*Am. J. Hum. Genet. 68:160–172, 2001*

# Lactase Haplotype Diversity in the Old World

Edward J. Hollox,[1] Mark Poulter,[1] Marek Zvarik,[2] Vladimir Ferak,[2] Amanda Krause,[3] Trefor Jenkins,[3] Nilmani Saha,[4] Andrew I. Kozlov,[5] and Dallas M. Swallow[1]

[1]MRC Human Biochemical Genetics Unit, Galton Laboratory, Department of Biology, University College London, London; [2]Department of Molecular Biology, Comenius University, Bratislava, Slovakia; [3]Department of Human Genetics, South African Institute for Medical Research and University of Witwatersrand, Johannesburg; [4]Department of Pediatrics, National University of Singapore, Singapore; and [5]ArctAn-C Innovative Research Laboratory, Moscow

**Lactase persistence, the genetic trait in which intestinal lactase activity persists at childhood levels into adulthood, varies in frequency in different human populations, being most frequent in northern Europeans and certain African and Arabian nomadic tribes, who have a history of drinking fresh milk. Selection is likely to have played an important role in establishing these different frequencies since the development of agricultural pastoralism ~9,000 years ago. We have previously shown that the element responsible for the lactase persistence/nonpersistence polymorphism in humans is *cis*-acting to the lactase gene and that lactase persistence is associated, in Europeans, with the most common 70-kb lactase haplotype, A. We report here a study of the 11-site haplotype in 1,338 chromosomes from 11 populations that differ in lactase persistence frequency. Our data show that haplotype diversity was generated both by point mutations and recombinations. The four globally common haplotypes (A, B, C, and U) are not closely related and have different distributions; the A haplotype is at high frequencies only in northern Europeans, where lactase persistence is common; and the U haplotype is virtually absent from Indo-European populations. Much more diversity is seen in sub-Saharan Africans than in non-Africans, consistent with an "Out of Africa" model for peopling of the Old World. Analysis of recent recombinant haplotypes by allele-specific PCR, along with deduction of the root haplotype from chimpanzee sequence, allowed construction of a haplotype network that assisted in evaluation of the relative roles of drift and selection in establishing the haplotype frequencies in the different populations. We suggest that genetic drift was important in shaping the general pattern of non-African haplotype diversity, with recent directional selection in northern Europeans for the haplotype associated with lactase persistence.**

## Introduction

Analysis of the distribution and nature of DNA variation in human populations is important for isolation of risk loci in multifactorial diseases and illumination of the evolutionary history of the species (Weiss 1993; Rogers and Jorde 1995). Most work has concentrated on nonrecombining DNA elements, such as mtDNA and Y-chromosomal DNA (Vigilant et al. 1991; Hammer 1995), and small autosomal regions with no significant recombination (Wainscoat et al. 1986; Tishkoff et al. 1996; Bergstrom et al. 1998; Jaruzelska et al. 1999), but, recently, large regions containing autosomal genes usually associated with diseases have been analyzed in a variety of populations (Clark et al. 1998; Kidd et al. 1998; Tishkoff et al. 1998; Yip et al. 1999). Here, we analyze haplotypes across 70 kb of the lactase gene (*LCT*), a region small enough for observation of linkage disequilibrium across the whole region (Harvey et al. 1995) yet large enough for observation of evidence of historical recombination events.

Population movement influences allelic frequencies and patterns of disequilibrium, but patterns of variation at certain loci can also be affected—directly, by selection, or indirectly, by selective sweep caused by a linked locus—and this has been investigated at several loci (Hill et al. 1992; Jobling et al. 1998). For example, analysis of diversity shows evidence of balancing selection by infectious pathogens at the major histocompatability complex (MIM 142857, MIM 604305), and recent work involving complete sequencing of genes in different individuals has shown evidence of selection by ultraviolet irradiation at the MC1R locus involved in melanin skin phenotype (Harding et al. 2000) and at the Duffy blood group locus by *Plasmodium vivax* malaria (Hamblin and Di Rienzo 2000). Phenotypic variation of digestive enzymes is a likely target of selection, and the classic case of this is the lactase-persistence/nonpersistence polymorphism. This is a genetic polymorphism controlled by an as-yet-unidentified *cis*-act-

Received August 10, 2000; accepted for publication October 27, 2000; electronically published November 28, 2000.

Address for correspondence and reprints: Dr. Dallas M. Swallow, MRC HBGU, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: dswallow@hgmp.mrc.ac.uk

© 2001 by The American Society of Human Genetics. All rights reserved.
0002-9297/2001/6801-0016$02.00

ing element (MIM 223100; Wang et al. 1995, 1998) within a 300-kb region containing the lactase gene (M. Poulter, E. Hollox, D. M. Swallow, unpublished data). The lactase-persistence allele (*LCT*P*), which causes high childhood levels of lactase to persist into adulthood rather than to decline, varies in frequency between populations. In northern Europeans and certain nomadic tribes of Africa and Asia (such as Bedouin and Beja), the frequency of *LCT*P* is high, whereas, in other populations, it is much lower. These populations also have a history of drinking fresh milk, which is rich in lactose, the disaccharide digested by lactase. Significant quantities of fresh milk were available only after the domestication of animals, which started ~9,000 years ago in the Fertile Crescent area of the Near East (Simoons 1971; Dudd and Evershed 1998). In some populations, strong selection may have operated after this time favoring those who could drink fresh milk (lactase persistent), and this may be an example of gene-culture coevolution (Feldman and Cavalli-Sforza 1989; Durham 1991*a*). Other populations have developed cultural adaptations to milk drinking which involve fermenting fresh milk to produce products such as cheese and yogurt, which have a lower level of lactose and hence are less likely to produce symptoms in nonpersistent individuals (Kolars et al. 1984). The selective factors involved in favouring fresh-milk drinkers are varied and are still debated but may include improved nutrition, prevention of dehydration, and improved calcium absorption (Holden and Mace 1997); for reviews, see (Flatz 1987; Durham 1991*b;* Swallow and Hollox 2000). Selection in favor of lactase nonpersistence in areas with endemic malaria has been proposed (Anderson and Vullo 1994) but has not been substantiated (Meloni et al. 1998).

Polymorphic sites across the 70-kb lactase gene have previously been described by us and other workers (Boll et al. 1991; Lloyd et al. 1992; Harvey et al. 1995), and, although none are causative of the phenotypic polymorphism, one in a conserved 5′ region appears to affect protein binding to the DNA (Hollox et al. 1999). We have previously described haplotypes of these polymorphic sites by analysis of families and cohorts of European individuals (Harvey et al. 1995, 1998), and these data suggested association of *LCT*P* with one haplotype (A) and a cline of this haplotype across Europe, mirroring the cline in *LCT*P* allele frequency across Europe (Harvey et al. 1998). Our hypothesis is that *LCT*P* is a mutation that occurred on the A haplotype and was subject to recent strong directional selection in northern Europeans after the development of pastoralism ~9,000 years ago.

