SUPPLEMENTARY METHODS

Table of Primers Used

Primers used to amplify introns 9 and 13 of MCM6

MCM6 Intron 9 Primers	**Sequence[1,2]**
MCM6-9-forward	CCGAGGGAGAGAAACCTTC
MCM6-9-reverse	TCAACAAGGCTATGGACGATG
PCR MCM6 Intron 13 Primers	**Sequence[1,2]**
First fragment	
MCM6-13-forward	ATCTCCGCCAGAGAGATGG
MCM6-13seq3-reverse	TCATAGATGTTTTCAATTCTTCAAGT
Second fragment	
MCM6-13seq4-forward	GGATCTCCTTTTGGACTTTCC
MCM6-13seq6-reverse	TGGACCTAAACCAATAATGATGAA

Primers used to sequence introns 9 and 13 of MCM6

MCM6 Intron 9 Primers	**Sequence[1]**
MCM6-9seq1-forward	ACCAGTGGTAAAGCGTCCAG
MCM6-9seq1-reverse	AACAGCAAACACACGTGCTC
MCM6-9seq2-forward	TGCATTGAGCCAAGATTGTG
MCM6-9seq2-reverse	TAGCCAGGTGTGGTGGTGTG
MCM6-9seq3-forward	TCCCTGTGGTAGCAGACTTTG
MCM6-9seq3-reverse	TCCCGCACGTCCATCTTATC
MCM6 Intron 13 Primers	**Sequence[1]**
MCM6-13seq1-forward	ATCTCCGCCAGAGAGATGG
MCM6-13seq1-reverse	GCTTTGGTTGAAGCGAAGAT
MCM6-13seq2-forward	GTTCTTTGAGCCCTGCATTC
MCM6-13seq2-reverse	AGGTTCGGGGGTACACATGC
MCM6-13seq3-forward	AGATACCCTGGGACAAGGTC
MCM6-13seq3-reverse	TCATAGATGTTTTCAATTCTTCAAGT
MCM6-13seq4-forward	GGATCTCCTTTTGGACTTTCC
MCM6-13seq4-reverse	TTCAACAAGAAACACTGAAAAACA
MCM6-13seq5-forward	GTGAGCCATGTGCTTTCTCC
MCM6-13seq5-reverse	GCACGGTGGCTCATGTCTAT
MCM6-13seq6-forward	TCTTCTTTCTCAGCCTCCTG
MCM6-13seq6-reverse	TGGACCTAAACCAATAATGATGAA

Primers used in vector construction

core promoter	**Sequence[1]**
F1_XhoI	ATTCGCTCGAGACTTTCCAAGG
R1_HindIII	ATTTTCTAGGAAGCTTTAGGAGGTATGTG
intronic regions	**Sequence[1]**
F1_SacI	ATATAGCTAGCTCATAGATGTTTTCAATTCTT
R1_NheI	TTATATAGAGCTCATCTCCGCCAG

1. all sequences are listed 5' to 3'
2. an annealing temperature of 61.6° was used.

Phenotype Test. 130 Tanzanian individuals consumed 0.068 ml alcohol/lb weight, in addition to the lactose load, in order to perform a Lactose Tolerance Test with Ethanol (LTTE)[1]. Ethanol blocks the conversion of galactose to glucose in the liver, resulting in more stable estimates of the levels of glucose and galactose in the blood. Because we were unable to accurately test for galactose in the field, and because some individuals did not consent to the test due to religious restrictions, we did not continue this test in Kenya or the Sudan. With a single factor ANOVA, only 0.15% of the phenotype variation is correlated with alcohol consumption, which is not significant (F=0.69, P=0.41). There are two cut-off values commonly given in the literature for classifying individuals as "Lactase Persistent". We used the more conservative definition: a rise of >1.7 mM was classified as "Lactase Persistent", a rise of < 1.1 mM was classified as "Lactase Non-Persistent", a rise of 1.1 – 1.7 mM is ambiguous and was classified as "Lactase Intermediate Persistent"[1]. The more liberal definition classifies individuals with a glucose rise of >1.4 mM as "Lactase Persistent"[2].

