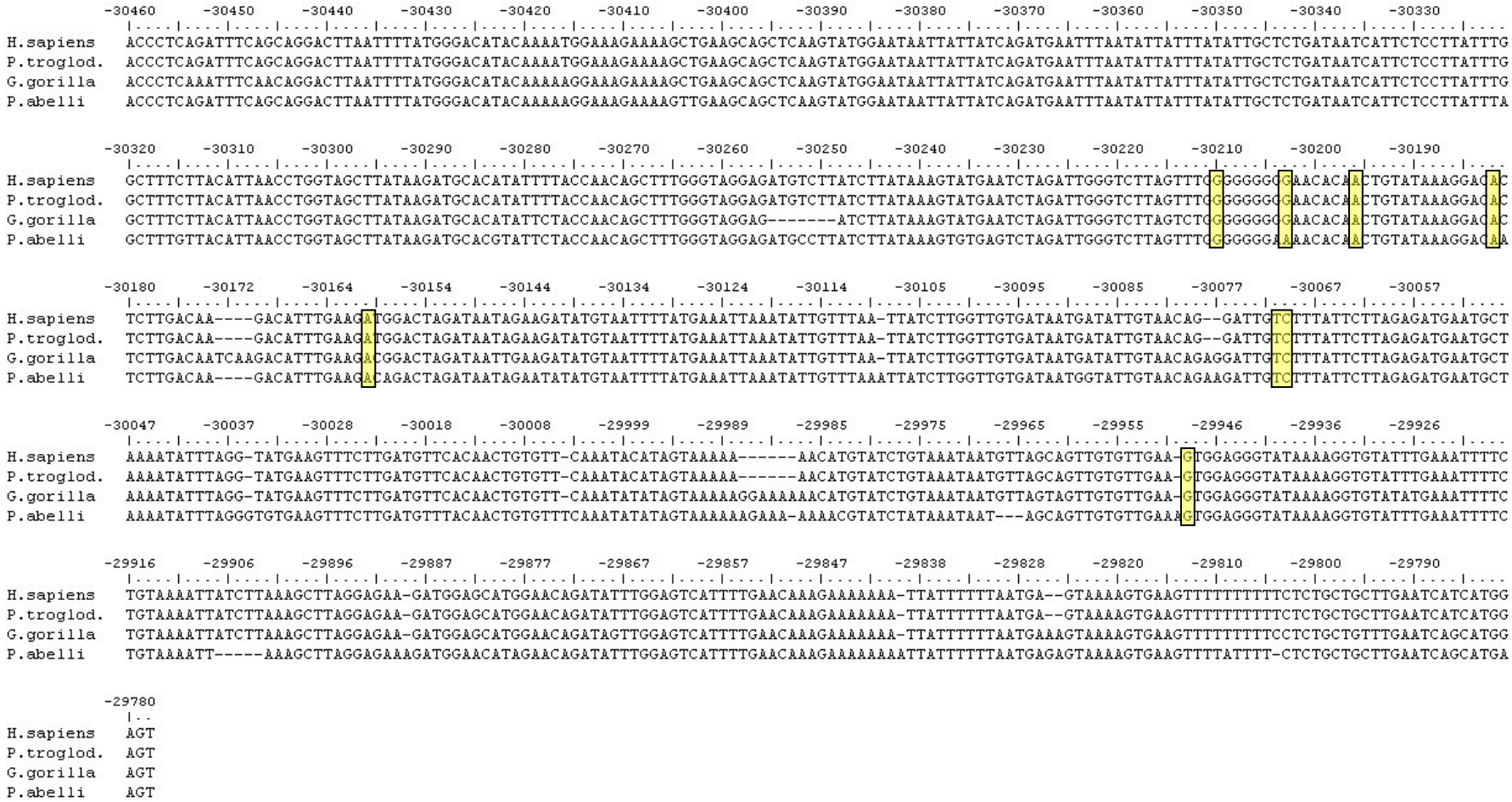
Appendices

Appendix A: Sequence alignments of the regions amplified during this thesis. Nucleotide positions in relation to the start of transcription of *LCT*. Positions of variants found are highlighted. Aligned species: Homo sapiens (H. sapiens), Pan troglodytes (P. troglod.), Gorilla gorilla (G. gorilla) and Pongo abelii (P. abelii).

