Supplementary Discussion

Identification of additional SNPs associated with the LTT phenotype

Although four additional SNPs were significantly correlated with the LTT phenotype in Kenyan NS populations (rs749017, rs3806502, rs3087343, rs3806502; **Fig. 3e, Supplementary Table 2** online), the alleles associated with the phenotype were in strong LD with the C -14010 allele in this population, and after removal of all individuals with C-14010 alleles, these SNPs were not significantly associated with the phenotype. Two additional SNPs (rs4988226 and rs3769005) showed a significant association in the Beja (**Fig. 3e**). However, these common SNPs are not associated with the phenotype in any other population, making them unlikely to be LCT regulatory mutations. In addition, no SNP other than C-14010 was significantly associated with the phenotype using a meta-analysis, which combines power from the individual population tests. (**Fig. 3f**)

Inference of historic migration events

Because the lactase persistence-associated mutations are geographically restricted, they are also particularly informative for inferring historic migrations. The absence of the European T- 13910 mutation in East Africa indicates low levels of recent gene flow to that region. The presence of the G-13907 and G-13915 mutations predominantly in the northern Cushitic-speaking Beja from northern Sudan and in eastern Cushitic-speaking northern Kenyans indicates common ancestry and/or gene flow between these groups. The C-14010, G-13915, and G-13907 alleles were absent in 19 Lebanese (one European T-13910 allele was observed only in this population), 16 Ethiopian Jews, and 24 Nigerian Yorubans. Only the G-13915 mutation was observed at 14% frequency in 22 Yemenite Jews, indicating historic gene flow between Semitic-speaking Arabian and Cushitic-speaking East African populations, possibly associated with the Axumite kingdom (*ca*. 1,900-1,300 years ago) that extended from modern Ethiopia and surrounding regions in Africa to the southern Arabian peninsula (Alison Brooks, personal communication). Genotyping of these SNPs in a broader sample of Middle Eastern, North African, and East African populations will be particularly informative for reconstructing historic migration events in these regions.

Adaptive significance of lactase persistence in the click-speaking hunter-gatherers

Surprisingly, the Hadza population of Tanzania, who speak a click-language and subsist by hunting and gathering, have the lactase persistence phenotype at ~50% frequency (**Fig. 2a**). The neighboring Sandawe population, also hunter-gatherers who speak a click language (although highly divergent from the Hadza language), have a low level of lactase persistence, as expected for non-pastoralists, which correlates with the frequency of the C-14010 allele. The latter observation suggests that the C-14010 allele was introduced to the Sandawe due to recent admixture with neighboring pastoralist/agro-pastoralist populations. By contrast, the C-14010 mutation is entirely absent in this sample of Hadza.

The high frequency of lactase persistence in the Hadza suggests 1) that these hunter-gatherers descend from a pastoralist population, 2) that there could be a selective role of lactase persistence to delay weaning and increase the timing between births[1] (but note that a mutation that allows an overall reduction in the number of offspring is expected to be under negative selection, unless it increases the likelihood of survivability of existing children), or

3) that the lactase persistence trait may be adaptive for something other than milk digestion. Indeed, the LPH enzyme also plays a role in hydrolysis of phlorizin, a bitter glycoside present in the roots and bark of plants in the *Rosaceae* family, some of which are indigenous to Tanzania[2]. Thus, it is possible that LPH activity is adaptive in the hunter-gatherers due to facilitating digestion of phlorizin.

Furthermore, the span of haplotype identity in the Sandawe (1.6 cM, Table 1) indicates that the C-14010 associated haplotype has had a long period of time to recombine relative to other populations (the estimated age of the C-14010 mutation is 5.7 or 6.9 kya, although there is a large CI of 1.3 – 23.3. kya (Table 1)). While it is likely that the C-14010 mutation was introduced into the Sandawe via gene flow from neighboring Cushitic populations who are thought to have migrated into the region ~5,000 years ago[3], we can't rule out the possibility that this mutation arose in the Sandawe (or their Khoisan ancestors) and spread to the AA and NS populations where it experienced even stronger selection. It is interesting to note a report of a moderate level of lactase persistence (~8%) in southern African Khoisan hunter-gatherers[4, 5]. These observations, together with the cultural, historical and archaeological evidence of several thousand years of pastoralism in some southern African Khoisan populations (e.g. the Khoikhoi or Nama) raises the possibility that these click-speaking modern-day hunter-gatherer groups may have practiced pastoralism in the past or fluctuated between pastoralism and foraging[5-7]. However, this is a subject of controversy[8] and it is also possible that the lactase persistence trait in these hunter-gatherer groups results from recent gene flow with neighboring pastoralist groups. Additional studies of lactase persistence in a larger number of Khoisan-speakers, including the Hadza, will be informative for addressing questions regarding ancient subsistence patterns in these populations.

