The genetic variation of lactase persistence alleles in northeast Africa

Nina Hollfelder,[*,1] Hiba Babiker,[2] Lena Granehäll,[1,3] Carina M Schlebusch,[1,4] and Mattias Jakobsson[*,1,4]

[1]Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany

[3]Institute for Mummy Studies, Eurac Research, Bolzano, Italy

[4]SciLifeLab, Uppsala University, Uppsala, Sweden

[*]Corresponding author: E-mail: nina.hollfelder@ebc.uu.se, mattias.jakobsson@ebc.uu.se.

Associate Editor:

Abstract

Lactase persistence (LP) is a well-studied example of a Mendelian trait under selection in some human groups due to gene-culture co-evolution. We investigated the frequencies of genetic variants linked to LP in Sudanese and South Sudanese populations. These populations have diverse subsistence patterns, and some are dependent on milk to various extents, not only from cows, but also from other livestock such as camels and goats. We sequenced a 316bp region involved in regulating the expression of the *LCT* gene on chromosome 2, which encompasses five polymorphisms that have been associated with LP. Pastoralist populations showed a higher frequency of LP-associated alleles compared to non-pastoralist groups, hinting at positive selection also in northeast African pastoralists. There was no incidence of the East African LP allele (-14010:C) in the Sudanese groups, and only one heterozygote individual for the European LP allele (-13910:T), suggesting limited recent admixture from these geographic regions. Among the LP variants, the -14009:G variant occurs at the highest frequency among the investigated populations, followed by the -13915:G variant, which is likely of Middle Eastern origin, consistent with Middle Eastern gene-flow to the Sudanese populations. The Beja population of the Beni Amer show three different LP-variants at substantial and similar levels, resulting in one of the greatest frequencies of LP-variants among all populations across the world.

Key words: Lactase persistence, Sudan, Northeast Africa, genetic diversity.

Introduction

Lactase persistence (LP) is the ability to digest the milk sugar, lactose, at an adult age. The phenotype is associated with several single nucleotide polymorphisms (SNPs) that are located 13.9kb upstream of the lactase gene (*LCT*) in an associated enhancer element. Currently, we know of at least five variants that are clearly associated with the LP phenotype (Ségurel and Bon, 2017). The best known case is the -13910:C>T polymorphism (rs4988235), which is strongly associated with LP in

populations of European ancestry (Enattah *et al.*, 2002) and has been under strong recent selection, likely co-evolving with dairy farming (Bersaglieri *et al.*, 2004).

The LP phenotype has been found at greater frequencies in milk-drinking pastoralist populations than non-pastoralist populations (Gerbault *et al.*, 2011; Holden and Mace, 1997; Itan *et al.*, 2010), and is also common in pastoralist African societies (Tishkoff *et al.*, 2007). However, LP occurs in populations that do not carry the derived -13910:T allele, specifically in the Middle East and Eastern Africa, therefore, the thoroughly investigated -13910:C>T polymorphism is not the causal variant in these populations (Myles *et al.*, 2005). Other SNPs have been identified to be the putative causal variants in these regions: -13907:C>G (rs41525747) in Ethiopia and Saudi-Arabia, -13915:T>G (rs41380347) in Saudi-Arabia, -14009:T>G (rs869051967) in African Arab groups, and -14010:G>C (rs145946881) in Kenya and Tanzania (Ingram *et al.*, 2009; Jones *et al.*, 2013; Liebert *et al.*, 2016; Priehodová *et al.*, 2014; Ranciaro *et al.*, 2014; Tishkoff *et al.*, 2007). These polymorphisms have been shown to increase *LCT* promoter expression in vitro (Enattah *et al.*, 2008; Jensen *et al.*, 2011; Jones *et al.*, 2013; Liebert *et al.*, 2016; Olds *et al.*, 2011; Tishkoff *et al.*, 2007), and the -13910:C>T variant was recently identified as the putative causal variant for LP in a GWAS study in the Fulani population of the African Sahel/Savannah belt (Vicente *et al.*, 2019). There is evidence for a selective sweep on -14010:G>C (Tishkoff *et al.*, 2007) that shows a stronger selection coefficient in the Massai (MKK) than the allele -13910:T shows in the European (CEU) population (Altshuler *et al.*, 2010; Schlebusch *et al.*, 2013), pointing to a strong increase in fitness for LP individuals in African pastoralist populations.

LP-associated SNPs have been reported in Northeast Africa (Enattah *et al.*, 2008; Hassan *et al.*, 2016; Tishkoff *et al.*, 2007) and there is linguistic and archaeological evidence that cow-herding has been practiced in northeast Africa for at least four thousand years (Ehret, 1979; Smith, 1992). The development of farming in northeast Africa depended on the climatic conditions. While the wetter conditions along the Nile allowed for crop farming and settlement, pastoralism with a semi-nomadic lifestyle was developed in the drier Savannah/Sahel regions (Haaland and Haaland, 2013). The pastoralist Beja populations of Sudan have been shown to have a high prevalence of LP (Bayoumi *et al.*, 1981; Tishkoff *et al.*, 2007) and moderately high allele frequencies of LP-associated alleles compared to neighboring populations, which could have arisen due to a selection event (Ranciaro *et al.*, 2014). The Nilotic populations of current-day South Sudan are dairy consuming pastoralists, which have been shown to be lactase persistent in low frequencies (Bayoumi *et al.*, 1981, 1982; Tishkoff *et al.*, 2007), but no

alleles associated with LP have this far been found (Hassan *et al.*, 2016; Tishkoff *et al.*, 2007).

To deepen our understanding of LP in Northeast Africa and the associated variants, we sequenced a 316bp region spanning all known SNPs associated with LP, in 221 individuals from 18 Sudanese and South Sudanese (SASS) populations. Combining this data with previously published high-density genome-wide genotyping data of the same individuals (Hollfelder *et al.*, 2017), we were able to investigate the allele frequencies of the LP-associated SNPs, their haplotype backgrounds and to scan for signals of selection.