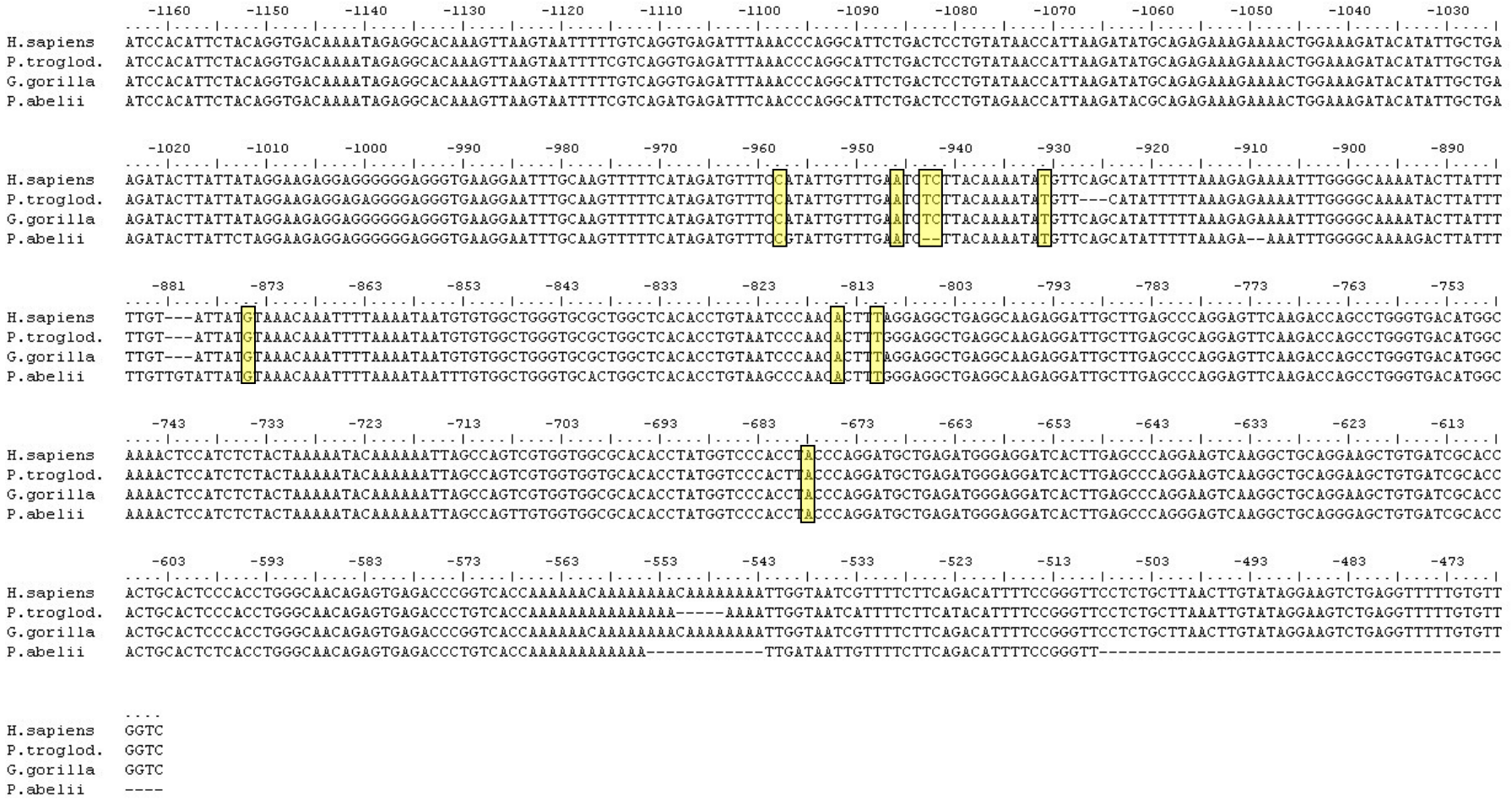
LCT enhancer, intron 13 of *MCM6*



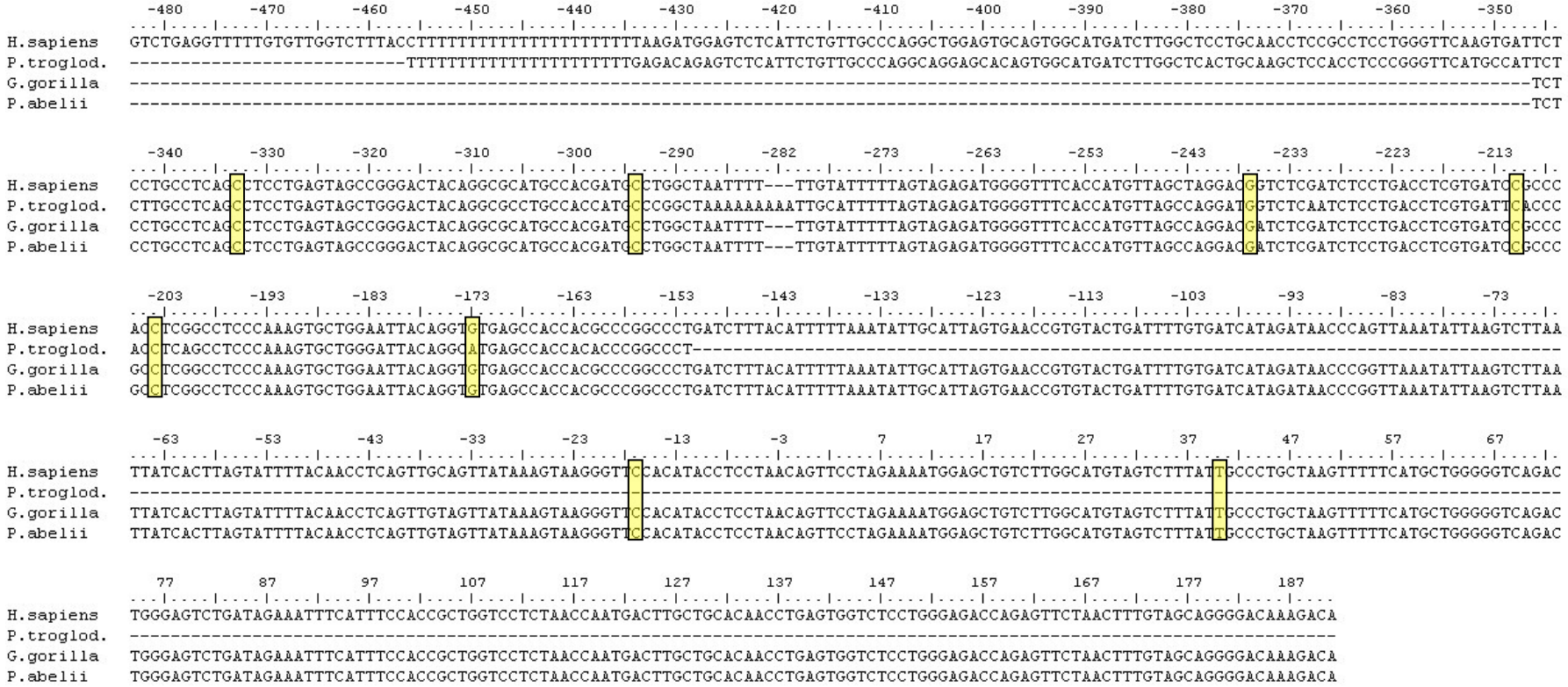
Intron 4*, MCM6*



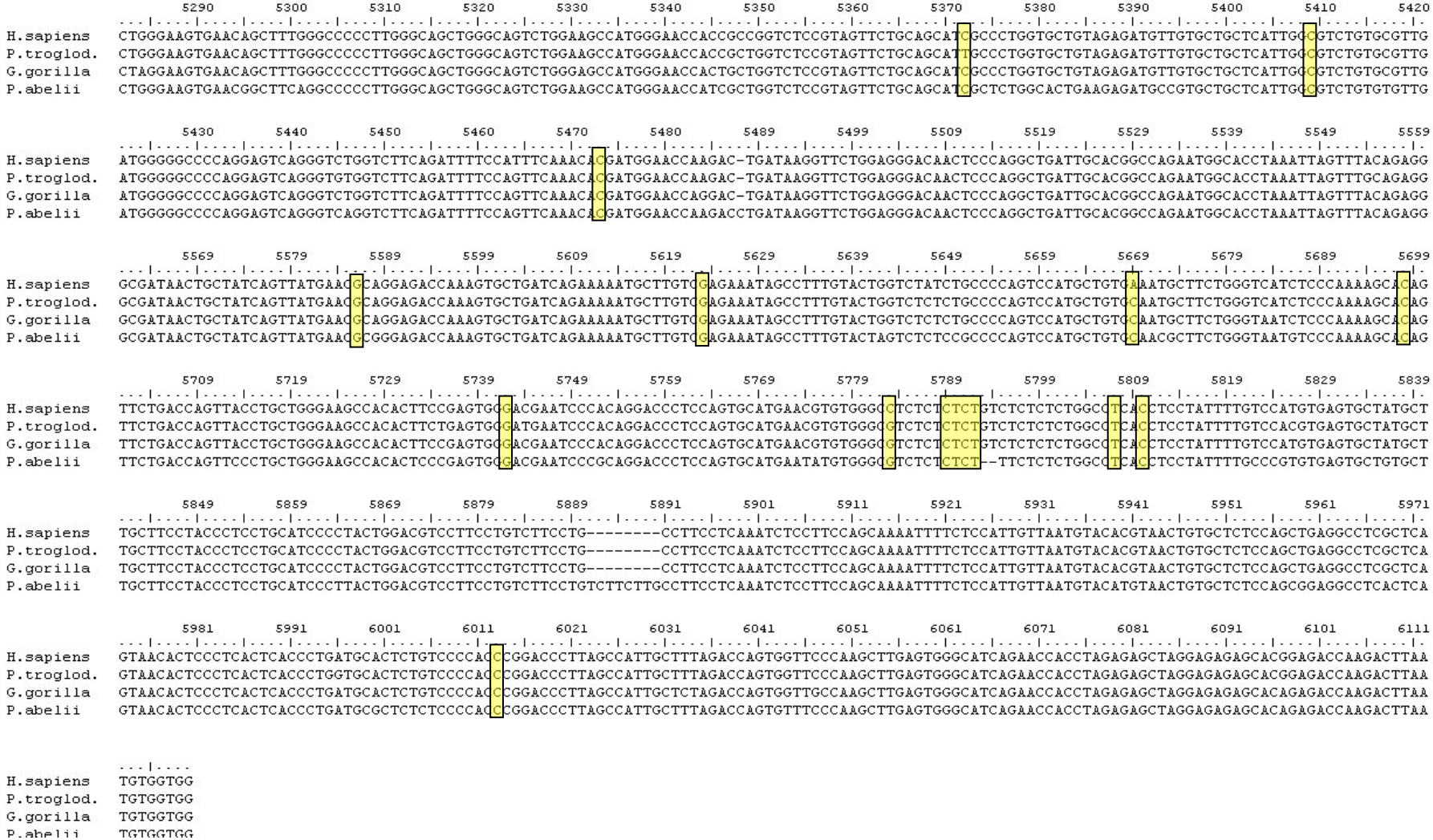
Hapdef region



Immediate promoter



Intron2, *LCT*


5290 5300 5310 5320 5330 5340 5350 5360 5370 5380 5390 5400 5410 5420
H.sapiens CTGGGAAGTGAACAGCTTTGGGCCCCCTTGGGCAGCTGGGCAGTCTGGAAGCCATGGGAACCACCGCCGGTCTCCGTAGTTCTGCAGCATCGCCCTGGTGCTGTAGAGATGTTGTGCTGCTCATTGGCGTCTGTGCGTTG
P.troglod. CTGGGAAGTGAACAGCTTTGGGCCCCCTTGGGCAGCTGGGCAGTCTGGAAGCCATGGGAACCACCGCTGGTCTCCGTAGTTCTGCAGCATTGCCCTGGTGCTGTAGAGATGTTGTGCTGCTCATTGGCGTCTGTGCGTTG
G.gorilla CTAGGAAGTGAACAGCTTTGGGCCCCCTTGGGCAGCTGGGCAGTCTGGGAGCCATGGGAACCACTGCTGGTCTCCGTAGTTCTGCAGCATCGCCCTGGTGCTGTAGAGATGTTGTGCTGCTCATTGGCGTCTGTGCGTTG
P.abelii CTGGGAAGTGAACGGCTTCAGGCCCCCTTGGGCAGCTGGGCAGTCTGGAAGCCATGGGAACCATCGCTGGTCTCCGTAGTTCTGCAGCATCGCTCTGGCACTGAAGAGATGCCGTGCTGCTCATTGGCGTCTGTGTGTTG
5430 5440 5450 5460 5470 5480 5489 5499 5509 5519 5529 5539 5549 5559
H.sapiens ATGGGGGCCCCAGGAGTCAGGGTCTGGTCTTCAGATTTTCCATTTCAAACACGATGGAACCAAGAC-TGATAAGGTTCTGGAGGGACAACTCCCAGGCTGATTGCACGGCCAGAATGGCACCTAAATTAGTTTACAGAGG
P.troglod. ATGGGGGCCCCAGGAGTCAGGGTGTGGTCTTCAGATTTTCCAGTTCAAACACGATGGAACCAAGAC-TGATAAGGTTCTGGAGGGACAACTCCCAGGCTGATTGCACGGCCAGAATGGCACCTAAATTAGTTTGCAGAGG
G.gorilla ATGGGGGCCCCAGGAGTCAGGGTCTGGTCTTCAGATTTTCCAGTTCAAACACGATGGAACCAGGAC-TGATAAGGTTCTGGAGGGACAACTCCCAGGCTGATTGCACGGCCAGAATGGCACCTAAATTAGTTTACAGAGG
P.abelii ATGGGGGCCCCAGGAGTCAGGGTCAGGTCTTCAGATTTTCCAGTTCAAACACGATGGAACCAAGACCTGATAAGGTTCTGGAGGGACAACTCCCAGGCTGATTGCACGGCCAGAATGGCACCTAAATTAGTTTACAGAGG
5569 5579 5589 5599 5609 5619 5629 5639 5649 5659 5669 5679 5689 5699
H.sapiens GCGATAACTGCTATCAGTTATGAACGCAGGAGACCAAAGTGCTGATCAGAAAAATGCTTGTCGAGAAATAGCCTTTGTACTGGTCTATCTGCCCCAGTCCATGCTGTCAAATGCTTCTGGGTCATCTCCCAAAAGCACAG
P.troglod. GCGATAACTGCTATCAGTTATGAACGCAGGAGACCAAAGTGCTGATCAGAAAAATGCTTGTCGAGAAATAGCCTTTGTACTGGTCTCTCTGCCCCAGTCCATGCTGTCCAATGCTTCTGGGTCATCTCCCAAAAGCACAG
G.gorilla GCGATAACTGCTATCAGTTATGAACGCAGGAGACCAAAGTGCTGATCAGAAAAATGCTTGTCGAGAAATAGCCTTTGTACTGGTCTCTCTGCCCCAGTCCATGCTGTCCAATGCTTCTGGGTAATCTCCCAAAAGCACAG
P.abelii GCGATAACTGCTATCAGTTATGAACGCGGGAGACCAAAGTGCTGATCAGAAAAATGCTTGTCGAGAAATAGCCTTTGTACTAGTCTCTCCGCCCCAGTCCATGCTGTCCAACGCTTCTGGGTAATGTCCCAAAAGCACAG
5709 5719 5729 5739 5749 5759 5769 5779 5789 5799 5809 5819 5829 5839
H.sapiens TTCTGACCAGTTACCTGCTGGGAAGCCACACTTCCGAGTGGGACGAATCCCACAGGACCCTCCAGTGCATGAACGTGTGGGCGTCTCTCTCTGTCTCTCTCTGGCGTCACCTCCTATTTTGTCCATGTGAGTGCTATGCT
P.troglod. TTCTGACCAGTTACCTGCTGGGAAGCCACACTTCTGAGTGGGATGAATCCCACAGGACCCTCCAGTGCATGAACGTGTGGGCGTCTCTCTCTGTCTCTCTCTGGCGTCACCTCCTATTTTGTCCACGTGAGTGCTATGCT
G.gorilla TTCTGACCAGTTACCTGCTGGGAAGCCACACTTCCGAGTGGGACGAATCCCACAGGACCCTCCAGTGCATGAACGTGTGGGCGTCTCTCTCTGTCTCTCTCTGGCGTCACCTCCTATTTTGTCCATGTGAGTGCTATGCT
P.abelii TTCTGACCAGTTCCCTGCTGGGAAGCCACACTCCCGAGTGGGACGAATCCCGCAGGACCCTCCAGTGCATGAATATGTGGGCGTCTCTCTCTCT--TTCTCTCTGGCGTCACCTCCTATTTTGCCCGTGTGAGTGCTGTGCT
5849 5859 5869 5879 5889 5891 5901 5911 5921 5931 5941 5951 5961 5971
H.sapiens TGCTTCCTACCCTCCTGCATCCCCTACTGGACGTCCTTCCTGTCTTCCTG--------CCTTCCTCAAATCTCCTTCCAGCAAAATTTTCTCCATTGTTAATGTACACGTAACTGTGCTCTCCAGCTGAGGCCTCGCTCA
P.troglod. TGCTTCCTACCCTCCTGCATCCCCTACTGGACGTCCTTCCTGTCTTCCTG--------CCTTCCTCAAATCTCCTTCCAGCAAAATTTTCTCCATTGTTAATGTACACGTAACTGTGCTCTCCAGCTGAGGCCTCGCTCA
G.gorilla TGCTTCCTACCCTCCTGCATCCCCTACTGGACGTCCTTCCTGTCTTCCTG--------CCTTCCTCAAATCTCCTTCCAGCAAAATTTTCTCCATTGTTAATGTACACGTAACTGTGCTCTCCAGCTGAGGCCTCGCTCA
P.abelii TGCTTCCTACCCTCCTGCATCCCTTACTGGACGTCCTTCCTGTCTTCCTGTCTTCTTGCCTTCCTCAAATCTCCTTCCAGCAAAATTTTCTCCATTGTTAATGTACATGTAACTGTGCTCTCCAGCGGAGGCCTCACTCA
5981 5991 6001 6011 6021 6031 6041 6051 6061 6071 6081 6091 6101 6111
H.sapiens GTAACACTCCCTCACTCACCCTGATGCACTCTGTCCCCACGCCGGACCCTTAGCCATTGCTTTAGACCAGTGGTTCCCAAGCTTGAGTGGGCATCAGAACCACCTAGAGAGCTAGGAGAGAGCACGGAGACCAAGACTTAA
P.troglod. GTAACACTCCCTCACTCACCCTGGTGCACTCTGTCCCCACGCCGGACCCTTAGCCATTGCTTTAGACCAGTGGTTCCCAAGCTTGAGTGGGCATCAGAACCACCTAGAGAGCTAGGAGAGAGCACGGAGACCAAGACTTAA
G.gorilla GTAACACTCCCTCACTCACCCTGATGCACTCTGTCCCCACGCCGGACCCTTAGCCATTGCTCTAGACCAGTGGTTGCCAAGCTTGAGTGGGCATCAGAACCACCTAGAGAGCTAGGAGAGAGCACAGAGACCAAGACTTAA
P.abelii GTAACACTCCCTCACTCACCCTGATGCGCTCTCTCCCCACGCCGGACCCTTAGCCATTGCTTTAGACCAGTGTTTCCCAAGCTTGAGTGGGCATCAGAACCACCTAGAGAGCTAGGAGAGAGCACAGAGACCAAGACTTAA
H.sapiens TGTGGTGG
P.troglod. TGTGGTGG
G.gorilla TGTGGTGG
P.abelii TGTGGTGG

Appendix B1: Frequencies of lactase persistence associated alleles, - indicates that not genotyped at that position.

Continent/Region	Country	Population/Region	Longitude	Latitude	Number of chromosomes	rs145946881 -14010 G>C	ss820486563 -14009 T>G	rs41380347 -13915 T>G	rs4988235 -13910 C>T	rs41525747 -13907 C>G	Sum of all LP assoicated alleles	Reference
Africa	Algeria	Algerian	-1.32	34.88	21	-	-	-	0.33	-	0.33	Mulcare (2006) PhD thesis, UCL
Africa	Algeria	Berber Mzab	3.68	32.49	66	0.00	0.00	0.00	0.17	0.00	0.17	Myles et al. (2005) Hum Genet. 117, 34.
Africa	Algeria	Mozabite	3.68	32.49	60	-	-	-	0.22	-	0.22	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Africa	Angola	Kuvale	12.15	-15.20	108	0.06	0.00	0.00	0.00	0.00	0.06	Coelho et al. (2009) BMC Evol Biol. 9, 80.
Africa	Angola	Nyaneka-Nkhumbi	12.15	-15.20	306	0.03	0.00	0.00	0.00	0.00	0.03	Coelho et al. (2009) BMC Evol Biol. 9, 80.
Africa	Angola	Ovimbundu	12.15	-15.20	192	0.01	0.00	0.00	0.00	0.00	0.01	Coelho et al. (2009) BMC Evol Biol. 9, 80.
Africa	Cameroon	Fulani	14.15	11.05	102	0.00	0.00	0.00	0.39	0.00	0.39	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Africa	Cameroon	Fulbe	14.15	11.05	102	-	-	-	0.21	-	0.21	Coelho et al. (2005) Hum Genet. 117, 329
Africa	Cameroon	Hausa	11.55	6.47	36	-	-	-	0.14	-	0.14	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Africa	Cameroon	Mambila	12.50	6.00	134	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Cameroon	Nso	10.67	6.20	252	-	-	-	0.00	-	0.00	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Africa	Cameroon	Pygmy	14.75	2.92	36	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Cameroon	Shuwa Arabs	14.50	13.00	124	0.00	0.00	0.08	0.02	0.00	0.10	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Cameroon	Yamba	11.55	6.47	42	-	-	-	0.00	-	0.00	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Africa	Congo	Brazzaville	15.28	-4.27	104	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Afar	41.36039	11.60212	152	0.01	0.01	0.18	0.01	0.24	0.45	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Amhara	38.65951	9.869192	152	0.00	0.03	0.04	0.00	0.06	0.13	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Anuak	34.41219	7.953241	138	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Maale	36.64333	5.714812	132	0.00	0.05	0.04	0.00	0.02	0.11	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Manjo	36.23	7.27	80	0.00	0.03	0.00	0.00	0.01	0.04	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Nuer	34.58	8.25	74	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Oromo	37.30817	7.83736	152	0.01	0.03	0.07	0.00	0.07	0.18	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Shabo	35.41	7.56	48	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.

												Jones (2012) PhD thesis, UCL, unpublished data,
Africa	Ethiopia	Somali	41.87	9.58	186	0.01	0.02	0.05	0.02	0.06	0.16	Ingram (2009) J Mol Evol. 69, 579.
Africa	Ethiopia	Somali (phenotyped)	41.87	9.58	218	0.00	0.01	0.05	0.02	0.06	0.15	Ingram (2009) J Mol Evol. 69, 579.
Africa	Ethiopia	Suri	35.59	7.00	100	0.04	0.00	0.00	0.00	0.00	0.04	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ethiopia	Amhara (phenotyped)	38.65951	9.869192	108	0.00	0.05	0.06	0.00	0.02	0.12	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Oromo (phenotyped)	37.30817	7.83736	150	0.01	0.03	0.13	0.00	0.05	0.23	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Tigray	39.47	13.50	88	0.00	0.03	0.11	0.00	0.19	0.34	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Wolayta	37.76	6.84	52	0.02	0.04	0.12	0.00	0.08	0.25	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ghana	Akan	-1.02	7.95	392	0.00	0.00	0.00	0.00	0.00	0.00	Torniainen et al. (2009) BMC Genet. 10, 31.
Africa	Ghana	Asante	-0.55	6.17	70	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Ghana	Builsa	-1.29	10.53	42	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Kenya	Borana	37.91	-0.02	16	0.13	-	0.19	0.00	0.13	0.44	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Burji	37.91	-0.02	16	0.06	-	0.00	0.00	0.00	0.06	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	El Molo	37.91	-0.02	18	0.11	-	0.00	0.00	0.00	0.11	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Gabra	37.91	-0.02	18	0.00	-	0.28	0.00	0.11	0.39	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Kikuyu	37.91	-0.02	4	0.75	-	0.00	0.00	0.00	0.75	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Konso	37.91	-0.02	12	0.08	-	0.08	0.00	0.00	0.17	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Maasai	37.91	-0.02	64	0.58	-	0.00	0.00	0.03	0.61	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Marakwet	37.91	-0.02	14	0.36	-	0.07	0.00	0.00	0.43	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Nandi	37.91	-0.02	8	0.25	-	0.00	0.00	0.00	0.25	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Ogiek	37.91	-0.02	22	0.36	-	0.00	0.00	0.00	0.36	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Pokot	37.91	-0.02	28	0.29	-	0.04	0.00	0.00	0.32	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Rendille	37.91	-0.02	16	0.13	-	0.13	0.00	0.06	0.31	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Sabaot	37.91	-0.02	12	0.17	-	0.00	0.00	0.00	0.17	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Samburu	37.91	-0.02	18	0.28	-	0.06	0.00	0.06	0.40	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Sengwer	37.91	-0.02	32	0.06	-	0.00	0.00	0.00	0.06	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Somali	37.91	-0.02	2	0.00	-	0.50	0.00	0.00	0.50	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Tugen	37.91	-0.02	32	0.19	-	0.00	0.00	0.00	0.19	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Turkana	37.91	-0.02	26	0.21	-	0.00	0.00	0.00	0.21	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Wata	37.91	-0.02	2	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Kenya	Yaaku	37.91	-0.02	28	0.54	-	0.00	0.00	0.04	0.58	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Malawi	Bantu	33.78	-13.98	310	-	-	-	0.00	-	0.00	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Africa	Malawi	Chewa	33.78	-13.98	100	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Morocco	Berber	-3.77	34.05	154	-	-	-	0.14	-	0.14	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Africa	Morocco	Moroccan	-6.84	34.03	180	-	-	-	0.18	-	0.18	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Africa	Morocco	Moroccan	-6.84	34.03	24	0.00	0.00	0.08	0.21	0.00	0.29	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Africa	Morocco	Saharawi	-6.84	34.03	114	-	-	-	0.26	-	0.26	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Africa	Morocco	Saharawi	-6.84	34.03	22	0.00	0.00	0.18	0.23	0.00	0.41	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Africa	Morocco	Berber Amizmiz (High-Atlas)	-8.23	31.21	78	0.00	0.00	0.00	0.14	0.00	0.14	Myles et al. (2005) Hum Genet. 117, 34.
Africa	Morocco	Berber Moyen-Atlas (Mid-Atlas)	-4.84	33.12	66	0.00	0.00	0.00	0.16	0.00	0.16	Myles et al. (2005) Hum Genet. 117, 34.

Africa	Mozambique	Mozambique (Maputo, Ronga Bantu)	32.58	-25.97	94	-	-	-	0.01	-	0.01	Coelho et al. (2005) Hum Genet. 117, 329
Africa	Mozambique	Sena	35.05	-17.44	136	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	N.E. Kenya	Bantu	35.49	3.32	24	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Africa	Namibia	San	18.19	-25.60	14	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Africa	Namibia	San	18.19	-25.60	34	0.06	0.00	0.00	0.00	0.00	0.06	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Nigeria	Yoruba	3.47	7.23	50	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Africa	São Tomé and Príncipe	Sao Tome	6.73	0.34	284	-	-	-	0.05	-	0.05	Coelho et al. (2005) Hum Genet. 117, 329
Africa	Senegal	Mandjak	-15.88	12.986	114	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Senegal	Wolof	-17.453	14.687	138	0.00	0.00	0.00	0.00	0.00	0.00	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Somalia	Somali	45.37	2.07	158	-	-	-	0.03	-	0.03	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Africa	South Africa	Bantu	28.08	-26.20	16	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Africa	South Africa	Mixed (for example with Europeans)	18.42	-33.92	124	0.07	0.00	0.00	0.22	0.00	0.28	Torniainen et al. (2009) BMC Genet. 10, 31.
Africa	South Africa	South Africans	28.23	-25.71	40	0.08	0.00	0.00	0.00	0.00	0.08	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	South Africa	Xhosa	20.66	-34.16	218	0.13	0.00	0.00	0.00	0.00	0.13	Torniainen et al. (2009) BMC Genet. 10, 31.
Africa	Sudan	Ama	30.22	12.86	4	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Beja (Banuamir)	30.22	12.86	12	0.00	-	0.17	0.00	0.25	0.42	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Beja (Hadandawa)	30.22	12.86	22	0.00	-	0.09	0.00	0.18	0.27	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Beni Amer	37.22	19.62	170	0.00	0.11	0.24	0.01	0.06	0.42	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Sudan	Dinka	30.22	12.86	18	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Dinka	30.22	12.86	68	-	-	-	0.00	-	0.00	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Africa	Sudan	Dunglawi	30.00	20.00	12	0.00	0.00	0.00	0.00	0.08	0.08	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Africa	Sudan	Fulani	30.00	20.00	88	-	-	-	0.48	-	0.48	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Africa	Sudan	Gaali	32.53	15.59	20	0.00	0.00	0.00	0.00	0.05	0.05	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Africa	Sudan	Jaali	33.43	16.69	172	0.00	0.06	0.13	0.01	0.01	0.21	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Africa	Sudan	Koalib	30.22	12.86	2	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Liguri/Logorik	30.22	12.86	2	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Mahas	32.53	15.59	30	0.00	0.00	0.17	0.00	0.00	0.17	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Africa	Sudan	Masalit	30.22	12.86	2	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Nuer	30.22	12.86	10	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Sudan	Shaigi	30.00	20.00	18	0.00	0.17	0.06	0.00	0.00	0.22	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Africa	Sudan	Shilook	30.22	12.86	16	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Akie	34.89	-6.37	28	0.25	-	0.00	0.00	0.00	0.25	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Burunge	34.89	-6.37	36	0.38	-	0.00	0.00	0.00	0.38	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Chagga	38.05	-5.38	92	0.14	0.00	0.00	0.00	0.00	0.14	Jones (2012) PhD thesis, UCL, unpublished data.
Africa	Tanzania	Datog	34.89	-6.37	8	0.63	-	0.00	0.00	0.00	0.63	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Dorobo	34.89	-6.37	20	0.40	-	0.00	0.00	0.00	0.40	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Fiome	34.89	-6.37	24	0.55	-	0.00	0.00	0.00	0.55	Tishkoff et al. (2007) Nat Genet. 39, 31

