

Report

Tracing Pastoralist Migrations to Southern Africa with Lactase Persistence Alleles

Enrico Macholdt,¹ Vera Lede,¹ Chiara Barbieri,^{1,5} Sununguko W. Mpoloka,² Hua Chen,³ Montgomery Slatkin,³ Brigitte Pakendorf,^{4,*} and Mark Stoneking^{1,*}

¹Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

²Department of Biological Sciences, University of Botswana, Gaborone, Botswana

³Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA

⁴Laboratoire Dynamique du Langage, UMR5596, CNRS and Université Lyon Lumière 2, 69007 Lyon, France

Summary

Although southern African Khoisan populations are often assumed to have remained largely isolated during prehistory, there is growing evidence for a migration of pastoralists from eastern Africa some 2,000 years ago [1–5], prior to the arrival of Bantu-speaking populations in southern Africa. Eastern Africa harbors distinctive lactase persistence (LP) alleles [6–8], and therefore LP alleles in southern African populations may be derived from this eastern African pastoralist migration. We sequenced the lactase enhancer region in 457 individuals from 18 Khoisan and seven Bantu-speaking groups from Botswana, Namibia, and Zambia and additionally genotyped four short tandem repeat (STR) loci that flank the lactase enhancer region. We found nine single-nucleotide polymorphisms, of which the most frequent is -14010^*C , which was previously found to be associated with LP in Kenya and Tanzania and to exhibit a strong signal of positive selection [8]. This allele occurs in significantly higher frequency in pastoralist groups and in Khoe-speaking groups in our study, supporting the hypothesis of a migration of eastern African pastoralists that was primarily associated with Khoe speakers [2]. Moreover, we find a signal of ongoing positive selection in all three pastoralist groups in our study, as well as (surprisingly) in two foraging groups.

Results and Discussion

Southern African populations that speak non-Bantu languages with click consonants (collectively referred to here as “Khoisan,” without implying genetic, linguistic, or cultural unity of Khoisan groups) harbor some of the deepest rooting lineages and greatest genetic diversity of all extant human populations [9–12]. Although the stereotypical view of the Khoisan is that they are prototypical hunter-gatherers who speak languages that must be related since they all contain

click sounds, in fact Khoisan populations exhibit considerable diversity in languages, subsistence, and phenotype [13–15].

While it has been commonly assumed that Khoisan groups diverged early in the history of modern humans and have since remained relatively isolated, there is growing evidence of multiple migrations that have contributed to the current gene pool of Khoisan groups. For example, Bantu-speaking populations arrived in southern Africa between 1,200 and 2,000 years ago [16–18], and there is substantial Bantu-related ancestry (up to 45%) in most Khoisan groups [9]. Moreover, there is increasing evidence for a migration of pastoralists from eastern Africa that preceded the Bantu migration [1–5], and, in particular, it has been hypothesized that the origins of one of the three southern African Khoisan language families, namely Khoe-Kwadi, traces to this pastoralist migration [2].

A further indication of a pastoralist migration from eastern to southern Africa might be expected from an analysis of lactase persistence (LP) alleles. Lactase persistence is the continued expression of lactase after weaning, and hence the continued ability to digest lactose in milk, and is conferred by mutations that occur in a lactase enhancer region about 14 kb upstream from the start of the lactase gene [19]. The LP trait, as well as LP alleles, is strongly associated with pastoralism worldwide [19–21], including in eastern Africa [8]. Eastern African pastoralist groups have relatively high frequencies of unique LP alleles that enhance lactase expression *in vitro* and exhibit strong signals of recent positive selection [6–8]. Thus, one would predict LP alleles of eastern African origin in southern African groups whose ancestors included immigrant eastern African pastoralists. This should be especially true for southern African pastoralist groups, as such groups should have experienced continued selection for LP. However, while an LP allele of likely eastern African origin, -14010^*C [8], has been found at low frequency in Bantu-speaking groups from Angola [22] and South Africa [23], to date LP alleles have not been investigated in southern African Khoisan groups, even though the LP phenotype is known to exist in some groups at frequencies up to 50% [24].