The objective of the present study was to examine the frequencies of these haplotypes worldwide, in relation to the known distribution of the *LCT*P* allele. We tested 669 individuals from a total of 11 populations and present these data together with *LCT*P* frequencies established in other studies. We also construct a phylogeny for the *LCT* haplotypes observed, which provides insight into the likely relative roles of selection and drift in establishing present-day haplotype frequencies.

## Methods

### *Samples*

Samples from the following human populations were tested: two sub-Saharan African populations (San from northwest Namibia and Bantu-speaking South Africans living in and around Johannesburg), five Indo-European language–speaking populations (northern Europeans, mainly from the United Kingdom; southern Europeans, mainly from Italy; Russians from Perm; Roma from Slovakia; and northern Indians), one Dravidian-speaking population (southern Indians from the Tamil Nadu and Kerala regions of India and Sri Lanka resident in Singapore), and one Finno-Ugric–speaking population (both Moksha and Erji groups of Mordavians from Mordavia in Russia). Four populations from East Asia were also tested: Malays and Chinese living in Singapore, Japanese from Osaka, and Papua New Guineans from both highland and lowland groups. All human DNA samples were collected with informed consent under the appropriate ethical procedures for the country of origin.

### *DNA from Blood*

DNA was prepared from whole blood by standard techniques, except for that from the Mordavian and Russian samples, which was extracted, by means of a simple alkaline lysis method, from blood dried on filter paper (Rudbeck and Dissing 1998).

### *Polymorphism Testing*

The polymorphisms were analyzed by denaturing gradient gel electrophoresis (DGGE), single-strand conformation analysis (SSCA), restriction-enzyme digest, and simple electrophoresis, as described elsewhere (Harvey et al. 1995; Hollox et al. 1999). The sites are all numbered with respect either to lactase cDNA or to the number of nucleotides from the start or end of transcription (Boll et al. 1991; Hollox et al. 1999). One previously unpublished polymorphic site (C458intT) was also tested in the population samples. This is within the F2A/F2S PCR product and is detected by the same SSCA method used for analysis of G666A. This new polymorphism is in intron 1 and is described as C458intT, because it is at base 458 of the genomic sequence accession no. M61835.

The likely original allelic state for each polymorphism was deduced by analysis of the equivalent position in

chimpanzee sequence, generated by amplification and direct sequencing of DNA from five unrelated animals using the primers designed on human sequence. As part of this process, 1 kb of contiguous chimpanzee sequence 5′ to the lactase gene was generated and has been submitted to EMBL/Genbank (AF282888).

T5579C was tested by using *Msp*1 digestion of a PCR product with PCR primers which included a nonpolymorphic *Msp*1 site (position 5779) as an internal control for digestion. Both T5579C and TG6236/7ΔΔ were also typed by an allele-specific PCR (AS-PCR) method, using primers specific for each allele, at both sites, that allowed phase of both alleles to be determined in individuals heterozygous for both sites. The nucleotide changes of representative samples from each population were confirmed by direct sequencing of the appropriate PCR product, and in some cases cloned products, as described elsewhere (Hollox et al. 1999). In a small subset of population samples, two further polymorphisms 3′ to exon 17 (CATT+225ΔΔΔΔ and C+658T, numbering relative to the end of exon 17) were analyzed by allele-specific PCR primers. The sequence of all AS-PCR primers can be found at our Web site.

### *Linkage Disequilibrium Analysis and Haplotype Estimation*

Pairwise linkage disequilibrium between all combinations of the six loci that we had tested previously (Harvey et al. 1995) was estimated by constructing 3 × 3 phenotype contingency tables (each locus having three phenotypes—A, AG, and G—or interpreted genotypes—AA, AG, and GG) and was analyzed using a maximum-likelihood method employed by the computer program ASSOCIATE (Ott 1991; Terwilliger and Ott 1994) available from the UK Human Genome Mapping Project Resource Centre. D′ was calculated using $D' = |D_{ij}/D_{\max}|$, where $D_{ij} = p_{ij} - p_i p_j$ and $D_{\max} = \min[p_i p_j,(1 - p_i)(1 - p_j)]$ if $D_{ij} < 0$ or $\min[(1 - p_i)p_j,p_i(1 - p_j)]$ if $D_{ij} > 0$. $p_i$ and $p_j$ are the frequencies of alleles $i$ and $j$, and $p_{ij}$ is the frequency of the haplotype having $i$ at the first locus and $j$ at the second locus. The ASSOCIATE program also calculates deviations of observed phenotype frequencies from those expected from a population in Hardy-Weinberg equilibrium.

Haplotypes across all 11 sites were inferred as another way of assessing linkage disequilibrium. This was made simpler by use of DGGE, SSCA, and AS-PCR, which allowed three partial, confidently assigned haplotypes to be determined and treated as multiallelic single loci in the complete haplotype estimation procedures. Haplotypes of sites C-958T, A-946G, C-942G, TC-942/3DD, and G-875A can be resolved by DGGE (Hollox et al. 1999). SSCA of the PCR product containing G666A reveals the allele C458intT in association with 666A as a novel band. AS-PCR of the exon 17 polymorphisms (T5579C and TG6236/7DD) allowed these to be haplotyped and treated as a one-locus, four-allele system.

Complete 70-kb haplotypes were deduced by two methods in all populations: haplotype counting and maximum-likelihood analysis. Haplotype counting involved recording the haplotypes of totally homozygous individuals and those heterozygous for only one site. In individuals heterozygous for more than one site, all possible haplotypes were considered, but the haplotypes that occur at the highest frequencies in the particular population were preferentially assumed. Maximum-likelihood estimates of haplotypes were carried out on all population samples using the computer program EH (Terwilliger and Ott 1994), which is available from the Web site of the United Kingdom Human Genome Mapping Project Resource Centre.

In all but two populations, haplotype frequencies deduced by the two methods were very similar. However, in Bantu and San, the high levels of heterozygosity at more than one site made haplotype counting more error prone. All haplotype frequencies shown are the maximum-likelihood estimates when EH is used.

Standard error of allele frequencies was calculated as $s = \sqrt{pq/n}$, where $n$ is the number of chromosomes tested. Significance of pairwise association and overall allelic association of the region was calculated using a $\chi^2$ test of the log likelihood statistics produced by ASSOCIATE and EH, respectively. All allelic associations detected by EH were significant ($P \leqslant .001$), except for the San population.

### *Heterozygosity*

Heterozygosity was calculated using the following statistic:

$$H = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_{i^2}\right),$$

and its corresponding variance was given by

$$V(H) = \frac{2}{n(n-1)} \times \left\{2(n-2)\left[\sum_{i=1}^{k} p_i^3 - \left(\sum_{i=1}^{k} p_i^2\right)^2\right] + \sum_{i=1}^{k} p_i^2 - \left(\sum_{i=1}^{k} p_i^2\right)^2\right\},$$

where $n$ is the number of chromosomes tested, $k$ the number of haplotypes, and $p_i$ the frequency of the $i$th haplotype (Nei 1987).

