Population genomics on the origin of lactase persistence in Europe and South Asia

YOKO SATTA[*1] and NAOYUKI TAKAHATA[*2]

School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa 240-0193, Japan

KEYWORDS: niche construction, hard selective sweep, 2D SFS, Steppe ancestry, ancient human genomes

ABSTRACT

The C to T mutation at rs4988235 located upstream of the lactase (*LCT*) gene is the primary determinant for lactase persistence (LP) that is prevalent among Europeans and South Asians. Here, we review evolutionary studies of this mutation based on ancient and present-day human genomes with the following concluding remarks: the mutation arose in the Pontic Steppe somewhere between 23,000 and 5000 years ago, emigrated into Europe and South Asia in the Bronze Age via the expansion of the Steppe ancestry, and experienced local hard sweeps with their delayed onsets occurring in the last 3600 years. We also argue that the G to A mutation at rs182549 arose earlier than 23,000 years ago, the intermediate CA haplotype ancestral to the LP-related TA haplotype is still represented by samples from Tuscans, admixed Americans and South Asians, and the great majority of G to A mutated descendants have hitchhiked since the C to T mutation was favored by local selection.

Corresponding author: Yoko Satta
e-mail address: *1: satta@soken.ac.jp, *2: takahata@soken.ac.jp

INTRODUCTION

Most mammals lose the ability to digest lactose after weaning, but some present-day humans continue to express the key digestion enzyme, lactase (LCT), in the small intestine throughout adult life (Enattah et al., 2002; Ségurel and Bon, 2017). This physiological change is referred to as lactase persistence (LP), and has attracted much attention particularly from the viewpoint of gene-culture coevolution, human self-domestication or niche construction (e.g., Aoki, 1986; Itan et al., 2009, 2010; Gerbault et al., 2011, 2013; O'Brien and Laland, 2012). Apart from LP in Africa (Ingram et al., 2007; Tishkoff et al., 2007), LP in Eurasia is tightly associated with derived alleles T at rs4988235 and A at rs182549 of single nucleotide polymorphisms (SNPs) that are located -13,910 bp and -22,018 bp, respectively, upstream of the *LCT* gene. The association of the C/T polymorphism at rs4988235 with non-LP/LP is perfect in Finnish and non-Finnish (Korean, Italian and German) samples, and that of the G/A polymorphism at rs182549 is nearly perfect in case-control study samples (Enattah et al., 2002). Both the T and A alleles enhance the *LCT* promoter activity, although the A allele results in minimal enhancement compared with the T allele (Olds and Sibley, 2003).

Aside from the adaptive significance, the dating of LP or T allele origin in Europe has been of particular interest in light of ancient genomes. Using early Holocene human remains, Burger et al. (2007) showed that most Meso- and Neolithic Europeans lacked the T and A alleles, modestly concluding that LP arose in the last 20,000 years – hereafter denoted as $t_{LCT} < 20{,}000$ for convenience (Bersaglieri et al., 2004; Coelho et al., 2005; Leonardi et al., 2012). Lack of the T allele in European Neolithic farmers was directly demonstrated by the Chalcolithic Tyrolean Iceman, a 5300-year-old natural mummy discovered in the Ötztal Alps (Keller et al., 2012) as well as a 7400-year-old Cardial individual from Cova Bonica in Barcelona (Olalde et al., 2015). Obviously, however, more individuals were needed to confirm this absence. A recent study of 400 Neolithic, Chalcolithic and Bronze Age Europeans (Olalde et al., 2018) observed that the T allele remained at a very low frequency across the transition from the Neolithic period to the Bell Beaker and Bronze Age periods, both in Britain and continental Europe, with a major increase in its frequency occurring only within the last 3500 years (Brace et al., 2018). Table 1 summarizes some such results of Allentoft et al. (2015), Mathieson et al. (2015), and Olalde et al. (2018) based on the Central European chronology in Haak et al. (2015); readers may also refer to Witas et al. (2015), Liebert et al. (2017) and Ségurel and Bon (2017).

The geographic location of the origin of European LP has been vague and contentious as well (Enattah et al., 2002, 2007 for the Pontic Steppe or Caucasian origin; Itan et al., 2009 and Gerbault et al., 2011 for theoretical modeling of LP/dairying coevolution) partly because the current geographic center of the LP distribution does not necessarily coincide with its location of origin when a population migrates and/or expands (Edmonds et al., 2004). Gerbault et al. (2011) provided an interpolated map of the distribution of the T allele in the Old World and summarized simulation results for the spread of European LP in time and space. Indeed, the T allele frequency varies considerably even within Europe (the 1000 Genomes Project Consortium, 2015; subsequently, we abbreviate the sequence dataset as 1000GD): 74% in Northern Europeans from Utah (CEU), 59% in Finnish from Finland (FIN), 72% in British from England and Scotland (GBR), 46% in Iberians from Spain (IBS), and 8.9% in Tuscans from Italy (TSI), with the overall frequency being 511/1006 = 51% in the European meta-population (EUR).

