Published in final edited form as:

Am J Phys Anthropol. 2015 April ; 156(4): 661–664. doi:10.1002/ajpa.22675.

New insights into the history of the C-14010 lactase persistence variant in Eastern and Southern Africa

Enrico Macholdt¹, Montgomery Slatkin², Brigitte Pakendorf^{3,*}, and Mark Stoneking^{1,*}

¹Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

²Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

³Laboratoire Dynamique du Langage, UMR5596, CNRS and Université Lyon Lumière 2, 69007 Lyon, France

Abstract

Lactase persistence (LP), the ability to digest lactose into adulthood, is strongly associated with the cultural traits of pastoralism and milk-drinking among human populations, and several different genetic variants are known that confer LP. Recent studies of LP variants in Southern African populations, with a focus on Khoisan-speaking groups, found high frequencies of an LP variant (the C-14010 allele) that also occurs in Eastern Africa, and concluded that the C-14010 allele was brought to Southern Africa via a migration of pastoralists from Eastern Africa. However, this conclusion was based on indirect evidence; to date no study has jointly analyzed data on the C-14010 allele from both Southern African Khoisan-speaking groups and Eastern Africa. Here, we combine and analyze published data on the C-14010 allele in Southern and Eastern African populations, consisting of haplotypes with the C-14010 allele and four closely-linked short tandem repeat (STR) loci. Our results provide direct evidence for the previously-hypothesized Eastern African origin of the C-14010 allele in Southern African Khoisan-speaking groups. In addition, we find evidence for a separate introduction of the C-14010 allele into the Bantu-speaking Xhosa. The estimated selection intensity on the C-14010 allele in Eastern Africa is lower than that in Southern Africa, which suggests that in Eastern Africa the dietary changes conferring the fitness advantage associated with LP occurred some time after the origin of the C-14010 allele. Conversely, in Southern Africa the fitness advantage was present when the allele was introduced, as would be expected if pastoralism was introduced concomitantly.

Keywords

pastoralism; selection; migration; Khoisan

Lactase persistence (LP), defined as the ability to digest lactose (the primary sugar found in mammalian milk) into adulthood, is a trait of considerable anthropological interest. LP is

*Correspondence to: Brigitte Pakendorf, Laboratoire Dynamique du Langage, UMR5596, CNRS and Université Lyon Lumière 2, 69007 Lyon, France; Mark Stoneking, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany brigitte.pakendorf@cnrs.fr; stoneking@eva.mpg.de.

strongly associated with pastoralism and milk-drinking in both Eurasia and Africa (Gerbault et al., 2011; Ingram et al., 2009; Itan et al., 2010; Tishkoff et al., 2007). Several genetic variants, located in a regulatory region about 14 kb upstream of the lactase gene, have been identified that increase expression of the lactase enzyme *in vitro* and/or are associated with the LP phenotype (Gerbault et al., 2011). Interestingly, different genetic variants are associated with LP in Eurasian vs. African populations, and several of these exhibit a signature of strong positive selection (Jensen et al., 2011; Jones et al., 2013; Tishkoff et al., 2007); LP variants thus provide an example of convergent evolution in human populations.

Recently, we (Macholdt et al., 2014) and another group (Breton et al., 2014) analyzed variation in the LP regulatory region in Southern African populations, with a focus on Khoisan-speaking groups. Both studies found elevated frequencies of one particular LP variant, the C-14010 allele, in pastoralist and Khoe-speaking groups, as well as evidence for positive selection on this allele in these groups. This allele was previously found at high frequencies in Eastern African populations, and was shown to increase lactase expression *in vitro* and to exhibit a signature of recent positive selection (Tishkoff et al., 2007). Both of the studies of Southern African populations concluded that the C-14010 allele was probably brought to Southern Africa from Eastern Africa via a migration of pastoralists. However, neither study obtained comparable data from Eastern African populations, and instead relied on indirect evidence for an Eastern African origin of the C-14010 variant in Southern Africa. For example, Breton et al. (2014) showed that the C-14010 variant in Southern Africa occurs on a haplotype background that is also found in the Maasai from Kenya. However, the Maasai data did not include information on the allelic state at the -14010 site (International HapMap et al., 2010); it is thus possible that the seemingly identical haplotypes actually differ at position -14010. Moreover, even if one assumes that the Maasai haplotype also carries the C-14010 variant, the existence of the same haplotype in Southern and Eastern Africa is not necessarily informative about the direction of migration. An origin of the C-14010 variant in Southern Africa followed by spread to Eastern Africa remains a formal possibility (albeit perhaps less likely than an Eastern African origin, given other evidence discussed below for a pastoralist migration from Eastern to Southern Africa).

