Am. J. Hum. Genet. 74:1111–1120, 2004

Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene

Todd Bersaglieri,[1] Pardis C. Sabeti,[3] Nick Patterson,[3] Trisha Vanderploeg,[1] Steve F. Schaffner,[3] Jared A. Drake,[1] Matthew Rhodes,[1,*] David E. Reich,[2,3] and Joel N. Hirschhorn[1,2,3]

[1]Divisions of Genetics and Endocrinology, Children's Hospital, and [2]Department of Genetics, Harvard Medical School, Boston; and [3]Program in Medical and Population Genetics, Whitehead/Massachusetts Institute for Technology Center for Genome Research, Cambridge, MA

In most human populations, the ability to digest lactose contained in milk usually disappears in childhood, but in European-derived populations, lactase activity frequently persists into adulthood (Scrimshaw and Murray 1988). It has been suggested (Cavalli-Sforza 1973; Hollox et al. 2001; Enattah et al. 2002; Poulter et al. 2003) that a selective advantage based on additional nutrition from dairy explains these genetically determined population differences (Simoons 1970; Kretchmer 1971; Scrimshaw and Murray 1988; Enattah et al. 2002), but formal population-genetics–based evidence of selection has not yet been provided. To assess the population-genetics evidence for selection, we typed 101 single-nucleotide polymorphisms covering 3.2 Mb around the lactase gene. In northern European–derived populations, two alleles that are tightly associated with lactase persistence (Enattah et al. 2002) uniquely mark a common (~77%) haplotype that extends largely undisrupted for >1 Mb. We provide two new lines of genetic evidence that this long, common haplotype arose rapidly due to recent selection: (1) by use of the traditional F_{ST} measure and a novel test based on p_{excess}, we demonstrate large frequency differences among populations for the persistence-associated markers and for flanking markers throughout the haplotype, and (2) we show that the haplotype is unusually long, given its high frequency—a hallmark of recent selection. We estimate that strong selection occurred within the past 5,000–10,000 years, consistent with an advantage to lactase persistence in the setting of dairy farming; the signals of selection we observe are among the strongest yet seen for any gene in the genome.

Introduction

Genes that have experienced recent positive selection offer a window into the evolutionary forces that shaped recent human history. For example, signatures of recent selection for resistance to malaria have been demonstrated around the HbS allele in the β-globin gene *HBB* (MIM 141900) (Pagnier et al. 1984), the A^- and Med alleles in *G6PD* (MIM 305900) (Tishkoff et al. 2001), the *O allele of the Duffy gene *FY* (MIM 110700) (Hamblin et al. 2002), and a promoter variant in the CD40 ligand gene *TNFSF5* (MIM 300386) (Sabeti et al. 2002). Other genes for which genetic data support a recent selective event include *CKR5* (MIM 601373) (Stephens et al. 1998), *HFE* (MIM 235200) (Toomajian et al. 2003), *ADH1B* (MIM 103720) (Osier et al. 2002), and possibly *CFTR* (MIM 602421) (Wiuf 2001 and references therein); the particular evolutionary advantage in these cases is less clear. Many of the selected alleles also contribute to or cause disease, indicating that identification of genes under selection may have significant consequences for medical genetics. Furthermore, once such genes have been definitively identified, characterizing the signatures of selection at these genes will guide the development of tools to search for other genes under selection.

One of the genes most frequently proposed to have experienced recent positive selection is *LCT* (MIM 603202), which encodes the enzyme lactase-phlorizin hydrolase. The epidemiologic data in favor of selection are quite strong: the ability to use this enzyme to digest lactose during adulthood varies dramatically across worldwide populations, with particularly high rates among northern Europeans (Bayless and Rosensweig 1966; Simoons 1969; Scrimshaw and Murray 1988). Furthermore, persistence of lactase activity into adulthood is genetically determined (Simoons 1970; Kretchmer 1971; Scrimshaw and Murray 1988; Enattah et al. 2002), and the geographic distribution of lactase persistence matches the distribution of dairy farming (Simoons 1969; Kretchmer 1971; Scrimshaw and Murray 1988). Because of these features, Cavalli-Sforza (1973) and others (Simoons 1970; Flatz 1987; Hollox et al. 2001; Poulter et al. 2003) proposed that the high rate of lactase persistence in European populations is explained by positive selection resulting from increased

Received December 18, 2003; accepted for publication March 10, 2004; electronically published April 26, 2004.

Address for correspondence and reprints: Dr. Joel N. Hirschhorn, Enders 561, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115. E-mail: joel.hirschhorn@childrens.harvard.edu.

* Current affiliation: Cornell University, Ithaca, NY.

0002-9297/2004/7406-0005$15.00

nutrition from dairy, the only dietary source of lactose. Despite these compelling epidemiologic data, neither formal population-genetics–based evidence of selection nor an estimate of the timing and magnitude of positive selection has been provided by analyzing genetic data at the *LCT* locus. In addition, many non-European populations show high rates of lactase persistence, raising questions about whether a single allele arose once and is shared by all lactase-persistent individuals or whether different alleles have arisen in human history.