To gain some additional insights into these problems regarding LP or the T allele at rs4988235 and the A allele at rs182549, we applied our inference method, briefly described below, to an *LCT* region of 100 kb for the five individual European populations and one South Asian population (PJL: Punjabi from Lahore), separately (Smith et al., 2018). We then summarize and discuss results of the application together with the phylogenetic analysis of *LCT* haplotypes in light of recent ancient genomic studies. In addition, we demonstrate that one variable used in the method can pinpoint rs4988235 as the causative mutation for LP.

INFERENCE OF A SELECTIVE SWEEP UNDER NONEQUILIBRIUM DEMOGRAPHY

To detect, classify and date an incomplete selective sweep for *LCT* in each population, we used the method of the two-dimensional site frequency spectrum $\{\varphi_{i,j}\}$, 2D SFS, by Satta et al. (2019). This method first seeks for a region under strong linkage disequilibrium (LD with $r^2 > 3/4$) surrounding a target site and divides a sample of size n into the derived and ancestral allele groups defined by the site. Each element of the $\{\varphi_{i,j}\}$ matrix then counts the number of segregating sites or SNPs within the region at each of which we find exactly i and j derived alleles in the derived and ancestral allele groups, respectively. It is to be noted that $\varphi_{i,j} > 0$ for possible positive i and j stems most likely from recombination between the two allele groups, although it is expected to be rare in a tightly linked region.

We may compare the performance of our inference method with that of recent studies. Such studies include Itan et al. (2009) who estimated $6256 < t_{LCT} < 8683$

using an approximate Bayesian computation (ABC) approach, Field et al. (2016) who developed the singleton density score method and estimated $t_{LCT} \geq 2000$ from the UK10K project, Smith et al. (2018) who developed a hidden Markov model and obtained $t_{LCT} = 9341$ (95% CI: 8688-19,989) for IBS and 9514 (95% CI: 8596-10,383) for PJL, Harris et al. (2018) who based their method on haplotype structure, Akbari et al. (2018) who claimed the high power of their iSAFE method to pinpoint the causative mutation of a single sweep, Nakagome et al. (2019) who developed an ABC method conditioned on the allele frequency in the past and applied it to four pigmentation SNPs, Satta et al. (2019) who used the 2D SFS method for EUR and estimated $t_{LCT} = 3280 \pm 480$, and Tournebize et al. (2019) who obtained $t_{LCT} = 4250$ (95% CI: 3700-17,680) for CEU. Readers may also refer to Tournebize et al. (2019) for various estimates of t_{LCT} either as the time to the most recent common ancestor (TMRCA) or as the age of the allele.

Here, we focus on the TMRCA within the derived allele group because our interest is in the time (t_{SEL}) of onset of positive selection rather than the origination time (t_{AGE}) of a mutation that produced the allele. For convenience, we use the symbol t_{LCT} to denote the TMRCA of the derived allele group. Obviously, t_{AGE} is no shorter than t_{LCT}, and t_{SEL} is no greater than t_{AGE}, although none of these times is directly observable. For instance, to infer t_{SEL}, Enattah et al. (2007) had to assume the initial frequency of the T allele and the selection intensity in a simple equation of describing an allele frequency change. Nevertheless, the distinction between t_{LCT} and t_{SEL} is important with respect to the mode of selective sweep. A necessary condition for *soft* sweeps (Hermisson and Pennings, 2005, 2017) is $t_{LCT} > t_{SEL}$. This is equivalent, though not exactly, to saying that multiple distinct ancestral lineages carrying a common mutation simultaneously undergo positive selection. An obvious caution is that positive selection acting on a standing variation does not necessarily classify the sweep as soft. To the contrary, a sufficient condition for *hard* sweeps is $t_{LCT} \leq t_{SEL}$. There are thus two distinguishable cases: hard ($t_{LCT} \leq t_{SEL} \leq t_{AGE}$) and *conditional* soft sweeps ($t_{SEL} < t_{LCT} \leq t_{AGE}$), the latter being related to the hardening of soft sweeps (Wilson et al., 2014). It seems that either origin of a target mutation or population structure is irrelevant or even confusing to this distinction.

To evaluate statistical significances of our estimates, we assumed a demographic model and neutrality of genome evolution (Kimura, 1968). We used a very simplified bottleneck and expansion model that incorporates some aspects of human nonequilibrium demographic history inferred thus far by SFS (Schaffner et al., 2005), PSMC (Li and Durbin, 2011), MSMC (Schiffels and Durbin, 2014), Stairway Plot (Liu

and Fu, 2015), SMC++ (Terhorst et al., 2017), or the level of heterozygosity estimated from present-day populations (e.g., Broushaki et al., 2016). In our simplified model, a single bottleneck is shared by all non-African populations from the exodus out of Africa (ca. 58,000 years ago) to the initial Eurasia-wide dispersal (ca. 45,000 years ago), and a population had since grown exponentially up to the population size 500 years ago (Figure 1, Table 2a). We assumed the size of the bottlenecked population to be 2000 and ignored the effect of population explosion during the last 500 years. The robustness of the method to nonequilibrium demography was discussed in Satta et al. (2019). For genome evolution, we assumed the infinite-sites model (Kimura, 1969), but with linkage.