At about the same time that the two Southern African LP studies appeared, a large study of LP variants in African populations was published (Ranciaro et al., 2014). The sampling in this study was mostly limited to Western and Eastern African populations, and Ranciaro et al. discussed in detail (pp.506–507) the need for more data from Southern Africa to address questions related to the origin, distribution, and spread of LP alleles in Southern Africa. Here, we combine our Southern African data (Macholdt et al. 2014) with comparable Eastern African data from Ranciaro et al. (2014), thereby providing a direct comparison of the variation associated with the C-14010 allele in Southern and Eastern Africa. This enables us to draw firm conclusions concerning the direction of migration, number of migrations, and relative strength of selection acting upon the C-14010 variant in Southern and Eastern Africa.

MATERIALS AND METHODS

In addition to the C-14010 allele, both Macholdt et al. and Ranciaro et al. genotyped four short tandem repeat (STR) loci that flank the LP enhancer region; analysis of Eastern African and Southern African haplotypes consisting of the four STR loci and the C-14010 variant forms the basis of our analysis. To combine the data, we first genotyped the STR loci (using the method described previously (Macholdt et al., 2014)), in six samples from the CEPH Human Genetic Diversity Panel that had been genotyped by Ranciaro et al., in order to reconcile the allelic terminology for the STR loci used by Macholdt et al. with that used by Ranciaro et al.

The combined STR genotypes were phased together with the program PHASE v2.1.1 (Stephens and Scheet 2005; Stephens et al., 2001). The frequency of the C-14010 allele across Africa was visualized with the program Surfer ver. 10.4.799 (Golden Software). A median-joining network (Bandelt et al., 1999) of the haplotypes was constructed using NETWORK v.4.611 (<http://www.fluxus-engineering.com>). The age and selection intensity for the C-14010 allele in Eastern Africa were jointly estimated from the STR genotype data by adapting the method of Slatkin (Slatkin 2001) as described in detail elsewhere (Macholdt et al., 2014).

RESULTS AND DISCUSSION

The combined dataset confirms the high frequency of the C-14010 variant in Southern and Eastern Africa (Figure 1A). A median-joining network of the combined haplotype data indicates greater diversity of Eastern African than Southern African haplotypes (Figure 1B). The Southern African haplotypes from our study are closely-related to one another, and one haplotype is shared with an Eastern African from Tanzania (Figure 1B). Overall, the Eastern and Southern African haplotypes with the C-14010 variant are more similar than haplotypes without the C-14010 variant (Macholdt et al. 2014; Ranciaro et al. 2014), suggesting a common origin rather than independent occurrence of the C-14010 variant in Eastern and Southern Africa. Moreover, the network analysis strongly supports an Eastern African source of the C-14010 variant in Southern Africa, as concluded previously from indirect evidence (Breton et al., 2014; Macholdt et al., 2014).

Somewhat surprisingly, the Southern African haplotypes from our study are not closely-related to the Southern African haplotypes from Ranciaro et al. (Figure 1B). The samples in our study come from Namibia, Botswana, and Zambia, and the C-14010 allele occurs at highest frequency in Khoe-speaking and pastoralist groups (Macholdt et al., 2014). The Southern African haplotypes from Ranciaro et al. - from five Bantu-speaking Xhosa individuals and one !Xun/Khoe individual from South Africa (Ranciaro et al., 2014) - are in many cases shared with Eastern African groups and overall appear to have a different origin than the Southern African haplotypes from Namibia/Botswana/Zambia (Figure 1B). Archaeological and linguistic evidence indicate that the ancestors of Nguni speakers (which include the Xhosa) migrated to Southern Africa from Eastern Africa about 1000 years ago (Mitchell 2002), which is more recent than the arrival of other Bantu-speaking groups in Southern Africa. Although Xhosa is a Bantu language, it has acquired several click

consonants, and it has been suggested that this was via intense contact with pastoralist Khoekhoe speakers in what is now South Africa (Herbert 1990). While the Khoekhoe-speaking pastoralists that were historically attested in the Cape have been absorbed into the so-called 'coloured' groups of South Africa, the Nama of Namibia represent a closely related group (Barnard 1992) who originated in what is now South Africa (Güldemann 2008). They therefore represent a good proxy for the hypothesized Khoekhoe-speaking population that the Xhosa are assumed to have admixed with. However, the closer relationship of the Xhosa C-14010 haplotypes to those in Eastern Africa rather than to those in the Nama indicate that the close contact between the Xhosa ancestors and indigenous Khoisan-speaking populations is more likely to have involved a non-Khoekhoe group (cf. Barnard 1992). In particular, the Xhosa C-14010 haplotypes reflect a different, later migration from Eastern Africa than the migration of pastoralists that introduced the C-14010 variant into Southern African Khoisan-speaking groups. However, these inferences concerning the Xhosa are based on just a few samples (Ranciaro et al., 2014); more in-depth studies of LP in South African populations are needed.