Africa	Tanzania	Hadza	34.89	-6.37	36	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Iraqw	34.89	-6.37	78	0.58	-	0.00	0.00	0.00	0.58	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Maasai	34.89	-6.37	38	0.45	-	0.00	0.00	0.00	0.45	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Mbugu	34.89	-6.37	60	0.31	-	0.00	0.00	0.00	0.31	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Mbugwe	34.89	-6.37	26	0.27	-	0.04	0.00	0.00	0.31	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Pare	34.89	-6.37	20	0.10	-	0.00	0.00	0.00	0.10	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Rangi	34.89	-6.37	70	0.27	-	0.00	0.00	0.00	0.27	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Samba'a	34.89	-6.37	6	0.00	-	0.00	0.00	0.00	0.00	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Tanzania	Sandawe	34.89	-6.37	62	0.13	-	0.00	0.00	0.00	0.13	Tishkoff et al. (2007) Nat Genet. 39, 31
Africa	Uganda	Bantu	32.98	0.43	44	-	-	-	0.00	-	0.00	Mulcare et al. (2004) Am J Hum Genet. 74, 1102.
Asia	Afghanistan	Pashtuns/Afghans	68.36	33.56	32	0.00	0.00	0.00	0.06	0.00	0.06	Liebert (2013) this thesis
Asia	Afghanistan	Tadjiks	68.71	36.13	28	0.00	0.00	0.00	0.25	0.00	0.25	Liebert (2013) this thesis
Asia	Afghanistan	Tadjiks	68.71	36.13	80	-	-	-	0.08	-	0.08	Mulcare (2006) PhD thesis, UCL
Asia	Afghanistan	Uzbeks	67.64	35.50	54	0.00	0.00	0.00	0.24	0.00	0.24	Liebert (2013) this thesis
Asia	Afghanistan	Uzbeks	67.64	35.50	30	-	-	-	0.03	-	0.03	Mulcare (2006) PhD thesis, UCL
Asia	Armenia	Armenians	45.04	40.07	88	-	-	-	0.01	-	0.01	Mulcare (2006) PhD thesis, UCL
Asia	Armenia	Armenians	45.04	40.07	102	0.00	0.00	0.00	0.18	0.00	0.18	Liebert (2013) this thesis
Asia	Azerbaijan	Azeri	47.58	40.14	80	0.00	0.00	0.00	0.01	0.00	0.01	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Asia	Cambodia	Cambodian	104.92	11.55	22	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Dai	102.83	24.88	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Daur	87.62	43.83	20	-	-	-	0.05	-	0.05	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Han	104.21	34.86	90	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Han	104.21	34.86	200	-	-	-	0.00	-	0.00	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	China	Hezhen	132.50	47.70	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Hezhen	132.50	47.70	154	-	-	-	0.00	-	0.00	Sun et al. (2007) Asia Pac J Clin Nutr. 16, 598.
Asia	China	Hezhen	132.50	47.70	196	0.00	-	0.00	-	0.00	0.00	Xu et al. (2010) Scand J Gastroentero. 45, 168.
Asia	China	Kazak	81.30	43.90	188	-	-	-	0.05	-	0.05	Sun et al. (2007) Asia Pac J Clin Nutr. 16, 598.
Asia	China	Kazak	81.30	43.90	194	0.00	-	0.00	-	0.00	0.00	Xu et al. (2010) Scand J Gastroentero. 45, 168.
Asia	China	Lahu	102.83	24.88	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Man	123.20	40.20	130	-	-	-	0.00	-	0.00	Sun et al. (2007) Asia Pac J Clin Nutr. 16, 598.
Asia	China	Man (Manchu)	123.20	40.20	216	0.00	-	0.00	-	0.00	0.00	Xu et al. (2010) Scand J Gastroentero. 45, 168.
Asia	China	Miaozu	106.63	26.65	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Mongols	119.70	29.20	164	-	-	-	0.02	-	0.02	Sun et al. (2007) Asia Pac J Clin Nutr. 16, 598.
Asia	China	Mongols	119.70	29.20	212	0.00	-	0.00	-	0.00	0.00	Xu et al. (2010) Scand J Gastroentero. 45, 168.
Asia	China	Mongols	119.70	29.20	20	-	-	-	0.10	-	0.10	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Naxi	100.23	26.86	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Northern Han	126.53	45.80	138	0.00	-	0.00	-	0.00	0.00	Xu et al. (2010) Scand J Gastroentero. 45, 168.
Asia	China	Oroqen	131.00	53.50	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Oroqen	131.00	53.50	90	-	-	-	0.01	-	0.01	Sun et al. (2007) Asia Pac J Clin Nutr. 16, 598.
Asia	China	Oroqen	131.00	53.50	136	0.00	-	0.00	-	0.00	0.00	Xu et al. (2010) Scand J Gastroentero. 45, 168.
Asia	China	She	119.65	27.09	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Tibetans (Nagqu)	92.07	31.48	418	0.00	0.00	0.00	0.00	0.00	0.00	Peng et al. (2012) J Hum Genet. 57, 394.

Asia	China	Tibetans (Shigatse)	88.88	29.27	572	0.00	0.00	0.00	0.00	0.00	0.00	Peng et al. (2012) J Hum Genet. 57, 394.
Asia	China	Tu	101.78	36.62	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Tujia	112.94	28.23	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Uygur	87.62	43.83	20	-	-	-	0.05	-	0.05	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Xibo	123.43	41.81	18	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	China	Yizu	103.77	29.55	20	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Georgia	Georgians	43.35	42.32	108	0.00	0.00	0.00	0.08	0.00	0.08	Liebert (2013) this thesis
Asia	India	Andaman and Nicobar Islands	92.77	11.67	68	0.00	0.00	0.00	0.02	0.00	0.02	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	India	Central	79.45	23.84	358	0.00	0.00	0.00	0.06	0.00	0.06	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	India	Dawoodi Bohra (Gujarat)	71.19	22.26	100	0.00	0.00	0.00	0.11	0.00	0.11	Easwaarkanth et al. (2009) Eur J Hum Genet. 18, 354
Asia	India	Dawoodi Bohra (Tamil Nadu)	78.66	11.13	124	0.00	0.00	0.00	0.14	0.00	0.14	Easwaarkanth et al. (2009) Eur J Hum Genet. 18, 354
Asia	India	East	85.02	20.80	620	0.00	0.00	0.00	0.02	0.00	0.02	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	India	Indian Shia	80.10	27.57	142	0.00	0.00	0.00	0.10	0.00	0.10	Easwaarkanth et al. (2009) Eur J Hum Genet. 18, 354
Asia	India	Indian Sunni	80.10	27.57	164	0.00	0.00	0.00	0.10	0.00	0.10	Easwaarkanth et al. (2009) Eur J Hum Genet. 18, 354
Asia	India	Iranian Shia	80.10	17.05	98	0.00	0.00	0.00	0.04	0.00	0.04	Easwaarkanth et al. (2009) Eur J Hum Genet. 18, 354
Asia	India	Mappla	78.66	11.13	124	0.00	0.00	0.00	0.02	0.00	0.02	Easwaarkanth et al. (2009) Eur J Hum Genet. 18, 354
Asia	India	North	76.78	30.75	580	0.00	0.00	0.00	0.16	0.00	0.16	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	India	North East	95.00	27.48	278	0.00	0.00	0.00	0.01	0.00	0.01	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	India	North	77.20	28.60	120	0.00	0.00	0.00	0.20	0.00	0.20	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Asia	India	North (Lucknow)	80.95	26.85	154	0.00	0.00	0.00	0.19	0.00	0.19	Babu et al. (2010) Am J Clin Nutr. 91, 140
Asia	India	South	77.57	12.97	1728	0.00	0.00	0.01	0.09	0.00	0.09	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	India	South	80.28	13.08	102	0.00	0.00	0.00	0.12	0.00	0.12	Liebert (2013) this thesis
Asia	India	South-Indian (Bangalore)	77.59	12.97	152	0.00	0.00	0.00	0.07	0.00	0.07	Babu et al. (2010) Am J Clin Nutr. 91, 140
Asia	India	West	73.68	24.58	936	0.00	0.00	0.00	0.21	0.00	0.21	Galego Romero et al. (2012) Mol Biol Evol. 29, 249.
Asia	Mongolia	Khalka	99.05	47.38	114	0.00	0.00	0.00	0.03	0.00	0.03	Liebert (2013) this thesis
Asia	Mongolia	Mongols	106.92	47.92	52	0.00	0.00	0.00	0.04	0.00	0.04	Liebert (2013) this thesis
Asia	Nepal	Nepalese	84.12	28.39	38	0.00	0.00	0.00	0.34	0.00	0.34	Liebert (2013) this thesis
Asia	Nepal	Tharu	80.90	28.83	80	0.00	0.00	0.00	0.18	0.00	0.18	Liebert (2013) this thesis
Asia	Pakistan	Balochi	67.02	30.21	50	-	-	-	0.36	-	0.36	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Balti	74.59	35.95	46	-	-	-	0.00	-	0.00	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Baluch	65.09	28.49	38	-	-	-	0.34	-	0.34	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Brahui	67.02	30.21	50	-	-	-	0.34	-	0.34	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Brahui	67.02	30.21	60	-	-	-	0.27	-	0.27	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Burusho	74.59	35.95	50	-	-	-	0.10	-	0.10	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Burusho	74.59	35.95	60	-	-	-	0.02	-	0.02	Enattah et al. (2007) Am J Hum Genet. 81, 615.

Asia	Pakistan	Hazara	67.02	30.21	28	-	-	-	0.04	-	0.04	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Hazara	67.02	30.21	50	-	-	-	0.08	-	0.08	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Kalash	71.78	35.84	60	-	-	-	0.00	-	0.00	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Kalash	71.78	35.84	50	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Kashmiri	73.94	33.89	40	-	-	-	0.12	-	0.12	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Makrani	67.02	30.21	50	-	-	-	0.34	-	0.34	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Makrani Baluch	68.03	26.27	58	-	-	-	0.17	-	0.17	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Mohannes	68.00	26.00	58	-	-	-	0.28	-	0.28	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Parsi	68.00	26.00	58	-	-	-	0.14	-	0.14	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Pathan	69.86	32.67	56	-	-	-	0.30	-	0.30	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Pakistan	Pathan	69.86	32.67	50	-	-	-	0.30	-	0.30	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Sindhi	67.03	24.89	50	-	-	-	0.32	-	0.32	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Pakistan	Sindhi	67.03	24.89	56	-	-	-	0.41	-	0.41	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Russia	Daghestans mixed	47.12	42.83	46	-	-	-	0.13	-	0.13	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Russia	Druss	47.12	42.83	34	-	-	-	0.12	-	0.12	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Russia	Nenets	49.02	67.66	178	-	-	-	0.07	-	0.73	Khabarova et al. (2012) Int J Circumpol Heal. 71, 1.
Asia	Russia	Nog	47.12	42.83	40	-	-	-	0.07	-	0.07	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Russia	North-Yakuts	146.19	70.77	22	0.00	0.00	0.00	0.14	0.00	0.14	Liebert (2013) this thesis
Asia	Russia	Ob-Ugric	68.00	65.00	40	-	-	-	0.03	-	0.03	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Russia	Udmurts	53.18	56.83	60	-	-	-	0.33	-	0.33	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Russia	Yakuts	129.73	62.03	50	-	-	-	0.06	-	0.06	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Asia	Russia	Yakuts	129.73	62.03	110	0.00	0.00	0.00	0.05	0.00	0.05	Liebert (2013) this thesis
Asia	Singapore	Han-Chinese	104.21	35.86	98	0.00	0.00	0.00	0.04	0.00	0.04	Liebert (2013) this thesis
Asia	Singapore	Japanese	138.24	36.21	84	0.00	0.00	0.00	0.00	0.00	0.00	Liebert (2013) this thesis
Asia	South Korea	South Korean	127.00	35.57	46	-	-	-	0.00	-	0.00	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Asia	Syria, Iraq, Lebanon, West Bank	Arabs	40.00	33.00	40	0.00	0.00	0.11	0.13	0.00	0.24	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Asia	Uzbekistan	Kazakh	63.46	40.13	166	-	-	-	0.16	-	0.16	Heyer et al. (2011) Hum Biol. 83, 379.
Asia	Uzbekistan	Tajiko-Uzbek	64.42	39.77	200	-	-	-	0.10	-	0.10	Heyer et al. (2011) Hum Biol. 83, 379.
Asia	Uzbekistan	Uzbeks	64.59	41.38	76	0.00	0.00	0.00	0.04	0.00	0.04	Liebert (2013) this thesis
Austalasia	Papua New Guinea	Papuan	140.71	-5.09	34	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Australasia	Solomon Islands	Melanesian (NAN)	155.19	-6.05	44	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Belarus	Belarusians	27.95	53.71	100	0.00	0.00	0.00	0.23	0.00	0.23	Liebert (2013) this thesis
Europe	England (Central)	British	-2.76	52.71	1368	-	-	-	0.76	-	0.76	Davey Smith et al. (2009) Eur J Hum Genet. 17, 357.
Europe	England (Northern)	British	-1.54	53.99	2336	-	-	-	0.75	-	0.75	Davey Smith et al. (2009) Eur J Hum Genet. 17, 357.
Europe	England (Southeastern)	British	-0.57	51.22	1894	-	-	-	0.70	-	0.70	Davey Smith et al. (2009) Eur J Hum Genet. 17, 357.
Europe	Estonia	Estonians	26.40	59.05	628	0.00	0.00	0.00	0.51	0.00	0.51	Lember (2006) World J Gastroentero. 12, 7329.
Europe	Finland	Finns	28.00	65.00	1876	0.00	0.00	0.00	0.58	0.00	0.58	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Europe	Finland	Saami	26.19	68.26	60	-	-	-	0.17	-	0.17	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Finland and Sweden	Scandinavians	18.05	59.33	360	-	-	-	0.82	-	0.82	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Finns	Finns, eastern	29.00	65.00	154	-	-	-	0.55	-	0.55	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Finns	Finns, western	24.00	64.00	308	-	-	-	0.62	-	0.62	Enattah et al. (2007) Am J Hum Genet. 81, 615.

Europe	France	Basques	-1.45	43.40	48	-	-	-	0.67	-	0.67	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	France	Basques	-1.45	43.40	170	-	-	-	0.66	-	0.66	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	France	French	2.22	46.23	58	-	-	-	0.43	-	0.43	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	France	French	2.22	46.23	34	-	-	-	0.34	-	0.34	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Germany	Germans	10.45	51.17	60	0.00	0.00	0.00	0.55	0.00	0.55	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Europe	Germany	Sorbs	14.31	51.26	64	0.00	0.00	0.00	0.50	0.00	0.50	Liebert (2013) this thesis
Europe	Greece	Greeks	21.83	39.07	200	-	-	-	0.09	-	0.09	Anagnostou et al. (2009) Am J Hum Biol. 21, 217.
Europe	Greece	Greeks	21.83	39.07	120	0.00	0.00	0.00	0.15	0.00	0.15	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Europe	Hungary	Hungarian	20.17	46.25	220	-	-	-	0.62	-	0.62	Nagy et al. (2009) Eur J Clin Nutr. 63, 909.
Europe	Hungary	Hungarian	20.17	46.25	362	-	-	-	0.36	-	0.36	Nagy et al. (2011) Am J Phys Anthr. 145, 262.
Europe	Ireland	Irish	-8.00	53.33	68	0.00	0.00	0.00	0.96	0.00	0.96	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Europe	Italy	Central	12.48	41.90	196	-	-	-	0.11	-	0.11	Anagnostou et al. (2009) Am J Hum Biol. 21, 217.
Europe	Italy	Central-North	11.25	43.77	412	-	-	-	0.13	-	0.13	Anagnostou et al. (2009) Am J Hum Biol. 21, 217.
Europe	Italy	Italians	13.91	42.22	134	-	-	-	0.13	-	0.13	Coelho et al. (2005) Hum Genet. 117, 329
Europe	Italy	North Italian	9.68	45.70	28	-	-	-	0.36	-	0.36	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Italy	North-Eastern	12.33	45.44	438	-	-	-	0.24	-	0.24	Anagnostou et al. (2009) Am J Hum Biol. 21, 217.
Europe	Italy	S. European	12.48	41.90	66	0.00	0.00	0.00	0.09	0.00	0.09	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Europe	Italy	Sardinian	9.12	39.22	306	-	-	-	0.07	-	0.07	Anagnostou et al. (2009) Am J Hum Biol. 21, 217.
Europe	Italy	Sardinian	9.12	39.22	56	-	-	-	0.07	-	0.07	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Italy	South Italians	16.25	39.30	200	-	-	-	0.05	-	0.05	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Italy	Southern	16.25	39.30	378	-	-	-	0.08	-	0.08	Anagnostou et al. (2009) Am J Hum Biol. 21, 217.
Europe	Italy	Tuscan	10.98	43.57	16	-	-	-	0.06	-	0.06	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Italy	Tyroleans Bozen	11.35	46.50	80	0.00	0.00	0.00	0.15	0.00	0.15	Liebert (2013) this thesis
Europe	Italy	Tyroleans Gadertal	11.92	46.72	76	0.00	0.00	0.00	0.42	0.00	0.42	Liebert (2013) this thesis
Europe	Italy	Tyroleans Vinschgau	10.77	46.63	102	0.00	0.00	0.00	0.36	0.00	0.36	Liebert (2013) this thesis
Europe	Macedonia	Macedonians	21.75	41.61	100	0.00	0.00	0.00	0.24	0.00	0.24	Liebert (2013) this thesis
Europe	Mixed	N. European	15.00	54.00	110	0.00	0.00	0.00	0.62	0.00	0.62	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Europe	Netherlands	Frisians	5.80	53.20	58	0.00	0.00	0.00	0.81	0.00	0.81	Liebert (2013) this thesis
Europe	Norway	Norwegians	8.47	60.47	88	0.00	0.00	0.00	0.84	0.00	0.84	Liebert (2013) this thesis
Europe	Poland	Ashkenazi-Jews	21.00	52.25	96	-	-	-	0.08	-	0.08	Mulcare (2006) PhD thesis, UCL
Europe	Poland	Polish	19.15	51.92	400	-	-	-	0.46	-	0.46	Madry et al. (2010) Acta Biochim Pol. 57, 585
Europe	Poland	Polish	19.15	51.92	446	-	-	0.00	0.30	0.00	0.30	Płoszaj et al. (2011) Cent Eur J Biol. 6, 176.
Europe	Portugal	Portuguese	-8.22	39.40	96	0.00	0.00	0.00	0.44	0.00	0.44	Liebert (2013) this thesis
Europe	Portugal	Portuguese-Centre	-7.50	40.28	140	-	-	-	0.39	-	0.39	Manco et al. (2012) Ann Hum Biol. 40, 205.
Europe	Portugal	Portuguese-North	-8.42	41.55	128	-	-	-	0.38	-	0.38	Manco et al. (2012) Ann Hum Biol. 40, 205.
Europe	Portugal	Portuguese-North	-8.42	41.55	180	-	-	-	0.37	-	0.37	Coelho et al. (2005) Hum Genet. 117, 329
Europe	Portugal	Portuguese-South	-7.83	38.52	130	-	-	-	0.27	-	0.27	Manco et al. (2012) Ann Hum Biol. 40, 205.
Europe	Romania	Romanians	24.97	45.94	118	0.00	0.00	0.00	0.17	0.00	0.17	Liebert (2013) this thesis