RESULTS

The hallmark of an incomplete selective sweep is reduced intra-allelic variability (IAV) within a derived allele group relative to the total variability in the entire sample (Fujito et al., 2018; Satta et al., 2019). IAV may be expressed by numerous variables that are useful for detecting and classifying a selective sweep as well as for dating the onset of positive selection. The significance test for individual variables F_c, L_{c0} and G_{c0} used to measure the level and pattern of IAV shows that the observed values are as extreme as 10^{-4} in all but the TSI population (Table 2a). To further strengthen the test, we may combine these mutually correlated variables. To avoid inflating false positives, we used the method by Brown (1975) that combines non-independent, one-sided tests of significance (Poole et al., 2016). The results of this test show that, in the European populations including TSI, the individual IAVs become highly incompatible with neutrality (Table 2b).

The 2D SFS method then infers that the selective sweep for the T allele is definitely hard. In addition to unusually reduced IAV levels measured by F_c and L_{c0}, this hardness is captured as low levels of multiplicity (G_{c0}) of derived alleles per SNP (Table 2a). Indeed, almost all alleles per SNP are singletons or doubletons, irrespective of populations, such that G_{c0} is close to 1 to 2. No other target sites that exhibit such low multiplicity in such a wide genomic region have yet to be revealed (Satta et al., 2019). The 2D SFS method also allows us to estimate the mutation-rate-scaled TMRCA (ut_{LCT}) from the total number ($C = \sum_{i=1}^{m-1} i\varphi_{i,0}/m$) of derived alleles per sample within the T allele group of size $m < n$. It is to be noted that recombinant SNP sites are automatically excluded from C. Using $ut_{LCT} = C$ or $t_{LCT} = C/u$ (Table 2a), we have $t_{LCT} = 3280 \pm 720$ in CEU of $m = 146$, 3420 ± 940 in FIN of $m = 117$, 3360 ± 840 in GBR of $m = 131$, 3380 ± 1320 in IBS of $m = 98$, and 2100 ± 1480 in

TSI of $m = 19$, if the above mutation rate (u) is 5×10^{-5} per 100 kb per year (Scally and Durbin, 2012).

We also conducted phylogenetic analysis to show that almost all T haplotypes (haplotypes within the T allele group defined in the LD region), including those in TSI and PJL, are tightly clustered so as to form a single clade (Figure 2). Unexpectedly, there is one T haplotype (HG03653) in PJL that is clustered in the ancestral C allele group. The DNA sequence of HG03653 suggests three possibilities: a mistyping of C to T at rs4988235, a parallel mutation in a divergent haplotype background resulting in the TG haplotype, or a migrant of the TG haplotype from the Urals and the Caucasus as discussed below. In general and more importantly, the ancestral C haplotype closest to the T cluster differs only by a single C to T mutation. This confirms the previous conclusion that the T haplotypes were descended from an ancestral C haplotype in the recent past by a single mutation (Enattah et al., 2007).

It is also interesting to examine how accurately some of our variables can pinpoint a true target site of positive selection (Figure 3). For this purpose, we selected F_c based on the fact that recombination plays essential roles in pinpointing a target site and that F_c includes recombinant SNPs, whereas both L_{c0} and G_{c0} do not. Nominating all candidate SNPs in a 2.5 Mb genomic region surrounding *LCT*, we measured F_c for each candidate within a window of 300 SNPs and then ranked all of these F_c values in the whole 2.5 Mb region. We also compared the result with that by the iSAFE with an "inserted" procedure, and found that this procedure does not improve the power of F_c. Most importantly, like iSAFE, F_c ranked the SNP at rs4988235 as 1.

DISCUSSION

Farming was introduced to most of Europe from Anatolia/Levant by 6000 to 5000 BC via two main routes: the Danube river line and the northern Mediterranean coast line. Farmers in Central Europe and Iberia were respectively associated with the LBK (Linearbandkeramik) tradition along the river line and the *Impresso*-Cardium culture along the coast line (Olalde et al., 2015; Hofmanová et al., 2016; Brace et al., 2018; Mathieson et al., 2018; Souilmi et al., 2020). In each route, farming was brought upon by demic diffusion with limited admixture from local hunter-gatherers and accordingly an apparent decrease in the Neolithic contribution with geographic distance from the Near East (Bramanti et al., 2009; Skoglund et al., 2014; Günther et al., 2015). Indeed, while Scandinavian farmers were intimately related to farmers in Southern Europe, such as the Tyrolean Iceman as well as Sardinians and many other present-day groups, they exhibited high levels of admixture from local hunter-gatherers (Skoglund et al., 2012,

2014). Rivollat et al. (2020) also revealed a higher proportion of western hunter-gatherer ancestry in Western European farmers than in Central European farmers, pointing to the complexity of their interactions in France where both routes converged.