## Sequencing

Direct sequencing of PCR products was performed either by use of the ThermoSequenase radiolabeled terminator kit (Amersham Pharmacia Biotech) and standard electrophoretic analysis or by use of the BigDye cycle sequencing kit (PE Biosystems) and electrophoresis on an Applied Biosystems 373A automatic sequencer.

## Results

The positions of the polymorphic sites analyzed are shown in figure 1, and the allele frequencies in each population are shown in table 1. All sites have been described elsewhere (Harvey et al. 1995; Hollox et al. 1999)—except C458intT, where the T allele is present at a significant frequency only in the San, although it was observed in Bantu-speaking South Africans (table 1) and several British of African or Afro-Caribbean ancestry (data not shown).

Pairwise linkage disequilibrium (D′) between alleles was determined for all populations except the San, who were not included because of the small sample size. These data were compared with our previously published linkage disequilibria calculated from the CEPH families and are represented in figure 2. The closely spaced loci at the 5′ end of lactase show either complete or virtually complete disequilibrium with each other in all populations. This is also the case with T5579C and TG6236/7ΔΔ loci, both in exon 17.

Haplotypes of the 11 polymorphic sites were deduced from diploid data from each population by maximum-likelihood statistics, and the allelic compositions of all haplotypes deduced is shown in figure 3, together with their overall frequency in all the samples used in this study. Only four haplotypes—A, B, C, and U—are present at a frequency of $\geq .05$. Haplotypes that are present at a frequency of $\leq .01$ in only one population are indicated by a hash mark.

The haplotype frequencies in the individual populations are shown in table 2, together with the haplotype heterozygosity. Haplotype A is the most common in all populations except the sub-Saharan African ones, and is particularly frequent (.86) in northern Europeans. Haplotype B, whose distribution reflects the component −958T allele, is at highest frequency in Papua New Guineans (.36) and is present in all populations except Bantu-speaking South Africans. Haplotype C is also present in all populations but is rare in northern Europeans and San. The frequency of the U haplotype closely reflects its component allele −942/943ΔΔ—except in Africans, in whom other haplotypes carry this allele—and, like its component allele, is only present in 2 of 384 individuals from Indo-European populations. The non-African populations show heterozygosity between .65 and .78, with the exception of northern Europeans, who show a heterozygosity of .26, reflecting the very high frequency of the A haplotype (other non-Africans ≠ northern Europeans, $P < .001$, Student's *t*-test). The San and the Bantu-speaking South Africans are the most diverse, with heterozygosities of .91 and .87 respectively, which are significantly higher than those of non-Africans (other non-Africans ≠ Africans, $P < .01$, Student's *t*-test).

The four common global haplotypes, A, B, C, and U, differ from each other at three or more sites (fig. 3), so the mutational relationship between them is not immediately clear. A coalescent approach for deducing the relationship between haplotypes, such as that used by Harding et al. (1997), was not appropriate, since this approach assumes no recombination—an assumption which cannot be made for this system, as will be demonstrated. Also, a reduced median approach to construct a network using the RM algorithm (Bandelt et al. 1995) was not appropriate, because this method assumes that common haplotypes are ancestral to rarer haplotypes, an assumption broken if selection or drift acts on these frequencies. An alternative approach was

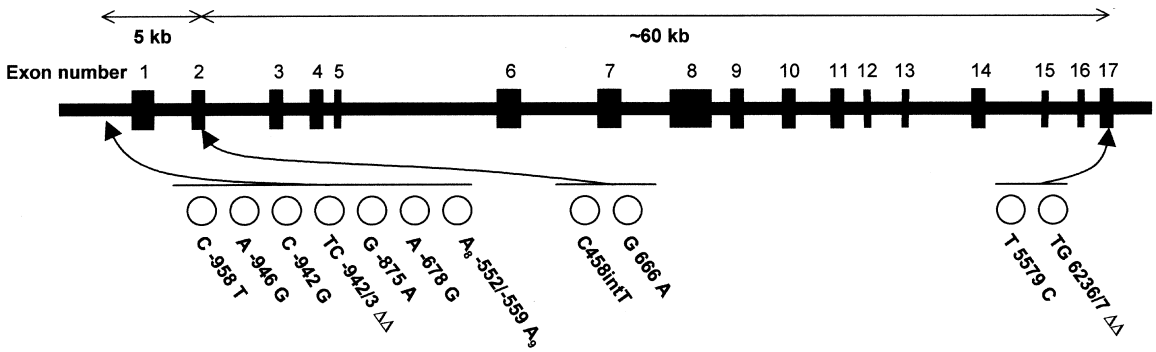


**Figure 1** *LCT* and polymorphic sites forming the lactase haplotype. Exons are shown as black bars. The series of circles underneath represent the polymorphic sites that form the haplotype.

## Table 1

**Frequencies of Alleles in Different Population Samples**

| Polymorphic Site | Frequency in Sample | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Northern European ($n = 104$) | Southern European ($n = 106$) | Northern Indian ($n = 156$) | Southern Indian ($n = 84$) | Malay ($n = 200$) | Chinese ($n = 102$) | Japanese ($n = 82$) | Papuan ($n = 72$) | Roma ($n = 170$) | Mordavian ($n = 68$) | Russian ($n = 80$) | Bantu ($n = 72$) | San ($n = 30$) |
| −958T | .08 ±.03 | .40 ±.05 | .28 ±.04 | .23 ±.05 | .17 ±.03 | .20 ±.04 | ***.11 ±.03*** | ***.42 ±.06*** | .35 ±.04 | .28 ±.05 | .27 ±.05 | ***0*** | ***.03 ±.03*** |
| −946G | 0 | 0 | 0 | 0 | 0 | 0 | ***0*** | ***0*** | 0 | 0 | 0 | ***.01 ±.01*** | ***.03 ±.03*** |
| −942G | 0 | 0 | 0 | 0 | .01 ±.01 | .03 ±.02 | ***.01 ±.01*** | ***0*** | 0 | 0 | 0 | ***0*** | ***0*** |
| −942/3DD | 0 | 0 | 0 | 0 | .10 ±.02 | .20 ±.04 | ***.28 ±.05*** | ***.04 ±.02*** | 0 | .04 ±.02 | .03 ±.02 | ***.28 ±.05*** | ***.57 ±.09*** |
| −875A | .03 ±.02 | .06 ±.02 | 0 | 0 | 0 | 0 | ***0*** | ***0*** | .01 ±.01 | .01 ±.01 | .01 ±.01 | ***0*** | ***0*** |
| −678G | .04 ±.02 | .12 ±.03 | .25 ±.03 | .32 ±.05 | .23 ±.03 | .12 ±.03 | .19 ±.04 | .17 ±.04 | .13 ±.03 | .13 ±.04 | .20 ±.04 | .42 ±.06 | .03 ±.03 |
| −552 to −559 $A_9$ | .89 ±.03 | .37 ±.05 | .44 ±.04 | .45 ±.05 | .61 ±.03 | .68 ±.05 | .68 ±.05 | .42 ±.06 | .49 ±.04 | .59 ±.06 | .50 ±.06 | .43 ±.05 | .70 ±.08 |
| 458intT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .03 ±.02 | .13 ±.06 |
| 666A | .08 ±.03 | .39 ±.05 | .29 ±.04 | .21 ±.04 | .27 ±.03 | .37 ±.05 | .41 ±.05 | .46 ±.06 | .34 ±.04 | .33 ±.06 | .30 ±.05 | .36 ±.06 | .80 ±.07 |
| 5579C | .90 ±.03 | .47 ±.05 | .47 ±.04 | .44 ±.05 | .52 ±.04 | .50 ±.05 | .40 ±.05 | .29 ±.05 | .53 ±.04 | .56 ±.06 | .50 ±.06 | .19 ±.05 | .23 ±.08 |
| 6236/7DD | .08 ±.03 | .38 ±.05 | .25 ±.03 | .19 ±.04 | .17 ±.03 | .15 ±.04 | .16 ±.04 | .36 ±.06 | .35 ±.04 | .25 ±.05 | .25 ±.05 | .19 ±.05 | .23 ±.08 |