Europe	Romania (Eastern Trasylvania)	Seklers	25.79	45.86	130	-	-	-	0.40	-	0.40	Nagy et al. (2011) Am J Phys Anthr. 145, 262.
Europe	Russia	Erzya	43.58	54.21	60	-	-	-	0.27	-	0.27	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Russia	Erzya	43.58	54.21	42	0.00	0.00	0.00	0.14	0.00	0.14	Liebert (2013) this thesis
Europe	Russia	Moksha	44.07	54.24	32	0.00	0.00	0.00	0.41	0.00	0.41	Liebert (2013) this thesis
Europe	Russia	Moksha	44.07	54.24	60	-	-	-	0.28	-	0.28	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Russia	Russian	37.62	55.75	50	-	-	-	0.24	-	0.24	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Russia	Russians (Northwest)	40.53	64.53	298	0.00	0.00	0.00	0.39	0.00	0.39	Khabarova et al. (2009) World J Gastroentero. 15, 1849.
Europe	Russia	Russians Perm	56.25	58.01	46	0.00	0.00	0.00	0.24	0.00	0.24	Liebert (2013) this thesis
Europe	Russia (Komi republic)	Komi	54.83	63.86	20	-	-	-	0.15	-	0.15	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Europe	Russian (Caucasus)	Adygei	40.08	44.60	34	-	-	-	0.12	-	0.12	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Scotland	British	-3.78	56.00	1032	-	-	-	0.82	-	0.82	Davey Smith et al. (2009) Eur J Hum Genet. 17, 357.
Europe	Slovakia	Roma	19.70	48.67	64	0.00	0.00	0.00	0.06	0.00	0.06	Liebert (2013) this thesis
Europe	Slovakia	Roma	19.70	48.67	108	-	-	-	0.11	-	0.11	Mulcare (2006) PhD thesis, UCL
Europe	Spain	Catalans	1.52	41.59	58	0.00	0.00	0.00	0.26	0.00	0.26	Liebert (2013) this thesis
Europe	Spain	Spanish	-3.70	40.51	1718	-	-	-	0.39	-	0.39	Agueda et al. (2010) Caclif Tissue Int. 87, 14.
Europe	Spain	Spanish	-3.70	40.51	62	0.00	0.00	0.00	0.40	0.00	0.40	Liebert (2013) this thesis
Europe	Sweden	Sami	20.87	67.87	60	0.00	0.00	0.00	0.29	0.00	0.29	Liebert (2013) this thesis
Europe	Sweden	Swedes	15.01	60.03	784	-	-	-	0.74	-	0.74	Almon et al. (2007) Scand J Gastroenterol. 42, 165.
Europe	Sweden	Swedes	15.01	60.03	74	0.00	0.00	0.00	0.78	0.00	0.78	Liebert (2013) this thesis
Europe	UK	Ashkenazi-Jews	-0.30	51.51	38	0.00	0.00	0.00	0.11	0.00	0.11	Liebert (2013) this thesis
Europe	UK	English	-0.12	51.51	102	0.00	0.00	0.00	0.75	0.00	0.75	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Europe	UK	Orcadian (Orkney Islands)	-3.15	59.04	32	-	-	-	0.69	-	0.69	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Europe	Ukraine	Ukrainians	31.17	48.38	92	-	-	-	0.22	-	0.22	Mulcare (2006) PhD thesis, UCL
Europe	Ukraine	Ukrainians	31.17	48.38	74	0.00	0.00	0.00	0.26	0.00	0.26	Liebert (2013) this thesis
Middle East	Cyprus	Greek-Cypriots	33.02	34.71	120	0.00	0.00	0.00	0.01	0.00	0.01	Liebert (2013) this thesis
Middle East	Iran	Iranians	53.69	32.43	42	0.00	0.00	0.00	0.10	0.00	0.10	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Middle East	Iran	Iranians	53.69	32.43	154	0.00	0.00	0.01	0.03	0.00	0.04	Liebert (2013) this thesis, Mulcare (2006) PhD thesis, UCL
Middle East	Iran	Iranians	53.69	32.43	42	-	-	-	0.10	-	0.10	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Middle East	Iran	Qashqai	52.53	29.62	20	-	-	-	0.05	-	0.05	Enattah et al. (2007) Am J Hum Genet. 81, 615.
Middle East	Israel	Ashkenazi-Jews	34.77	32.07	192	-	-	-	0.09	-	0.09	Raz et al. (2013) Gene. 519, 67.
Middle East	Israel	Bedouin	34.88	30.71	98	-	-	-	0.03	-	0.03	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Middle East	Israel	Bedouin Israeli	34.77	32.07	38	0.00	0.00	0.13	0.03	0.00	0.16	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Middle East	Israel	Bedouin Arabs	34.77	32.07	302	-	-	0.28	0.02	-	0.02	Raz et al. (2013) Gene. 519, 67.
Middle East	Israel	Druze	34.77	32.07	96	-	-	-	0.02	-	0.02	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Middle East	Israel	Druze	34.77	32.07	28	0.00	0.00	0.11	0.04	0.00	0.14	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Middle East	Israel	Iraqi Jews	34.77	32.07	192	-	-	-	0.04	-	0.04	Raz et al. (2013) Gene. 519, 67.

Middle East	Israel	Moroccan	34.77	32.07	192	-	-	-	0.09	-	0.09	Raz et al. (2013) Gene. 519, 67.
Middle East	Israel	Palestinian Arabs	35.20	31.90	102	-	-	-	0.04	-	0.04	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Middle East	Israel	Urban Arabs	34.77	32.07	84.00	0.00	0.01	0.01	0.02	0.00	0.05	Liebert (2013) this thesis
Middle East	Israel	Urban Arabs	34.77	32.07	162	0.00	0.01	0.06	0.00	0.00	0.06	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Middle East	Israel/PAA	Palestinian Arabs	35.20	31.90	36	0.00	0.00	0.00	0.00	0.00	0.00	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Middle East	Japan	Japanese	138.24	37.21	62	-	-	-	0.00	-	0.00	Bersaglieri et al. (2004) Am J Hum Genet. 74, 1111.
Middle East	Jordan	Jordanian	35.93	31.95	112	0.00	0.00	0.05	0.05	0.00	0.11	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Middle East	Jordan	Jordanian	36.24	30.59	46	0.00	0.00	0.35	0.00	0.00	0.35	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Middle East	Kurwait	Ajman	48.00	28.83	74	0.00	0.00	0.06	0.00	0.00	0.06	Hill et al. (2013) Am J Phys Anthr. (Epub ahead of print)
Middle East	Kuwait	Kuwaiti	47.48	29.31	66	0.00	0.02	0.25	0.03	0.00	0.30	Liebert (2013) this thesis
Middle East	Kuwait	Kuwaiti	47.48	29.31	28	-	-	-	0.00	-	0.00	Mulcare (2006) PhD thesis, UCL
Middle East	Kuwait	Mutran	47.94	29.24	58	0.00	0.00	0.55	0.00	0.03	0.58	Hill et al. (2013) Am J Phys Anthr. (Epub ahead of print)
Middle East	Oman	Arabs of Northern Oman	57.01	23.88	684	0.00	0.00	0.14	0.01	0.00	0.15	Al-Abri et al. (2012) Hum Biol. 84, 271.
Middle East	Oman	Dhofaris Arabs of Southern Oman	54.43	18.40	420	0.00	0.00	0.72	0.00	0.00	0.72	Al-Abri et al. (2012) Hum Biol. 84, 271.
Middle East	Oman	Omani	58.50	23.53	104	0.00	0.00	0.77	0.03	0.00	0.80	Al-Abri et al. (2013) Oman Med J. 28, 341.
Middle East	Saudi Arabia	Bedouin	45.00	23.00	94	0.00	0.00	0.48	0.00	0.00	0.48	Ingram et al. (2007) Hum Genet. 120, 779, Ingram (2008) PhD thesis UCL
Middle East	Saudi Arabia	Central	46.72	24.71	180	0.00	0.00	0.61	0.00	0.00	0.61	Imtiaz et al. (2007) J Med Genet. 44, e89.
Middle East	Saudi Arabia	Eastern	50.68	22.30	164	0.00	0.00	0.62	0.00	0.00	0.62	Imtiaz et al. (2007) J Med Genet. 44, e89.
Middle East	Saudi Arabia	Northern	39.32	29.89	164	0.00	0.00	0.52	0.01	0.00	0.53	Imtiaz et al. (2007) J Med Genet. 44, e89.
Middle East	Saudi Arabia	Southern	44.13	17.49	184	0.00	0.00	0.58	0.00	0.00	0.58	Imtiaz et al. (2007) J Med Genet. 44, e89.
Middle East	Saudi Arabia	Western	39.82	21.42	172	0.00	0.00	0.65	0.01	0.00	0.65	Imtiaz et al. (2007) J Med Genet. 44, e89.
Middle East	Saudi Arabia	Arabs	45.00	23.00	248	0.00	0.00	0.57	0.00	0.01	0.58	Enattah et al. (2008) Am J Hum Genet. 82, 57.
Middle East	Syria	Assyrians	36.30	33.50	80	-	-	-	0.04	-	0.04	Mulcare (2006) PhD thesis, UCL
Middle East	Syria	Syrians	39.00	34.80	140	0.00	0.00	0.03	0.02	0.00	0.05	Liebert (2013) this thesis
Middle East	Turkey	Anatolian-Turks	31.33	38.99	98	-	-	-	0.03	-	0.03	Mulcare (2006) PhD thesis, UCL
Middle East	Turkey	Anatolian-Turks	31.33	38.99	116	0.00	0.00	0.00	0.08	0.00	0.08	Liebert (2013) this thesis
Middle East	Yemen	Yemeni_Hadramaut	49.37	16.93	166	0.01	0.00	0.24	0.02	0.01	0.28	Liebert (2013) this thesis
Middle East	Yemen	Yemeni_Sena	44.21	15.35	68	0.00	0.00	0.29	0.15	0.00	0.44	Liebert (2013) this thesis
Middle East	Yemen	Yemeni (Sana'a University)	44.21	15.35	478	0.00	0.00	0.55	0.00	0.00	0.55	Al-Abri et al. (2012) Hum Biol. 84, 271.

Appendix B2: Lactase persistence phenotype data. Test methods: BH: breath hydrogen, BG: blood glucose, UG: urine galactose.

Continent/ Region	Country	Population/ Region	Longitude	Latitude	N	Frequency of digesters	Test method	Reference
Africa	Botswana	Shua	28.63	-23.04	22.00	0.09	BG	Nurse & Jenkins (1974) Br Med J. 2, 728.
Africa	Egypt	Cairo and Giza	31.25	30.05	67.00	0.33	BH	Hussein et al. (1982) Hum Hered. 32, 94.
Africa	Egypt	Nile Delta	32.00	31.50	291.00	0.27	BH	Hussein et al. (1982) Hum Hered. 32, 94.
Africa	Egypt	Suez Canal Zone	32.50	31.00	16.00	0.31	BH	Hussein et al. (1982) Hum Hered. 32, 94.
Africa	Egypt	Upper Egypt, North	28.00	31.00	111.00	0.15	BH	Hussein et al. (1982) Hum Hered. 32, 94.
Africa	Egypt	Upper Egypt, South	28.00	28.00	85.00	0.40	BH	Hussein et al. (1982) Hum Hered. 32, 94.
Africa	Ethiopia	Amhara (phenotyped)	38.66	9.87	94.00	0.34	BH	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Ethiopians	42.13	9.31	294.00	0.45	BH	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Oromo (phenotyped)	37.31	7.84	132.00	0.41	BH	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Somali	41.86	9.59	90.00	0.24	BH	Ingram (2009) J Mol Evol. 69, 579.
Africa	Ethiopia	Somali	41.87	9.58	130.00	0.25	BH	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Tigray	39.47	13.50	78.00	0.62	BH	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Ethiopia	Wolayta	37.76	6.84	40.00	0.40	BH	Jones et al. (2013) Am J Hum Genet. 93, 538.
Africa	Gabon	Bantu	9.45	0.38	20.00	0.40	BH	Gendrel et al. (1989) J Pediatr Gastroenterol Nutr. 8, 545.
Africa	Kenya	Borana	37.91	-0.02	7.00	0.71	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Burji	37.91	-0.02	6.00	0.50	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	El Molo	37.91	-0.02	6.00	0.67	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Gabra	37.91	-0.02	8.00	1.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Kikuyu	37.91	-0.02	2.00	0.50	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Konso	37.91	-0.02	4.00	0.50	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Maasai	37.91	-0.02	26.00	0.88	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Marakwet	37.91	-0.02	5.00	0.60	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Nandi	37.91	-0.02	2.00	0.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Ogiek	37.91	-0.02	11.00	0.55	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Pokot	37.91	-0.02	10.00	0.60	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Rendille	37.91	-0.02	7.00	0.71	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Sabaot	37.91	-0.02	4.00	0.75	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Samburu	37.91	-0.02	9.00	0.89	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Sengwer	37.91	-0.02	12.00	0.17	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Somali	37.91	-0.02	1.00	1.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Tugen	37.91	-0.02	11.00	0.73	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Turkana	37.91	-0.02	8.00	0.50	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Wata	37.91	-0.02	1.00	0.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Kenya	Yaaku	37.91	-0.02	11.00	0.73	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.

Africa	Namibia	!Kung	20.50	-19.60	40.00	0.03	BG	Jenkins et al. (1974) Br Med J. 2, 23.
Africa	Namibia	Herero	13.70	-18.30	39.00	0.03	BG	Currie et al. (1978) S Afr J Sci. 74, 227.
Africa	Niger	Tuareg	2.12	13.52	118.00	0.87	BH	Flatz et al. (1986) Am J Hum Genet. 38, 515.
Africa	Nigeria	Hausa/Fulani	3.47	7.23	15.00	0.40	BG	Olatunbosun et al. (1971) Am J Dig Dis. 16, 909.
Africa	Nigeria	Ibo	3.47	7.23	11.00	0.18	BG	Olatunbosun et al. (1971) Am J Dig Dis. 16, 909.
Africa	Nigeria	Yoruba	3.47	7.23	48.00	0.17	BG	Olatunbosun et al. (1971) Am J Dig Dis. 16, 909.
Africa	Rwanda	Hutu-Hutu	29.74	-2.60	36.00	0.58	UG	Cox et al. (1974) Am J Dig Dis. 19, 714.
Africa	Rwanda	Hutu-Tutsi	29.74	-2.60	11.00	0.45	UG	Cox et al. (1974) Am J Dig Dis. 19, 714.
Africa	Rwanda	Shi	29.74	-2.60	28.00	0.04	UG	Cox et al. (1974) Am J Dig Dis. 19, 714.
Africa	Rwanda	Tussi-Tutsi	29.74	-2.60	27.00	0.93	UG	Cox et al. (1974) Am J Dig Dis. 19, 714.
Africa	Senegal	Diolas	-16.25	12.60	40.00	0.73	BG	Arnold et al. (1980) C R Seances Soc Biol Fil. 174, 983.
Africa	Senegal	Peuhls	-15.12	15.40	29.00	1.00	BG	Arnold et al. (1980) C R Seances Soc Biol Fil. 174, 983.
Africa	Senegal	Sereres	-17.43	14.67	38.00	0.71	BG	Arnold et al. (1980) C R Seances Soc Biol Fil. 174, 983.
Africa	Senegal	Toucouleurs	-17.43	14.67	40.00	0.90	BG	Arnold et al. (1980) C R Seances Soc Biol Fil. 174, 983.
Africa	Senegal	Wolof	-17.43	14.67	53.00	0.51	BG	Arnold et al. (1980) C R Seances Soc Biol Fil. 174, 983.
Africa	South Africa	Shangaan	28.08	-26.20	7.00	0.14	BH	Segal et al. (1983) Am J Clin Nutr. 38, 901.
Africa	South Africa	Sotho	28.08	-26.20	23.00	0.35	BH	Segal et al. (1983) Am J Clin Nutr. 38, 901.
Africa	South Africa	Swazi	28.08	-26.20	12.00	0.25	BH	Segal et al. (1983) Am J Clin Nutr. 38, 901.
Africa	South Africa	Tswana	28.08	-26.20	24.00	0.17	BH	Segal et al. (1983) Am J Clin Nutr. 38, 901.
Africa	South Africa	Xhosa	28.08	-26.20	17.00	0.18	BH	Segal et al. (1983) Am J Clin Nutr. 38, 901.
Africa	South Africa	Zulu	31.02	-29.85	47.00	0.11	BG	O'Keefe & Adam (1983) S Afr Med J. 63, 778.
Africa	South Africa	Zulu	28.08	-26.20	32.00	0.19	BH	Segal et al. (1983) Am J Clin Nutr. 38, 901.
Africa	Sudan	Ama	30.22	12.86	2.00	0.50	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Amarar	37.22	19.62	82.00	0.87	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Artega	36.22	19.62	22.00	0.82	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Bedja	30.95	18.05	9.00	0.89	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Beja Banuamir	30.22	12.86	6.00	1.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Beja Hadandawa	30.22	12.86	11.00	0.82	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Beni Amir	37.22	19.62	40.00	0.88	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Bisharin	37.22	19.62	22.00	0.86	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Dinka	30.22	12.86	7.00	0.86	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Dinka	33.63	7.67	208.00	0.25	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Dongolawi	30.95	18.05	16.00	0.19	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Gomoeia	30.95	18.05	31.00	0.68	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Habbani	30.35	13.08	19.00	0.47	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Haddendoa	37.22	19.62	137.00	0.80	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Jaali	32.53	15.59	113.00	0.53	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Jaali	33.43	16.69	94.00	0.48	BH	Ingram et al. (2007) Hum Genet. 120, 779.
Africa	Sudan	Kahli	30.95	18.05	21.00	0.62	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Koalib	30.22	12.86	1.00	1.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Liguri/Logorik	30.22	12.86	1.00	0.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Masalit	30.22	12.86	1.00	1.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Misseri	30.35	13.08	20.00	0.40	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.