We therefore analyzed the lactase enhancer region in detail in 457 individuals from 18 Khoisan and seven Bantu-speaking groups from Namibia, Botswana, and Zambia (Figure 1 and Table S1 available online). Sequence analysis of a 342 bp region (Figure S1) that encompasses previously identified LP-associated alleles revealed nine single-nucleotide polymorphisms (SNPs) (Table S1). Observed genotype frequencies did not deviate from Hardy-Weinberg expectations for any SNP in any population. We also genotyped four linked short tandem repeat (STR) loci (Figure S1) that have previously been shown to be informative [25] and inferred a total of 429 haplotypes (Table S2).

Only one copy of the common European LP allele (-13910^*T) was found, in one Nama. Given the high frequency of this allele in European populations [21], this result suggests a low amount of recent European ancestry in these samples. Other alleles found at low frequency include -13913^*C , -14011^*T , -14044^*T , -14091^*T , -14107^*A , -14176^*C , and -14156^*A (Table S1); all of these (except -14011^*T) have been reported previously in other populations from Africa and/or elsewhere [23, 26–29].

⁵Present address: Department of Biological, Geological and Environmental Sciences, Laboratory of Molecular Anthropology, University of Bologna, 40126 Bologna, Italy

*Correspondence: brigitte.pakendorf@cnrs.fr (B.P.), stoneking@eva.mpg.de (M.S.)





Figure 1. Map of the Approximate Locations of the Groups Included in This Study

The most frequent derived allele in our sample is -14010°C , which occurs in 15 of the Khoisan populations and four of the Bantu-speaking groups, at an overall frequency of 7.4% (Table S1). This allele was first reported in Kenya and Tanzania, at overall frequencies of 28% and 32%, respectively [8], but is rare or absent in other populations (Figure 2A). The -14010°C allele is associated with the lactase persistence phenotype in eastern Africa, where it shows a strong signal of recent positive selection [8], and it significantly increases lactase expression in vitro [6, 8].

The -14010°C allele occurs at significantly higher frequency ($p < 0.001$) in the Khoe speakers (11.3%) than in Tuu speakers

(2.4%), Kx'a speakers (4.1%), or Bantu speakers (3.9%). Moreover, all STR haplotypes with the -14010°C allele in non-Khoe groups also occur in Khoe groups, and all haplotypes in non-pastoral groups are either shared with pastoral groups or occur in Khoe-speaking foragers (Figures 2B and 2C and Table S2). This allele also occurs at significantly higher frequency ($p < 0.001$) in pastoralists (20.2%) than in foragers (6.7%) or in agriculturalists (1.3%). The highest frequency is in the Nama, a pastoralist Khoe-speaking group, where it attains a frequency of 36% (Table S1). These results suggest that the -14010°C allele was brought to southern Africa via a migration of pastoralists from eastern Africa who either interacted predominantly with Khoe speakers or perhaps even spoke languages which were ancestors of the Khoe languages [1, 2].

Further information can be obtained from the haplotypes defined by four STR loci; we observed seven different STR haplotypes associated with the -14010°C allele (Table S2 and Figures 2B and 2C), which are all linked by single-step mutations. We cannot directly test for a relationship between -14010°C in eastern Africa and southern Africa on the basis of the STR haplotypes because data for these STR loci are lacking in the relevant eastern African populations. However, we can use the associated STR variation to estimate the age and selection intensity for this allele in the southern African data and thereby gain more insights into its history. The method used assumes that -14010°C arose once by mutation and experienced the continuing action of genic selection with selection intensity s [30]. The allele age is then estimated as a function of s . This analysis assumes that the founding population was relatively large and carried -14010°C in roughly the frequency that it had in the eastern African source population; this assumption is supported by recent analyses of genome-wide data from the same southern African populations that indicate large influxes from eastern African populations [4].