Olalde et al. (2018) showed that Neolithic Anatolian and Aegean farmers were indeed non-LP despite archaeological and genetic evidence for their taurine cattle domestication and milk use around 6500 BC (Evershed et al., 2008). Allentoft et al. (2015) found the T allele at low frequency in the Bronze Age of Eurasia (ca. 3000-1000 BC) and showed a possible Steppe origin of LP from the highest occurrence of the T allele in the Yamnaya and Corded Ware cultures (CWCs) between 3000/2800 and 2300/2000 BC. Haak et al. (2015) also observed that the T allele increased in frequency only after the Yamnaya Steppe ancestry became ubiquitous in Central and Western Europe. Consistent with these findings, it was recently identified that the T allele is possessed by only one CWC individual from Sweden (Malmström et al., 2019), only one Final Neolithic individual from Switzerland (Furtwängler et al., 2020), and none of 16 CWC individuals from Poland (Witas et al., 2012; Linderholm et al., 2020), to mention a few. Using 230 ancient Eurasians, Mathieson et al. (2015) found that the earliest appearance of the T allele is in a Central European Bell Beaker sample dated to between 2450 and 2140 BC, and the selective sweep dated to the last 4000 years (Burger et al., 2007; Gamba et al., 2014; Haber et al., 2016 for a review). Bell Beaker culture (BBC) was widespread in Western and Central Europe from 2750/2500 to 2200/1800 BC, and associated with Steppe-related ancestry with the replacement in gene pool being most pronounced in Britain (Olalde et al., 2018). Interestingly, however, Northern Italy and Sicily Bell Beakers had no sign of LP. Raveane et al. (2019) pointed out that they can be modeled almost exclusively as Anatolia Bronze Age, with less than 5% Steppe Bronze Age. Olalde et al. (2018) also argued that there is limited genetic affinity between Beaker-complex-associated individuals from Iberia and Central Europe. Furthermore, focusing on the 8000-year history of Iberia, Olalde et al. (2019) concluded that in Iberia, the T allele continued to occur at low frequency in the Iron Age, pointing to recent strong positive selection, and that present-day Basques inherited substantial levels of Steppe ancestry (cf. Plantinga et al., 2012 for Late Neolithic Basques with 27% LP). In short, although estimates of the ancient T allele frequency still vary, neither hunter-gatherers nor early farmers in Anatolia and Europe possessed the T allele (Table 1) and the earliest carriers came from CWC and BBC individuals in Central Europe, Iberia and Scandinavia after ca. 2500 BC.

It is therefore reasonable to assume that the T allele arose somewhere in the Pontic Steppe and emigrated into Central Europe via the Yamnaya expansion in ca. 3000 BC.

Yet, positive selection began to operate for the immigrated T allele somewhat later. Literally speaking, as positive selection acted on such a standing variation ($t_{SEL} < t_{AGE}$), the selective sweep might be defined as soft. Nonetheless, the level and pattern of IAV in the T allele group are best described as hard. In this context, it is not really important to distinguish between hard and soft sweeps based on the origin of a selected mutation. What can be done is to practically classify the mode based on IAV in a linked genomic region. The delayed onset of positive selection for European LP suggests the initial operation on genetically indistinguishable T haplotypes ($t_{LCT} \leq t_{SEL}$). This may result from either where selection operated on a single ancestral lineage or where multiple descendants, if involved, did not have enough time to differentiate from each other by mutations. In any event, both cases lead to $t_{LCT} \leq t_{SEL} < t_{AGE}$ and look like hard sweeps. Evaluation of t_{AGE} may be directly made from the age estimate of the T allele, but here we used the TMRCA ($t_{LCT(A)}$) of the derived A allele group defined at rs182549 as a surrogate of the age of the T allele ($t_{AGE(T)}$). The widespread T allele occurred in the background of the A allele and $t_{AGE(T)}$ is bounded by $t_{LCT(A)} = 23{,}100 \pm 20{,}000$ in EUR. Thus, $t_{AGE(T)}$ must be somewhere between $t_{LCT(T)} = 3280$ and $t_{LCT(A)} = 23{,}100$. As $t_{LCT(A)}$ is well after non-African populations differentiated from each other, it is likely that both the T and A alleles originated in a single locally differentiated Eurasian population. However, there is an exception to this conclusion. From genotyping a ~30 kb *LCT* region in >1600 samples from 37 global populations, Enattah et al. (2007) found that the T mutation occurred twice independently on highly divergent haplotype backgrounds and that one mutation is geographically restricted to populations living in west of the Urals and north of the Caucasus.

An interesting caveat is related to Tuscans. Clearly, the present-day TSI is an outlier in terms of European LP, exhibiting an exceptionally low T allele frequency in a large sample. By comparing TSI genomes with other Europeans and Middle Easterners, Pardo-Seco et al. (2014) revealed that admixture took place between 1100 and 600 BC. Although this admixture event may have been much older and even recurrent, it implies the partial eastern origin of the non-Indo-European Etruscans, ancestral to Tuscans in the Bronze/Iron Age. Intriguingly, some cattle breeds (*Bos taurus*) were also simultaneously imported from the east Mediterranean area (Pellecchia et al., 2007). However, the T allele in TSI suggests that this cattle import did not act as a selective agency for LP due presumably to lack of relevant culture or absence of the T allele *per se* at that time. We emphasize that, although evidence for positive selection is relatively weak for TSI, the Brown's combining method could detect a hard sweep for the T allele

group (Table 2b). One possibility is that the T allele in TSI resulted from recent gene flow from neighboring LP areas. Coelho et al. (2005) estimated the LP frequency to be 24% and the T allele frequency to be 13% in Central Italy (37 individuals from Tocco da Causaria and 30 from Rome). While these frequencies in Central Italy are much lower than those in Portugal, they are much higher than those in Tuscany and Central Italian genomes would show a hard sweep if statistically tested.