Africa	Sudan	Nilotic	27.67	7.77	18.00	0.33	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Nuba	29.68	16.80	58.00	0.21	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Nubians	30.95	18.05	21.00	0.33	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Nuer	30.22	12.86	2.00	1.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Sudan	Nuer	33.63	7.67	23.00	0.22	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Shaygi	30.95	18.05	42.00	0.38	BH	Bayoumi et al. (1981) Hum Genet. 57, 279.
Africa	Sudan	Shilluk	33.63	7.67	8.00	0.38	BH	Bayoumi et al. (1982) Am J Phys Anthropol. 58, 173.
Africa	Sudan	Shilook	30.22	12.86	4.00	0.75	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Akie	34.89	-6.37	11.00	0.55	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Burunge	34.89	-6.37	16.00	0.38	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Datog	34.89	-6.37	1.00	0.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Dorobo	34.89	-6.37	6.00	0.67	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Fiome	34.89	-6.37	7.00	0.14	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Hadza	34.89	-6.37	15.00	0.60	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Iraqw	34.89	-6.37	19.00	0.95	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Maasai	34.89	-6.37	15.00	0.67	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Mbugu	34.89	-6.37	23.00	0.43	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Mbugwe	34.89	-6.37	8.00	0.50	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Pare	34.89	-6.37	8.00	0.75	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Rangi	34.89	-6.37	26.00	0.65	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Samba'a	34.89	-6.37	2.00	0.00	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tanzania	Sandawe	34.89	-6.37	23.00	0.35	BG	Tishkoff et al. (2007) Nat Genet. 39, 31.
Africa	Tunisia	Tunisian	10.18	36.80	43.00	0.16	BH	Filali et al. (1987) Gastroenterol Clin Biol. 11, 554.
Africa	Uganda	Baganda	32.57	0.32	12.00	0.00	BG	Cook & Dahlquist (1968) Gastroenterol. 55, 328.
Africa	Uganda	Batutsi	32.57	0.32	5.00	1.00	BG	Cook & Dahlquist (1968) Gastroenterol. 55, 328.
Africa	Uganda	Nilotic	32.57	0.32	9.00	0.56	BG	Cook et al. (1966) Lancet. 1, 725.
Africa	Uganda	Ugandan Bantu	32.57	0.32	17.00	0.06	BG	Cook et al. (1966) Lancet. 1, 725.
Africa	Zambia	Bantu of Zambia	28.11	-15.28	26.00	0.00	BG	Cook et al. (1973) Gastroenterol. 64, 405.
Asia	Afghanistan	Hazara	69.18	34.52	10.00	0.20	BG	Rahimi et al. (1976) HumGenet. 34, 57.
Asia	Afghanistan	Mixed urban	69.18	34.52	34.00	0.24	BG	Rahimi et al. (1976) HumGenet. 34, 57.
Asia	Afghanistan	Pasha-I	71.00	36.00	60.00	0.13	BG	Rahimi et al. (1976) HumGenet. 34, 57.
Asia	Afghanistan	Pashtun	69.18	34.52	71.00	0.21	BG	Rahimi et al. (1976) HumGenet. 34, 57.
Asia	Afghanistan	Tajik	69.18	34.52	79.00	0.18	BG	Rahimi et al. (1976) HumGenet. 34, 57.
Asia	Afghanistan	Uzbek	69.18	34.52	16.00	0.00	BG	Rahimi et al. (1976) HumGenet. 34, 57.
Asia	China	Kazakh	87.58	43.80	195.00	0.24	BH	Yongfa et al. (1984) Hum Genet. 67, 103.
Asia	China	Mongols	111.65	40.81	198.00	0.12	BH	Yongfa et al. (1984) Hum Genet. 67, 103.
Asia	China	Northern Han	116.39	39.93	248.00	0.08	BH	Yongfa et al. (1984) Hum Genet. 67, 103.
Asia	India	Indians	72.83	18.98	100.00	0.36	BG	Desai et al. (1970) Indian J Med Sci. 24, 729.
Asia	India	Indians	78.47	17.38	18.00	0.39	BG	Reddy & Pershad (1972) Am J Clin Nutr. 25, 114.
Asia	India	Indians	80.28	13.80	38.00	0.00	Biopsy	Swaminathan et al. (1970) Clin Chim Acta. 30, 707.
Asia	India	Northern Indians	77.20	28.60	70.00	0.73	BG	Gupta et al. (1971) J Trop Med Hyg. 74, 225.
Asia	Japan	Japanese	140.47	40.59	40.00	0.28	BG	Yoshida et al. (1975) Gastroenterol Jpn. 10, 29.
Asia	Myanmar	Burmese	95.96	21.91	50.00	0.08	BG	Aung-Than-Batu et al. (1972) Union Burma J Life Sci. 5, 133.

Asia	Pakistan	Baloochi	67.05	24.87	4.00	1.00	BG	Rab et al. (1976) Br Med J. 1, 436.
Asia	Pakistan	Baluchistani	71.92	32.27	32.00	0.38	BH	Ahmad & Flatz (1984) Hum Hered. 34, 69.
Asia	Pakistan	Kashmiri	71.92	32.27	27.00	0.30	BH	Ahmad & Flatz (1984) Hum Hered. 34, 69.
Asia	Pakistan	Mohajir	67.05	24.87	15.00	0.80	BG	Rab et al. (1976) Br Med J. 1, 436.
Asia	Pakistan	Pathan	67.05	24.87	15.00	1.00	BG	Rab et al. (1976) Br Med J. 1, 436.
Asia	Pakistan	Punjabi	71.92	32.27	322.00	0.41	BH	Ahmad & Flatz (1984) Hum Hered. 34, 69.
Asia	Pakistan	Punjabi	67.05	24.87	9.00	1.00	BG	Rab et al. (1976) Br Med J. 1, 436.
Asia	Pakistan	Punjabi	73.07	33.60	53.00	0.55	BG	Abbas & Ahmad (1983) Hum Genet. 64, 277.
Asia	Pakistan	Sindhi	71.92	32.27	33.00	0.42	BH	Ahmad & Flatz (1984) Hum Hered. 34, 69.
Asia	Pakistan	Sindhi	67.05	24.87	12.00	1.00	BG	Rab et al. (1976) Br Med J. 1, 436.
Asia	Russia	Khanty (Northern)	69.02	61.04	115.00	0.29	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Asia	Russia	Komi-Izhems	50.81	61.67	56.00	0.38	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Asia	Russia	Mansi	66.95	56.96	81.00	0.28	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Asia	Russia	Nenets (West Siberia)	80.86	59.06	9.00	0.22	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Asia	Russia	Udmurtians	53.23	56.85	30.00	0.60	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Asia	Russia	West-Siberian	80.86	59.06	47.00	0.51	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Asia	Sri Lanka	Sri Lankan	80.64	7.30	135.00	0.29	BG	Thomas et al. (1990) J Trop Pediatr. 36, 80.
Asia	Sri Lanka	Sri Lankans ("Ceylonese")	80.60	7.26	200.00	0.28	BG	Senewiratne et al. (1977) Gastroenterol. 72, 1257.
Asia	Taiwan	Chinese	121.45	25.02	50.00	0.12	BG	Sung et al. (1972) Asian J Med. 8, 149.
Asia	Thailand	Thai	100.52	13.75	140.00	0.03	BG	Keusch et al. (1969) Am J Clin Nutr. 22, 638.
Asia	Thailand	Thai	100.49	13.45	40.00	0.00	BG	Troncale et al. (1967) Br Med J. 4, 578.
Australasia	Australia	Aboriginal	123.97	-17.30	45.00	0.16	BH	Brand et al. (1983) Am J Clin Nutr. 37, 449.
Australasia	New Zealand	Maori	174.77	-36.87	28.00	0.36	BH	Abbott & Tasman-Jones (1985) N Z Med J. 10, 228.
Australasia	Papua New Guinea	Central (inc. Port Moresby)	147.19	-9.46	14.00	0.07	BG	Cook (1979) Ann Hum Biol. 6, 55.
Australasia	Papua New Guinea	E and W Sepik provinces	143.52	-4.18	35.00	0.23	BH	Arnhold et al. (1981) Ann Hum Biol. 5, 481
Australasia	Papua New Guinea	Gulf & Western	147.19	-9.46	13.00	0.08	BG	Cook (1979) Ann Hum Biol. 6, 55.
Australasia	Papua New Guinea	Highlands	147.19	-9.46	13.00	0.00	BG	Cook (1979) Ann Hum Biol. 6, 55.
Australasia	Papua New Guinea	Huli, Mendi, and Dunai	142.95	-5.70	30.00	0.10	BG	Jenkins et al. (1981) Ann Hum Biol. 8, 447.
Australasia	Papua New Guinea	Milne Bay	150.60	-10.51	2.00	0.00	BG	Cook (1979) Ann Hum Biol. 6, 55.
Australasia	Papua New Guinea	Morobe & Northern	147.19	-9.46	5.00	0.00	BG	Cook (1979) Ann Hum Biol. 6, 55.
Australasia	Papua New Guinea	N. Solomons & E. New Britain	147.19	-9.46	3.00	0.00	BG	Cook (1979) Ann Hum Biol. 6, 55.
Europe	Austria	Austrian	14.00	47.75	118.00	0.75	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Austria	Austrian	14.00	47.75	57.00	0.79	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Austria	Austrian	14.00	47.75	88.00	0.80	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Austria	Austrian	14.00	47.75	32.00	0.81	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Austria	Karnten Austrian	14.31	46.62	46.00	0.80	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Austria	Oberosterreich Austrian	14.30	48.30	45.00	0.84	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Austria	Tirol Austrian	9.77	47.50	124.00	0.83	BH	Rosenkranz et al. (1982) Hum Genet. 62, 158.
Europe	Cyprus	Greek Cypriots	33.37	35.17	50.00	0.34	BG	Kanaghinis et al. (1974) Am J Dig Dis. 19, 1021.
Europe	Czech Republic	Czech	15.50	49.00	17.00	0.82	BG	Leichter (1972) Am J Dig Dis. 17, 73.
Europe	Denmark	Danes	12.58	55.73	91.00	0.96	BG	Busk et al. (1975) Ugeskr Laeger. 137, 2062.
Europe	Estonia	Estonian	26.40	59.05	112.00	0.75	BG	Lember et al. (1991) Eur J Gastroenterol Hepatol. 3, 479.
Europe	Estonia	Setus	27.64	57.96	100.00	0.51	UG	Lember et al. (1991) Eur J Gastroenterol Hepatol. 3, 479.

Europe	Finland	Finnish-speaking Finns	21.43	60.60	91.00	0.92	BG	Sahi (1974) Scand J Gastroenterol. 9, 303.
Europe	Finland	Finns	27.68	62.90	638.00	0.83	Biopsy	Jussila (1969) Ann Clin Res. 1, 199.
Europe	Finland	Rural Finn	21.93	60.41	159.00	0.83	BG	Jussila et al. (1970) Scand J Gastroenterol. 5, 49.
Europe	Finland	Swedish-speaking Finns	21.43	60.60	156.00	0.83	BG	Sahi (1974) Scand J Gastroenterol. 9, 303.
Europe	France	French	5.82	44.93	102.00	0.76	BH	Cloarec et al. (1991) Gastroenterol Clin Biol. 15, 588.
Europe	France	French	6.63	49.75	85.00	0.71	BH	Cuddenec et al. (1982) Gastroenterol Clin Biol. 6, 776.
Europe	France	Maghrebins (Northern African Muslims)	7.25	43.70	55.00	0.22	BG	O'Morain et al. (1978) Acta Gastroenterol Belg. 41, 56-63.
Europe	France	Northern French	6.63	49.75	76.00	0.78	BH	Cuddenec et al. (1982) Gastroenterol Clin Biol. 6, 776.
Europe	France	Southern French	6.63	49.75	40.00	0.43	BH	Cuddenec et al. (1982) Gastroenterol Clin Biol. 6, 776.
Europe	France	Southern French	7.25	43.70	55.00	0.58	BG	O'Morain et al. (1978) Acta Gastroenterol Belg. 41, 56-63.
Europe	Germany	Baden-Wurttemberg Germans	9.50	48.40	136.00	0.76	BH	Flatz et al. (1982) Hum Genet. 62, 152.
Europe	Germany	Bayern Germans	12.53	47.80	221.00	0.86	BH	Flatz et al. (1982) Hum Genet. 62, 152.
Europe	Germany	Eastern Germans	13.75	51.05	246.00	0.78	BH	Flatz et al. (1982) Hum Genet. 62, 152.
Europe	Germany	Germans	8.52	53.18	60.00	0.87	Biopsy	Howell et al. (1980) Hepatogastroenterology 27, 208.
Europe	Germany	Northwest Germans	8.80	53.08	341.00	0.91	BH	Flatz et al. (1982) Hum Genet. 62, 152.
Europe	Germany	Rheinland and Pfalz Germans	8.27	50.00	182.00	0.86	BH	Flatz et al. (1982) Hum Genet. 62, 152.
Europe	Germany	Schleswig-Holstein Germans	9.55	54.52	100.00	0.94	BH	Flatz et al. (1982) Hum Genet. 62, 152.
Europe	Greece	Continental Greeks	23.73	37.98	600.00	0.55	BG	Kanaghinis et al. (1974) Am J Dig Dis. 19, 1021.
Europe	Greece	Cretan Greek	25.13	35.33	50.00	0.44	BG	Kanaghinis et al. (1974) Am J Dig Dis. 19, 1021.
Europe	Greece	Greek	23.73	37.98	16.00	0.63	BG	Spanidou & Petrakis(1972) Lancet. 2, 872.
Europe	Greece	Greeks	23.73	37.98	200.00	0.25	BH	Ladas et al. (1982) Gut. 23, 968.
Europe	Greece	Greeks	23.73	37.98	250.00	0.77	BG	Zografos et al. (1973) Lancet. 301,367.
Europe	Hungary	Eastern Hungarian	20.08	47.50	70.00	0.71	BH	Czeizel et al. (1983) Hum Genet. 64, 398.
Europe	Hungary	Hungarian	19.08	47.50	262.00	0.59	BH	Czeizel et al. (1983) Hum Genet. 64, 398.
Europe	Hungary	Matyo	20.58	47.82	172.00	0.63	BH	Czeizel et al. (1983) Hum Genet. 64, 398.
Europe	Hungary	Northeastern Hungarian	21.73	47.98	103.00	0.58	BH	Czeizel et al. (1983) Hum Genet. 64, 398.
Europe	Hungary	Romai	21.72	47.95	113.00	0.44	BH	Czeizel et al. (1983) Hum Genet. 64, 398.
Europe	Hungary	Western Hungarian	19.08	47.50	100.00	0.72	BH	Czeizel et al. (1983) Hum Genet. 64, 398.
Europe	Ireland	Native Irish	-6.25	53.33	50.00	0.96	BG	Fielding et al. (1981) Ir J Med Sci. 150, 276.
Europe	Italy	Italians	9.20	45.47	42.00	0.38	BH	Bozzani et al. (1986) Dig Dis Sci. 31, 1313.
Europe	Italy	Italians	9.20	45.47	89.00	0.48	BG	Cavalli-Sforza et al. (1987) Am J Clin Nutr. 45, 748.
Europe	Italy	Italians	12.48	41.90	65.00	0.82	BG	Cavalli-Sforza et al. (1987) Am J Clin Nutr. 45, 748.
Europe	Italy	Italians	14.25	40.83	51.00	0.59	BG	Cavalli-Sforza et al. (1987) Am J Clin Nutr. 45, 748.
Europe	Italy	Italians	14.25	40.83	44.00	0.23	Biopsy	Rossi et al. (1997) Gastroenterology. 112, 1506.
Europe	Italy	Italians	9.20	45.47	20.00	0.25	BH	Zuccato et al. (1983) Eur J Clin Invest. 13, 261.
Europe	Italy	Napolitans	14.25	40.83	99.00	0.46	BH	Rinaldi et al. (1984) Lancet. 1, 355.
Europe	Italy	Neapolitan	14.25	40.83	9.00	0.00	BG	De Ritis et al. (1970) Enzymol Biol Clin. 11, 263.
Europe	Italy	Northern Italians	7.67	45.05	208.00	0.49	BH	Burgio et al. (1984) Am J Clin Nutr. 39, 100.
Europe	Italy	Sardinians	8.56	40.73	50.00	0.14	BH	Meloni et al. (2001) Am J Clin Nutr. 73, 582.
Europe	Italy	Sardinians	9.00	39.40	47.00	0.15	BH	Meloni et al. (1998) Ital J Gastroenterol Hepatol. 30, 490.
Europe	Italy	Sardinians	9.00	40.10	53.00	0.11	BH	Meloni et al. (1998) Ital J Gastroenterol Hepatol. 30, 490.
Europe	Italy	Sardinians	9.00	40.30	38.00	0.18	BH	Meloni et al. (1998) Ital J Gastroenterol Hepatol. 30, 490.
Europe	Italy	Sicilians	13.37	38.12	100.00	0.29	BH	Burgio et al. (1984) Am J Clin Nutr. 39, 100.