The results (Figure 3) indicate that the selection coefficient is at least 0.05, and with this selection coefficient the age of

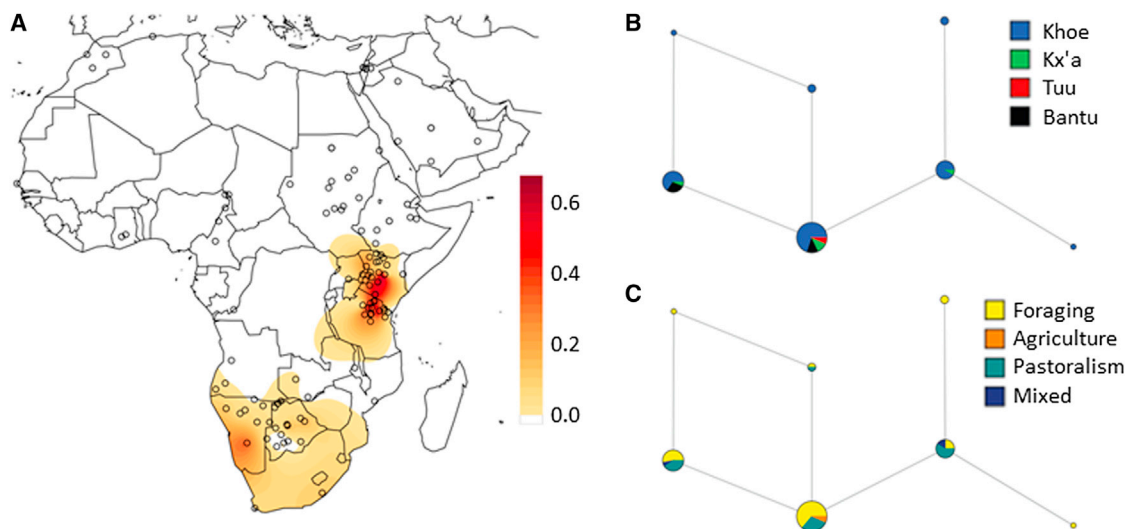


Figure 2. Frequency Distribution of the -14010°C Allele and Networks of Associated Haplotypes

(A) Surfer map of the -14010°C allele frequency. Circles indicate sample locations.

(B) Network of STR haplotypes associated with the -14010°C allele, colored according to language family.

(C) Network of STR haplotypes associated with the -14010°C allele, colored according to subsistence.

Circles denote haplotypes. The size of the circle indicates the number of times that haplotype was observed, and each branch connecting two haplotypes consists of a single step-mutation at one STR locus.

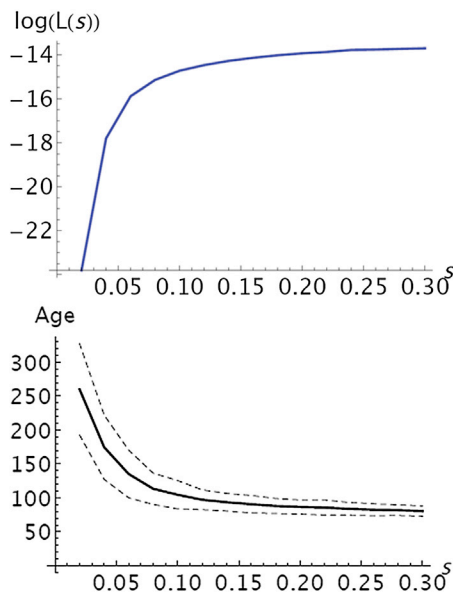


Figure 3. Estimated Selection Coefficient and Age of The -14010°C Allele in Southern African Populations

Results are shown for a population growth rate of $r = 0.01$; see also Figure S2 for comparable results assuming no population growth. Top: a graph of the composite log likelihood $[\log(L)]$ for the four STR loci on the y axis versus the selection coefficient (s) on the x axis. Bottom: a graph of the age of the -14010°C allele (in generations) on the y axis versus the selection coefficient (s) on the x axis. The solid line shows the expected age given s , and the dashed lines show the expectation ± 2 SDs of the posterior distribution of age, given s .

the -14010°C allele is about 150 generations, or 4,500 years. These results assume a past population growth rate of $r = 0.01$, but essentially the same results are obtained when $r = 0$ (Figure S2). Higher selection coefficients, which are also compatible with the data (Figure 3), would result in more recent ages for the allele, e.g., a selection coefficient of 0.1 results in an age of about 100 generations ($\sim 3,000$ years) ago. Our estimates of both allele age and the intensity of selection are comparable to previous estimates for this allele in eastern Africa [8]. This result provides further evidence for an eastern African origin of the -14010°C allele and suggests that it continued to be favored by positive selection in southern Africa.