Note.—The standard error, *s*, of each allele frequency was estimated by assumption that the allele frequencies were binomially distributed (data in bold italics from Hollox et al. [1999]; northern and southern European data from Harvey et al. [1998]). Underlined frequencies show that alleles were not in Hardy-Weinberg proportions ($\chi^2$ tests, $P < .05$, not corrected for multiple observations). $n$ = number of chromosomes.

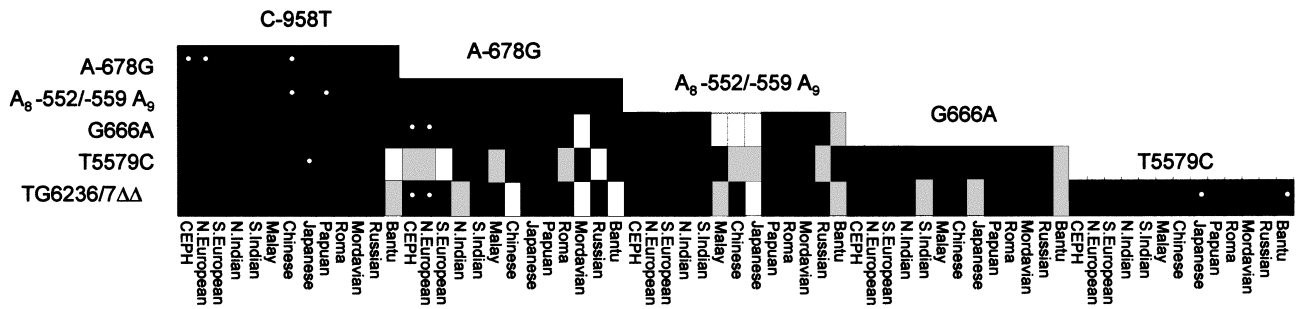


**Figure 2** Pairwise linkage disequilibrium analysis across the lactase gene. Shown are 195 pairwise D′ values for all possible combinations of six polymorphic loci in 13 populations. D′ values <0.5 are shown as a white boxes, D′ values between 0.5 and 0.8 are shown as gray boxes, and D′ values >0.8 are shown as black boxes. D′ values >0.8 differ significantly from D′ = 0 ($\chi^2$ test, $P < .05$, no correction for multiple observations) except in those boxes labeled with a white spot, in which statistical significance is not reached because of the low the numbers of the rare allele.

used to construct a meaningful relationship between the haplotypes that did not rely heavily on assumptions based on modern haplotype frequencies, which have been affected by selection and genetic drift. The procedure involved connecting each haplotype that differs by one position, and simplifying the resulting complex network by stepwise removal of haplotypes that fell into certain categories.

Haplotypes that are likely to be recent recombinations of the four common haplotypes were first identified and removed from the network. These were identified by analysis of the nature of the two polymorphic sites 3′ to exon 17. Allele-specific PCR on samples from different populations containing the common haplotypes allowed determination of the allelic state and phase of these polymorphisms. Of the two sites tested, CATT+225 ΔΔΔΔ was in complete disequilibrium with the haplotype in 123 out of 124 chromosomes carrying the four common haplotypes (A, B, C, and U). At the C+658T site, 108 of the 124 chromosomes carried the most frequent allele for that haplotype. Eleven haplotypes from the 33 further chromosomes tested can be inferred to be likely simple recombinants of the four common haplotypes (fig. 4) and were removed.

The next step was to remove rare haplotypes (≤.01 in any population, labeled as "#" in fig. 3) which could be artifacts of maximum-likelihood analysis. The resulting network is shown in figure *5a* and links the common haplotypes with the root haplotype, which is deduced by analysis, in five unrelated chimpanzees, of the nucleotides present at the 11 sites that are polymorphic in humans. This network suggests a direction for the evolution of at least some of the haplotypes. However, even after simplification, it is clear that at least one recombination must have been involved in the evolution of the common haplotypes, unless mutations recurred at seven of the sites.

Graphical networks that include haplotype-frequency information are shown for northern Europeans, other non-Africans, and Africans (figs. 5b–5d). The African network shows all the predicted linking haplotypes, with only haplotype C at a reasonably high frequency. The networked haplotypes represent 79% of the total African haplotype diversity. In contrast, the A, B, and C haplotypes are dominant in the non-African networks, whereas five of the networked haplotypes—two of which are direct precursors to the observed haplotypes—are not observed.

## Discussion

Lactase persistence/nonpersistence is an example of a polymorphic human trait for which allele frequencies clearly have been affected by selection. Although the molecular basis of this phenotypic polymorphism is not known, lactase persistence is associated with one 70-kb haplotype in Europeans, and the locus is *cis*-acting (Wang et al. 1995). Thus, this large haplotype would show evidence of a selective sweep if selection were acting on a locus within the region of linkage disequilibrium. Although still debated, the nature of the selective advantage of lactase persistence is clearer than many other postulated examples of natural selection in humans (Holden and Mace 1997; Swallow and Hollox 2000). The consensus view is that selection has only acted since the domestication of mammals 9,000 years ago, which, in relation to the history of the species, is very recent. Thus, interpretations of patterns of haplotype diversity can be apportioned into that which is attributed to recent events, which affect a single population or a subset of populations and probably generated by selection for lactase persistence; and older events affecting a large subset of populations and probably generated by drift.