Europe	Poland	Eastern Polish	23.13	52.03	35.00	0.63	BH	Socha et al. (1984) Ann Hum Biol. 11, 311.
Europe	Poland	Northeastern Polish	22.35	53.83	34.00	0.59	BH	Socha et al. (1984) Ann Hum Biol. 11, 311.
Europe	Poland	Polish	21.00	52.25	21.00	0.71	BG	Leichter (1972) Am J Dig Dis. 17, 73.
Europe	Poland	Polish	19.00	51.73	29.00	0.62	BH	Socha et al. (1984) Ann Hum Biol. 11, 311.
Europe	Poland	Polish	19.37	52.23	92.00	0.63	BH	Socha et al. (1984) Ann Hum Biol. 11, 311.
Europe	Poland	Polish	19.37	52.23	85.00	0.64	BH	Socha et al. (1984) Ann Hum Biol. 11, 311.
Europe	Russia	Kildin Saami	34.30	69.29	50.00	0.52	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Europe	Russia	Komi-Permiaks	37.59	55.75	112.00	0.50	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Europe	Russia	Udmurtians	32.00	68.00	75.00	0.41	BG	Kozlov (1998) Int J Circumpol Heal. 57, 18.
Europe	Spain	Galician	-7.28	42.37	338.00	0.66	BH	Leis et al. (1997) J Pediatr Gastroenterol Nutr. 25, 296.
Europe	UK	British	-3.20	55.95	150.00	0.95	BG	Ferguson et al. (1984) Gut. 25, 163.
Europe	UK	British natives	-1.25	51.75	75.00	0.95	Biopsy	Ho et al. (1982) Am J Hum Genet. 34, 650.
Europe	UK	White British	-1.92	52.47	67.00	0.97	Biopsy	Iqbal et al. (1993) Br Med J. 306, 1303.
Europe	Uzbekistan	Kazakh	63.46	40.13	83.00	0.25	BG+BH	Heyer et al. (2011) Hum Biol. 83, 379
Europe	Uzbekistan	Tajiko-Uzbek	64.42	39.77	100.00	0.11	BG+BH	Heyer et al. (2011) Hum Biol. 83, 379
Middle East	Iran	Iranian	51.42	35.67	21.00	0.14	BG	Sadre et al. (1979) Am J Clin Nutr. 32, 1948.
Middle East	Israel	Arabs	34.95	32.23	67.00	0.19	BG	Gilat et al. (1971) Dig Dis. 16, 203
Middle East	Jordan	Jordanian Arabs	35.93	31.95	148.00	0.25	BH	Hijazi et al. (1983) Trop Geogr Med. 35, 157.
Middle East	Jordan	Mediterranean origin Jordanian Arabs	35.93	31.95	56.00	0.23	BG	Snook et al. (1976) Trop Geogr Med. 28, 333.
Middle East	Jordan	Urban/agricultural Jordanian Arabs	35.93	31.95	162.00	0.76	BH	Hijazi et al. (1983) Trop Geogr Med. 35, 157.
Middle East	Kuwait	Arab Kuwaiti	47.98	29.37	70.00	0.53	BH	Sanae et al. (2003) Med Princ Pract. 12, 160.
Middle East	Kuwait	Asian Kuwaiti	47.98	29.37	79.00	0.42	BH	Sanae et al. (2003) Med Princ Pract. 12, 160.
Middle East	Lebanon	Lebanese	35.51	33.87	74.00	0.22	BG	Nasrallah (1979) Am J Clin Nutr. 32, 1994.
Middle East	Oman	Omani	58.50	23.53	50.00	0.16	BH	Al-Abri et al. (2013) Oman Med J. 28, 341.
Middle East	Saudi Arabia	Arabs	50.11	26.43	109.00	0.43	BH	Dissanayake et al. (1990) Ann Saudi Med. 10, 598.
Middle East	Saudi Arabia	Bedouin	50.11	26.43	21.00	0.81	BH	Dissanayake et al. (1990) Ann Saudi Med. 10, 598.
Middle East	Saudi Arabia	Beduin and Urban Saudi	46.77	24.64	14.00	0.86	BG	Cook & Al Torki (1975) Br Med J. 3, 135.
Middle East	Saudi Arabia	Yemenites	50.11	26.43	17.00	0.53	BH	Dissanayake et al. (1990) Ann Saudi Med. 10, 598.
Middle East	Turkey	Central Anatolia	39.50	34.00	104.00	0.29	BH	Flatz et al. (1986) Am J Hum Genet. 38, 515.
Middle East	Turkey	Eastern Anatolia	39.50	40.00	122.00	0.26	BH	Flatz et al. (1986) Am J Hum Genet. 38, 515.
Middle East	Turkey	North Coast of Turkey	34.00	41.50	64.00	0.31	BH	Flatz et al. (1986) Am J Hum Genet. 38, 515.
Middle East	Turkey	South Coast of Turkey	33.00	36.50	54.00	0.28	BH	Flatz et al. (1986) Am J Hum Genet. 38, 515.
Middle East	Turkey	Turks	32.86	39.93	30.00	0.63	BG	Tuncbilek et al. (1973) Lancet. 21, 151.
Middle East	Turkey	Western Anatolia and European Turkey	28.96	41.02	126.00	0.30	BH	Flatz et al. (1986) Am J Hum Genet. 38, 515.

Appendix C1: Diversity measures and tests of neutrality for the regions included in haplotype analysis for phenotyped samples. N: number of chromosomes.

Population	N	Sequence region	Segregating sites	Haplotypes	Haplotype diversity, *H* (±SD)	Nucleotide diversity, π	Tajima's *D*	Fu & Li's *D**	Fu & Li's *F**
Europeans D	40	Intron 4	4	5	0.459 (0.008)	0.001	-1.095	-1.103	-1.282
		Enhancer	3	4	0.500 (0.006)	0.002	0.554	-0.350	-0.096
		Hapdef	2	3	0.349 (0.008)	0.001	-0.406	0.771	0.498
		Total	9	10	0.542 (0.009)	0.001	-0.488	-0.520	-0.598
Europeans ND	30	Intron 4	4	5	0.752 (0.002)	0.003	0.087	0.045	0.066
		Enhancer	2	3	0.480 (0.005)	0.001	0.088	-0.738	-0.583
		Hapdef	3	4	0.690 (0.003)	0.002	0.504	0.950	0.952
		Total	9	9	0.839 (0.001)	0.002	0.298	0.204	0.272
Amhara D	26	Intron 4	2	3	0.665 (0.002)	0.002	1.178	0.826	1.065
		Enhancer	6	7	0.806 (0.003)	0.002	-0.689	0.476	0.157
		Hapdef	2	3	0.665 (0.002)	0.002	1.178	0.826	1.065
		Total	10	8	0.855 (0.002)	0.002	0.263	0.891	0.819
Amhara ND	54	Intron 4	2	3	0.653 (0.001)	0.002	1.832	0.740	1.236
		Enhancer	5	6	0.470 (0.005)	0.001	-1.267	-1.835	-1.940
		Hapdef	2	3	0.662 (0.001)	0.002	1.723	0.740	1.199
		Total	9	10	0.760 (0.002)	0.002	0.430	-0.662	-0.358
Oromo D	50	Intron 4	3	4	0.690 (0.001)	0.003	1.110	0.892	1.115
		Enhancer	8	9	0.722 (0.002)	0.002	-0.898	-1.554	-1.578
		Hapdef	3	4	0.688 (0.000)	0.002	0.953	-0.413	-0.001
		Total	14	12	0.776 (0.002)	0.002	0.079	-0.876	-0.655
Oromo ND	74	Intron 4	3	5	0.559 (0.002)	0.002	0.223	0.853	0.769
		Enhancer	6	7	0.397 (0.005)	0.001	-1.449	-0.690	-1.103
		Hapdef	3	4	0.584 (0.001)	0.001	0.210	-0.515	-0.339
		Total	12	14	0.738 (0.002)	0.001	-0.692	-0.298	-0.516
Jaali D	50	Intron 4	3	4	0.664 (0.001)	0.002	0.477	-0.413	-0.168
		Enhancer	8	9	0.786 (0.001)	0.002	-0.930	-1.554	-1.590
		Hapdef	3	4	0.611 (0.002)	0.002	0.228	-0.413	-0.256
		Total	14	14	0.900 (0.000)	0.002	-0.369	-1.360	-1.213
Jaali ND	26	Intron 4	3	4	0.717 (0.001)	0.003	0.639	-0.216	0.030
		Enhancer	5	6	0.686 (0.008)	0.002	-0.744	0.308	0.003
		Hapdef	3	4	0.717 (0.001)	0.002	0.639	-0.216	0.030
		Total	11	9	0.862 (0.002)	0.002	0.075	-0.001	0.026

Appendix C2: Diversity measures and tests of neutrality for the regions included in haplotype analysis for the combined dataset of 28 populations. N: number of chromosomes, significant values (p<0.05) for neutrality tests are shaded.

Population	N	Sequencing region	Segregating sites	Haplotypes	Haplotype diversity, *H* (±SD)	Nucleotide diversity, π	Tajima's *D*	Fu & Li's *D**	Fu & Li's *F**
Afar	124	Intron_4	2	3	0.653 (0.000)	0.002	1.752	0.669	1.180
		Enhancer	7	9	0.754 (0.000)	0.002	0.066	-0.622	-0.459
		Hapdef	2	3	0.644 (0.000)	0.002	1.651	0.669	1.143
		Total	11	12	0.810 (0.000)	0.002	1.108	0.039	0.502
Amhara	80	Intron_4	2	3	0.662 (0.000)	0.002	1.857	0.704	1.231
		Enhancer	7	8	0.600 (0.003)	0.001	-1.154	0.365	-0.165
		Hapdef	2	3	0.661 (0.000)	0.002	1.765	0.704	1.198
		Total	11	12	0.794 (0.001)	0.002	0.291	0.794	0.735
Anatolian Turks	40	Intron_4	4	5	0.694 (0.003)	0.003	0.565	1.027	1.034
		Enhancer	2	3	0.232 (0.007)	0.001	-0.416	0.771	0.495
		Hapdef	3	4	0.679 (0.001)	0.002	0.649	-0.350	-0.063
		Total	9	9	0.771 (0.003)	0.002	0.454	0.733	0.757
Asante	40	Intron_4	3	4	0.714 (0.002)	0.002	0.665	0.916	0.978
		Enhancer	2	3	0.273 (0.008)	0.001	-0.745	0.771	0.384
		Hapdef	2	3	0.472 (0.006)	0.001	0.171	0.771	0.692
		Total	7	7	0.833 (0.001)	0.001	0.122	1.256	1.056
Beni Amer	128	Intron_4	5	5	0.654 (0.000)	0.002	-0.198	-2.266	-1.867
		Enhancer	5	6	0.786 (0.000)	0.002	1.026	1.016	1.205
		Hapdef	3	4	0.612 (0.001)	0.002	0.591	0.807	0.868
		Total	13	14	0.855 (0.000)	0.002	0.614	-0.364	-0.008
Chagga	82	Intron_4	3	4	0.703 (0.001)	0.003	1.014	0.844	1.050
		Enhancer	4	5	0.459 (0.004)	0.001	-0.753	-0.220	-0.457
		Hapdef	2	3	0.417 (0.003)	0.001	0.185	0.702	0.635
		Total	9	9	0.863 (0.000)	0.001	0.114	0.603	0.517
Chewa	40	Intron_4	5	5	0.733 (0.001)	0.003	-0.266	-0.736	-0.692
		Enhancer	3	4	0.273 (0.008)	0.001	-1.177	-1.616	-1.727
		Hapdef	2	3	0.512 (0.002)	0.001	0.243	-0.828	-0.599
		Total	10	10	0.829 (0.001)	0.001	-0.562	-1.506	-1.415
Iranians	76	Intron_4	4	5	0.675 (0.001)	0.003	0.628	0.965	1.007
		Enhancer	2	3	0.362 (0.003)	0.001	-0.017	0.708	0.569
		Hapdef	3	4	0.674 (0.000)	0.002	1.038	-0.521	-0.044
		Total	9	8	0.726 (0.001)	0.002	0.819	0.617	0.811
Israeli-Arabs	40	Intron_4	6	6	0.736 (0.002)	0.003	-0.457	-0.456	-0.533
		Enhancer	7	8	0.555 (0.007)	0.001	-1.632	-2.464	-2.5812
		Hapdef	2	3	0.656 (0.001)	0.002	1.407	0.771	1.107
		Total	15	13	0.837 (0.002)	0.002	-0.778	-1.460	-1.457
Israeli-Bedouin	32	Intron_4	3	4	0.720 (0.001)	0.003	0.860	0.942	1.064
		Enhancer	5	6	0.641 (0.006)	0.002	-0.836	-1.526	-1.537
		Hapdef	2	3	0.677 (0.001)	0.002	1.515	0.798	1.158
		Total	10	10	0.861 (0.001)	0.002	0.333	-0.240	-0.072
Italians	32	Intron_4	4	5	0.752 (0.001)	0.003	0.171	1.051	0.920
		Enhancer	3	4	0.504 (0.006)	0.001	-0.478	-1.508	-1.402
		Hapdef	3	4	0.702 (0.002)	0.002	0.641	0.942	0.991
		Total	10	12	0.867 (0.001)	0.002	0.151	0.309	0.304
Jaali	76	Intron_4	3	4	0.682 (0.000)	0.002	0.825	0.851	0.986
		Enhancer	8	9	0.752 (0.001)	0.002	-0.829	-1.039	-1.144
		Hapdef	3	4	0.654 (0.001)	0.002	0.692	0.851	0.938
		Total	14	14	0.891 (0.000)	0.002	-0.042	-0.057	-0.061
Jordanian Bedouin	44	Intron_4	2	3	0.672 (0.001)	0.003	1.919	0.761	1.275
		Enhancer	4	5	0.725 (0.002)	0.002	0.781	1.017	1.103
		Hapdef	3	4	0.707 (0.001)	0.002	1.170	0.905	1.145
		Total	9	7	0.775 (0.002)	0.002	1.616	1.357	1.692
Khalka	38	Intron_4	3	4	0.585 (0.005)	0.002	-0.110	0.922	0.718
		Enhancer	3	4	0.542 (0.003)	0.001	-0.334	-1.592	-1.419
		Hapdef	2	3	0.519 (0.006)	0.001	0.403	0.777	0.775
		Total	8	8	0.795 (0.001)	0.001	-0.066	-0.030	-0.048
Kuwaiti	56	Intron_4	3	5	0.647 (0.002)	0.003	0.992	-0.444	-0.008
		Enhancer	6	7	0.639 (0.003)	0.002	-0.867	-1.464	-1.495
		Hapdef	3	4	0.662 (0.001)	0.002	0.940	-0.444	-0.027
		Total	12	12	0.822 (0.001)	0.002	0.190	-1.300	-0.942

Mambila	40	Intron_4	3	4	0.509 (0.006)	0.002	-0.425	0.916	0.603
		Enhancer	3	4	0.355 (0.009)	0.001	-0.745	0.916	0.493
		Hapdef	1	2	0.328 (0.006)	0.001	0.565	0.564	0.651
		Total	7	7	0.731 (0.004)	0.001	-0.480	1.256	0.836
Northern Europeans	40	Intron_4	3	4	0.458 (0.008)	0.001	-0.615	0.916	0.538
		Enhancer	3	4	0.522 (0.005)	0.002	0.681	-0.350	-0.052
		Hapdef	2	3	0.349 (0.008)	0.001	-0.406	0.771	0.498
		Total	8	8	0.569 (0.008)	0.001	-0.121	0.631	0.463
Norwegians	40	Intron_4	2	3	0.229 (0.007)	0.001	-0.946	-0.828	-0.999
		Enhancer	2	3	0.268 (0.007)	0.001	0.063	0.771	0.656
		Hapdef	2	3	0.229 (0.007)	0.001	-0.946	-0.828	-0.999
		Total	6	4	0.273 (0.008)	0.001	-0.862	-0.456	-0.680
Oromo	124	Intron_4	3	5	0.643 (0.001)	0.002	1.041	0.809	1.039
		Enhancer	9	10	0.598 (0.002)	0.001	-1.214	0.529	-0.097
		Hapdef	4	5	0.655 (0.000)	0.002	0.295	-1.545	-1.116
		Total	16	18	0.798 (0.001)	0.002	-0.355	-0.005	-0.162
Palestinians	38	Intron_4	6	6	0.758 (0.001)	0.003	-0.293	-1.250	-1.118
		Enhancer	3	4	0.422 (0.007)	0.001	-0.735	-1.592	-1.555
		Hapdef	3	4	0.684 (0.001)	0.002	0.533	-0.335	-0.093
		Total	12	13	0.862 (0.001)	0.002	-0.240	-1.541	-1.321
Romanians	40	Intron_4	4	5	0.713 (0.001)	0.003	0.047	-1.103	-0.882
		Enhancer	2	3	0.344 (0.007)	0.001	0.480	0.771	0.796
		Hapdef	3	4	0.704 (0.002)	0.002	0.887	0.916	1.054
		Total	9	9	0.859 (0.001)	0.002	0.587	0.107	0.304
Saudi Bedouin	40	Intron_4	2	3	0.640 (0.002)	0.002	1.814	0.771	1.244
		Enhancer	5	6	0.674 (0.002)	0.002	-0.555	-1.662	-1.546
		Hapdef	2	3	0.653 (0.001)	0.002	1.711	0.771	1.209
		Total	9	10	0.773 (0.003)	0.002	0.858	-0.520	-0.097
Shabo	42	Intron_4	5	6	0.791 (0.001)	0.003	0.042	0.178	0.159
		Enhancer	3	4	0.297 (0.007)	0.001	-1.196	-1.639	-1.752
		Hapdef	2	3	0.487 (0.003)	0.001	0.153	-0.843	-0.641
		Total	10	11	0.854 (0.001)	0.001	-0.419	-0.953	-0.919
Syrians	82	Intron_4	3	4	0.656 (0.001)	0.003	1.035	0.844	1.058
		Enhancer	2	3	0.427 (0.003)	0.001	0.171	0.702	0.630
		Hapdef	4	5	0.680 (0.001)	0.002	0.371	-0.220	-0.038
		Total	9	11	0.736 (0.001)	0.002	0.744	0.603	0.771
Tharu	36	Intron_4	4	5	0.668 (0.003)	0.002	-0.362	-0.008	-0.132
		Enhancer	2	3	0.341 (0.009)	0.001	-0.012	0.783	0.642
		Hapdef	2	3	0.595 (0.003)	0.001	0.807	0.783	0.914
		Total	8	8	0.825 (0.001)	0.001	0.083	0.653	0.557
Ukrainians	40	Intron_4	3	4	0.664 (0.002)	0.002	0.344	0.916	0.868
		Enhancer	3	4	0.627 (0.003)	0.002	0.887	-0.350	0.018
		Hapdef	3	4	0.558 (0.004)	0.001	-0.195	-0.350	-0.353
		Total	9	7	0.795 (0.001)	0.002	0.463	0.107	0.258
Yemeni Hadramaut	154	Intron_4	4	5	0.660 (0.000)	0.003	0.587	-0.356	-0.055
		Enhancer	9	10	0.675 (0.001)	0.002	-1.041	-1.907	-1.906
		Hapdef	4	5	0.680 (0.000)	0.002	0.645	0.905	0.969
		Total	17	20	0.867 (0.000)	0.002	-0.209	-1.056	-0.878
Yemeni Sena	66	Intron_4	4	5	0.716 (0.001)	0.003	0.331	0.978	0.907
		Enhancer	6	7	0.770 (0.001)	0.002	0.195	1.153	0.992
		Hapdef	2	3	0.610 (0.001)	0.002	1.163	0.721	0.994
		Total	12	13	0.892 (0.000)	0.002	0.587	1.476	1.387