To further investigate whether this allele was positively selected in southern African populations, we compared the current frequency of the -14010°C allele to the estimated ancestral frequency of this allele in each population immediately after the eastern African admixture. For five populations, the current frequency of the -14010°C allele is significantly higher than the estimated ancestral frequency after admixture (Table 1; comparable data for the nonsignificant comparisons are shown in Table S3), suggesting that there has been ongoing positive selection to increase the frequency of the -14010°C allele in these southern African populations. These five populations include all three of the pastoralist groups in the study (Nama, Himba, and Herero), as well as two Khoe-speaking foraging groups (Tshwa and G||ana).

Under a model of positive selection, the increase in the frequency of the -14010°C allele from the estimated value in the ancestral southern African population (immediately after admixture with the eastern African pastoralists) to its current frequency can be used to estimate the selection intensity

[29]. These estimates are shown in Table 1 and range from 0.02 to 0.08, within the range of selection coefficients estimated from the STR data (Figure 3), and are also in agreement with previous estimates for the selection coefficient associated with this allele in eastern Africa [8].

It should be noted that this analysis depends on a number of assumptions. First, the amount of eastern African ancestry in each population is inferred from the amount of Eurasian ancestry; the most likely explanation for Eurasian ancestry in the southern African populations in this study is via gene flow from an eastern African population that carried Eurasian ancestry, as discussed in detail elsewhere [4], so this assumption seems reasonable. Second, we used current estimates of the -14010°C allele frequency in eastern African populations to estimate the allele frequency in the ancestral population that migrated from eastern Africa. This is likely to be an overestimate because the demonstrated positive selection for this allele in eastern Africa [8] means that the frequency of the -14010°C allele 2,000 years ago was lower than it is today. Hence, our analysis is conservative, in that with lower estimates of the ancestral frequency of the -14010°C allele we would detect even more significant increases in frequency in southern Africa, not fewer.

In addition, random genetic drift could in principle also cause elevated frequencies of the -14010°C allele. However, with genetic drift there should be an equal chance for an allele to either increase or decrease in frequency with respect to the ancestral frequency, whereas we observed a significant increase in the frequency of the -14010°C allele in five populations, but no significant decrease in any population (Tables 1 and S3). Moreover, there is a nonrandom distribution with respect to subsistence strategy: all three pastoralist populations, but only two of 22 nonpastoralist populations, exhibited a significant increase in the frequency of the -14010°C allele, which is unlikely to occur by chance (Fisher's exact test, $p = 0.004$). Thus, it is unlikely that genetic drift alone can explain these results. However, it should be emphasized that we cannot distinguish between selection favoring -14010°C that is actually ongoing in southern African populations from selection that stopped in the very recent past because of changes in diet or other social conditions.

The fact that all three pastoralist groups in our study show a signal of continued selection on the -14010°C allele is in keeping with the hypothesis that it is advantageous to be able to consume milk after weaning. Milk is an important component of the diet of the Nama and Herero [24], and presumably also the Himba, who are closely related both culturally and genetically to the Herero. The frequency of the lactase persistence phenotype in the Nama is about 50%, higher than in the Herero ($\sim 5\%$), or indeed any other southern African group tested [24], in keeping with the higher frequency of the -14010°C allele in the Nama in our study (Table S1). Given the lower frequency of lactase persistence and the -14010°C allele in the Herero and Himba, these results would suggest that they adopted pastoralism more recently than the Nama, and more likely as a result of cultural diffusion accompanied by only a small amount of gene flow, in keeping with their very low amount of inferred eastern African ancestry [4].