Dominance Estimates. C-14010 and G-13907 both behave as incomplete dominant alleles in regard to the LTT phenotype, whereas G-13915 behaves as an overdominant allele[3]. Thus, individuals heterozygous for C-14010 and G-13907 tend to have an intermediate rise in blood glucose relative to homozygous individuals, whereas G-13915 heterozygotes have a slightly higher rise in blood glucose then either homozygote, although there is considerable variance within each genotype class and our sample size for G-13907 and G-13915 is small (**Fig. 2b).**

To estimate the degree of dominance from the LTT phenotype data, the heterozygous genotype's x-axis value was adjusted and the point of maximum r^2 recorded, which was used as an estimate of the dominance parameter, h. To assess a significant elevation above additive co-dominance (h=0.5) a linear regression was performed with phenotype values pooled for the homozygote class at one x-axis value and heterozygote phenotypes assigned to another x-axis value. Dominance was also calculated following the framework of Falconer and MacKay[3]; here, the average deviation of the heterozygotes from the average of the averages of the two homozygote classes is estimated (d) and this is scaled by half the differences of the two homozygote phenotypes (a). We observed that heterozygotes have a significantly elevated phenotype distribution relative to the average of all homozygotes (r^2=0.18, P=1.9x10^{-12} for G/C-14010 and r^2=0.031, P=0.00022 for C/G-13907 and r^2=0.099, P=4.6x10^{-7} for T/G-13915), providing statistical support for a dominant (or overdominant) mode of inheritance. The degree of dominance in a quantitative genetics sense (d/a) was converted to h (the population genetic dominance parameter) according to the following formula,

$$h = \frac{1 + d/a}{2}.$$

Both methods resulted in essentially identical estimates of h for G/C-14010 and C/G-13907, h=0.62 and h=0.61 for G/C-14010 respectively, h=0.81 and h=0.81 for C/G-13907. However for T/G-13915 h=1.73 in the preceding "regression" method and h=1.25 in the latter "quantitative" method. Note that some of the T/G-13915 heterozygotes have flanking lactase persistence-associated mutations which may contribute to a false overdominant pattern and for both T/G-13915 and C/G-13907, the sample size for derived homozygotes is very small (see Figure 2).

LD plot shown in Supplementary Figure 2. A plot was constructed to visualize the extent of LD in the region using Haploview 3.1.1 software[4].

Vector Construction, transfection and expression assay. The lactase "core" promoter, starting 3083 bp upstream of *LCT* at position -3 of the transcription start site, was PCR amplified using high-fidelity Phusion polymerase (Finnzyme, Espoo, Finland). PCR products were then cloned and ligated into a pGL3-Basic luciferase reporter (Promega, Madison, Wisconsin, United States) using primers containing *Xho*I and *Hind*III restriction cutsites (see **Table of Primers Used** above). Five variant constructs containing intronic regions were constructed by cloning 2035 bp of the 13th intron of *MCM6*, beginning at position -14,354 bp relative to *LCT,* 5' of the "core" promoter using primers containing *Sac*I and *Nhe*I restriction cutsites (see **Table of Primers Used** above). All vectors were confirmed by bidirectional sequencing. Caco-2 cells were cultured at 37°C and 5.5% CO_2 in Eagle's Minimum Essential Medium with Earles Balanced Salt Solution, non-essential amino acids, 2 mM L-glutamine, 1 mM sodium pyruvate, and 1500 mg/L sodium bicarbonate, and supplemented with 20% FBS. 48 hours prior to transfection, $3x10^4$ cells per well were seeded in 500 µl of media in a 24 well plate. Transfections were performed at 45-55% confluency following the manufacturer's protocol with a mixture of 0.6 µl FuGENE 6 (Roche Applied-Science, Indianapolis, Indiana, United States), 19.4 µl OPTI-MEM, 0.029 pmol of pGL3basic vector, and 10 ng of *Renilla*-TK (Promega) as a co-reporter. 48 hours after transfection each well (85-95% confluent) was lysed with 100 µl 1x PLB (Promega). Luciferase activity was measured using the Dual-Luciferase Reporter Assay (DLR) System (Promega) and a Veritas Microplate Luminometer (Turner BioSystems, Sunnyvale, California, United States) using a two second measurement delay and a ten second measurement. Measurements are reported as ratios of Luciferase:*Renilla* and expression was assessed using paired t-tests. Transfections of cells were performed six times for control and "core" promoters and 12 times for vectors with the intron from *MCM6*. Note that although the magnitude of the difference in expression driven by the derived and the ancestral haplotypes was consistent across eight replicate assays (range ~15-20%), a difference in expression of this range approaches the boundary of sensitivity for the DLR assay.