Given this distinction between C-14010 haplotypes in the Xhosa vs. other Southern African groups, in the following discussion, “Southern Africa” refers only to the C-14010 haplotypes found in Namibia/Botswana/Zambia, from the Macholdt et al. study. The network (Figure 1B) indicates a single source for the C-14010 variant in Southern Africa, which means that the estimated age of the C-14010 variant in Southern African populations (based on the associated STR variation) can be used as an estimate of when the variant arrived in this region. The estimated age also depends on the strength of the selection coefficient, s ; using a method to estimate both s and the age of the variant (Slatkin 2001), we previously estimated s to be 0.05–0.10 (Figure 2A), with associated ages for the C-14010 variant in Southern Africa of 100–150 generations (Macholdt et al., 2014).

A similar analysis of s for the Eastern African data indicates that the strength of positive selection for the C-14010 variant is somewhat weaker in Eastern Africa than in Southern Africa, with a maximum s of about 0.03–0.04 (Figure 2A). Population expansions can mimic the signature of recent positive selection; however the results shown in Figure 2A were obtained assuming a population growth rate of $r=0.01$ (meaning that the population would double in size every 2,000 years). Virtually identical results were obtained for a constant population size model and for $r=0.02$ (corresponding to a population doubling time of ~900 years), indicating that the inference of selection is robust to reasonable estimates of population growth. Indeed, Bayesian skyline plots based on mtDNA genome sequences indicate constant population sizes or even recent population size decreases for most of the Khoisan-speaking groups in this study (Barbieri et al., 2014). The estimated age of the C-14010 variant in Eastern Africa, given these values of s , is about 200–250 generations (Figure 2B), in keeping with previous estimates (Tishkoff et al., 2007). The older age of the C-14010 allele in Eastern than in Southern Africa thus further supports an Eastern African origin of the allele in Southern Africa.

Overall, these results suggest the following scenario: the C-14010 variant arose somewhere in Eastern Africa some 200–250 generations (6000–7500 years) ago. Selection on this variant was initially weak, which suggests that it may have taken some time for the dietary

changes to occur that would allow the populations to fully realize the fitness benefits of lactase persistence. This is consistent with dates of about 5000 years ago, based on archaeological evidence, for the first presence of domestic cattle in Eastern Africa (Phillipson 2005). Approximately 100–150 generations ago, pastoralists began migrating from Eastern Africa to Southern Africa, bringing with them the C-14010 allele. The comparatively stronger signal of selection on the C-14010 variant in Southern Africa suggests that pastoralism was already entrenched in the migrating population and spread with the C-14010 variant into Southern Africa, and so the fitness benefit of the C-14010 mutation was present from the outset.

Additional genetic support for this putative Eastern African migration comes from a specific Y-chromosome lineage (Henn et al., 2008) as well as from a signal of Eurasian ancestry in Southern African Khoisan-speaking groups that appears to have come from Eastern Africa (Pickrell et al., 2014). Archaeological data also support an introduction of pastoralism from Eastern to Southern Africa (Mitchell 2002; Pleurdeau et al., 2012), while linguistic analyses further indicate that this migration was associated with and/or influenced Khoe-speaking groups (Güldemann 2008). Overall, the direct comparison of haplotypes associated with the C-14010 allele in Eastern African and in Southern African Khoisan-speaking populations provides strong support for the spread of pastoralism from Eastern to Southern Africa via demic diffusion; i.e. via a migration of pastoralists that primarily influenced Khoe-speakers, as concluded previously from indirect evidence (Breton et al. 2014; Macholdt et al. 2014). In addition to this primary migration, subsequent migrations from Eastern Africa brought additional haplotypes with the C-14010 allele to other Southern African populations, such as the Xhosa.

Acknowledgments

We thank Alessia Ranciaro and Sarah Tishkoff for providing their data on the C-14010 allele.

Funding: The Max Planck Society; the Deutsche Forschungsgemeinschaft as part of the European Science Foundation EUROCORES Programme EuroBABEL (B. Pakendorf); US NIH grant R01-GM40282 (M. Slatkin).