Results and Discussion

Allele frequencies

In total, we identified ten different polymorphisms in this study (Table 1). We detected four of the five LP-associated alleles (-13907:G, -13910:T, -13915:G, and -14009:G) and their frequencies per population are shown in Table 2. None of the LP-associated SNPs are significantly deviating from Hardy-Weinberg equilibrium in the investigated SASS populations.

We found a Bataheen individual with a derived adenine allele at position -14010 (5.6% allele frequency in the population, Table 2). While this is not a novel allele, it occurs at very low frequencies and has not been associated with LP. The LP-associated allele -14010:C has been shown to occur at its highest frequency in the Afro-Asiatic and Nilosaharan pastoralist populations of East Africa (Schlebusch *et al.*, 2013; Tishkoff *et al.*, 2007; Wagh *et al.*, 2012), and was also found in southern Africa, where it was introduced through gene flow from East Africa (Breton *et al.*, 2014; Coelho *et al.*, 2009; Macholdt *et al.*, 2014; Torniainen *et al.*, 2009).

Table 1. SNPs identified on the targeted sequences

Locus	alleles	SNP-ID	bp pos. (hg19)
-13907	C>G	rs41525747	136608643
-13910	C>T	rs4988235	136608646
-13913	C>T	rs41456145	136608649
-13915	T>G	rs41380347	136608651
-14009	T>G	rs869051967	136608745
-14010	G>T/A	rs145946881	136608746
-14011	G>A	rs4988233	136608747
-14107	C>T	rs574071884	136608843
-14108	G>A	rs56150605	136608844

The allele associated with LP in Europeans, -13910:T, was almost completely absent from the investigated populations, except for one heterozygote Gaalien individual (Table 2). The -13910:T allele has previously been detected in African populations, as a result of European gene flow, and has also been reported to occur in populations of Sudan in low frequencies (∼0.01 allele frequency) such as the Shokrya (Hassan *et al.*, 2016), the Gaalien (Ingram *et al.*, 2007), and the Beni Amer (Jones *et al.*, 2015), as well as in higher frequency in the Fulani population that spread across the Sahel/Savannah belt (0.23–0.48 allele frequency, Enattah *et al.*, 2007; Hassan *et al.*, 2016; Ingram *et al.*, 2007; Lokki *et al.*, 2011; Ranciaro *et al.*, 2014; Vicente *et al.*, 2019).

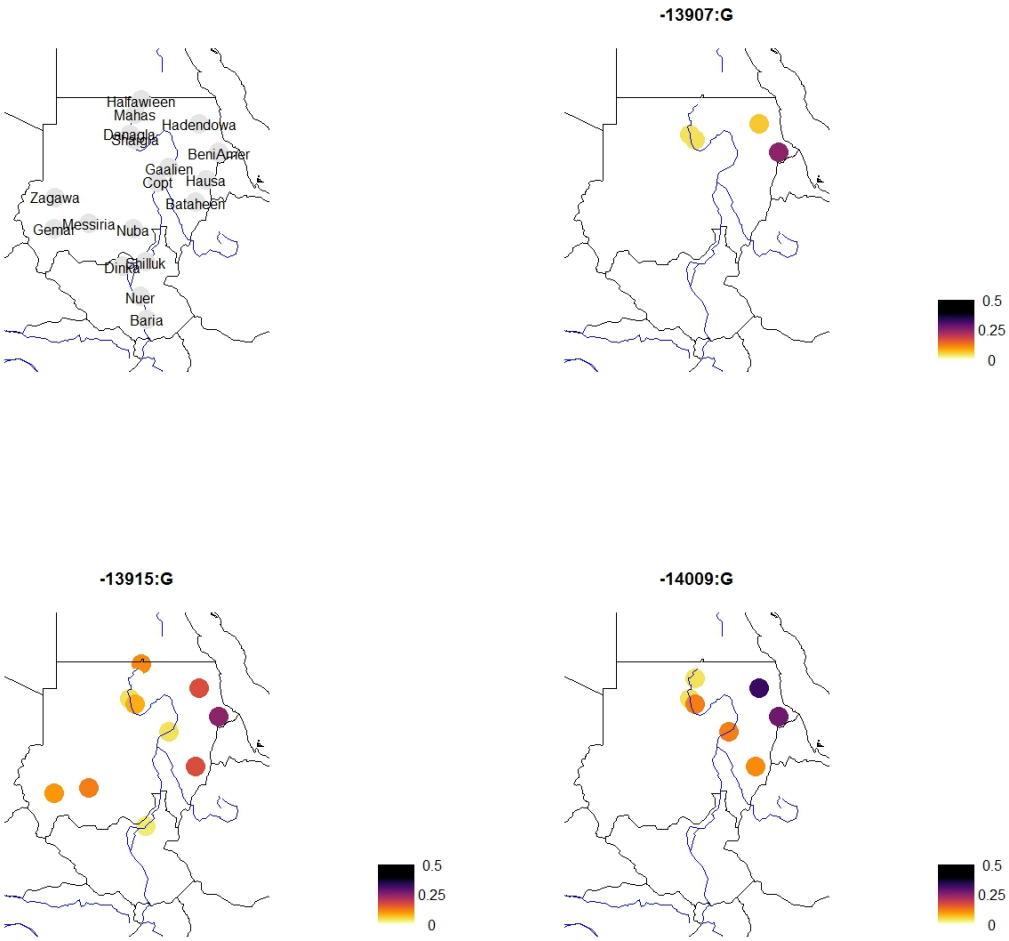

FIG. 1. Allele frequency distribution of the three LP-associated alleles found in multiple SASS populations. A distribution of these alleles (including -13010:T and -14010:C) in Africa can be seen in figure S6.