Recently, new tools to study selection at *LCT* have become available. In particular, Enattah et al. (2002) demonstrated that two polymorphisms upstream of *LCT* are tightly associated with lactase persistence. In that study, the persistence-associated alleles were found primarily on a single 250-kb microsatellite haplotype in the Finnish population. By use of 18 SNPs spanning 1 Mb, Swallow and colleagues also recently reported a long haplotype around these alleles (Poulter et al. 2003). However, the mere presence of a long haplotype, although consistent with selection, does not by itself constitute a signature of a selective event (Sabeti et al. 2002).

A variety of genetic signatures of positive selection have been described (reviewed in Bamshad and Wooding 2003). These include an excess of rare variants (indicating a selective sweep followed by the accumulation of new, rare mutations), large allele-frequency differences among populations (indicating differential effects of selection that cause alleles to rise dramatically in frequency in some but not all of the populations), or a common haplotype that remains intact over unusually long distances (indicating an allele that rose rapidly to high frequency before recombination could disrupt the haplotype on which the allele lies). The last two signatures are particularly appealing because they can be detected by genotyping common polymorphisms in one or more populations and may have better power for identifying recent positive selection (Sabeti et al. 2002). Large differences in allele frequencies between populations have traditionally been detected by use of the population-genetics measure F_{ST} (e.g., Akey et al. 2002), whereas demonstration that a common haplotype is unexpectedly long requires application of the recently described long-range haplotype test (Sabeti et al. 2002).

In this study, we analyze genotypes for >100 SNPs in multiple populations, and we demonstrate two striking signatures of selection at the *LCT* gene. First, SNPs near *LCT* show large differences in allele frequencies among populations, demonstrated not only with the traditional F_{ST} measure but also with a more informative metric, p_{excess}. In addition, we show that the long (1 Mb) haplotype carrying the persistence-associated alleles is much longer and more common than would be expected in the absence of selection. We are also able to estimate from these genetic data the time period during which selection occurred, and we show that the selective pressure at *LCT* was comparable to the strongest selection yet documented in the genome.

Subjects and Methods

DNA Samples

DNA samples for European American, African American, and East Asian populations were obtained from the Coriell Institute (Coriell Institute for Medical Research Web site); a complete list of these samples and geographic origins is given in table A1 (online only). The Scandinavian population, which has been described elsewhere (Altshuler et al. 2000), is a subset of 379 normal glucose-tolerant trios from Finland and Sweden, and the samples we typed represent 360 independent chromosomes. The remaining populations listed in table 1 have also been described elsewhere (Rosenberg et al. 2002). This project was approved by the appropriate local institutional review boards, and subjects gave informed consent.

Selection and Genotyping of SNPs

SNPs were selected from dbSNP (dbSNP Home Page), preferentially choosing the SNP Consortium (TSC) and BAC overlap SNPs (submitter handles: TSC, SC_JCM, and KWOK) and genotyping SNPs at a greater density closer to the *LCT* gene. In addition, we intentionally genotyped the two SNPs reported to be associated with *LCT* persistence (Enattah et al. 2002). A complete list is given in table A2 (online only). SNPs were genotyped by use of the mass-spectrometry–based MassArray platform provided by Sequenom, implemented as described elsewhere (Gabriel et al. 2002). Primers were designed by use of Spectrodesigner software (Sequenom), and sequences are available on request.

Statistical Analysis

F_{ST} was calculated as described by Akey et al. (2002), with Nei's correction for sample size (Nei and Chesser 1983). To generate a genomewide distribution for F_{ST} and p_{excess}, allele frequencies at markers throughout the genome were downloaded from the SNP Consortium (TSC) Web site, by use of data from the Whitehead Institute Center for Genome Research (WICGR), Celera, Motorola, and Orchid. We excluded data from pooled samples, since the F_{ST} distribution was different for pooled data (Akey et al. 2002 and data not shown). In total, data from 28,440 markers were used to generate a genomewide F_{ST} distribution. To compare the F_{ST} at markers around *LCT* with the genomewide distribution, we applied the Wilcoxon rank-sum test (Rosner 1982), limiting our analysis to markers separated by at least 20

Table 1

Frequencies in Different Populations of Two Alleles Associated with Lactase Persistence

Population Group (Region and/or Country)	No. of Chromosomes	Frequency (%) for	
		−13910T	−22018A
European American	48	77.2	77.1
African American	100	14.0	13.3
East Asian	35	0	0
Yoruba (Nigeria)	50	0	0
Bantu Northeast (Kenya)	24	0	0
San (Namibia)	14	0	0
Bantu (South Africa)	16	0	0
Mozabite (Mzab, Algeria)	60	21.7	21.7
Bedouin (Negev, Israel)	98	3.1	4.1
Druze (Carmel, Israel)	96	2.1	2.1
Palestinian (Central Israel)	102	3.9	3.9
Brahui (Pakistan)	50	34.0	38.0
Balochi (Pakistan)	50	36.0	42.0
Hazara (Pakistan)	50	8.0	12.0
Makrani (Pakistan)	50	34.0	36.0
Sindhi (Pakistan)	50	32.0	30.0
Pathan (Pakistan)	50	30.0	32.0
Kalash (Pakistan)	50	0	0
Burusho (Pakistan)	50	10.0	12.0
Han (China)	90	0	0
Tujia (China)	20	0	0
Yizu (China)	20	0	0
Miaozu (China)	20	0	0
Oroqen (China)	20	0	0
Daur (China)	20	5.0	5.0
Mongola (China)	20	10.0	10.0
Hezhen (China)	20	0	0
Xibo (China)	18	0	0
Uygur (China)	20	5.0	10.0
Dai (China)	20	0	0
Lahu (China)	20	0	0
She (China)	20	0	0
Naxi (China)	20	0	0
Tu (China)	20	0	0
Yakut (Siberia)	50	6.0	6.0
Japanese (Japan)	62	0	0
Cambodian (Cambodia)	22	0	0
Papuan (New Guinea)	34	0	0
Melanesian[a] (Bougainville)	44	0	0
French (France)	58	43.1	44.8
French Basque (France)	48	66.7	66.7
Sardinian (Italy)	56	7.1	7.1
Tuscan (Italy)	16	6.3	6.3
North Italian (Bergamo, Italy)	28	35.7	35.7
Orcadian (Orkney Islands)	32	68.8	68.8
Adygei (Russian Caucasus)	34	11.8	11.8
Russian (Russia)	50	24.0	24.0
Swedish and Finnish (Scandinavia)	360	81.5	ND
Pima (Mexico)	50	0	0
Maya (Mexico)	50	2.0	2.0
Colombian (Colombia)	26	0	0
Karitiana (Brazil)	48	0	0
Surui (Brazil)	42	0	0