Regardless of the population origins of the C-14010 mutation, it has not reached fixation in any population, either because it is so recent and/or because cultural practices, such as fermentation of milk, which facilitate digestion of dairy products, has relaxed the intensity of selection on these alleles in the more recent past.

Evidence for a secondary selective sweep in the LCT gene region

There are a number of genes within 500 kb of *MCM6/LCT* including *CXCR4*, *DARS*, *CXCR4*, *UBXD2*, and *R3HDM* which could be target of a secondary selective sweep. In support of a secondary sweep in this region, we find evidence for a tract of EHH (on the same background as haplotype E, **Fig. 7a**) centered on rs2322813, located ~12 kb downstream of the *LCT* start codon, and extending ~1.3 Mbp (~0.5 cM) on chromosomes with the ancestral G -14010 allele in several populations (**Fig. 6; Supplementary Fig 2** online). Although this mutation is not significantly associated with lactase persistence in the current study ($P = 0.87$ after Bonferroni correction for 123 SNPs), it is possible that further resequencing studies will identify an *LCT* regulatory mutation on that haplotype background.

Effect of C-14010, G-13915, and G-13907 on transcript expression from the *LCT* promoter

It is likely that *in vivo* expression levels associated with the derived C-14010, G-13915, and G-13907 alleles are even higher then the expression levels observed *in vitro* (~18% – 30% greater then for the ancestral alleles). Previous studies have shown undifferentiated Caco-2 cells having lower levels of transcription of LCT reporter constructs

then differentiated Caco-2 cells. For example, the European T-13910 allele was even more active (3-6 fold) relative to the ancestral allele in these differentiated cells[9]. Also, other important components of regulation of lactase gene expression present *in vivo* may be absent *in vitro* because the flanking sequences outside of intron 13 are not included in the construct and because this is a transient transfection which may be lacking regulatory mechanisms based on chromatin structure[10]. Furthermore, LCT expression is very localized and tissue specific, being highly expressed only in the brush border cells of the small intestine[11]. While the Caco-2 cell line is derived from the small intestine, and was the cell type used in previous studies of LCT transcription[10], it is not the cell type in which LCT is most highly expressed *in vivo*.

1. Hollox, E. & Swallow, D. M. in The Genetic Basis of Common Diseases (eds. King, R. A., Rotter, J. I. & Motulsky, A. G.) 250 - 265 (Oxford University Press, Oxford, 2002).
2. Swallow, D. M., Poulter, M. & Hollox, E. J. Intolerance to lactose and other dietary sugars. Drug Metab Dispos 29, 513-6 (2001).
3. Newman, J. The Peopling of Africa (Yale University Press, New Haven and London, 1995).
4. Nurse, G. T. & Jenkins, T. Lactose intolerance in San populations. British Medical Journal 2, 728 (1974).
5. Casimir, M. J. On Milk-drinking San and the "Myth of the Primitive Isolate." 31, 551-554 (1990).
6. Denbow, J. R. & Wilmsen, E. N. Advent and Course of Pastoralism in the Kalahari. . Science 234, 1509 - 1515 (1986).
7. Smith, A. B. Pastoralism in Africa: Origins and Development Ecology. (Hurst and Company, London, 1992).
8. Sadr, K. Kalahari Archaeology and the Bushman Debate. Current Anthropology 38, 104-112 (1997).
9. Troelsen, J. T., Olsen, J., Moller, J. & Sjostrom, H. An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology 125, 1686-94 (2003).
10. Olds, L. C. & Sibley, E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet 12, 2333-40 (2003).
11. Swallow, D. M. Genetics of lactase persistence and lactose intolerance. Annu Rev Genet 37, 197-219 (2003).