Surprisingly, in addition to the three pastoralist groups, two Khoe-speaking foraging groups (the G||ana and the Tshwa) also exhibited a significant signal of recent positive selection (Table 1). It is not clear why this is the case since positive selection would not be expected to act on an allele obtained by foragers via gene flow from pastoralists. A reversion to

Table 1. Populations with Significantly More -14010°C Alleles than Expected

Population	n	Eastern African		E(-14010°C) LP = 0.2	E(-14010°C) LP = 0.6	Observed	p (LP = 0.2)	p (LP = 0.6)	s (LP = 0.2)	s (LP = 0.6)
		Ancestry								
Nama	50	0.4	4.0	12.0	18	<0.001	0.016	0.054	0.019	
Tshwa	30	0.1	0.6	1.8	5	<0.001	0.020	0.057	0.030	
G ana	20	0.06	0.2	0.7	4	<0.001	0.007	0.076	0.049	
Himba	32	0.01	0.1	0.2	4	<0.001	<0.001	0.103	0.077	
Herero	42	0.01	0.1	0.3	3	<0.001	0.002	0.087	0.061	

Expected values are based on the amount of inferred eastern African ancestry [4]. n, sample size (number of alleles); E(-14010°C), expected number of -14010°C alleles assuming the frequency in the eastern African migrating population was 20% or 60% respectively; observed, observed number of -14010°C alleles; p, empirical probability of the observed number of -14010°C alleles given the expected numbers for ancestral frequencies of 20% or 60%, respectively; s, the estimated selection coefficient assuming ancestral frequencies of 20% or 60%, respectively. See also Table S3.

foraging from an erstwhile pastoralist lifestyle might explain this unexpected signal of ongoing selection. Indeed, such a reversion has been suggested for Khoe-speaking foragers who phenotypically resemble Bantu speakers [2, 31], which would include the Tshwa (as well as the Shua, ||Xo, ||Ani, and Buga, none of whom show a significant increase in the frequency of the -14010°C allele), but not the G||ana. However, it has also been suggested that other Khoisan groups may have switched from pastoralism to foraging at different times, perhaps in relation to different geographic and/or climatic conditions [24]. Further investigation into the signal of recent positive selection in these foraging groups is needed.

In conclusion, the -14010°C allele is a further indication of a pre-Bantu migration of eastern African pastoralists to southern Africa, as postulated previously from linguistic and archaeological evidence [2, 5], as well as analyses of Y chromosome [3] and genome-wide data [4]. Thus, contrary to the stereotypical view of Khoisan groups as autochthonous foragers living in splendid isolation for a long period of time, there have been at least two prehistoric migrations just within the past 2,000 years or so (eastern African pastoralists, followed by the Bantu migration) that have had a detectable impact on their genetics, subsistence, and languages.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, two figures, and three tables and can be found with this article online at <http://dx.doi.org/10.1016/j.cub.2014.03.027>.

Acknowledgments

We are grateful to all sample donors for participating in this study, and we thank the governments of Botswana, Namibia, and Zambia for supporting this research. We thank C. de Filippo, T. Güldemann, H. Ono, and H. Nakagawa for helpful discussion; D. Luiselli and S. de Fanti for providing control samples for the STR assays; and S. Oliveira for assistance with figures. This work, as part of the European Science Foundation EUROCORES Programme EuroBABEL, was supported by a grant from the Deutsche Forschungsgemeinschaft (to B.P.) and by the Max Planck Society; H.C. and M. Slatkin acknowledge support from NIH R01-GM40282.