Estimating selection intensity and sweep ages. We applied a rejection-sampling approach using the centiMorgan (cM) span surrounding the selected site to estimate selection intensity and ages of the candidate lactase persistence-associated mutations for each population[5]. The algorithm has the following form: (1) calculate the cM span (distance at which EHH = 0.25) for a population, (2) draw a selection parameter, σ=2Ns, from its prior distribution, (3) simulate data under a coalescent framework using the chosen value of σ[6] and compute the value of the cM span for each simulated data set (4) if the simulated cM span was within 0.005 of the observed data, accept the simulated value of sigma as well as the age of the mutation; otherwise, reject the simulated data set. 100,000 simulations for each population were generated from a uniform prior of the selection parameter (0 to 3000). The model assumes that a new mutation at a specified position in the sequence experiences a constant selection pressure σ=2Ns, where N is the population size and ***s*** is the selective advantage per copy per generation. Point estimates (and confidence intervals) for the selection intensity and ages are presented, assuming an additive or fully dominant fitness effect Although our model assumes constant population size, previous studies have demonstrated that for an

allele that rapidly increases in frequency, population demographic history has only a modest effect on allele age estimates[7, 8].

Due to the way that SNPs were ascertained, the allele frequency spectrum departs from the expectation for DNA sequence data. To model the effect of ascertainment bias of SNPs selected for genotyping, we followed the approach in Voight et al.[9]. This procedure simulates data with complete ascertainment and then uses rejection sampling to obtain the observed frequency spectrum (separately for each population sample). The procedure can be thought of as an empirical model of the unknown ascertainment process. In addition, the observed data varies in terms of SNP density: a dense central core region is flanked by regions with lower SNP density (on average). To match this feature of the data, a secondary rejection step was applied such that the average SNP density for central and flanking regions (both left and right) matched the observed density. With respect to recombination, for each simulation we chose to exactly match the recombination map estimated from the data, using the Li and Stephens algorithm[10] For all populations, we calculated cM spans assuming the estimated population genetic map for the Yoruba Hapmap dataset[11], and calculated those distances assuming the rates estimated from the deCODE genetic map across 40Mb flanking this region on chromosome 2[12].

References

1. Arola, H. Diagnosis of hypolactasia and lactose malabsorption. Scand J Gastroenterol Suppl 202, 26-35 (1994).
2. 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).
3. Falconer, D. S. & MacKay, T. F. C. Introduction to Quantitative Genetics (Prentice Hall, New York, 1996).
4. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263-5 (2005).
5. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution 16, 1791-1798 (1999).
6. Spencer, C. C. & Coop, G. SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics 20, 3673-5 (2004).
7. Tishkoff, S. A. et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293, 455-62 (2001).
8. Wiuf, C. Recombination in human mitochondrial DNA? Genetics 159, 749-56 (2001).
9. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. PLoS Biol 4, e72 (2006).
10. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213-33 (2003).
11. The International HapMap Consortium. A haplotype map of the human genome. Nature 437, 1299-1320 (2005).
12. Kong, A. et al. A high-resolution recombination map of the human genome. Nat Genet 31, 241-7 (2002).