Appendix D: Pairwise F_{ST} values (lower triangle) and corresponding *p*-values (upper triangles) for the 52 populations tested in chapter 3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	*	0.045	0.559	0.036	0.000	0.000	0.000	0.270	0.000	0.991	0.000	0.000	0.135	0.297	0.117	0.000	0.009	0.099	0.000	0.000	0.000	0.459	0.000	0.000	0.459
2	0.035	*	0.261	0.000	0.387	0.225	0.000	0.793	0.000	0.090	0.000	0.000	0.739	0.000	0.000	0.000	0.000	0.000	0.000	0.396	0.009	0.018	0.000	0.784	0.721
3	-0.009	0.002	*	0.000	0.171	0.036	0.000	0.676	0.000	0.550	0.000	0.000	0.550	0.216	0.027	0.000	0.009	0.027	0.000	0.090	0.000	0.225	0.027	0.153	0.991
4	0.033	0.124	0.072	*	0.000	0.000	0.000	0.009	0.000	0.072	0.000	0.991	0.000	0.432	0.586	0.000	0.955	0.559	0.000	0.000	0.000	0.477	0.000	0.000	0.135
5	0.077	-0.001	0.029	0.169	*	0.748	0.000	0.234	0.000	0.018	0.000	0.000	0.189	0.000	0.000	0.000	0.000	0.000	0.000	0.991	0.099	0.000	0.153	0.586	0.333
6	0.117	0.007	0.048	0.245	-0.011	*	0.000	0.342	0.000	0.009	0.000	0.000	0.144	0.000	0.000	0.000	0.000	0.000	0.000	0.865	0.117	0.009	0.468	0.396	0.333
7	0.625	0.479	0.528	0.692	0.401	0.367	*	0.000	0.342	0.000	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.008	-0.013	-0.017	0.129	0.006	0.019	0.498	*	0.000	0.387	0.000	0.000	0.991	0.045	0.009	0.000	0.000	0.018	0.000	0.351	0.045	0.171	0.072	0.514	0.991
9	0.727	0.570	0.632	0.813	0.484	0.459	-0.003	0.605	*	0.000	0.009	0.000	0.000	0.000	0.000	0.036	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	-0.009	0.030	-0.011	0.043	0.070	0.110	0.621	0.004	0.727	*	0.000	0.018	0.234	0.234	0.045	0.000	0.018	0.072	0.000	0.009	0.000	0.360	0.000	0.027	0.730
11	0.452	0.266	0.320	0.554	0.185	0.147	0.068	0.283	0.129	0.446	*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.270	0.000	0.045	0.000	0.000
12	0.049	0.158	0.111	-0.010	0.208	0.307	0.736	0.180	0.851	0.063	0.621	*	0.009	0.135	0.225	0.000	0.631	0.405	0.000	0.000	0.000	0.189	0.000	0.000	0.054
13	0.015	-0.006	-0.007	0.085	0.010	0.024	0.507	-0.016	0.592	0.011	0.301	0.108	*	0.018	0.009	0.000	0.000	0.000	0.000	0.144	0.000	0.063	0.009	0.441	0.991
14	0.001	0.074	0.016	0.004	0.118	0.170	0.651	0.049	0.755	0.007	0.492	0.014	0.045	*	0.865	0.000	0.162	0.757	0.000	0.000	0.000	0.856	0.000	0.000	0.144
15	0.011	0.102	0.033	-0.002	0.156	0.224	0.703	0.078	0.801	0.015	0.565	0.002	0.066	-0.005	*	0.000	0.081	0.991	0.000	0.000	0.000	0.991	0.000	0.000	0.099
16	0.863	0.742	0.835	0.944	0.668	0.681	0.129	0.816	0.087	0.867	0.363	0.956	0.749	0.888	0.906	*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	0.048	0.139	0.084	-0.017	0.183	0.256	0.695	0.140	0.808	0.057	0.558	-0.005	0.099	0.016	0.010	0.934	*	0.153	0.000	0.000	0.000	0.108	0.000	0.000	0.036
18	0.016	0.107	0.043	-0.005	0.156	0.226	0.694	0.091	0.802	0.021	0.554	-0.003	0.070	-0.006	-0.007	0.920	0.007	*	0.000	0.000	0.000	0.991	0.000	0.000	0.081
19	0.183	0.177	0.130	0.197	0.187	0.192	0.556	0.143	0.607	0.186	0.386	0.242	0.169	0.180	0.216	0.749	0.197	0.211	*	0.000	0.000	0.000	0.000	0.000	0.018
20	0.088	0.002	0.037	0.186	-0.010	-0.013	0.394	0.011	0.479	0.082	0.178	0.228	0.016	0.133	0.173	0.667	0.200	0.173	0.195	*	0.090	0.000	0.243	0.468	0.514
21	0.310	0.115	0.172	0.464	0.051	0.026	0.195	0.135	0.279	0.302	0.015	0.548	0.146	0.363	0.449	0.558	0.464	0.447	0.265	0.046	*	0.000	0.532	0.027	0.081
22	-0.002	0.067	0.012	-0.001	0.106	0.154	0.628	0.047	0.747	0.002	0.457	0.013	0.040	-0.011	-0.012	0.912	0.015	-0.011	0.155	0.120	0.334	*	0.000	0.027	0.216
23	0.221	0.058	0.111	0.369	0.014	-0.004	0.264	0.076	0.350	0.214	0.061	0.449	0.085	0.275	0.353	0.610	0.374	0.352	0.225	0.009	-0.018	0.250	*	0.018	0.198
24	0.052	-0.007	0.014	0.139	-0.006	-0.003	0.450	-0.005	0.536	0.047	0.235	0.172	0.000	0.093	0.123	0.706	0.154	0.125	0.183	-0.005	0.089	0.084	0.038	*	0.541
25	-0.006	-0.022	-0.030	0.156	-0.003	0.011	0.489	-0.036	0.601	-0.010	0.265	0.227	-0.026	0.038	0.069	0.843	0.157	0.090	0.121	0.003	0.116	0.042	0.063	-0.014	*
26	0.745	0.608	0.671	0.817	0.530	0.513	0.016	0.646	-0.011	0.745	0.179	0.848	0.627	0.770	0.808	0.053	0.814	0.808	0.651	0.526	0.345	0.764	0.414	0.576	0.648
27	0.598	0.427	0.479	0.683	0.342	0.304	-0.005	0.446	0.021	0.594	0.026	0.736	0.458	0.629	0.690	0.203	0.684	0.682	0.506	0.335	0.131	0.602	0.196	0.395	0.432
28	0.045	-0.021	-0.002	0.210	-0.016	-0.010	0.442	-0.021	0.548	0.039	0.216	0.281	-0.015	0.100	0.146	0.788	0.211	0.162	0.146	-0.013	0.074	0.098	0.028	-0.020	-0.030
29	0.163	0.020	0.072	0.327	-0.009	-0.020	0.329	0.038	0.425	0.156	0.109	0.406	0.042	0.224	0.292	0.682	0.332	0.299	0.200	-0.012	0.001	0.208	-0.019	0.006	0.029
30	0.003	0.040	-0.001	0.028	0.073	0.099	0.572	0.015	0.665	0.004	0.383	0.046	0.023	0.008	0.011	0.830	0.038	0.019	0.087	0.083	0.240	0.002	0.173	0.055	0.001
31	-0.019	0.030	-0.016	0.025	0.064	0.096	0.577	0.006	0.693	-0.018	0.381	0.054	0.011	-0.014	-0.009	0.892	0.039	-0.003	0.127	0.075	0.242	-0.019	0.171	0.045	-0.007
32	0.294	0.141	0.195	0.385	0.081	0.053	0.167	0.158	0.237	0.286	0.012	0.442	0.171	0.335	0.397	0.447	0.396	0.386	0.286	0.074	-0.019	0.306	0.001	0.116	0.147
33	-0.011	0.043	-0.006	0.023	0.080	0.122	0.611	0.018	0.725	-0.010	0.433	0.042	0.019	-0.006	-0.009	0.887	0.029	-0.003	0.125	0.094	0.295	-0.012	0.216	0.059	0.006
34	0.028	-0.009	0.000	0.107	0.002	0.011	0.485	-0.014	0.572	0.024	0.275	0.135	-0.006	0.064	0.089	0.735	0.122	0.092	0.175	0.006	0.123	0.057	0.065	-0.005	-0.024
35	0.099	-0.003	0.034	0.244	-0.016	-0.019	0.388	0.007	0.485	0.093	0.165	0.311	0.011	0.155	0.209	0.718	0.253	0.217	0.178	-0.016	0.037	0.143	0.002	-0.011	-0.002
36	0.158	0.031	0.077	0.276	-0.001	-0.011	0.313	0.046	0.397	0.150	0.101	0.338	0.051	0.204	0.266	0.621	0.287	0.264	0.209	-0.002	-0.002	0.186	-0.015	0.017	0.036
37	0.391	0.210	0.265	0.495	0.136	0.101	0.110	0.227	0.177	0.383	-0.011	0.562	0.244	0.433	0.506	0.412	0.501	0.495	0.348	0.128	-0.005	0.398	0.028	0.181	0.211
38	0.001	0.006	-0.012	0.059	0.031	0.050	0.539	-0.011	0.627	0.000	0.341	0.081	-0.002	0.026	0.041	0.783	0.073	0.046	0.159	0.038	0.189	0.020	0.121	0.016	-0.024
39	0.280	0.113	0.169	0.397	0.055	0.030	0.204	0.132	0.282	0.272	0.026	0.465	0.143	0.329	0.400	0.520	0.405	0.392	0.278	0.049	-0.022	0.298	-0.013	0.088	0.120
40	0.687	0.530	0.590	0.768	0.445	0.417	-0.010	0.562	-0.011	0.686	0.094	0.809	0.554	0.716	0.764	0.127	0.766	0.762	0.584	0.440	0.237	0.701	0.309	0.496	0.557
41	0.023	0.110	0.043	0.003	0.160	0.221	0.690	0.086	0.785	0.029	0.548	0.009	0.076	0.004	0.000	0.894	0.006	0.002	0.168	0.176	0.429	-0.001	0.339	0.130	0.077
42	0.115	-0.002	0.036	0.286	-0.021	-0.024	0.363	0.008	0.459	0.108	0.138	0.368	0.011	0.175	0.238	0.719	0.286	0.247	0.166	-0.020	0.014	0.162	-0.010	-0.011	-0.002
43	0.035	-0.011	0.001	0.133	-0.002	0.006	0.475	-0.015	0.570	0.029	0.261	0.173	-0.009	0.076	0.107	0.755	0.147	0.113	0.168	0.001	0.112	0.069	0.056	-0.008	-0.024
44	0.016	-0.008	-0.009	0.100	0.008	0.021	0.500	-0.018	0.594	0.011	0.290	0.132	-0.010	0.048	0.074	0.771	0.114	0.080	0.160	0.014	0.139	0.045	0.079	-0.002	-0.029
45	0.289	0.128	0.183	0.392	0.069	0.042	0.185	0.146	0.259	0.281	0.018	0.455	0.158	0.334	0.401	0.483	0.402	0.391	0.291	0.062	-0.020	0.303	-0.006	0.103	0.134
46	0.210	0.075	0.128	0.305	0.031	0.011	0.246	0.093	0.321	0.202	0.056	0.356	0.101	0.254	0.309	0.525	0.317	0.301	0.257	0.025	-0.016	0.230	-0.016	0.056	0.084
47	0.106	0.008	0.045	0.212	-0.010	-0.015	0.365	0.018	0.449	0.098	0.148	0.263	0.022	0.151	0.200	0.654	0.225	0.199	0.189	-0.011	0.026	0.136	-0.002	-0.001	0.009
48	0.001	0.076	0.018	0.002	0.119	0.174	0.652	0.056	0.768	0.005	0.492	0.013	0.045	-0.008	-0.008	0.912	0.017	-0.008	0.170	0.133	0.374	-0.016	0.284	0.093	0.050
49	0.099	-0.001	0.036	0.231	-0.014	-0.017	0.388	0.009	0.482	0.092	0.166	0.293	0.013	0.152	0.204	0.705	0.242	0.209	0.182	-0.014	0.038	0.139	0.003	-0.009	0.000
50	-0.005	0.063	0.006	0.014	0.108	0.162	0.657	0.036	0.764	-0.003	0.498	0.027	0.035	-0.004	-0.003	0.893	0.027	-0.001	0.194	0.123	0.369	-0.012	0.277	0.081	0.025
51	0.137	0.154	0.111	0.131	0.174	0.183	0.547	0.124	0.588	0.138	0.390	0.152	0.145	0.130	0.145	0.698	0.132	0.141	-0.007	0.182	0.271	0.113	0.228	0.167	0.107
52	0.175	0.143	0.110	0.199	0.138	0.124	0.414	0.110	0.439	0.179	0.247	0.242	0.149	0.184	0.231	0.584	0.204	0.216	0.015	0.141	0.148	0.150	0.129	0.139	0.087