TSI is exceptional also in terms of pairwise LD between C/T at rs4988235 and G/A at rs182549. While the TA haplotype at these SNPs is in complete LD ($r^2 = 1$) in CEU, FIN, GBR, and IBS, it is not ($r^2 = 0.95$) in TSI, owing to the presence of a CA haplotype in addition to derived TA and ancestral CG haplotypes (Figure 2). As a number of CA haplotypes can be also found in MXL (Mexican Ancestry from Los Angeles) and PEL (Peruvian in Lima) from admixed Americans (AMR), as well as BEB (Bengali in Bangladesh), ITU (Indian Telugu in the UK), PJL, and STU (Sri Lankan Tamil from the UK) from South Asian Ancestry (SAS), the age of the A allele must be older than that of the T allele as aforementioned. Interestingly, the T allele frequency is quite high in some of these populations: 24% in MXL, 10% in PEL, and 26% in PJL. Yet, IAV is substantially reduced in MXL or even non-existent in PEL (i.e., $t_{LCT} = 0$). These features are compatible with very recent ongoing selective sweeps, but a more likely cause would be admixture, founder effects, or a combination of these factors after Columbus. Similarly, IAV in PJL is significantly reduced ($\chi_f^2 = 23.3$ with $c = 2.01$ and $f = 2.98$; $P = 4 \times 10^{-5}$) with $t_{LCT} = 3600 \pm 1320$ (Figure 2). Although admixture also occurred in PJL, we consider a prehistoric event as a more likely cause: if the T allele arose in the Pontic Steppe, it must have spread not only into Europe from 3000 BC, but also into Turan by 2100 BC and further into South Asia after the decline of the Indus Valley Civilization in ca. 1800 BC (Damgaard et al., 2018; Narasimhan et al., 2019; Shinde et al., 2019). As the time elapsed in Punjabi was as long as 3800 years, it is tempting to speculate the presence of an independent selective sweep for the T allele in this locality too. Thus, it appears that LP in Europe and South Asia shares the early history of the expanding Steppe ancestry in the Bronze Age, and offers strong selective advantages to its local bearers in lactose-relevant niche construction.

ACKNOWLEDGEMENTS

We thank Drs. Seiji Kadowaki, Quintin Lau and Kenichi Aoki for their interest and comments. This work was supported in part by the Scientific Research on Innovative Areas, a MEXT Grant-in-Aid Project FY2016-2020.

REFERENCES

1000 Genomes Project Consortium; Auton, A., Brooks, L. D. , Durbin, R. M., Garrison, E. P., Kang, H. M., et al., 2015. A global reference for human genetic variation, *Nature* 526: 68-74.

Akbari, A., Vitti, J. J., Bakhtiari, M., Sabeti, P. C., Mirarab, S., and Bafna, V., 2018. Identifying the favored mutation in a positive selective sweep, *Nature Methods*, 15: 279-282.

Allentoft, M. E., Sikora, M., Sjögren, K-G., Rasmussen, S., Rasmussen, M., et al., 2015. Population genomics of Bronze Age Eurasia, *Nature*, 522: 167-172.

Aoki, K., 1986. A stochastic model of gene-culture coevolution suggested by the "culture historical hypothesis" for the evolution of adult lactose absorption in humans, *Proceedings of National Academy of Sciences of the USA*, 83(9): 2929-2933.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., et al., 2004. Genetic signatures of strong recent positive selection at the lactase gene, *American Journal of Human Genetics*, 74: 1111–1120.

Brace, S., Diekmann, Y., Booth, T. J., Faltyskova, Z, Rohland, N., et al., 2018. Population replacement in early Neolithic Britain, *bioRχiv,* preprint doi: https://doi.org/10.1101/267443.

Bramanti, B., Thomas, M. G., Haak, W., Unterlander, M., Jores, P., et al., 2009. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers, *Science*, 326: 137-140.

Broushaki, F., Thomas, M. G., Link, V., López, S., van Drop, L., et al., 2016. Early Neolithic genomes from the eastern Fertile Crescent, *Science*, 353: 499-503.

Brown, M. B., 1975. A method for combining non-independent, one-sided tests of significance, *Biometrics*, 31: 987-992.

Burger, J., Kirchner, M., Bramanti, B., Haak, W., and Thomas M. G., 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans, *Proceedings of National Academy of Sciences of the USA*, 104(10): 3736-3741.

Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A. I., Seixas, S., et al., 2005. Microsatellite variation and evolution of human lactase persistence, *Human Genetics,* 117: 329-339.

Damgaard, P. B., Martiniano, R., Kamm, J., Moreno-Mayar, J. V., Kroonen, G., et al., 2018. The first horse herders and the impact of early Bronze Age Steppe

expansions into Asia, *Science*, 360: eaar7711.