## Table 2

**Frequencies of Haplotypes in Different Population Samples**

| Haplotype | Frequency in Sample[a] | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Northern European ($n$ = 104) | Southern European ($n$ = 106) | Northern Indian ($n$ = 156) | Southern Indian ($n$ = 84) | Malay ($n$ = 200) | Chinese ($n$ = 102) | Japanese ($n$ = 82) | Papuan ($n$ = 72) | Roma ($n$ = 170) | Mordavian ($n$ = 68) | Russian ($n$ = 80) | Bantu ($n$ = 72) | San ($n$ = 30) |
| Persistence allele[b] | .75 | .26 | .48 | .18 | … | .06 | .10 | .05 | .25 | .28 | .37 | .13 | .03 |
| A | .86 | .36 | .44 | .43 | .49 | .47 | .37 | .29 | .49 | .56 | .44 | .10 | .06 |
| B | .06 | .32 | .22 | .19 | .13 | .10 | .10 | .36 | .30 | .20 | .21 | … | .03 |
| C | .03 | .12 | .23 | .31 | .18 | .09 | .15 | .18 | .09 | .07 | .15 | .31 | .03 |
| D | .03 | .03 | … | … | … | … | … | … | .01 | .02 | .01 | … | … |
| E | .02 | .06 | .01 | … | .02 | .02 | .02 | … | .02 | .01 | .05 | … | … |
| F | … | … | … | … | … | … | … | … | … | … | … | … | … |
| G | … | .01 | .04 | .01 | .02 | .06 | … | .06 | .02 | .03 | .04 | … | … |
| H | … | … | .01 | … | … | … | .01 | … | … | … | .01 | … | .04 |
| I | … | .02 | .01 | .01 | … | … | .02 | … | … | … | .01 | … | … |
| J | … | .03 | … | … | .01 | .01 | .01 | … | .02 | … | … | … | … |
| K | … | … | … | .01 | … | … | … | … | .01 | … | .03 | .08 | … |
| L | … | .01 | … | … | … | … | … | … | … | … | … | … | … |
| M | … | .01 | .01 | … | .01 | .01 | … | … | .01 | .02 | … | .05 | … |
| N | … | … | … | … | … | .02 | .01 | … | … | … | … | … | … |
| O | … | … | … | … | … | … | … | … | … | … | … | .06 | .20 |
| P | … | … | .01 | .01 | … | … | … | … | … | … | … | .04 | .10 |
| Q | … | … | .01 | .02 | .02 | .02 | .01 | .07 | .01 | .02 | .03 | .04 | … |
| R | … | … | … | … | … | … | … | … | … | … | … | … | … |
| S | … | … | … | … | .02 | … | .04 | … | … | … | … | .02 | … |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | … | … | … | … | … | … | … | … | … | … | … | .01 | .03 |
| U | … | … | … | … | .08 | .17 | .24 | .04 | … | .02 | .01 | .07 | .17 |
| V | … | … | … | … | … | … | … | … | … | … | … | .01 | .03 |
| W | … | … | … | … | … | … | .02 | … | … | … | … | … | … |
| X | … | … | … | … | … | … | … | … | … | … | … | .07 | … |
| Y | … | … | … | … | … | … | … | … | … | … | … | … | .13 |
| Z | … | … | … | … | … | … | … | … | … | … | … | … | … |
| a | … | … | .01 | … | … | … | … | … | .01 | … | … | … | … |
| b | … | … | … | … | … | … | … | … | … | … | … | … | … |
| c | … | … | … | … | … | … | … | … | … | … | … | … | … |
| d | … | … | … | … | … | … | … | … | … | … | … | … | .04 |
| e | … | … | … | … | … | .01 | … | … | … | … | … | … | … |
| f | … | … | … | … | … | … | … | … | … | … | … | .01 | … |
| g | … | … | … | … | … | … | … | … | … | … | … | .01 | … |
| h | … | … | … | … | … | … | … | … | … | … | … | … | … |
| i | … | … | … | … | … | … | … | … | … | .02 | … | … | … |
| j | … | … | … | … | … | … | … | … | … | … | … | .05 | … |
| k | … | … | … | … | … | … | … | … | .01 | … | … | .01 | … |
| l | … | … | … | … | … | … | … | … | … | … | … | .06 | … |
| m | … | … | … | … | … | … | … | … | … | … | … | … | .03 |
| n | … | … | … | … | … | … | … | … | … | … | … | … | .10 |
| o | … | … | … | … | … | … | … | … | … | .03 | … | … | … |
| p | … | … | … | … | .01 | … | … | … | … | … | … | … | … |
| Heterozygosity[c] | .26 ± .10 | .75 ± .09 | .65 ± .08 | .69 ± .10 | .71 ± .07 | .73 ± .09 | .78 ± .10 | .75 ± .10 | .67 ± .08 | .65 ± .13 | .74 ± .10 | .87 ± .08 | .91 ± .12 |

[a] Underlined values are the frequencies of the four common haplotypes.
[b] Frequency of persistence allele, from independent studies referenced by Flatz (1987).
[c] Heterozygosity over the whole haplotype.

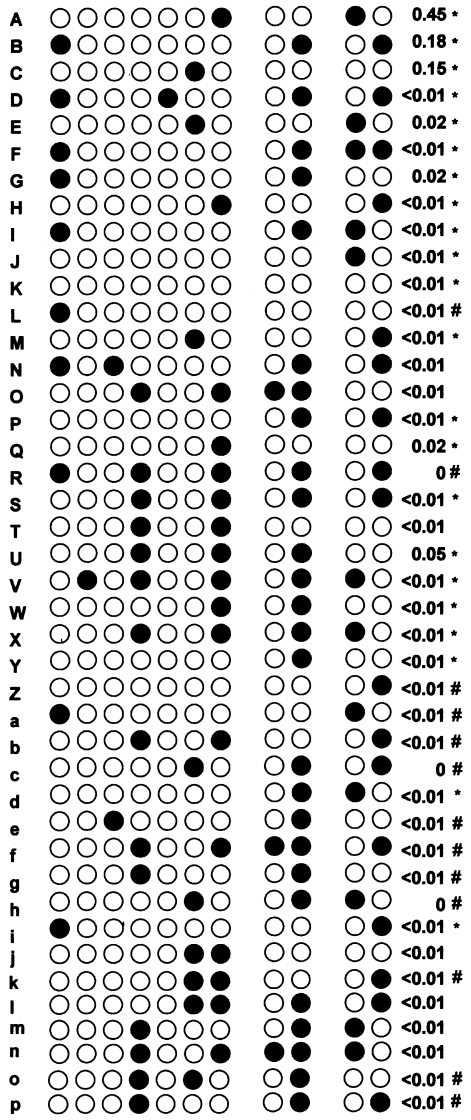


**Figure 3** List of all deduced LCT haplotypes. Complete list of LCT haplotypes, showing their allelic composition and frequency in the total sample set used in this study. The sites are as in figure 1 and are represented by either a blackened or an unblackened circle, which indicates whether the allele is the new or ancestral allele, respectively. Each haplotype is therefore represented as the combination of its component alleles. The ancestral allele at each site is defined by analysis of the analogous sequence of five unrelated chimpanzees. One chimpanzee was heterozygous for TG6236/7ΔΔ, so the commonest allele in the chimpanzee (6236/7TG) is taken as ancestral. The ancestral allele at $A_8$-552/-559$A_9$ cannot be deduced, because the chimpanzee sequence at this point is very different and so is arbitrarily taken as $A_8$. Asterisks (*) indicate definite haplotype, either by homozygosity observed in an individual, heterozygosity at only one site, or deduced from previous family studies. Number signs (#) indicate that the haplotype was present only at .01 frequency in any population. Haplotype R was only observed in one British individual of African descent but is not included in the data set. Haplotype c was deduced by the visual inspection of the genotypes of two individuals but was not identified by maximum likelihood analysis. Both haplotypes are included for the sake of completeness.