Note.—The European American, African American, and East Asian samples are from the Coriell Institute Cell Repository (Coriell Institute for Medical Research Web site). The Scandinavian (Altshuler et al. 2000) and the remaining (Rosenberg et al. 2002) populations have been described elsewhere. ND = not done.

[a] Non-Austronesian.

kb to minimize correlation between markers. To eliminate artifactual effects at the lower end of the F_{ST} distribution (which can be due, in part, to the correction for sample size), we treated all F_{ST} values below the population mean as ties. Applying this test to the markers around *LCT* yields a *P* value of .002. However, because we cannot fully correct for the correlation between markers, this *P* value may overestimate the significance of the excess markers with high F_{ST} values.

To understand the rationale for using the p_{excess} statistic, consider the scenario where positive selection rapidly introduces a single haplotype at frequency h into a population. Under the model of strong selection, a particular long-range haplotype will rapidly rise from a single copy (frequency near 0) to a frequency of h in the selected population. Consider now a marker within the long-range haplotype with an allele of frequency p prior to the selective event. If there has been little opportunity for recombination, nearly all copies of the selected haplotype will carry the same allele at this marker. For the allele that lies on the selected haplotype, the allele frequency will increase to $p_1 = p(1 - h) + h$ after selection; for an allele that does not lie on the selected haplotype, the allele frequency will decrease to $p_1 = p(1 - h)$. Solving for h, $h = (p_1 - p)/(1 - p)$ if $p_1 > p$ and $h = (p - p_1)/p$ if $p > p_1$. This is algebraically identical to p_{excess} (Hastbacka et al. 1994); here, p_1 is the allele frequency in the population under consideration, and p is the ancestral allele frequency, which we estimate by the average allele frequency in the populations that have not experienced selection (in this case, the East Asian and African American populations). To maximize the chance that the variant predates the selective event (essential for using p_{excess} to estimate h), we only calculate p_{excess} for polymorphisms in which the allele frequencies in all populations are between 10% and 90%. Similar results were obtained whether or not we corrected the allele frequencies in African Americans for the estimated 21% European admixture (Parra et al. 1998). Of the markers from the SNP Consortium (TSC) Web site, 13,696 have allele frequencies between 10% and 90% for all three populations, and these were used for calculating the genomewide characteristics of p_{excess}. For comparison, we identified 952 regions with at least 5 markers spanning 50 kb–100 kb. We found that none of these 952 regions contains runs of ≥ 5 consecutive markers that span at least 50 kb and have p_{excess} values above the 90th percentile; the *LCT* region has 16 consecutive markers spanning 800 kb with p_{excess} values above the 95th percentile.

The long-range haplotype test, the calculation of relative extended haplotype homozygosity (REHH), and the assessment of the significance of REHH by use of simulations were performed as described elsewhere (Sabeti et al. 2002). In brief, a core region was defined as a block of linkage disequilibrium with little evidence of recombination (Gabriel et al. 2002). The genotype data was converted to inferred, fully phased haplotype data, and, within the core region, each common haplotype (>5% frequency) was analyzed separately. At each marker, a chromosome was considered intact if, from the core through that marker, the chromosome was identical to all other intact chromosomes carrying the same core haplotype. For *LCT*, the core region was chosen to contain the persistence-associated markers. For the simulations, cores and genotypes extending outward from the cores were generated as described elsewhere (Sabeti et al. 2002). The empirical *P* value for the 5′ markers was .012. For the 3′ markers, 10,000 simulations generated ~25,000 core haplotypes, of which ~2,500 had a frequency similar to that of the *LCT* core; none of these had an REHH near that seen for *LCT* (empirical $P < .0004$). To better estimate the *P* value for the 3′ markers, the REHH distribution from the simulated data was log-transformed to achieve normality, and the mean, median, and SD were used to estimate *P* values for the actual REHH value observed in *LCT*. The estimation of dates was performed according to methods described elsewhere (Reich and Goldstein 1998; Stephens et al. 1998).