Edmonds, C. A., Lillie, A. S., and Cavalli-Sforza, L. L., 2004. Mutations arising in the wave front of an expanding population, *Proceedings of the National Academy of Sciences of the USA,* 101:975–979.

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: 233– 237.

Enattah, N. S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., et al., 2007. Evidence of still-ongoing convergence evolution of the lactase persistence T_{-13910} alleles in humans, *The American Journal of Human Genetics*, 81: 615-625.

Evershed, R. P., Payne, S., Sherratt, A. G., Copley, M. S., Coolodge, J., et al., 2008. Earliest date for milk use in the Near East and southern Europe linked cattle herding, *Nature*, 455: 528-531.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., et al., 2016. Detection of human adaptation during the past 2000 years, *Science*, 354: 760-764.

Fujito, N. T., Satta, Y., Hayakawa, T., and Takahata, N., 2018. A new inference method for ongoing selective sweep, *Genes and Genetics Systems*, 93: 149-161.

Furtwängler, A., Rohrlach, A. B., Lamnidis, T. C., Papac, L., Neumann, G. U., et al., 2020. Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland, *Nature Communications,* 11: 1915.

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., et al., 2014. Genome flux and stasis in a five millennium transect of European prehistory, *Nature Communications,* 5: 5257.

Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, P., et al., 2011. Evolution of lactase persistence: an example of human niche construction, *Proceedings of the Royal Society B*, 366: 863-877.

Gerbault, P., Roffet-Salque, M. , Evershed, R. P., and Thomas, M. G., 2013. How long have adult humans been consuming milk? *IUBMB Life*, 65: 983-990.

Günther, T., Valdiosera, C., Malmström, H., Ureña, I., Rodriguez-Varela, R., et al., 2015. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques, *Proceedings of the National Academy of Sciences of the USA,* doi/10.1073/pnas.1509851112.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., et al., 2016. Massive migration from the steppe was a source for Indo-European languages in Europe, *Nature*, 522: 207-211.

Haber, M., Mezzavilla, M., Xue, Y., and Tyler-Smith, C., 2016. Ancient DNA and the rewriting of human history: Be sparing with Occam's razor, *Genome Biology*, 17: 1. https://doi.org/10.1186/s13059-015-0866-z.

Harris, A. M., Garud, N. R., and DeGiorgio, M., 2018. Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity, *Genetics*, 210: 1429-1452.

Hermisson, J., and Pennings, P. S., 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation, *Genetics*, 169: 2335-2352.

Hermisson, J., and Pennings, P. S., 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation, *Methods in Ecology and Evolution*, 8: 700–716.

Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., et al., 2016. Early farmers from across Europe directly descended from Neolithic Aegeans, *Proceedings of the National Academy of Sciences of the USA,* 113:6886-6891.

Hudson, R. R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation, *Bioinformatics*, 18: 337-338.

Ingram, C. J., Elamin M. F., Mulcare, C. A., Weale, M. E., Tarekegn, A., et al., 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Human Genetics*, 120: 779-788.

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 genotype, *MBC Evolutionary Biology*, 10:36, doi:10.1186/1471-2148-10-36.

Itan, Y., Powell, A., Beaumont, M. A., Burger, J., and Thomas, M. G., 2009. The origins of lactase persistence in Europe, *PLoS Computational Biology*, 5: e1000491.

Keller, A., Graefen, A., Ball, M., Matzas, M., Boisquerin, V., et al., 2012. New insights into the Tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing, *Nature Communications*, 3: 698.

Kimura, M., 1968. Evolutionary rate at the molecular level, *Nature*, 217: 624-626.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics*, 61: 893-903.

Kost, J. T., and McDermott, M. P., 2002, Combining dependent p-values, *Statistics and Probability Letters*, 60: 183-190.

Leonardi, M., Gerbault, P., Thomas, M. G., and Burger, J., 2012. The evolution of lactase persistence in Europe. A synthesis of archaeological and genetic evidence, *International Dairy Journal*, 22: 88-97.

Li, H., and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences, *Nature*, 475: 493-496.

Liebert, A., López, S., Jones, B. L., Montalva, N., Gerbault, P., et al., 2017. World-wide distribution of lactase persistence alleles and the complex effects of recombination and selection, *Human Genetics*, 136: 1445-1453.

Linderholm, A., Kılınç, G. M., Szczepanek, A., Włodarczak, P., Jarosz, P., et al., 2020. Corded Ware cultural complexity uncovered using genomic and isotopic analysis from south-eastern Poland, *Scientific Reports*, 10: 6885, doi:10.1038/s41598-020-63138-w.

Liu, X., and Fu, Y-X., 2015. Exploring population size changes using SNP frequency spectra, *Nature Genetics*, 47: 555-559.

Malmström, H., Günther, T., Svensson, E. M., Juras, A., Fraser, M., et al., 2019. The genomic ancestry of the Scandinavian Battle Axe Culture people and their relation to the broader Corded Ware horizon, *Proceedings of the Royal Society B*, 286: 20191528.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A.,Rohlannd, N., et al., 2018. The genomic history of Southeastern Europe. Nature 555: 197-203.