The LP-associated alleles -13907:G, -13915:G, and -14009:G appear in frequencies up to 0.34 in SASS, mainly in Arab, Nubian and Beja populations of Sudan (Table 2). The LP-associated allele -13915:G has previously been found in the Middle East, where it likely originated (Priehodová *et al.*, 2017), and in East Africa (Enattah *et al.*, 2008; Imtiaz *et al.*, 2007; Ingram *et al.*, 2007, 2009; Priehodová *et al.*, 2017; Tishkoff *et al.*, 2007). It has been shown that -13915:G appears at higher frequencies in nomadic populations compared to sedentary populations (Priehodová *et al.*, 2017). The allele likely spread from the Middle East to Africa through the nomadic Bedouin population, where it occurs at high frequencies (Ingram *et al.*, 2007; Priehodová *et al.*, 2014). The derived allele of -13915 is

Table 2.: Population information and allele frequencies of alleles at positions associated with LP.

Population	Ethnic group	n	Linguistic group	Subsistence pattern	LP associated alleles					Predicted LP Phenotype
					-13907:G	-13910:T	-13915:G	-14009:G	-14010:A*	
Bataheen	Arab	9	Afro-Asiatic	Agro-pastoralist	0	0	0.2778	0.1667	0.0556	66.7%
Gaalien	Arab	12	Afro-Asiatic	Agriculturalist	0	0.0417	0.0417	0.125	0	41.7%
Messiria	Arab	8	Afro-Asiatic	Pastoralist	0	0	0.125	0	0	12.5%
Shaigia	Arab	12	Afro-Asiatic	Agriculturalist	0.0417	0	0.0836	0.125	0	41.7%
Copt	Copt	11	Afro-Asiatic	Agriculturalist	0	0	0	0	0	0%
Hausa	Hausa	5	Afro-Asiatic	Agriculturalist	0	0	0	0	0	0%
Beni Amer	Beja	16	Afro-Asiatic	Pastoralist	0.25	0	0.25	0.2813	0	87.5%
Hadendowa	Beja	9	Afro-Asiatic	Pastoralist	0.0556	0	0.2222	0.3333	0	88.9%
Danagla	Nubian	12	Nilo-Saharan	Agriculturalist	0.0417	0	0.0417	0.0417	0	25%
Halfawieen	Nubian	9	Nilo-Saharan	Agriculturalist	0	0	0.1111	0	0	22.2%
Mahas	Nubian	14	Nilo-Saharan	Agriculturalist	0	0	0	0.0357	0	7.1%
Baria	Nilotic	5	Nilo-Saharan	Agro-pastoralist	0	0	0	0	0	0%
Dinka	Nilotic	14	Nilo-Saharan	Agro-pastoralist	0	0	0	0	0	0%
Nuer	Nilotic	15	Nilo-Saharan	Agro-pastoralist	0	0	0	0	0	0%
Shilluk	Nilotic	16	Nilo-Saharan	Agro-pastoralist	0	0	0.0313	0	0	6.3%
Gemar	Gemar	5	Nilo-Saharan		0	0	0.1	0	0	20%
Zaghawa	Zaghawa	15	Nilo-Saharan	Agro-pastoralist	0	0	0	0	0	0%
Nuba	Nuba	16	Nilo-Saharan & Niger-Congo		0	0	0	0	0	0%
European (Bersaglieri *et al.*, 2004)	European			Agropastoralist		0.815				
Saudi (Imtiaz *et al.*, 2007)	Arab			Agropastoralist			0.59375			

NOTE.—*The allele -14010:A is not the derived variant previously associated with LP and has not been added to the calculation of predicted lactase-persistence

found in all Arab populations of Sudan (4.2-27.8%, Table 2). The observed allele frequencies are similar to previously reported values for the Gaalien and Shaigia (Enattah *et al.*, 2008; Hassan *et al.*, 2016; Ingram *et al.*, 2007). This is the only LP-variant present in the Messiria population (12.5%), an Arab population from southwest Sudan. It is also found in the Beja and Nubian populations, concurrent with previous results (Hassan *et al.*, 2016). We furthermore find -13915:G at 3.1% in the Shilluk and at 10% in the Gemar. The allele frequency of -13915:G correlates significantly (ρ=0.5880782, p=0.01026) with the Middle Eastern admixture proportions of the carrier populations (Figure 2).

The derived allele of -13907 has been found in populations of the Sudan and East Africa (Ingram *et al.*, 2007; Jones *et al.*, 2013; Ranciaro *et al.*, 2014; Tishkoff *et al.*, 2007). The Beja populations have been shown to carry the highest frequencies of this allele, along with Afro-Asiatic Kenyans (Tishkoff *et al.*, 2007). The -13907:G allele has also been observed in low frequency in sedentary Arab populations of Sudan (Enattah *et al.*, 2008; Ingram *et al.*, 2007; Ranciaro *et al.*, 2014) and in the Danagla (<10%) (Ingram *et al.*, 2007). We observed this allele only at low frequency (<5%) in the Shaigia Arab population and the Nubian Danagla. A previous study of LP in Sudan (Hassan *et al.*, 2016), also found only low frequencies (<5%) of this allele in populations

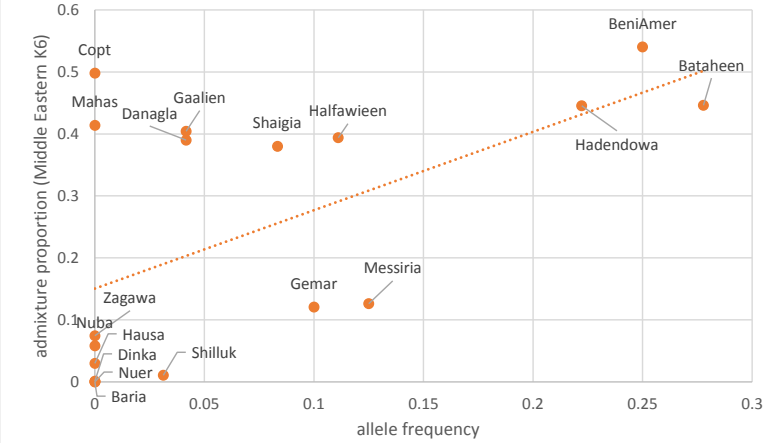

FIG. 2. Comparison of allele frequency of -13915:G to the Middle Eastern ancestry component assuming 6 clusters in an ADMIXTURE analyses with worldwide populations (Hollfelder *et al.*, 2017, Fig. S3).

other than the Beja, where we observed allele frequencies up to 25%.