For these analyses, fully phased haplotype data were required. We used two phasing programs: PHASE, a Bayesian method for phasing diploid genotype data (Stephens and Donnelly 2003; PHASE Web site), and also a similar program (wphase) that we developed for this purpose. Similar results were obtained from the two phasing algorithms. The mathematical models underlying the two programs are similar, but PHASE performs a Markov Chain–Monte Carlo procedure, whereas wphase carries out a hill climb, (approximately) maximizing the likelihood. We estimated REHH and dates at distances on either side of the core region, where approximately one recombination per chromosome had occurred on the persistence-associated haplotype (that is, ~1/e chromosomes carrying the persistence-associated haplotype remained unrecombined).

We estimated the coefficient of selection, *s*, by applying a formula (Hartl and Clark 1997) that relates the frequency in generation $t + 1 (p_{t+1})$ to the frequency in generation $t (p_t)$:

$$p_{t+1} = \frac{p_t q_t([p_t][w_{11} - w_{12}] + [q_t][w_{12} - w_{22}])}{(p_t^2 \times w_{11} + 2p_t q_t \times w_{12} + q_t^2 \times w_{22})} .$$

In this formula, $q_t = 1 - p_t$,w_{11} is the relative fitness of individuals homozygous for the selected allele, w_{12} is the relative fitness of heterozygous individuals, and w_{22} is the relative fitness of individuals homozygous for the unselected allele. We assumed a dominant model for lactase persistence—that is, $w_{11} = w_{12} = 1$ and $w_{22} = 1 - s$. We also assumed the initial frequency p_0 to be

between 1/1,000 and 1/10,000 (corresponding to a new mutation in a population with an effective size between 500 and 5,000; larger population sizes yield even higher coefficients of selection). Starting from these initial frequencies, we calculated values of w_{22} that would yield a frequency of $p = 0.77$ after 2,188–20,650 years of selective pressure for the United States population and 1,625–3,188 years for the Scandinavian population, assuming 25 years/generation.

Results

To examine the evidence for selection, we began by genotyping the two SNPs that were recently reported to be very tightly associated with lactase persistence (Enattah et al. 2002): rs4988235 (−13910C→T) and rs182549 (−22018G→A). We determined the frequencies of the persistence-associated alleles (T and A, respectively) in three populations for which many thousands of markers have been genotyped (European Americans, African Americans, and East Asians), thereby permitting comparison of our results to a genomewide background distribution (Akey et al. 2002). The persistence-associated alleles occur with a frequency of 77% in European Americans, 13% and 14% in African Americans, and 0% in East Asians (table 1), broadly consistent with the rates of lactase persistence in these populations (Scrimshaw and Murray 1988). Large differences in allele frequencies across populations, such as we observe at these markers, are suggestive of selective pressure that differed among the populations (Lewontin and Krakauer 1973; Bowcock et al. 1991; Akey et al. 2002). The unusually large magnitude of the population frequency differences for these two markers is reflected in their values of F_{ST}, a traditional measure of population differentiation—the F_{ST} values (0.53 for both markers) exceed 99.9% of the F_{ST} values from a genomewide set of >28,000 SNPs (see the "Subjects and Methods" section). We also genotyped these two associated SNPs in a more diverse set of samples (Altshuler et al. 2000; Rosenberg et al. 2002); the frequencies of the persistence-associated alleles were much lower in southern European than in northern European or Basque populations, and the persistence-associated alleles were rare or absent in almost all non-European–derived populations tested, except Algerians and Pakistanis (table 1). The wide range of allele frequencies among European populations is consistent with selective pressure that postdates the colonization of Europe, resulting in different prevalences of lactase-persistence alleles in northern and southern European populations.

To extend these results, we genotyped an additional 99 markers in 3.2 Mb flanking the *LCT* locus, again looking for high degrees of population differentiation. In response to strong positive selection, a selected allele rises rapidly in frequency. The frequency of the haplotype on which the allele occurs will increase correspondingly, because there is insufficient time for recombination to disrupt the haplotype while it becomes more common. Thus, allele frequencies at flanking markers on the haplotype will be altered. To measure this effect, we used two metrics of allele-frequency differences: the traditional F_{ST} and a newer metric, p_{excess}. F_{ST} has limited utility when the flanking allele on the selected haplotype was already fairly common prior to selection, because, in this case, the F_{ST} value will be quite low; thus, only a fraction of flanking markers are expected to show elevated F_{ST} values within a region of selection. Consistent with this expectation, there was an excess of high F_{ST} values among the 99 markers, but F_{ST} values varied widely from marker to marker (fig. 1*a*; see the "Subjects and Methods" section for additional details). The excess elevation of F_{ST} is predominantly derived from markers located in the vicinity of the *LCT* gene (fig. 1*a*), with allele frequencies that are generally different in Europeans than in the other two populations (table A2 [online only]). This elevated F_{ST} in markers flanking *LCT* confirms the signal of selection seen with the −13910C→T and −22018G→A variants. However, as expected, only some of the markers near *LCT* have elevated F_{ST} values. Accordingly, we sought an alternative measure of population differentiation that would reveal a more consistent signal in the vicinity of a selected allele.