This study confirms the strong linkage disequilibrium across the lactase gene that we reported previously in Europeans (Harvey et al. 1995). In some cases, pairwise linkage disequilibrium values are lower and appear to reflect a particular haplotype frequency. For example, in southern Europeans the reasonably high frequency of haplotype E (.06, table 2), a probable recombinant between haplotypes A and C (fig. 4), is reflected in a $D'_{A\text{-}678G/T5579C}$ value of .06, or almost complete linkage equilibrium, between A-678G and T5579C. The same, if less dramatic, observation can be made in Russians (frequency of haplotype E = .05, $D'_{A\text{-}678G/T5579C}$ = .23,). The frequency of the U haplotype also varies between populations, and this is reflected in pairwise linkage disequilibrium scores between $A_8$-552/-559$A_9$ and G666A. For example, in Japanese, the frequency of the U haplotype is .24 and the $D'_{A8\text{-}552\ to\text{-}559A9/A666G}$ value is .35, whereas in Papua New Guineans (frequency of haplotype U = .04) and in south Indians (frequency of haplotype U = 0) the $D'_{A8\text{-}552\ to\text{-}559A9/G666A}$ values are .82 and 1, respectively.

Construction of the *LCT* haplotype network reveals the relationships of the common haplotypes to each other and to the root haplotype and helps in the interpretation of the present-day distributions by emphasizing that the A, B, and U haplotypes are not directly connected to the root and must have been generated by intermediate haplotypes that are now rare or absent from the non-African populations.

The sub-Saharan African samples show much greater haplotype diversity than other populations (tables 1 and 2) with many extra haplotypes not seen in the other population samples. Analysis of the unweighted means of *LCT* haplotype-class frequencies shows that the four common haplotypes and their simple derivatives (haplotypes A, B, C, U, O, D, N, and F) comprise 89% of total haplotype diversity in non-Africans, but only 51% in sub-Saharan Africans. Both the San and Bantu have high frequency of haplotypes (30% and 28%, respectively) that are the intermediate steps between the four common haplotypes, but these are almost absent in non-African populations (≤3%). This pattern of higher diversity in Africans has been observed at other loci and has been interpreted as evidence for an "out of Africa" model for peopling of the non-African continents (Ayala and Escalante 1996) though the role of selection and population size in altering diversity within populations is debated (Harding et al. 1997; Jobling et al. 1998). Some data have been interpreted as favoring a "multiregional hypothesis" with large amounts of gene flow between populations (Xiong et al. 1991; Harding et al. 1997), although, in one study, the data set was rather small (Harris and Hey 1999).

The *LCT* data reported here suggest that much of the present-day haplotype distribution and loss of haplo-

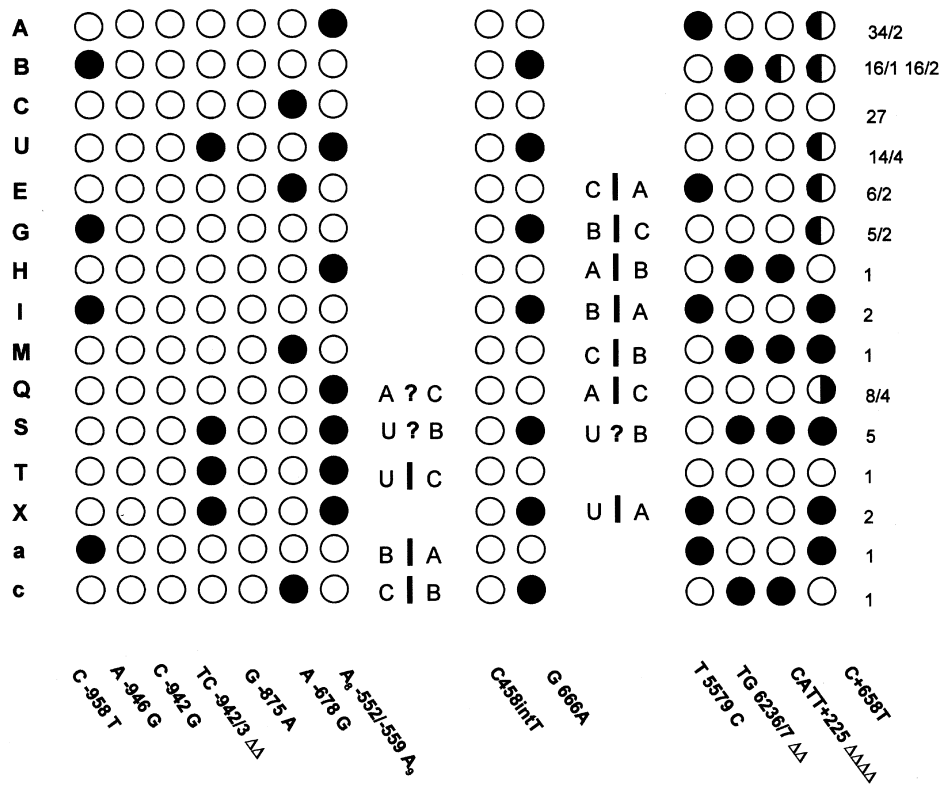


**Figure 4** Simple recombinants of the four common haplotypes. The alleles at each polymorphic site are represented as in figures 1 and 2. The two additional sites tested are CATT+225ΔΔΔΔ and C+658T. The numbers of chromosomes of each haplotype analyzed by this method are shown on the right. When a circle is half blackened, both alleles were found on the haplotype defined by the 11 sites, and the two numbers shown are the numbers with each allele at the new (12th or 13th) site. The first number represents the number of chromosomes with the allele indicated by the first half of the circle, and the second number represents the number of chromosomes with the allele indicated by the second half of the circle. The probable position of the recombination is shown by a vertical bar, with the progenitor haplotypes indicated on either side. In some cases, the recombination point was localized more precisely by means of an exon 6 polymorphism (Boll et al. 1991; data not shown). For example both forms of haplotype E (+658C and +658T) were shown to be like haplotype C at exon 6, suggesting recombination between exon 6 and exon 17. The vertical bar is replaced by a question mark (?) where the position of recombination is ambiguous.

type diversity outside Africa resulted from genetic drift. Direct selection for particular alleles is unlikely to account for the loss of diversity, since there are no known alleles shared between the four common haplotypes that are not present in other haplotypes, and their distant relationship means that a linked allele would have been shared by other intermediate haplotypes.