The LP-associated polymorphism -14009:G has its highest reported occurrence in the Beja populations of Sudan, but is also found in smaller frequencies in other populations such as African Arab groups and the populations of the Middle East and East Africa (Ingram *et al.*, 2009; Jones *et al.*, 2013, 2015; Liebert *et al.*, 2016; Priehodová *et al.*, 2014; Ranciaro *et al.*, 2014). In this study, the highest frequency of this allele is also found in the Beja populations, but it is also the most common occurring LP-associated allele found in SASS populations.

The Beni Amer of the Beja show the highest frequencies of LP-associated alleles among the tested populations (Figure 1). Both -13907:G and -13915:G appear at 25%, and -14009:G has a frequency of 28.1%. The Hadendowa have a higher frequency of the derived allele -14009:G (33.3%), but lower frequency of -13915:G (22.2%) and -13907:G (5.6%) than the Beni Amer. The -13907:G variant was previously reported at even higher frequencies in the Hadendowa than observed here (Ranciaro *et al.*, 2014). The comparatively high allele frequencies of LP-associated alleles leads to the highest prediction of LP-phenotype of close to 90% in the Beja populations (Table 2). Previous studies have registered the LP-phenotype to be 64%-100% in the Beni Amer and 82% in the Hadendowa (Bayoumi *et al.*, 1981; Holden and Mace, 1997; Tishkoff *et al.*, 2007).

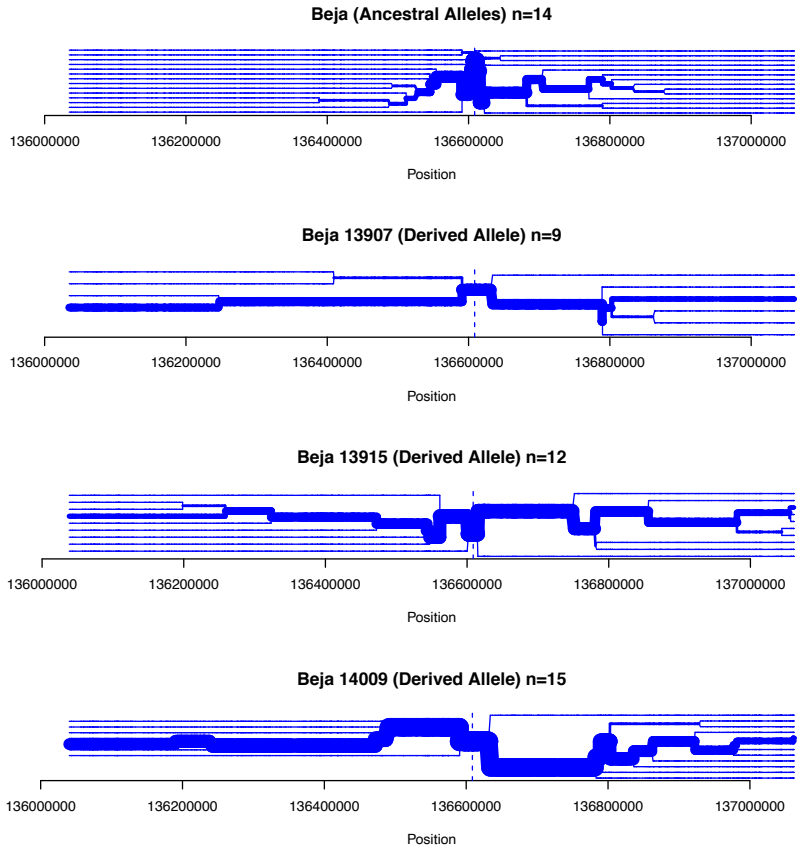

FIG. 3. Bifurcation plots for the Beja populations. The topmost plot is centered around position -13907 and contains only haplotypes that have non of the derived LP-associated alleles.

The genetic differentiation, that has been identified previously between the Arabs of central/north Sudan and the Messiria (Babiker *et al.*, 2011; Hollfelder *et al.*, 2017), is also seen in the frequency of the LP alleles. The derived allele for -14009 is found in the Bataheen, Gaalien and Shaigia at 0.125-0.167 frequency but not in the Messiria (site specific $F_{ST}^{Messiria;X}$=0.062167 to 0.113881, where X is one of the other Sudanese Arab populations). Both groups have previously been shown to be genetically close to surrounding populations and resulted through an admixture process between local populations and migrating Middle Eastern populations leading to the emergence of an Arab ethnic affiliation (Hollfelder *et al.*, 2017). The Messiria are part of the Baggara Arabs, a collective term for nomadic tribes of Kordofan that are dairy farming pastoralists (Bayoumi *et al.*, 1981). Priehodová *et al.* (2017) hypothesized that there were two directions of Middle Eastern gene flow, one entered along the Nile giving rise to the Arab populations that reside along the Nile, while the other came from Lake Chad, forming the Baggara

Arabs. This could potentially explain the absence of the other LP-associated alleles, other than -13915:G, in the Messiria. Alternatively, through the lower levels of admixture seen in the Messiria, only -13915:G might have persisted.