We chose to study the p_{excess} statistic, which has previously been used to localize disease-causing alleles in founder populations and is a measure of differences in haplotype frequencies across long distances (Hastbacka et al. 1994). p_{excess} is also equivalent to the measure of linkage disequilibrium, δ (Devlin and Risch 1995). If a single haplotype differs in frequency across a long region, p_{excess} will be elevated and relatively constant across multiple markers within that region, with values approximately equal to the increase in frequency of the haplotype (see the "Subjects and Methods" section for details). We observed a consistent, marked elevation of p_{excess} in the *LCT* region: 17 consecutive markers in a region spanning 500 kb around *LCT* have nearly identical, very high values of p_{excess} that approximate the frequency of the persistence-associated haplotype (0.77) (fig. 1*b*). Furthermore, the elevation in p_{excess} extends for at least 1,500 kb (fig. 1*b*; table A2 [online only]). To provide a framework for comparison, we calculated p_{excess} values for marker pairs and the correlation between pairs as a function of distance for >13,000 SNPs throughout the genome; we found that the correlation is normally minimal at distances of as little as 100 kb ($r^2 = 0.002$). Indeed, in this genomewide data set, none of 952 comparison regions had a consistent elevation in p_{excess} values approaching that seen around *LCT* (see

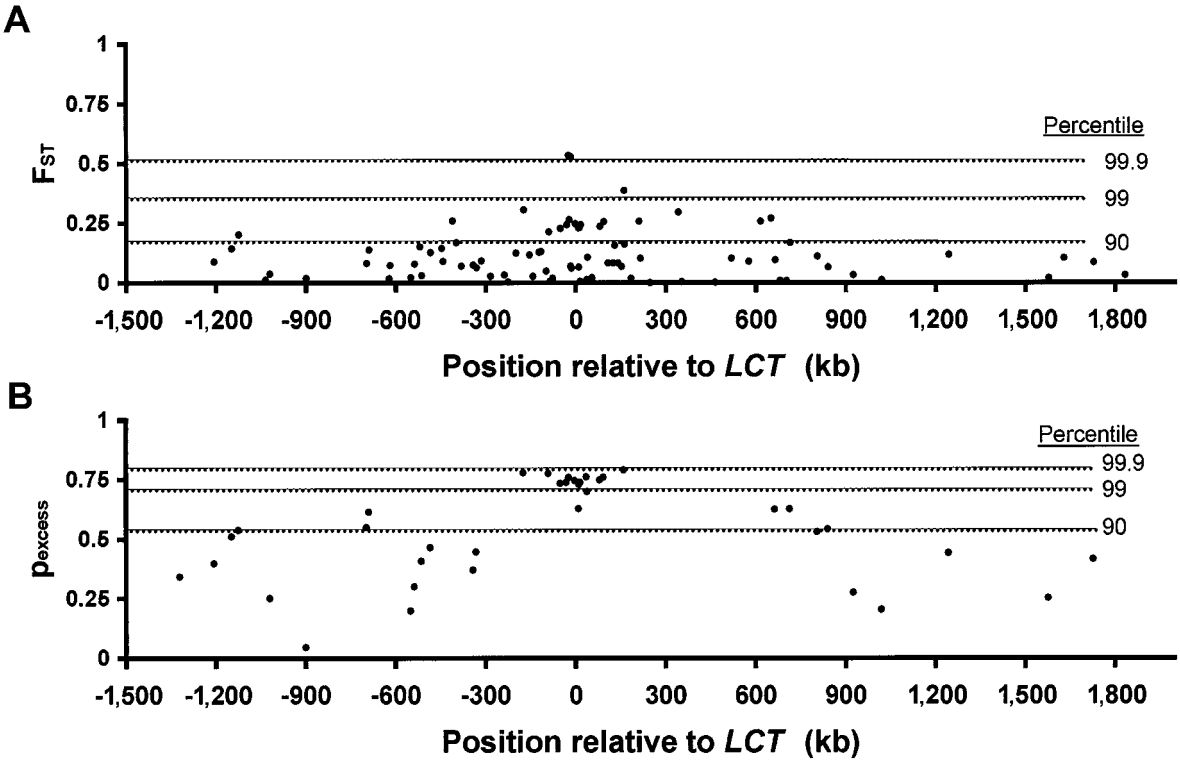

Figure 1 Elevation in (*a*) F_{ST} and (*b*) p_{excess} at multiple SNPs in a 3.2-Mb region around the *LCT* gene. Position in kb relative to the start of transcription of *LCT* is on the *X*-axis. The 90th, 99th, and 99.9th percentiles for F_{ST} and p_{excess} are indicated by dashed lines and are based on 28,440 and 13,696 markers, respectively, throughout the genome (see the "Subjects and Methods" section).

the "Subjects and Methods" section for details). These results further mark the *LCT* region as very unusual when compared with the remainder of the genome, and they strongly suggest that genetic hitchhiking due to selection has occurred: that is, a selected allele rose in frequency over such a short time period that the frequencies of linked alleles on the surrounding >1 Mb haplotype were dragged up as well (Braverman et al. 1995).