Mathieson. I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., et al., 2015. Genome-wide patterns of selection in 230 ancient Eurasians, *Nature*, 528: 499-503.

Nakagome, S., Hudson, R. R., and Di Rienzo, A., 2019. Inferring the model and onset of natural selection under varying population size from the site frequency spectrum and haplotype structure, *Proceedings of the Royal Society B*, 286: 20182541.

Narashimhan, V. M., Patterson, N., Moorjani, P., Rohland, N., Bernardos, R., et al., 2019. The formation of human populations in South and Central Asia, *Science*, 365: eaat7489.

O'Brien, M. J., and Laland, K. N., 2012. Genes, culture, and agriculture: An example of human niche construction. *Current Anthropology*, 53(4): 434-170.

Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., et al., 2018. The Beaker phenomenon and the genomic transformation of northwest Europe, *Nature*, 555: 190-196.

Olalde, L., Mallick, S., Patterson, N., Rohland, N., Villalba-Mouco, V., et al., 2019. The genetic history of the Iberian Peninsula over the past 8000 years, *Science*, 363: 1230-1234.

Olalde, I., Schroeder H., Sandoval-Velasco, M., Vinner, L., Lobón, I., et al., 2015. A common genetic origin for early farmers from Mediterranean Cardial and central

European LBK cultures, *Molecular Biology and Evolution*, 32(12): 3132-3142.
Olds, L. C., and Sibley, E., 2003. Lactase persistence DNA variant enhances lactase promoter activity *in vitro*: functional role as a *cis* regulatory element, *Human Molecular Genetics*, 12 (18): 2333-2340.
Pardo-Seco, J., Gómez-Carballa, A., Amigo, J., Martinón-Torres, F., and Salas, A., 2014. A genome-wide study of modern-day Tuscans: Revisiting Herodotus's theory on the origin of the Etruscans, *PLoS ONE*, 9: e105920.
Pellecchia, M., Negrini, R., Colli, I., Patrini, M., Milanesi, E., et al., 2007. The mystery of Etruscan origins: novel clues from *Bos taurus* mitochondrial DNA, *Proceedings of the Royal Society B*, 274: 1175-1179.
Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B., Knijnenberg, T. A., 2016. Combining dependent P-values with an empirical adaptation of Brown's method, *Bioinformatics*, 32: i430-i436.
Plantinga, T. S., Alonso, S., Izagirre, N., Hervella, M., Fregel, R., et al., 2012. Low prevalence of lactase persistence in Neolithic South-West Europe, *European Journal of Human Genetics*, 20: 778-782.
Raveane, A., Aneli, S., Montinaro, F., Athanasiadis, G., Barlera, S., et al., 2019. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe, *Science Advances*, 5: eaaw3492.
Rivollat, M., Jeong, C., Schiffels, S., Küçükkalıpçı, İ., Pemonge, M-H., et al., 2020. Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers, *Scientific Advances*, 6: eaaz5344.
Saitou, N., and Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, 4: 406-425.
Satta, Y., Zheng, W., Nishiyama, K. V., Iwasaki, R. L., Hayakawa, T., et al., 2019. Two-dimensional site frequency spectrum for detecting, classifying and dating incomplete selective sweeps, *Genes and Genetic Systems*, 94: 283-300.
Scally, A., and Durbin, R., 2012. Revising the human mutation rate: implications for understanding human evolution, *Nature Reviews Genetics*, 1: 745-753.
Schaffner, S. F., Foo. C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D., 2005. Calibrating a coalescent simulation of human genome sequence variation, *Genome Research*, 15: 1576-1583.
Schiffels, S., and Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences, *Nature Genetics*, 46: 919-925.
Ségurel, L., and Bon, C., 2017. On the evolution of lactase persistence in humans,

Annual Review of Genomics and Human Genetics, 18: 297-319.

Shinde, V., Narasimhan, V. M., Rohland, N., Mallick, S., Mah, M., et al., 2019. An ancient Harappan genome lacks ancestry from Steppe pastoralists or Iranian farmers, *Cell*, 179: 729-735.

Skoglund, P., Malmstrom, M., Raghavan, J., Stora, P., Hall, E., et al., 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe, *Science*, 336: 466-469.

Skoglund, P., Malmstrom, H., Omrak, A., Raghavan, M., Valdiosera, C., et al., 2014. Genetic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers, *Science*, 344: 747-750.

Smith, J., Coop, G., Stephens, M., and Novembre, J., 2018. Estimating time to the common ancestor for a beneficial allele, *Molecular Biology and Evolution*, 35(4): 1003-1017.

Souilmi, Y, Tobler, R., Johar, A., Williams, M., Shane T. Grey, S. T., et al., 2020. Ancient human genomes reveal a hidden history of strong selection in Eurasia, *bioRχiv,* preprint doi: https://doi.org/10.1101/2020.04.01.021006.

Terhorst, J., Kamm, J. A., and Song, Y. S., 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes, *Nature Genetics*, 49: 303-309.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., et al., 2007. Convergent adaptation of human lactase persistence in Africa and Europe, *Nature Genetics*, 39: 31-40.