The Nubians show low frequencies of the LP-associated alleles. The Danagla have three individuals with one heterozygote derived LP-allele each (4.2% frequency of each -13907:G, -13915:G, and -14009:G). A previous study has not observed -13915:G in the Danagla (Ingram *et al.*, 2007), but -13907:G was observed at 8% (Ingram *et al.*, 2007). The Halfawieen only carry derived alleles of -13915 (11.1%), concurrent with previous results (Hassan *et al.*, 2016), and the Mahas have one individual with heterozygote state of the -14009 polymorphism (3.6%). The -13915:G allele was previously observed at 0.038-0.167 allele frequency in the Mahas (Enattah *et al.*, 2008; Hassan *et al.*, 2016), but no derived allele was found in the Mahas in this study. The predicted frequency of lactose digesters (Table 2) is in agreement with frequencies found in a study identifying lactose digesters through the hydrogen breath test (Bayoumi *et al.*, 1981). The Nubians and Sudanese Arab populations have similar levels of Middle Eastern admixture, however, the Nubians show lower frequencies of the LP-associated alleles. The genetic differentiation of the LP-associated alleles between Nubians and central Sudanese Arabs is higher than 0.05 in three of the nine pairwise comparisons, both pairing a Nubian with the Bataheen population. The Bataheen also show differentiation in the LP-associated alleles to the Gaalien ($F_{ST} > 0.05$). The Bataheen show the highest frequencies of LP-associated alleles and have the highest predicted LP phenotype of the Nubian and Arab populations. Assuming that the non-African admixture into all Arab and Nubian population occurred in the same event, it is possible that there is a stronger selective pressure on the LP-associated region in the camel-breeding Bataheen than the other populations, that are more reliant on agriculture (Bayoumi *et al.*, 1981; Hassan *et al.*, 2016).

No LP-associated alleles were found in the Nilotic populations of South Sudan. Due to the close proximity of SASS populations to East Africa it is surprising that there is no evidence of the derived -14010:C allele in the SASS populations. This absence in Nilotic South Sudanese populations, despite the occurrence in Nilotic Tanzanians and Kenyans, where it is significantly associated with LP, has previously been noted (Tishkoff *et al.*, 2007), and is in agreement with a previous study, that found the South Sudanese Nilotes to have remained largely isolated (Hollfelder *et al.*, 2017). The lack of LP-associated alleles in the agro-pastoralist Nilotic populations has been observed before (Hassan *et al.*, 2016; Tishkoff *et al.*, 2007) despite the intermediate prevalence of lactose-digesters (>20%) in tested Nilotic populations (Bayoumi

et al., 1981, 1982; Tishkoff *et al.*, 2007). This might be indicative of unknown LP associated variants in the Nilotic populations. LP associated alleles are also absent in the Hausa of Sudan, although a Hausa population of Cameroon had previously shown 0.139 allele frequency of -13910:T (Mulcare *et al.*, 2004). In an early study of lactose digesters in Sudan (Bayoumi *et al.*, 1981) the Nuba and the Messiria also showed higher LP-phenotypes than predicted in this study. These populations are genetically close to the Nilotic populations (Hollfelder *et al.*, 2017) and LP might be driven by the same unknown mechanism/mutations as in the Nilotes.

Additional SNPs were found within the 316 bp region that have not been associated with LP (Table 1). The -13913:C>T (rs41456145) polymorphism was found in heterozygote state in one Mahas and one Copt individual (allele frequencies: 0.0357 and 0.0454). Although this SNP is inside the Oct-1 binding site (Ingram *et al.*, 2007), it does not appear to have an effect on LP (Jones *et al.*, 2013). This SNP has previously been found in the Gaalien of Sudan and Fulani of Cameroon (Ingram *et al.*, 2007), Khoe-San populations at frequencies up to 0.075 (Breton *et al.*, 2014; Macholdt *et al.*, 2014), and Ethiopian populations up to 0.09 (Jones *et al.*, 2013). One Bataheen individual was found to be heterozygote for -14011:G>A (rs4988233) (0.0556). This SNP has been shown to influence promoter activity in vitro (Liebert *et al.*, 2016) and has previously been observed in European and Middle Eastern populations (Lember *et al.*, 2006; Liebert *et al.*, 2016), Bantu-speaking populations of southern Africa (Macholdt *et al.*, 2014), as well as Ethiopian populations (Jones *et al.*, 2013). The allele -14107:T (rs574071884) was found in one instance in a Beni Amer individual (allele frequency: 0.03125). This SNP has previously been found in Xhosa and Ghana populations (Torniainen *et al.*, 2009), the Fulani of Mali (Lokki *et al.*, 2011), Shuwa Arabs of Chad (Priehodová *et al.*, 2014), and Bantu populations of Southern Africa (Macholdt *et al.*, 2014). One allele of -14108:A (rs56150605) has been found in the Danagla. This allele has previously been encountered in the Gaalien of Sudan in very low frequency (Enattah *et al.*, 2008).

Haplotype Structure

We created a plot showing the allelic state of each SNP in the populations containing the three LP-associated alleles found in moderate frequencies in the investigated populations: -13907, -13915, and -14009 (Figure 5). As observed before (Tishkoff *et al.*, 2007), the LP-associated SNPs are found in distinct haplotype blocks, and have therefore evolved independently. Furthermore, bifurcation plots visualize the extension of the haplotypes surrounding the LP-associated alleles in the Beja population, who carry the highest number of LP-associated alleles (Figure 3). These plots might over-represent haplotypes due to the allelic

drop-out, as might be the cause for a particular long run of homozygosity around the position of the LP alleles in a Hadendowa individual, who is homozygous for -14009:G (Figures 3 and S10).