In addition to the tests above, which are measures of differentiation between populations, we also employed the recently described long-range haplotype test of Sabeti et al. (2002), which detects selection by measuring the characteristics of haplotypes within a single population. A recent haplotype should be surrounded by long stretches of homozygosity, since recombination will have had few opportunities to juxtapose adjacent segments from other chromosomes with the selected haplotype. The evidence for selection is a haplotype that arose recently—as evidenced by long flanking stretches of homozygosity—but is so common that the haplotype could not have risen quickly to such high frequency without the aid of selection. We observed precisely this pattern at the haplotype containing the lactase-persistence–associated alleles −13910T and −22018A. The haplotype containing these alleles was very common (77% in European Americans) but also largely identical over nearly 1 cM (>800 kb), indicating a recent origin (*red bars* in fig. 2). This long stretch of homozygosity was not simply due to a low local recombination rate—the other haplotypes in this region show shorter extents of homozygosity, indicating abundant historical recombination (*blue bars* near the bottom of fig. 2), and the recombination rate in this region is typical of that in the genome as a whole (Kong et al. 2002).

To formally assess the significance of these results, we focused on the REHH statistic (Sabeti et al. 2002); REHH values much greater than 1 indicate increased homozygosity of a haplotype compared with other haplotypes in the region. For the lactase-persistence–associated haplotype, REHH was 13.2 in the region 3′ to *LCT,* indicating much less breakdown of homozygosity at the persistence-associated haplotype than at haplotypes not carrying the persistence-associated alleles. We compared the *LCT* data to data from coalescent population-genetics simulations analogous to those in Sabeti et al. (2002), and the empirical *P* value for excess homozygosity 3′ to *LCT* was .0004 (fig. 3 and the "Subjects and Methods" section); other estimates of significance suggest a *P* value closer to 10^{-7} (see the "Subjects

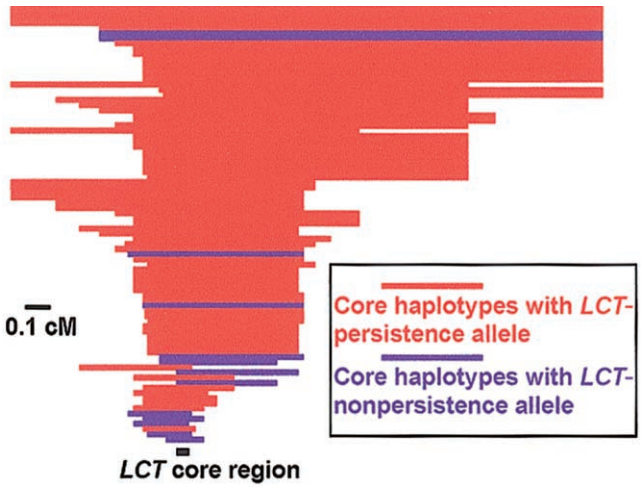

Figure 2 Long-range extended homozygosity for the core haplotype containing the persistence-associated alleles at *LCT* at various distances from *LCT*. The extent to which the common core haplotypes remains intact is shown for each chromosome in cM. The core region containing −13910C/T is shown as a black bar, and the *LCT* gene is oriented from left to right. Core haplotypes containing the persistence-associated allele (−13910T) are shown in red, and those containing the non-persistence–associated allele (−13910C) are shown in blue. Haplotypes are from European-derived U.S. pedigrees; all chromosomes with core haplotypes having a frequency ⩾5% in this population are depicted.

and Methods" section). As confirmation, we compared the *LCT* haplotype to actual genotype data from 12 control regions spanning 500 kb each. The distribution of REHH was similar for the control regions and the simulations, and the *LCT* haplotype had a higher REHH than any of the matched control haplotypes. It is notable that the signal for selection is much stronger for *LCT* than for the well-established case of *G6PD*—although higher haplotype frequencies are in general associated with lower REHH values (Sabeti et al. 2002) (fig. 3), we observe a larger REHH statistic for the 77% *LCT* haplotype (REHH = 13.2) than for the 18% *G6PD* haplotype (REHH = 7) (see Sabeti et al. 2002). Although we cannot rule out the possibility that the extended homozygosity of the high-frequency *LCT* haplotype is due to dominant suppression of recombination over Mb distances because of an allele on this haplotype, positive selection seems to be a more biologically plausible phenomenon, especially since the haplotype has such a strikingly wide spread of frequencies across European populations. Furthermore, the parental core haplotype on which the persistence-associated alleles arose is present in Asian and African American populations, and it does not have an elevated REHH value (data not shown).

We next estimated the age of the lactase-persistence–associated haplotype, on the basis of the decay of haplotypes in either direction from the *LCT* core region (Reich and Goldstein 1998; Stephens et al. 1998). On the basis of our analysis of European-derived U.S. pedigrees, the best estimates of the time at which the persistence-associated haplotype began to rise rapidly in frequency are between 2,188 and 20,650 years ago, consistent with the estimated origin of dairy farming in northern Europe ~9,000 years ago (Simoons 1970; Kretchmer 1971; Scrimshaw and Murray 1988). Even more recent estimates (1,625–3,188 years ago) were obtained by analyzing a Scandinavian population of parent-offspring trios, suggesting stronger and more recent selection in this population. On the basis of these ranges of ages, we estimate the coefficient of selection associated with carrying at least one copy of the lactase-persistence allele to be between 0.014 and 0.15 for the CEPH population and between 0.09 and 0.19 for the Scandinavian population (see the "Subjects and Methods" section for details). By comparison, the selective advantage in a region endemic for malaria has been estimated at 0.02–0.05 for G6PD deficiency (Tishkoff