	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
1	0.000	0.000	0.090	0.000	0.270	0.991	0.000	0.757	0.045	0.018	0.000	0.000	0.225	0.000	0.000	0.027	0.018	0.045	0.126	0.000	0.000	0.000	0.369	0.000	0.703	0.000
2	0.000	0.000	0.991	0.216	0.027	0.216	0.000	0.117	0.991	0.532	0.045	0.000	0.225	0.009	0.000	0.000	0.324	0.991	0.541	0.000	0.027	0.198	0.027	0.315	0.027	0.000
3	0.000	0.000	0.477	0.054	0.441	0.649	0.009	0.342	0.387	0.144	0.009	0.000	0.694	0.000	0.000	0.045	0.090	0.279	0.432	0.000	0.000	0.072	0.108	0.036	0.297	0.000
4	0.000	0.000	0.009	0.000	0.045	0.189	0.000	0.171	0.000	0.000	0.000	0.000	0.036	0.000	0.000	0.288	0.000	0.000	0.000	0.000	0.000	0.000	0.360	0.000	0.396	0.000
5	0.000	0.000	0.658	0.559	0.000	0.036	0.000	0.000	0.333	0.991	0.342	0.000	0.081	0.072	0.000	0.000	0.883	0.441	0.225	0.009	0.045	0.811	0.000	0.991	0.009	0.000
6	0.000	0.000	0.649	0.802	0.018	0.054	0.045	0.000	0.279	0.991	0.622	0.000	0.036	0.162	0.000	0.000	0.892	0.198	0.162	0.081	0.297	0.991	0.000	0.991	0.000	0.000
7	0.108	0.514	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.739	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.838	0.135	0.252	0.505	0.000	0.243	0.874	0.324	0.099	0.000	0.811	0.009	0.000	0.018	0.315	0.811	0.991	0.000	0.018	0.144	0.099	0.279	0.135	0.000
9	0.667	0.162	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.568	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.099	0.000	0.378	0.712	0.000	0.658	0.126	0.027	0.000	0.000	0.243	0.000	0.000	0.027	0.009	0.054	0.162	0.000	0.000	0.000	0.333	0.009	0.468	0.000
11	0.000	0.153	0.000	0.000	0.000	0.000	0.306	0.000	0.000	0.000	0.036	0.676	0.000	0.162	0.018	0.000	0.009	0.000	0.000	0.189	0.027	0.000	0.000	0.000	0.000	0.000
12	0.000	0.000	0.000	0.000	0.000	0.090	0.000	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.153	0.000	0.000	0.000	0.000	0.000	0.000	0.279	0.000	0.171	0.000
13	0.000	0.000	0.766	0.081	0.099	0.234	0.000	0.162	0.757	0.252	0.027	0.000	0.423	0.000	0.000	0.000	0.279	0.793	0.991	0.000	0.000	0.117	0.018	0.162	0.045	0.000
14	0.000	0.000	0.054	0.000	0.270	0.676	0.000	0.631	0.000	0.000	0.000	0.000	0.081	0.000	0.000	0.288	0.000	0.000	0.027	0.000	0.000	0.000	0.892	0.000	0.559	0.000
15	0.000	0.000	0.018	0.000	0.189	0.685	0.000	0.775	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.369	0.000	0.000	0.000	0.000	0.000	0.000	0.991	0.000	0.586	0.000
16	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	0.000	0.000	0.000	0.000	0.000	0.081	0.000	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.171	0.000	0.000	0.000	0.000	0.000	0.000	0.126	0.000	0.207	0.000
18	0.000	0.000	0.000	0.000	0.090	0.595	0.000	0.387	0.000	0.000	0.000	0.000	0.036	0.000	0.000	0.207	0.000	0.000	0.000	0.000	0.000	0.000	0.847	0.000	0.459	0.000
19	0.000	0.000	0.000	0.000	0.018	0.018	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.793
20	0.000	0.000	0.640	0.649	0.000	0.054	0.018	0.000	0.324	0.991	0.441	0.000	0.009	0.027	0.000	0.000	0.820	0.387	0.198	0.018	0.126	0.739	0.000	0.991	0.000	0.000
21	0.000	0.000	0.072	0.252	0.000	0.000	0.829	0.000	0.000	0.153	0.405	0.306	0.000	0.838	0.000	0.000	0.315	0.009	0.000	0.865	0.604	0.126	0.000	0.162	0.000	0.000
22	0.000	0.000	0.036	0.000	0.523	0.991	0.000	0.739	0.000	0.000	0.000	0.000	0.162	0.000	0.000	0.450	0.000	0.027	0.108	0.000	0.000	0.000	0.991	0.009	0.766	0.000
23	0.000	0.000	0.171	0.784	0.000	0.009	0.351	0.000	0.045	0.324	0.541	0.090	0.009	0.586	0.000	0.000	0.468	0.009	0.009	0.432	0.739	0.351	0.000	0.297	0.000	0.000
24	0.000	0.000	0.991	0.333	0.009	0.153	0.000	0.063	0.649	0.685	0.099	0.000	0.099	0.000	0.000	0.000	0.459	0.667	0.450	0.000	0.045	0.396	0.009	0.568	0.000	0.000
25	0.000	0.000	0.775	0.315	0.432	0.685	0.036	0.450	0.811	0.351	0.081	0.009	0.910	0.045	0.000	0.000	0.550	0.757	0.991	0.000	0.135	0.279	0.261	0.486	0.586	0.000
26	*	0.036	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.207	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
27	0.050	*	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.000	0.063	0.000	0.000	0.315	0.000	0.000	0.000	0.000	0.009	0.000	0.000	0.000	0.000	0.000	0.000
28	0.599	0.382	*	0.405	0.081	0.333	0.054	0.099	0.991	0.757	0.261	0.000	0.450	0.036	0.000	0.000	0.649	0.991	0.730	0.045	0.108	0.477	0.009	0.703	0.063	0.000
29	0.485	0.262	-0.001	*	0.000	0.018	0.117	0.000	0.153	0.649	0.775	0.036	0.036	0.324	0.000	0.000	0.721	0.144	0.036	0.216	0.468	0.829	0.000	0.640	0.000	0.000
30	0.697	0.530	0.040	0.129	*	0.838	0.000	0.910	0.009	0.027	0.000	0.000	0.270	0.000	0.000	0.297	0.000	0.018	0.108	0.000	0.000	0.000	0.333	0.000	0.324	0.000
31	0.722	0.537	0.039	0.132	-0.011	*	0.000	0.991	0.180	0.144	0.009	0.000	0.505	0.000	0.000	0.225	0.081	0.090	0.324	0.000	0.000	0.009	0.649	0.027	0.991	0.009
32	0.290	0.110	0.105	0.027	0.249	0.246	*	0.000	0.000	0.018	0.099	0.541	0.000	0.640	0.000	0.000	0.108	0.000	0.000	0.856	0.396	0.036	0.000	0.027	0.000	0.000
33	0.745	0.582	0.058	0.168	-0.019	-0.024	0.282	*	0.036	0.009	0.000	0.000	0.297	0.000	0.000	0.514	0.000	0.072	0.144	0.000	0.000	0.000	0.694	0.000	0.991	0.000
34	0.609	0.434	-0.019	0.026	0.035	0.023	0.149	0.035	*	0.486	0.054	0.000	0.234	0.000	0.000	0.000	0.342	0.991	0.838	0.000	0.009	0.207	0.018	0.252	0.027	0.000
35	0.539	0.325	-0.019	-0.018	0.083	0.082	0.065	0.107	0.000	*	0.523	0.000	0.117	0.117	0.000	0.000	0.856	0.577	0.288	0.054	0.252	0.991	0.000	0.991	0.000	0.000
36	0.454	0.248	0.009	-0.020	0.131	0.125	0.025	0.151	0.037	-0.009	*	0.027	0.009	0.189	0.000	0.000	0.829	0.072	0.027	0.090	0.414	0.622	0.000	0.378	0.000	0.000
37	0.231	0.059	0.164	0.067	0.329	0.324	-0.005	0.373	0.219	0.117	0.064	*	0.000	0.306	0.000	0.000	0.018	0.000	0.000	0.414	0.081	0.000	0.000	0.000	0.000	0.000
38	0.660	0.494	0.000	0.074	0.011	-0.004	0.206	0.006	0.004	0.035	0.083	0.284	*	0.000	0.000	0.000	0.225	0.333	0.441	0.000	0.000	0.018	0.108	0.081	0.099	0.000
39	0.341	0.142	0.078	0.007	0.232	0.227	-0.011	0.270	0.121	0.041	0.008	0.003	0.182	*	0.000	0.000	0.162	0.000	0.009	0.820	0.595	0.108	0.000	0.072	0.000	0.000
40	0.004	0.002	0.504	0.382	0.628	0.648	0.199	0.681	0.533	0.443	0.358	0.139	0.590	0.242	*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
41	0.795	0.674	0.147	0.282	0.006	0.004	0.386	-0.004	0.097	0.207	0.262	0.491	0.047	0.388	0.749	*	0.000	0.000	0.000	0.000	0.000	0.000	0.342	0.000	0.207	0.000
42	0.521	0.295	-0.020	-0.025	0.088	0.089	0.048	0.122	0.002	-0.026	-0.022	0.094	0.039	0.025	0.418	0.232	*	0.270	0.234	0.126	0.225	0.955	0.000	0.820	0.000	0.000
43	0.610	0.423	-0.022	0.019	0.038	0.029	0.137	0.043	-0.010	-0.004	0.030	0.205	0.004	0.110	0.529	0.114	-0.004	*	0.847	0.000	0.027	0.189	0.000	0.333	0.009	0.000
44	0.632	0.450	-0.017	0.039	0.023	0.011	0.163	0.019	-0.008	0.008	0.046	0.234	-0.004	0.136	0.554	0.083	0.008	-0.010	*	0.000	0.009	0.117	0.063	0.243	0.036	0.000
45	0.315	0.125	0.092	0.017	0.244	0.238	-0.011	0.277	0.136	0.053	0.017	-0.002	0.196	-0.014	0.219	0.391	0.037	0.124	0.151	*	0.514	0.045	0.000	0.045	0.000	0.000
46	0.374	0.184	0.048	-0.006	0.184	0.176	0.001	0.205	0.082	0.019	-0.004	0.026	0.135	-0.010	0.282	0.306	0.007	0.073	0.096	-0.004	*	0.180	0.000	0.207	0.000	0.000
47	0.501	0.302	-0.010	-0.017	0.093	0.086	0.055	0.106	0.011	-0.017	-0.011	0.103	0.048	0.032	0.410	0.200	-0.024	0.007	0.020	0.044	0.013	*	0.000	0.991	0.000	0.000
48	0.780	0.632	0.113	0.235	0.006	-0.017	0.332	-0.009	0.063	0.163	0.210	0.432	0.025	0.329	0.724	0.003	0.188	0.078	0.052	0.332	0.252	0.154	*	0.000	0.784	0.000
49	0.535	0.326	-0.017	-0.017	0.085	0.081	0.066	0.106	0.002	-0.021	-0.007	0.118	0.037	0.042	0.440	0.202	-0.024	-0.002	0.011	0.054	0.020	-0.015	0.158	*	0.000	0.000
50	0.776	0.637	0.089	0.219	0.005	-0.020	0.334	-0.012	0.053	0.146	0.201	0.437	0.017	0.329	0.723	0.011	0.169	0.064	0.038	0.333	0.249	0.143	-0.009	0.143	*	0.000
51	0.623	0.501	0.136	0.195	0.071	0.101	0.287	0.090	0.153	0.169	0.206	0.352	0.130	0.279	0.568	0.113	0.164	0.148	0.138	0.293	0.255	0.183	0.124	0.172	0.137	*
52	0.497	0.351	0.098	0.117	0.096	0.115	0.176	0.126	0.147	0.114	0.132	0.219	0.148	0.170	0.422	0.196	0.098	0.128	0.133	0.180	0.170	0.125	0.162	0.120	0.193	0.034

Populations: 1: Anatolian Turks, 2: Armenians, 3: Ashkenazi Jews, 4: Azeri, 5: Belarusians, 6: Catalans, 7: English, 8: Erzja, 9: Frisians, 10: Georgians, 11: Germans, 12: Greek Cypriots, 13: Greeks, 14: Han Chinese, 15: Iranians, 16: Irish, 17: Japanese, 18: Khalka, 19: Kuwaiti, 20: Macedonians, 21: Moksha, 22: Mongolians, 23: Nepalese, 24: North Indians, 25: Northern Yakuts, 26: Norwegians, 27: other C/NW Europeans, 28: other C/South Asians, 29: other E/SE Europeans, 30: other ME, 31: Pashtuns/Afghans, 32: Portuguese, 33: Roma, 34: Romanians, 35: Russians Perm, 36: Sami, 37: Sorbs, 38: South Indians, 39: Spanish, 40: Swedes, 41: Syrians, 42: Tadjiks Afghanistan, 43: Tharu, 44: Tyroleans Bozen, 45: Tyroleans Gadertal, 46: Tyroleans Vinschgau, 47: Ukrainians, 48: Uzbeks, 49: Uzbeks Afghanistan, 50: Yakuts, 51: Yemeni Hadramaut, 52: Yemeni Sena

Appendix E: Summary of the functional studies conducted during this thesis. Abbreviations: Comp: Competitior, AB: Antibody. +: has an effect, bold crosses: strong effect, - no effect, (+) weak effect in competitions or slight inhibition in supershift.

		-14028*T	-14028*C	-14011*C	-14011*T	-14009*G	-13779*G	-13779*C
Transfections								
	2 days		+		+	+		+
	9 days		-		+	+		-
EMSAs								
Cdx-2	Comp	**+**	+	+	+	+	(+)	(+)
	AB	**+**	-	(+)	(+)	(+)	-	-
Oct-1	Comp	+	(+)	+	+	+	(+)	-
	AB	-	-	+	**+**	+		
HNF-1α	Comp	**+**	+	**+**	**+**	+	(+)	-
	AB	-	-	(+)	(+)	(+)		
HNF-4α	Comp	-	**+**				**+**	**+**
	AB	-	+				+	+
GATA 3, 4	Comp			+	+			
c-Ets-1,	Comp			-		**+**		
Aff_ELF	Comp			+		**-**		
Aff_ELK-1	Comp			+		**+**		
Aff_ETS (1)	Comp			(+)		**+**		
Ets/Tel2, Aff_c-Ets-1	Comp			-		-		
Ets-1, Ets 1/2	AB			-		-		
NF_kappaB, Pax, Pax4-8, Pbx	Comp			(+)		(+)		
LEF-1,	Comp						-	-
GKLF-1, AML-1	Comp						(+)	(+)
Bioinf. Pred.								
TRANSFAC		CdxA	HNF4a	Oct, CdxA	Oct, CdxA, GATA4	Oct, CdxA, c-Ets-1, Ets	-	HNF4
MatInspector		Cdx2, others		Oct1, HNF1	GATA3, Oct1, HNF1	ELF5, Oct1, HNFP	-	RU49, LEF1, GKLF
TFSEARCH		CdxA, others	HNF4a	Oct1, CdxA, others	Oct1, CdxA, others	ELK1, ELK4, c-Ets, Oct1, CdxA, others	-	AML-1a

Appendix F: Pairwise linkage disequilibrium *D'* across the 80kb haplotype region of the Middle Eastern groups tested. Monomorphic markers were excluded from analysis. Statistically significant values after Bonferroni correction for multiple testing for the respective amount of populations per group are shaded in green.

Middle East, Afro-Asiatic

	-30210	-30203	-30196	-30182	-30160	-30071/70	-29949	-14011	-14010	-14009	-13915	-13913	-13910	-13907	-13806	-13779	-13744	-13730	-13603	-13495	-958	-943	-942	-931	-875	-678	666
-30203	1																										
-30196	1	1																									
-30182	0.98	1	1																								
-30160	1	1	1	1																							
-30071/70	1	1	1	1	1																						
-29949	1	1	1	1	1	1																					
-14011	1	1	1	1	1	1	1																				
-14010	1	1	1	1	1	1	1	1																			
-14009	1	1	1	1	1	1	1	1	1																		
-13915	1	1	1	1	1	1	1	1	1	1																	
-13913	1	1	1	1	1	1	1	1	1	1	1																
-13910	1	1	1	1	1	1	1	1	1	1	1	1															
-13907	1	1	1	1	1	1	1	1	1	1	1	1	1														
-13806	1	1	1	1	1	1	1	1	1	1	1	1	1	1													
-13779	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1												
-13744	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1											
-13730	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1										
-13603	0.70	1	1	0.77	1	1	1	1	1	1	1	1	1	1	1	1	1	1									
-13495	0.95	1	1	0.77	1	1	0.78	1	1	1	1	1	1	1	1	1	1	1	1								
-958	0.96	1	1	0.97	1	1	1	1	1	1	0.97	1	1	1	1	1	1	1	1	1							
-943	1	1	1	0.88	1	1	0.12	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
-942	0.86	1	1	0.88	1	1	0.12	1	1	1	1	1	1	1	1	1	1	1	1	1	0.85	1					
-931	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1				
-875	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
-678	1	1	1	1	1	1	1	1	1	1	0.97	1	1	1	1	1	0.69	1	1	0.96	1	1	1	1	1		
666	0.97	1	1	0.97	1	1	0.57	1	1	1	1	1	1	1	1	1	1	1	1	0.98	0.99	1	1	1	1	1	
5579	0.81	1	1	0.64	1	1	1	1	1	1	0.97	1	1	1	1	1	1	1	1	0.92	0.86	0.10	0.08	1	1	0.77	0.75

Africa, Niger-Congo

	-30210	-30182	-30071/70	-29949	-14010	-13910	-13800	-13730	-13603	-13495	-943/42	-678	666
-30182	1												
-30071/70	1	1											
-29949	1	1	1										
-14010	1	1	1	1									
-13910	1	1	1	1	1								
-13800	1	1	1	1	1	1							
-13730	1	1	1	1	1	1	1						
-13603	1	1	1	1	1	1	1	1					
-13495	1	1	1	1	1	1	1	1	1				
-943/42	0.94	1	1	0.43	1	1	1	1	1	0.79			
-678	1	1	1	1	1	1	1	1	1	1	1		
666	1	0.96	1	0.14	1	1	1	1	1	1	0.97	0.96	
5579	1	0.91	1	0.73	1	1	0.69	1	0.92	1	0.16	0.92	0.18

Africa, Nilo-Saharan

	-30210	-30196	-30182	-30160	-29949	-13800	-13730	-13495	-958	-943/42	-678	666
-30196	1											
-30182	1	1										
-30160	1	1	1									
-29949	1	1	1	1								
-13800	1	1	1	1	1							
-13730	1	1	1	1	1	1						
-13495	0.07	1	1	1	1	1	1					
-958	1	1	1	1	1	1	1	1				
-943/42	1	1	1	1	1	1	1	1	1			
-678	1	1	1	1	1	1	1	1	1	1		
666	0.53	1	1	1	1	1	1	0.30	1	1	1	
5579	0.22	1	1	1	1	1	1	0.65	1	0	1	0.42

Middle East, Indo-European

	-30210	-30182	-30160	-29949	-13910	-13495	-958	-875	-678	666
-30182	1									
-30160	1	1								
-29949	1	1	1							
-13910	1	1	1	1						
-13495	1	0.34	1	1	1					
-958	1	1	1	1	1	1				
-875	1	1	1	1	1	1	1			
-678	1	1	1	1	1	1	1	1		
666	1	1	1	1	1	1	1	1	1	
5579	0.91	0.21	1	1	1	1	0.91	1	0.60	0.91

Europe/Asia, Indo-European

	-30210	-30182	-30160	-29949	-14028	-14011	-13910	-13495	-958	-943	-942	-875	-678	666
-30182	1													
-30160	1	1												
-29949	1	1	1											
-14028	1	1	1	1										
-14011	1	1	1	1	1									
-13910	1	1	1	1	1	1								
-13495	1	0.88	0.48	1	1	1	1							
-958	0.98	1	1	1	1	1	1	1						
-943	1	1	1	1	1	1	1	1	0.60					
-942	0.65	1	1	1	1	1	1	1	0.29	1				
-875	1	1	1	1	1	1	1	1	1	1	1			
-678	0.87	0.93	1	1	1	1	1	1	1	1	1	1		
666	0.97	1	1	1	1	1	1	1	1	1	1	1	1	
5579	0.97	0.58	1	0.79	1	1	1	1	0.97	1	1	1	0.70	0.97

Middle East/Asia, Altaic

	-30210	-30182	-30160	-29949	-14011	-13910	-13495	-958	-943/42	-875	-678	666
-30182	1											
-30160	1	1										
-29949	1	1	1									
-14011	1	1	1	1								
-13910	1	1	1	1	1							
-13495	1	1	1	0.47	1	1						
-958	1	1	1	1	1	1	1					
-943/42	1	1	1	1	1	1	1	1				
-875	1	1	1	1	1	1	1	1	1			
-678	0.80	1	1	1	1	1	1	1	1	1		
666	0.87	1	1	1	1	1	1	1	1	1	1	
5579	0.76	0.49	1	1	1	1	1	0.83	1	1	0.52	0.87

Appendix G: Publications

(Jones et al. 2013)
(Gallego Romero et al. 2012)
(Jensen et al. 2011)
(Gerbault et al. 2011)