Selection Scan

We computed the LSBL-statistic (Shriver *et al.*, 2004) across chromosome 2 and for each SASS population, as well as the MKK and CEU population from the 1000 Genomes Project dataset (1000 Genomes Project Consortium, 2015) (Figures 4, S1, S2, S3, S4, and S5), to search for signals of positive selection. In both datasets, the area around the LP-associated polymorphisms is a clear outlier in MKK and CEU, which have previously been shown to be subjected to strong positive selection (Bersaglieri *et al.*, 2004; Schlebusch *et al.*, 2013). Both Beja populations show increased LSBL-signals in one of the neighboring windows. It is likely that selective event is responsible for the high frequency of lactose digesters. The increase of the frequency of associated alleles surrounding the target of selection is consistent with a selective sweep. We might be observing three independent sweeps affecting the three LP-associated alleles in the same region, which appears similar to a soft sweep and is difficult to detect. This sweep might, however, have differentiated the regions surrounding the LP-associated alleles enough to increase the LSBL above the threshold. Two other regions on chromosome 2 are distinguished from the comparative populations and affect more than four populations (Table S1).

Conclusion

LP-associated persistence alleles from Europe (-13910:T) and East Africa (-14010:C) have been used to track migration patterns of African populations (Ben Halima *et al.*, 2017; Breton *et al.*, 2014; Enattah *et al.*, 2007; Myles *et al.*, 2005). Sudanese populations have been shown to be recipients of non-African gene flow, likely from a Middle Eastern source (Hollfelder *et al.*, 2017). The absence of the European and East African LP-allele (-13910:T and -14010:C) suggests negligible amounts of recent gene-flow from these regions into the populations of Sudan and South Sudan, while the occurrence of the allele associated with LP in the Middle East (-13915:G) is consistent with recent gene-flow from the Middle East into Sudan.

Even though this study investigated a range of Nilotic populations, no LP-associated SNPs were detected in these agropastoralist populations, that have been shown to be able to digest milk in hydrogen breath tests (Bayoumi *et al.*, 1981, 1982; Tishkoff *et al.*, 2007). Further studies on Nilotic populations will reveal more about the substructure in northeast Africa (Hollfelder *et al.*, 2017) and can be informative about the underlying biology of LP in Nilotic populations. The traditionally pastoralist Beja people have been shown to have among the highest number

of lactose digesters in the population (Bayoumi *et al.*, 1981; Holden and Mace, 1997; Tishkoff *et al.*, 2007). Both -13907:G and -14009:G appear at their highest frequency in the Beja and it is possible that they emerged among these Sudanese populations. Another LP-associated SNP, -13915:G also appears at high frequency in the Beja population. The three alleles found in the Beja populations are on different haplotype backgrounds driving the frequency of putative lactose digesters to the highest seen in the area (Table 2). There is a clear extension of the haplotypes surrounding the derived alleles of the SNPs associated with LP (Figure 3). There is also an increase in LSBL-values close to the LP-associated region. Both of these signals suggest a selective event that drove the LP phenotype to a high frequency in the Beja population, that harbor several LP-associated alleles. The similar frequency of these three alleles in the Beni-Amer suggests that these SNPs are of similar age.

Materials and Methods

A total of 221 individuals from 18 Sudanese and South Sudanese populations were selected for sequencing. These individuals have previously been investigated using microsatellites (Babiker *et al.*, 2011) and dense SNPs (Hollfelder *et al.*, 2017). A 316 base pair (bp) region of intron 13 of the *MCM6* gene was targeted for sequencing, encompassing all variants associated with LP (-13907, -13910, -13915, -14009, and -14010). Primer sequences were obtained from Coelho *et al.* (2009). DNA was extracted from Whatman FTA cards using Whatman protocol BD09 and BD01. PCR was performed using 0.625U AmpliTaq Gold DNA Polymerase, 1x Gold Buffer, 0.5mM dNTP mix, 2.5mM $MgCl_2$, and 0.2µM of each primer per reaction in 30 cycles of 95°C at 15s, 55°C at 30s, and 72°C at 45s, with an initial deamination step of 10 minutes at 95°C and a final elongation of 5 minutes at 72°C. Sequencing was performed by the SNP&Seq Centre, SciLifeLab, Uppsala.

The obtained electropherograms were visually checked using GeneStudio and aligned to hg19 using MEGA7 (Kumar *et al.*, 2016). Of the 221 individuals sequenced, 203 individuals gave successful sequencing results. All polymorphic sites were covered by concordant forward and reverse strands except for two individuals (one from each the Shaigia and the Bataheen populations) who had a successful result only with the forward-primer. All polymorphism peaks were unambiguous.

Phasing and imputation to analyze haplotype structure

The sequencing results were added to 323,726 additional SNPs from chromosome 2, obtained from a filtered dataset of 3.9 million SNPs, typed on an Illumina HumanOmni5M Exome SNP array in a previous study (Hollfelder *et al.*, 2017). This combined dataset was imputed and phased using

fastPHASE version 1.4.0 (Scheet and Stephens, 2006). The number of haplotype clusters was set to 25, with 25 runs of the EM algorithm. The number of haplotypes sampled from the posterior distribution obtained from a particular random start of the EM algorithm was set to 100. We used the phase information to create a visualization of the haplotypes surrounding the LP control region (Figure 5). The R-package 'rehh' (Gautier and Vitalis, 2012) was used to create bifurcation plots visualizing the haplotype structure surrounding the LP-associated alleles.

Locus specific branch length (LSBL)

Regions of extreme genetic differentiation on chromosome 2 were detected using LSBL (Shriver *et al.*, 2004). LSBL estimates the branch length per locus by comparing pairwise F_{ST} values of three populations. This allows to detect in which of the three populations the genetic differentiation took place.

LSBL was calculated on the SASS populations from the Hollfelder *et al.* (2017) dataset as well as populations from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). The Hollfelder *et al.* (2017) dataset experienced a degree of allelic drop-out, which excludes the possibility of selection scans using haplotype based methods for this dataset. It was, however, shown that F_{ST} estimates on this diploid data set correlate strongly with a randomly haploidized version of the data set, therefore, measures such as LSBL can be used safely on the fully diploid data set (Hollfelder *et al.*, 2017, SI).