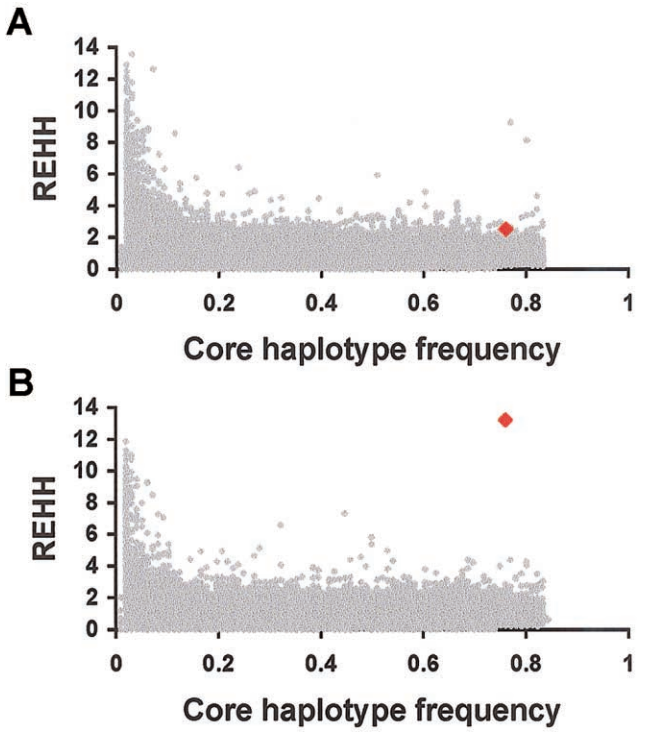

Figure 3 REHH, a measure of extended haplotype homozygosity, plotted for the persistence-associated haplotype at *LCT*, in comparison with REHH from haplotypes in 10,000 sets of simulated data (Sabeti et al. 2002). Data are shown using markers (*a*) 5′ and (*b*) 3′ to the core region. Data for the *LCT*-persistence-associated haplotype are indicated by red symbols, and data from simulations are indicated by gray symbols. REHH distributions from actual genotypes for 12 control regions were consistent with the simulated distributions (data not shown).

et al. 2001) and 0.05–0.18 for the sickle-cell trait (Li 1975). Thus, the added nutrition from dairy appears to have provided a selective advantage in northern Europe comparable to that provided by resistance to malaria in malaria-endemic regions.

Discussion

We have now demonstrated, on the basis of three different analytic methods (elevated F_{ST} at markers associated with lactase persistence, runs of elevated p_{excess} at flanking markers, and extended haplotype homozygosity), that strong positive selection occurred in a large region that includes the *LCT* gene. This selection occurred after the separation of European-derived populations from Asian- and African-derived populations, and it likely occurred after the colonization of Europe. The high frequency and young age of this haplotype, the high estimated coefficient of selection, and the very high REHH value all suggest that *LCT* represents one of the strongest signals of recent positive selection yet documented in the genome. Our results strongly support the hypothesis that the additional nutrition provided by dairy was very important for survival in the recent history of Europe and perhaps in other regions of the world as well.

Our results show that chromosomes carrying the allele associated with lactase persistence (−13910T) share a very long haplotype around this allele. We and others have noted that the presence of this long haplotype raises the possibility that a variant located somewhere in this large region, other than −13910C→T, could be the cause of lactase persistence (Grand et al. 2003; Poulter et al. 2003). Indeed, Swallow and colleagues have identified an individual who is homozygous for the nonpersistence-associated allele at −13910C→T but retains lactase activity (Poulter et al. 2003). Recently, Olds and Sibley (2003) demonstrated differential in vitro transcriptional activity between short segments of DNA carrying the C and T alleles, but the predictive value of such in vitro data for the in vivo phenotype remains uncertain. A comprehensive assessment of variation throughout this long haplotype may be required to determine if −13910C→T is truly the causal polymorphism. Of course, it is also possible that the strong signature of selection is not due to variation at *LCT* but rather to a coincidental selective event acting on a nearby unrelated gene. However, the striking geographic correlation of lactase persistence with dairy farming (Simoons 1969; Kretchmer 1971; Scrimshaw and Murray 1988) and the recently described evidence of selection on cattle-milk protein genes in regions of Europe with a high prevalence of lactase persistence (Beja-Pereira et al. 2003) lend strong support to the dairy hypothesis.

The −13910T allele was rare or absent in the sub-Saharan African populations we tested, indicating that the presence of the T allele in African Americans that we and Enattah et al. (2002) observed is probably explained by admixture of European-derived chromosomes into the African American population (Parra et al. 1998). Thus, our data do not provide evidence that the −13910T allele predates the differentiation of European and African populations. The absence of the T allele in African populations also suggests that either −13910C/T is not the causal allele or that lactase persistence arose multiple times, because lactase persistence is prevalent in a number of African populations (Scrimshaw and Murray 1988). Consistent with these suggestions, the study by Mulcare and colleagues (in this issue of the *Journal*) showed that the −13910T allele was absent from several African populations known to have high rates of lactase persistence (Mulcare et al. 2004 [in this issue]). We did not specifically survey these populations, but such surveys will help determine whether lactase persistence arose multiple times in human history or whether a single very old polymorphism rose independently to high frequencies in multiple populations, as has been suggested (Enattah et al. 2002). Finally, the T allele was present at high frequencies in Pakistan and at somewhat lower frequencies in Middle Eastern populations (table 1) and was found on the same local haplotype in these populations as in Europeans (data not shown). These data suggest that individuals carrying the lactase-persistence allele might have migrated between populations (perhaps along with dairy farming), and their descendants may be responsible for the increased allele frequencies in diverse populations in Europe and neighboring regions.