Another noteworthy aspect of the non-African populations is that, with the exception of the northern Europeans, they show rather similar levels of haplotype diversity. This contrasts with a study of the pattern of variation at the phenylalanine hydroxylase (PAH) locus that used a four-site haplotype spanning 75 kb across the gene (Kidd et al. 2000). The PAH locus shows interpopulation differences in haplotype composition and diversity. For example, the heterozygosities for the Russians and Japanese are .76 and .51, respectively, which is a statistically significant difference ($P = .05$, Student's t-test), whereas the two heterozygosities at the *LCT* locus for a similar-sized sample set from the same populations are .74 and .78 ($P = .4$). The similarity of the *LCT* locus in non-African populations suggests that much of the diversity was determined before the populations divided and hence before domestication of mammals. Therefore, the bottleneck that caused the predominance of four haplotypes in non-Africans occurred before modern humans were geographically spread over the Old World. This occurred recently enough for linkage disequilibrium to be observed across the region in modern non-African populations.

Despite the overall similarity, there are some differences in the frequencies of haplotypes among non-sub-Saharan African populations that deserve comment. Haplotype U is found at reasonable frequencies in all populations except Indo-Europeans, in whom it is very rare or absent. This haplotype may have originated in Asia and moved back into Africa, a movement which has been suggested by Y-chromosome analysis (Ham-

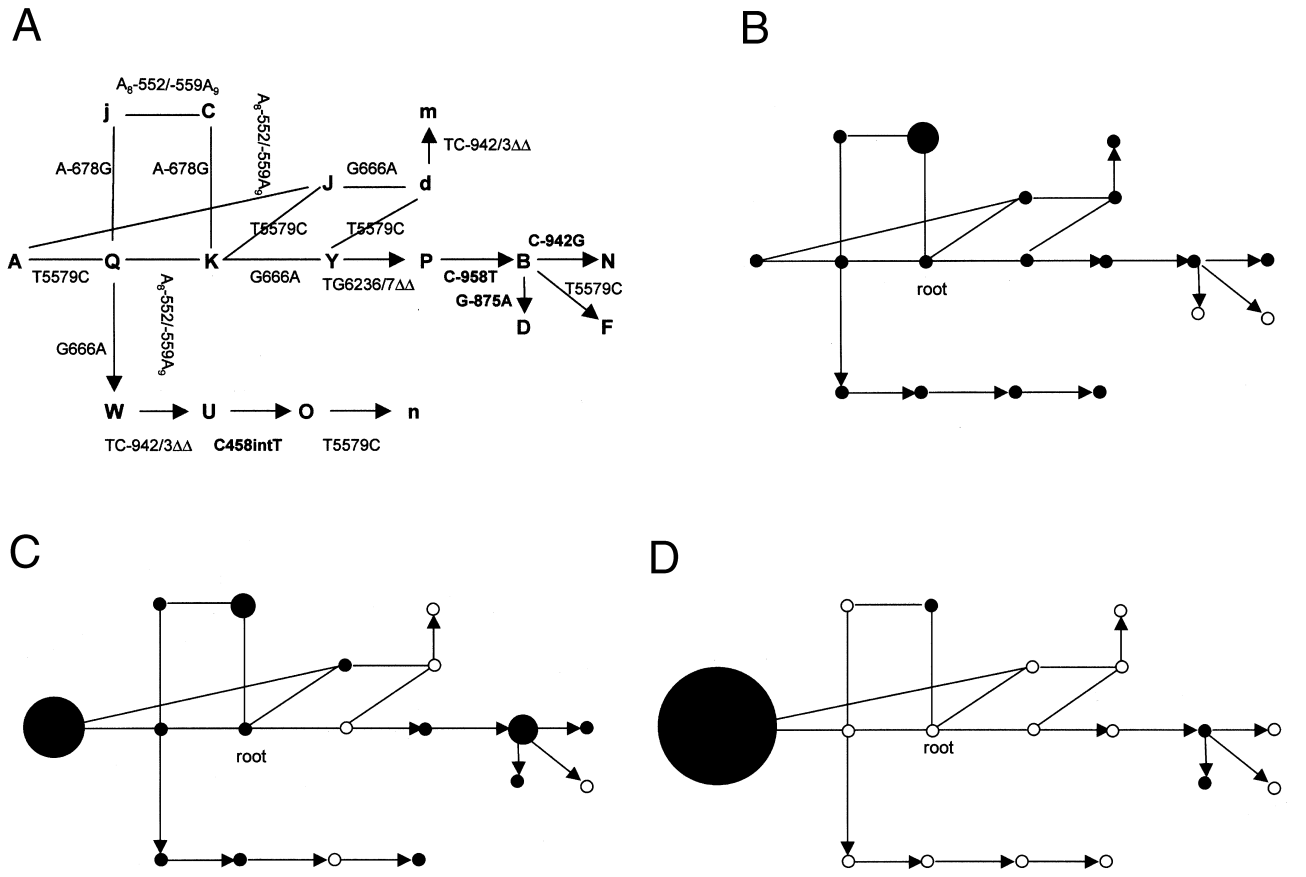


**Figure 5** Lactase haplotype networks. *A,* Haplotype network showing probable phylogeny of the four common haplotypes (A, B, C, and U). Each line is annotated with its corresponding mutational change, and an arrow is shown where the directionality of the mutation is known. Mutational changes shown in bold are changes that occur only once in the network. *B,* Haplotype network, based on the framework of *A,* with circle size corresponding to the frequency of the haplotype in the population. An unblackened circle shows that none of that haplotype was observed in the population, and the smallest blackened circle represents frequencies of ⩽.1. The sub-Saharan African populations are grouped and shown here, with 79% of total haplotype diversity represented in the diagram. *C,* As *B,* with non-African populations showing 92% of total non-African haplotype diversity represented in the diagram. Non-African excludes northern European. *D,* As *B,* with northern European populations showing 98% of total northern European haplotype diversity represented in the diagram.

mer et al. 1998). The distribution may also be an indirect consequence of expansion of Indo-European populations, or perhaps selection acting on a linked locus. The observation of U haplotypes in the Russian population in the Urals may be a result of gene flow from a neighboring Siberian population, since the Yakut have a high frequency of U haplotypes (E. J. Hollox, A Kozlov and S. Markova, unpublished data).

The A haplotype is present in all populations, whatever their lactase-persistence allele frequency. We have previously suggested that lactase persistence arose on an A haplotype background (Harvey et al. 1998), and, although lactase persistence is associated with the A haplotype in Europeans, there are clearly many non-persistent A haplotypes in other populations. This study emphasizes the unusually high frequency of the A haplotype in northern Europeans, the only population tested in which lactase persistence is at very high frequency, and suggests that this high frequency is caused by selection for the linked allele for lactase persistence. The low heterozygosity in this group (caused by the high incidence of A chromosomes) contrasts with the high heterozygosity value at the PAH locus, emphasizing the different histories of the two loci. However, a change in the frequency of haplotype A is not evident for the less-marked cline in lactase persistence from north to south India. This could be a sampling artifact caused by population sub-structure within the Indians tested, or a high frequency of non-persistent A haplotypes in the south Indians.