Literature Cited

- Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 1999; 16:37–48. [PubMed: 10331250]
- Barbieri C, Güldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, Mpoloka SW, Stoneking M, Pakendorf B. Unraveling the complex maternal history of Southern African Khoisan populations. *American journal of physical anthropology.* 2014; 153:435–448. [PubMed: 24323467]
- Barnard, A. *Hunters and Herders of Southern Africa.* Cambridge University Press; Cambridge: 1992.
- Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. Lactase persistence alleles reveal partial East african ancestry of southern african Khoe pastoralists. *Current biology : CB.* 2014; 24:852–858. [PubMed: 24704072]
- Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow DM, Thomas MG. Evolution of lactase persistence: an example of human niche construction. *Philos T R Soc B.* 2011; 366:863–877.
- Güldemann T. A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *South Afr Humanit.* 2008; 20:93–132.
- Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *P Natl Acad Sci USA.* 2008; 105:10693–10698.

- Herbert RK. The sociohistory of clicks in Southern Bantu. *Anthropological Linguistics*. 1990; 32:295–315.
- Ingram CJE, Mulcare CA, Itan Y, Thomas MG, Swallow DM. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet*. 2009; 124:579–591. [PubMed: 19034520]
- The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–58. [PubMed: 20811451]
- Itan Y, Jones BL, Ingram CJE, Swallow DM, Thomas MG. A worldwide correlation of lactase persistence phenotype and genotypes. *Bmc Evol Biol*. 2010; 10:36. [PubMed: 20144208]
- Jensen TKG, Liebert A, Lewinsky R, Swallow DM, Olsen J, Troelsen JT. The -14010*C variant associated with lactase persistence is located between an Oct-1 and HNF1 alpha binding site and increases lactase promoter activity. *Hum Genet*. 2011; 130:483–493. [PubMed: 21327791]
- Jones BL, Raga TO, Liebert A, Zmarz P, Bekele E, Danielsen ET, Olsen AK, Bradman N, Troelsen JT, Swallow DM. Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. *Am J Hum Genet*. 2013; 93:538–544. [PubMed: 23993196]
- Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, Pakendorf B, Stoneking M. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. *Current biology : CB*. 2014; 24:875–879. [PubMed: 24704073]
- Mitchell, P. *The Archaeology of Southern Africa*. Cambridge University Press; Cambridge: 2002.
- Phillipson, DW. *African Archaeology*. Cambridge University Press; Cambridge: 2005.
- Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A*. 2014; 111:2632–2637. [PubMed: 24550290]
- Pleurdeau D, Imalwa E, Detroit F, Lesur J, Veldman A, Bahain JJ, Marais E. “Of sheep and men”: earliest direct evidence of caprine domestication in southern Africa at Leopard Cave (Erongo, Namibia). *PloS one*. 2012; 7:e40340. [PubMed: 22808138]
- Ranciaro A, Campbell MC, Hirbo JB, Ko WY, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, Tishkoff SA. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *Am J Hum Genet*. 2014; 94:496–510. [PubMed: 24630847]
- Slatkin M. Simulating genealogies of selected alleles in a population of variable size. *Genet Res*. 2001; 78:49–57. [PubMed: 11556137]
- Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*. 2005; 76:449–462. [PubMed: 15700229]
- Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. 2001; 68:978–989. [PubMed: 11254454]
- Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*. 2007; 39:31–40. [PubMed: 17159977]

Figure 1.

Analyses of the C-14010 LP variant and associated STR haplotypes in Eastern and Southern African populations. (A) Surfer map of the C-14010 allele frequency. Red circles denote sampling locations from Macholdt et al.; blue squares denote sampling locations from Ranciaro et al.; black crosses denote data from other published studies (taken from Macholdt et al. and Ranciaro et al.). (B) Median-joining network of haplotypes associated with the C-14010 variant, based on four STR loci that flank the LP enhancer region. The M superscript denotes data from Macholdt et al., and the R superscript denotes data from Ranciaro et al.

Figure 2.

Estimating the selection index s and the age of the C-14010 allele. (A) Plot of the composite log-likelihood (based on the four STR loci) for different values of s for the Southern African (red) and Eastern African (blue) data. The plot assumes a population growth rate of $r=0.01$; similar results were obtained for a constant population size model and for $r=0.02$. (B) Estimated age of the C-14010 variant in Eastern Africa vs. the selection coefficient s . Solid line, the expected age given s ; dashed lines, expected age bounded by two standard deviations of the posterior distribution of the allele age given s .