More generally, we have implemented two methods of detecting signatures of positive selection: runs of consecutive markers with elevated p_{excess} and the long-range haplotype test. It is important to note that these two tests identified *LCT* as strikingly unusual because *LCT* was at the far extreme of the genomewide distribution. With the availability of data for loci throughout the genome, empirical comparisons of individual loci to the genomewide distribution will distinguish other genes that are in the extreme tail of the distribution and, thus, are likely to have experienced selection. Ideally, the metrics will be compared not only to an empirical distribution but also to a simulated distribution derived from an appropriate model of recent human evolution that is consistent with empirical data. As models that incorporate more-complete descriptions of human history are developed, such simulations will become more useful.

Both of these methods should be readily applicable to genomewide SNP genotype data being generated by the haplotype map of the human genome (HapMap

Project Web site). In particular, runs of markers with consistently elevated p_{excess} should be detectable once an adequate number of SNPs have been genotyped in multiple populations; our experience with *LCT* suggests that these runs of elevated p_{excess} may be more informative than signals from individual markers with high F_{ST} values, particularly where selection has dramatically increased the frequency of a single haplotype. The long-range haplotype test should also be useful, even in studies of a single population. Thus, it should be possible in the near future to identify many other loci that have undergone recent positive selection, leading to new insights into recent human evolution and also human disease.

Acknowledgments

D.E.R. and J.N.H. are recipients of Burroughs Wellcome Career Awards in Biomedical Sciences. We thank Richard Grand, Robert Montgomery, Eric Lander, David Altshuler, Helen Lyon, and members of the Hirschhorn Lab for useful comments and discussion.

Electronic-Database Information

The URLs for data presented herein are as follows:

Coriell Institute for Medical Research, http://locus.umdnj.edu/ccr/
dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/
HapMap Project, http://www.hapmap.org
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *HBB, G6PD, FY, TNFSF5, CKR5, HFE, ADH1B, CFTR,* and *LCT*)
PHASE, http://www.stat.washington.edu/stephens/phase.html
SNP Consortium (TSC) Web Site, http://snp.cshl.org/allele_frequency_project/
UCSC Genome Bioinformatics, http://genome.ucsc.edu

References

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

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES (2000) The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 26:76–80

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111

Bayless TM, 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: 968–972

Beja-Pereira A, Luikart G, England PR, Bradley DG, Jann OC, Bertorelle G, Chamberlain AT, Nunes TP, Metodiev S, Ferrand N, Erhardt G (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. Nat Genet 35:311–313

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–96

Cavalli-Sforza L (1973) Analytic review: some current problems of population genetics. Am J Hum Genet 25:82–104

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237

Flatz G (1987) Genetics of lactose digestion in humans. In: Harris H, Hirschhorn K (eds) Advances in human genetics. Vol 16. Plenum Press, New York, pp 1–77

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229

Grand RJ, Montgomery RK, Chitkara DK, Hirschhorn JN (2003) Changing genes; losing lactase. Gut 52:617–619

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70:369–383

Hartl D, Clark A (1997) Principles of population genetics. Sinauer Associates, Sunderland, MA

Hastbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Trivedi B, Weaver A, Coloma A, Lovett M, Buckler A, Kaitila I, Lander ES (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–87

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM (2001) Lactase haplotype diversity in the Old World. Am J Hum Genet 68: 160–172

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kretchmer N (1971) Memorial lecture: lactose and lactase—a historical perspective. Gastroenterology 61:805–813

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–195

Li WH (1975) The first arrival time and mean age of a deleterious mutant gene in a finite population. Am J Hum Genet 27:274–286

Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, 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 74:1102–1110 (in this issue)

Nei M, Chesser R (1983) Estimation of fixation indices and gene diversities. Ann Hum Genet 47:253–259

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

Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi A, Okonofua F, Parnas J, Schulz LO, Bertranpetit J, Bonne-Tamir B, Lu RB, Kidd JR, Kidd KK (2002) A global perspective on genetic variation at the *ADH* genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet 71:84–99

Pagnier J, Mears JG, Dunda-Belkhodja O, Schaefer-Rego KE, Beldjord C, Nagel RL, Labie D (1984) Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc Natl Acad Sci USA 81:1771–1773

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63: 1839–1851

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, 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 67:298–311

Reich DE, Goldstein DB (1998) Estimating the age of mutations using the variation at linked markers. In: Goldstein DB, Schlotter C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Rosner B (1982) Fundamentals of biostatistics. Duxbury Press, Boston, MA

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Scrimshaw N, Murray E (1988) The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. Am J Clin Nutr 48:1079–1159

Simoons F (1969) Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. I. Review of the medical research. Am J Dig Dis 14: 819–836

——— (1970) 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: 695–710

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, et al (1998) Dating the origin of the *CCR5-Δ32* AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001) Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. Science 293:455–462

Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165:287–297

Wiuf C (2001) Do ΔF508 heterozygotes have a selective advantage? Genet Res 78:41–47