Several authors have attempted to explain present-day lactase-persistence frequencies, in relation to the short time period available for selection, using standard selective (Bodmer and Cavalli-Sforza 1976) and gene-

culture coevolutionary models (Aoki 1986; Feldman and Cavalli-Sforza 1989). Here, we show that the A haplotype is a useful "molecular handle" on a process that has played a unique role in the evolution of humans. Subdivision of A haplotypes using associated novel microsatellites is underway, and these subdivided haplotypes may provide the necessary resolution for human genetic-history studies within closely related populations differing in milk-drinking culture. These future studies, together with the identification of the molecular basis, will clarify the interaction between genetic and cultural diversity in the evolution of this fascinating polymorphism.

## Acknowledgments

We thank David Whitehouse, David Hopkinson, Sir Walter Bodmer, and the anonymous reviewers for helpful comments on the manuscript. We would also like to thank Philip Johnson for Japanese and Papua New Guinean DNA samples.

## Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Authors' "Lactase" page, http://www.gene.ucl.ac.uk/mucin/lactase.html (for AS-PCR primers)

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/index.html (for 1 kb of contiguous chimpanzee sequence 5′ to the lactase gene)

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for major histocompatability complex [MIM 142857, MIM 604305] and as-yet-unidentified *cis*-acting element [MIM 223100])

UK-MRC HGMP, http://www.hgmp.mrc.ac.uk/Registered/Option/eh.html (for EH and ASSOCIATE programs)

## References

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

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

Ayala FJ, Escalante AA (1996) The evolution of human populations: a molecular perspective. Mol Phylogenet Evol 5: 188–201

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Bergstrom T, Josefsson A, Erlich HA, Gyllensten U (1998) Recent origins of HLA-DRB1 alleles and implications for human evolution. Nat Genet 18:237–242

Bodmer WF, Cavalli-Sforza LL (1976) Genetics, evolution, and man. W. H. Freeman, San Francisco

Boll W, Wagner P, Mantei N (1991) Structure of the chromosomal gene and cDNAs coding for lactase phlorizin hydrolase in humans with adult-type hypolactasia or persistence of lactase. Am J Hum Genet 48:889–902

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartianen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612

Dudd SN, Evershed RP (1998) Direct demonstration of milk as an element of archaeological economies. Science 282: 1478–1481

Durham W (1991*a*) Coevolution: genes, culture, and human diversity. Stanford University Press, Stanford, CA

Durham W (1991*b*) Cultural mediation: the evolution of adult lactose absorption. In: Durham W (ed) Coevolution: genes, culture, and human diversity. Stanford University Press, Stanford, CA, pp 226–285

Feldman MW, Cavalli-Sforza LL (1989) On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem. In: Feldman MW (ed) Mathematical evolutionary theory. Princeton University Press, Princeton, NJ, pp 145–173

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

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy Blood Group locus. Am J Hum Genet 66:1669–1679

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

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15: 427–441

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–789

Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, Rees JL (2000) Evidence for variable selective pressures at MC1R. Am J Hum Genet 66:1351–1361

Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA 96:3320–3324

Harvey CB, Hollox EJ, Poulter M, Wang Y, Rossi M, Auricchio S, Iqbal TH, Cooper BT, Barton R, Sarner M, Korpela R, Swallow DM (1998) Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. Ann Hum Genet 62:215–223

Harvey CB, Pratt W, Islam I, Whitehouse DB, Swallow DM (1995) DNA polymorphisms in the lactase gene: linkage disequilibrium across the 70 kb region. Eur J Hum Genet 3:27–41

Hill AV, Allsopp CE, Kwiatkowski D, Taylor TE, Yates SN, Anstey NM, Wirima JJ, Brewster DR, McMichael AJ, Molyneux ME, Greenwood BM (1992) Extensive genetic diversity in the HLA class II region of Africans, with a focally predominant allele, DRB1*1304. Proc Natl Acad Sci USA 89:2277–2281

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

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

Jaruzelska J, Zietkiewicz E, Labuda D (1999) Is selection responsible for the low level of variation in the last intron of the ZFY locus? Mol Biol Evol 16:1633–1640

Jobling MA, Williams GA, Schiebel GA, Pandya GA, McElreavey GA, Salas GA, Rappold GA, Affara NA, Tyler-Smith C (1998) A selective difference between human Y-chromosomal DNA haplotypes. Curr Biol 8:1391–1394

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet 66: 1882–1899

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, Lu RB, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, Kidd JR (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227

Kolars JC, Levitt MD, Aouji M, Savaiano DA (1984) Yogurt—an autodigestive source of lactase. N Engl J Med 310: 1–3

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

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

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

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

Rogers AR, Jorde LB (1995) Genetic evidence on modern human origins. Hum Biol 67:1–36

Rudbeck L, Dissing J (1998) Rapid, simple alkaline extraction of human genomic DNA from whole blood, buccal epthelial cells, semen and forensic stains for PCR. Biotechniques 25: 588–590

Simoons FJ (1971) The antiquity of dairying in Asia and Africa. Geographical Rev 61:431–439

Swallow DM, Hollox EJ (2000) The genetic polymorphism of intestinal lactase activity in adult humans. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular bases of inherited disease. McGraw-Hill, New York

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

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Wainscoat JS, Hill AVS, Boyce AL, Flint J, Hernandez M, Thein SL, Old JM, Lynch JR, Falusi AG, Weatherall DJ, Clegg JB (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature 319:491–493

Wang Y, Harvey CB, Hollox EJ, Phillips AD, Poulter M, Clay P, Walker-Smith JA, Swallow DM (1998) The genetically programmed down-regulation of lactase in children. Gastroenterology 114:1230–1236

Wang Y, Harvey CB, Pratt WS, Sams VR, Sarner M, Rossi M, Auricchio S, Swallow DM (1995) The lactase persistence/non-persistence polymorphism is controlled by a *cis*-acting element. Hum Mol Genet 4:657–662

Weiss KM (1993) Genetic variation and human disease: principles and evolutionary approaches. Cambridge University Press, Cambridge

Xiong WJ, Li WH, Posner I, Yamamura T, Yamamoto A, Gotto AMJ, Chan L (1991) No severe bottleneck during human evolution: evidence from two apolipoprotein C-II deficiency alleles. Am J Hum Genet 48:383–389

Yip SP, Lovegrove JU, Rana NA, Hopkinson DA, Whitehouse DB (1999) Mapping recombination hotspots in human phosphoglucomutase (PGM1). Hum Mol Genet 8: 1699–1706