We calculated Weir and Cockerham's F_{ST} as implemented in plink v1.90 (Chang *et al.*, 2015). LSBL was calculated for each locus on the SASS populations using two comparative non-LP populations (YRI and CHB), one African and one non-African to account for admixture in the SASS populations.

$$LSBL_{pop} = \frac{F_{ST}^{YRI;pop} + F_{ST}^{CHB;pop} - F_{ST}^{YRI;CHB}}{2}$$

where *pop* is the test population. LSBL is calculated for each of the three combined populations. All SASS populations were tested, as well as MKK and CEU, which have been subjected to strong selection in the genomic region of the LP-associated alleles (Bersaglieri *et al.*, 2004; Schlebusch *et al.*, 2013). We computed the mean LSBL in non-overlapping 500 kilo base (kb) windows containing at least 50 SNPs and highlighted areas that are more than 3 standard deviations higher than the mean (Figure 4, S1, S2, S4, S5). A control was performed where negative F_{ST} estimates were exchanged to 0 (Hider *et al.*, 2013). The treatment of negative F_{ST} estimates did not have an impact on the results (Table S1).

Supplementary Material

Supplementary table S1 and figures S1–S10 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank the volunteering participants of this project that provided DNA samples. Sanger sequencing was performed at the Uppsala Genome Center, which is part of the Swedish National Genomics Infrastructure. The computations were performed on a high performance compute cluster at Uppsalas Multidisciplinary Center for Advanced Computational Science (UPPMAX). This work was supported by the Swedish research council (grant number 2018-05537 to MJ, 621-2014-5211 to CS), the European Research council (ERC 759933 to CS), and the Knut and Alice Wallenberg foundation.

References

1000 Genomes Project Consortium 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74.

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Marie Muzny, D., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarroll, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Cristina Manca, M., Marshall, P. A., Matsuda, I., Ngare, D., Ota Wang, V., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311): 52–58.

Babiker, H. M. A., Schlebusch, C. M., Hassan, H. Y., and Jakobsson, M. 2011. Genetic variation and population structure of Sudanese populations as indicated by 15 Identifiler sequence-tagged repeat (STR) loci. *Investigative genetics*, 2(1): 1.

Bayoumi, R., Saha, N., Salih, A., Bakkar, A., and Flatz, G. 1981. Distribution of the lactase phenotypes in the population of the Democratic Republic of the Sudan. *Human Genetics*, 57(3): 279–281.

Bayoumi, R. A. L., Flatz, S. D., Kühnau, W., and Flatz, G. 1982. Beja And Nilotes: Nomadic pastoralist groups in the Sudan with opposite distributions of the adult lactase phenotypes. *American Journal of Physical Anthropology*, 58(2): 173–178.

Ben Halima, Y., Kefi, R., Sazzini, M., Giuliani, C., De Fanti, S., Nouali, C., Nagara, M., Mengozzi, G., Elouej, S., Abid, A., Jamoussi, H., Chouchane, L., Romeo, G., Abdelhak, S., and Luiselli, D. 2017. Lactase persistence in Tunisia as a result of admixture with other Mediterranean populations. *Genes & Nutrition*, 12(1): 20.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, 74(6): 1111–1120.

Breton, G., Schlebusch, C. M., Lombard, M., Sjödin, P., Soodyall, H., and Jakobsson, M. 2014. Lactase persistence alleles reveal partial East African ancestry

of southern African Khoe pastoralists. *Current Biology*, 24(8): 852–858.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(7).

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 Evolutionary Biology*, 9(1): 80.

Ehret, C. 1979. On the Antiquity of Agriculture in Ethiopia. *The Journal of African History*, 20(02): 161.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2): 233–237.

Enattah, N. S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., Greco, L., Rossi, M., Lentze, M., Seo, J., Rahgozar, S., Khalil, I., Alifrangis, M., Natah, S., Groop, L., Shaat, N., Kozlov, A., Verschubskaya, G., Comas, D., Bulayeva, K., Mehdi, S. Q., Terwilliger, J. D., Sahi, T., Savilahti, E., Perola, M., Sajantila, A., Järvelä, I., and Peltonen, L. 2007. Evidence of Still-Ongoing Convergence Evolution of the Lactase Persistence T-13910 Alleles in Humans. *The American Journal of Human Genetics*, 81(3): 615–625.

Enattah, N. S., Jensen, T. G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J. K., Alifrangis, M., Khalil, I. F., Natah, A., Ali, A., Natah, S., Comas, D., Mehdi, S. Q., Groop, L., Vestergaard, E. M., Imtiaz, F., Rashed, M. S., Meyer, B., Troelsen, J., and Peltonen, L. 2008. Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture. *The American Journal of Human Genetics*, 82(1): 57–72.

Gautier, M. and Vitalis, R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, 28(8): 1176–1177.

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. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1566): 863–877.

Haaland, R. and Haaland, G. 2013. Early farming societies along the Nile. In P. Mitchell and P. Lane, editors, *The Oxford Handbook of African Archaeology*, chapter 37. OUP Oxford.

Hassan, H. Y., Erp, A., Jaeger, M., Tahir, H., Oosting, M., Joosten, L. A. B., and Netea, M. G. 2016. Genetic diversity of lactase persistence in East African populations. *BMC research notes*, 9(1): 1.

Hider, J. L., Gittelman, R. M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J. M., and Parra, E. J. 2013. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evolutionary Biology*, 13(1): 150.

Holden, C. and Mace, R. 1997. Phylogenetic analysis of the evolution of lactose digestion in adults. *Human biology*, 69(5): 605–28.

Hollfelder, N., Schlebusch, C. M., Gu, T., Babiker, H., Hassan, H. Y., and Jakobsson, M. 2017. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. *PLOS Genetics*, 13(8): 1–17.