Received: January 14, 2014

Revised: February 17, 2014

Accepted: March 11, 2014

Published: April 3, 2014

References

- Ehret, C. (1967). Cattle-keeping and milking in eastern and southern African history - linguistic evidence. *J. Afr. Hist.* 8, 1–17.
- Güldemann, T. (2008). A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *South Afr. Humanit.* 20, 93–132.
- Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S.A., Mountain, J.L., and Underhill, P.A. (2008). Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc. Natl. Acad. Sci. USA* 105, 10693–10698.
- Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. USA* 111, 2632–2637. <http://dx.doi.org/10.1073/pnas.1313787111>.
- Pleurdeau, D., Imalwa, E., Détroit, F., Lesur, J., Veldman, A., Bahain, J.J., and Marais, E. (2012). "Of sheep and men": earliest direct evidence of caprine domestication in southern Africa at Leopard Cave (Erongo, Namibia). *PLoS ONE* 7, e40340.
- Jensen, T.G.K., Liebert, A., Lewinsky, R., Swallow, D.M., Olsen, J., and Troelsen, J.T. (2011). The -14010°C variant associated with lactase persistence is located between an Oct-1 and HNF1 α binding site and increases lactase promoter activity. *Hum. Genet.* 130, 483–493.
- Jones, B.L., Raga, T.O., Liebert, A., Zmarz, P., Bekele, E., Danielsen, E.T., Olsen, A.K., Bradman, N., Troelsen, J.T., and Swallow, D.M. (2013). Diversity of lactase persistence alleles in Ethiopia: signature of a soft selective sweep. *Am. J. Hum. Genet.* 93, 538–544.
- Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40.
- Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. *Nat. Commun.* 3, 1143.
- Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G.B., et al. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338, 374–379.
- Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardina, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F., Qi, J., et al. (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463, 943–947.
- Tishkoff, S.A., Gonder, M.K., Henn, B.M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T.B., Ramakrishnan, U., et al. (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.* 24, 2180–2195.
- Güldemann, T., and Stoneking, M. (2008). A Historical Appraisal of Clicks: A Linguistic and Genetic Population Perspective. *Annu. Rev. Anthropol.* 37, 93–109.
- Barnard, A. (1992). *Hunters and Herders of Southern Africa* (Cambridge: Cambridge University Press).
- Nurse, G.T., and Jenkins, T. (1977). *Health and the Hunter-Gatherer. Biomedical Studies on the Hunting and Gathering Populations of Southern Africa* (Basel: S. Karger).
- Ehret, C. (2001). Bantu expansions: Re-envisioning a central problem of early African history. *Int. J. Afr. Hist. Stud.* 34, 5–41.
- Pakendorf, B., Bostoen, K., and de Filippo, C. (2011). Molecular perspectives on the Bantu expansion: a synthesis. *Lang. Dyn. Change* 1, 50–88.
- Mitchell, P. (2002). *The Archaeology of Southern Africa* (Cambridge: Cambridge University Press).
- Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D.M., and Thomas, M.G. (2011). Evolution of lactase persistence: an example of human niche construction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 863–877.

20. Ingram, C.J.E., Mulcare, C.A., Itan, Y., Thomas, M.G., and Swallow, D.M. (2009). Lactose digestion and the evolutionary genetics of lactase persistence. *Hum. Genet.* 124, 579–591.
21. Itan, Y., Jones, B.L., Ingram, C.J.E., Swallow, D.M., and Thomas, M.G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol. Biol.* 10, 36.
22. Coelho, M., Sequeira, F., Luiselli, D., Beleza, S., and Rocha, J. (2009). On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC Evol. Biol.* 9, 80.
23. Tormiainen, S., Parker, M.I., Holmberg, V., Lahtela, E., Dandara, C., and Jarvela, I. (2009). Screening of variants for lactase persistence/non-persistence in populations from South Africa and Ghana. *BMC Genet.* 10, 31.
24. Nurse, G.T., Weiner, J.S., and Jenkins, T. (1985). *The Peoples of Southern Africa and Their Affinities* (Oxford: Clarendon Press).
25. Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A.I., Seixas, S., Destro-Bisol, G., and Rocha, J. (2005). Microsatellite variation and evolution of human lactase persistence. *Hum. Genet.* 117, 329–339.
26. Gallego Romero, I., Basu Mallick, C., Liebert, A., Crivellaro, F., Chaubey, G., Itan, Y., Metspalu, M., Easwarkhanth, M., Pitchappan, R., Villems, R., et al. (2012). Herders of Indian and European cattle share their predominant allele for lactase persistence. *Mol. Biol. Evol.* 29, 249–260.
27. Lember, M., Tormiainen, S., Kull, M., Kallikorm, R., Saadla, P., Rajasalu, T., Komu, H., and Järvelä, I. (2006). Lactase non-persistence and milk consumption in Estonia. *World J. Gastroenterol.* 12, 7329–7331.
28. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.
29. Enattah, N.S., Jensen, T.G.K., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J.K., Alifrangis, M., Khalil, I.F., et al. (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am. J. Hum. Genet.* 82, 57–72.
30. Slatkin, M. (2001). Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.* 78, 49–57.
31. Cashdan, E. (1986). Hunter-gatherers of the Northern Kalahari. In *Contemporary Studies on Khoisan: In Honour of Oswin Köhler on the Occasion of his 75th Birthday*, R. Vossen and K. Keuthmann, eds. (Hamburg: Helmut Buske Verlag), pp. 145–180.