Imtiaz, F., Savilahti, E., Sarnesto, A., Trabzuni, D., Al-Kahtani, K., Kagevi, I., Rashed, M. S., Meyer, B. F., and Jarvela, I. 2007. The T/G 13915 variant upstream of the lactase gene (LCT) is the founder allele of lactase persistence in an urban Saudi population. *Journal of Medical Genetics*, 44(10): e89–e89.

Ingram, C. J. E., Elamin, M. F., Mulcare, C. A., Weale, M. E., Tarekegn, A., Raga, T. O., Bekele, E., Elamin, F. M., Thomas, M. G., Bradman, N., and others 2007. A novel polymorphism associated with lactose tolerance in

Africa: multiple causes for lactase persistence? *Human genetics*, 120(6): 779–788.

Ingram, C. J. E., Raga, T. O., Tarekegn, A., Browning, S. L., Elamin, M. F., Bekele, E., Thomas, M. G., Weale, M. E., Bradman, N., and Swallow, D. M. 2009. Multiple Rare Variants as a Cause of a Common Phenotype: Several Different Lactase Persistence Associated Alleles in a Single Ethnic Group. *Journal of Molecular Evolution*, 69(6): 579–588.

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 evolutionary biology*, 10(1): 1.

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 HNF1α binding site and increases lactase promoter activity. *Human Genetics*, 130(4): 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. *The American Journal of Human Genetics*, 93(3): 538–544.

Jones, B. L., Oljira, T., Liebert, A., Zmarz, P., Montalva, N., Tarekeyn, A., Ekong, R., Thomas, M. G., Bekele, E., Bradman, N., and others 2015. Diversity of lactase persistence in African milk drinkers. *Human genetics*, 134(8): 917–925.

Kumar, S., Stecher, G., Tamura, K., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glockner, F. O. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33(7): 1870–1874.

Lember, M., Torniainen, S., Kull, M., Kallikorm, R., Saadla, P., Rajasalu, T., Komu, H., and Järvelä, I. 2006. Lactase non-persistence and milk consumption in Estonia. *World journal of gastroenterology*, 12(45): 7329–31.

Liebert, A., Jones, B. L., Danielsen, E. T., Olsen, A. K., Swallow, D. M., and Troelsen, J. T. 2016. In Vitro Functional Analyses of Infrequent Nucleotide Variants in the Lactase Enhancer Reveal Different Molecular Routes to Increased Lactase Promoter Activity and Lactase Persistence. *Annals of Human Genetics*, 80(6): 307–318.

Lokki, A., Järvelä, I., and Israelsson, E. 2011. Lactase persistence genotypes and malaria susceptibility in Fulani of Mali. *Malaria*.

Macholdt, E., Lede, V., Barbieri, C., Mpoloka, S. W., Chen, H., Slatkin, M., Pakendorf, B., and Stoneking, M. 2014. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. *Current Biology*, 24(8): 875–879.

Mulcare, C. A., Weale, M. E., Jones, A. L., Connell, B., Zeitlyn, D., Tarekegn, A., Swallow, D. M., Bradman, N., and Thomas, M. G. 2004. The T Allele of a Single-Nucleotide Polymorphism 13.9 kb Upstream of the Lactase Gene (LCT) (C13.9kbT) Does Not Predict or Cause the Lactase-Persistence Phenotype in Africans. *The American Journal of Human Genetics*, 74(6): 1102–1110.

Myles, S., Bouzekri, N., Haverfield, E., Cherkaoui, M., Dugoujon, J.-M., and Ward, R. 2005. Genetic evidence in support of a shared Eurasian-North African dairying origin. *Human Genetics*, 117(1): 34–42.

Olds, L. C., Ahn, J. K., and Sibley, E. 2011. 13915*G DNA polymorphism associated with lactase persistence in Africa interacts with Oct-1. *Human Genetics*, 129(1): 111–113.

Priehodová, E., Abdelsawy, A., Heyer, E., and Černý, V. 2014. Lactase persistence variants in Arabia and in the African Arabs. *Human biology*, 86(1): 7–18.

Priehodová, E., Austerlitz, F., Čížková, M., Mokhtar, M. G., Poloni, E. S., and Černý, V. 2017. The historical spread of Arabian Pastoralists to the eastern African Sahel evidenced by the lactase persistence 13,915*G allele and mitochondrial DNA. *American Journal of*

Human Biology, 29(3): e22950.

Ranciaro, A., Campbell, M. C., Hirbo, J. B., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze, M. J., Ibrahim, M., Nyambo, T., Omar, S. A., and others 2014. Genetic origins of lactase persistence and the spread of pastoralism in Africa. *The American Journal of Human Genetics*, 94(4): 496–510.

Scheet, P. and Stephens, M. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, 78(4): 629–644.

Schlebusch, C. M., Sjödin, P., Skoglund, P., and Jakobsson, M. 2013. Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. *European Journal of Human Genetics*, 21(5): 550–553.

Ségurel, L. and Bon, C. 2017. On the Evolution of Lactase Persistence in Humans. *Annual Review of Genomics and Human Genetics*, 18(1): 091416–035340.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M., and Jones, K. W. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics 2004 1:4*, 72(4): 1492–1504.

Smith, A. B. 1992. Origins and Spread of Pastoralism in Africa. *Annual Review of Anthropology*, 21(1): 125–141.

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., and others 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics*, 39(1): 31–40.

Torniainen, 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 Genetics*, 10(1): 31.

Vicente, M., Priehodová, E., Diallo, I., Podgorná, E., Poloni, E. S., Černý, V., and Schlebusch, C. M. 2019. Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. *BMC Genomics*, 20(915).

Wagh, K., Bhatia, A., Alexe, G., Reddy, A., Ravikumar, V., Seiler, M., Boemo, M., Yao, M., Cronk, L., Naqvi, A., Ganesan, S., Levine, A. J., and Bhanot, G. 2012. Lactase Persistence and Lipid Pathway Selection in the Maasai. *PLoS ONE*, 7(9): e44751.