Tournebize, R., Poncet, V., Jakobsson, M., Vigouroux, Y., and Manel, S., 2019. McSwan: A joint site frequency spectrum method to detect and date selective sweeps across multiple population genomes, *Molecular Ecology Resources*, 19: 283-295.

Wilson, B. A., Petrov, D. A., Messer, P. W., 2014. Soft selective sweeps in complex demographic scenarios, *Genetics*, 198: 669-684.

Witas, H. W., Płoszaj, T., Jędrychowska-Dańska, K., Witas, P. J., Masłowska, A., et al., 2015. Hunting for the LCT-13910*T allele between the Middle Neolithic and the Middle Ages suggests its absence in dairying LBK people entering the Kuyavia region in the 8th millennium, *PLOS ONE*, 10: 1-24.

Table 1. *Frequencies of the T allele at rs4988235 in ancient and present-day genomes in Central Europe.*

Central Europe	Cultures in Europe and Steppe	a	b	c
Paleolithic 43,000-10,000 BC	Pleistocene hunter-gatherer	0	-	0
Mesolithic 9700-6200 BC	Mesolithic hunter-gatherer 8.2 event [d]	0.17 [e]	0	0
Early Neolithic 6000-4000 BC	Holocene hunter-gatherer LBK (5500 -4800BC)	0	0	0
Middle Neolithic 4000-3000 BC	Late Copper Age Yamnaya (3000-2400 BC)	0	0	0
Late Neolithic 2600-2200 BC	CWC (2800-2300 BC) BBC (2600-2000 BC)	~0.20	~0.05	~0.05 [f]
Bronze Age 2200-1000 BC	Sintashta (2100-1800 BC) Andronovo (1700-1500 BC)	0.05	~0.05	-
Iron Age 900 BC		-	-	~0.05 [f]
Modern Europeans [g]		0.67	0.74	0.46

a, Allentoft et al. (2015); b, Mathieson et al. (2015); c, Olalde et al. (2018); d, time of an abrupt climate change caused by a glacier meltwater outburst 8200 years ago; e, after Figure 4a and Supplementary Table 1, but 0 when inferred from imputation of ancient genomes (Figure 4b); f, extended data Figure 7 in Olalde et al. (2018); g, North Europe, CEU, and IBS, respectively. The chronology of Central Europe (left) is taken from Haak et al. (2015).

Table 2a. *IAV variables in the 100 kb LD region surrounding rs4988235 in five European populations in the 1000GD (sample size* n *and derived allele frequency* $f_r = m/n$*).*

Population	CEU	FIN	GBR	IBS	TSI
α	1.3×10^{-4}	3.6×10^{-5}	5.1×10^{-5}	6.0×10^{-5}	5.1×10^{-5}
m/n	146/198	117/198	131/182	98/214	19/214
S_ξ	213	218	204	287	351
F_c	0.0141	0.0076	0.0135	0.0017	0.0059
(P)	(< 10^{-4})	(< 10^{-4})	(< 10^{-4})	(< 10^{-4})	(0.015)
L_{c0}	0.0026	0.0021	0.0026	0.0003	0.0002
(P)	(< 10^{-4})	(< 10^{-4})	(< 10^{-4})	(< 10^{-4})	(0.013)
G_{c0}	1.09	1.25	1.16	1.50	1.00
(P)	(< 10^{-4})	(< 10^{-4})	(< 10^{-4})	(< 10^{-4})	(0.036)
C	0.164	0.171	0.168	0.184	0.105
(S. E.)	(0.036)	(0.047)	(0.042)	(0.066)	(0.074)

The P values in parentheses were computed from simulation results of 10^4 replications of *ms* (Hudson, 2002) under the bottleneck and exponential growth model with the per-year rate α (Figure 1). The variables are defined as $F_c = \frac{\sum i\varphi_{i,j}}{\sum (i+j)\varphi_{i,j}}$, $L_{c0} = L_{\zeta 0}/L_\xi$, $G_{c0} = L_{\zeta 0}/S_{\zeta 0}$ and $C = L_{\zeta 0}/m$ where Σ in F_c indicates that the summation is taken over a certain range of i and j (see Fujito et al., 2018 and Satta et al., 2019 for details), $S_{\zeta 0} = \sum_{i=1}^{m-1} \varphi_{i,0}$, $L_{\zeta 0} = \sum_{i=1}^{m-1} i\varphi_{i,0}$, $S_\xi = \sum_{i=0}^{m} \sum_{j=0}^{n-m} \varphi_{i,j}$, and $L_\xi = \sum_{i=0}^{m} \sum_{j=0}^{n-m} (i+j)\varphi_{i,j}$.

Table 2b. P values obtained by the method that combines non-independent, one-sided F_c, L_{c0} and G_{c0} variables.

Population	CEU	FIN	GBR	IBS	TSI
$\rho(F_c \cdot L_{c0})$	0.383	0.539	0.470	0.551	0.594
$\rho(F_c \cdot G_{c0})$	0.124	0.422	0.132	0.560	0.357
$\rho(L_{c0} \cdot G_{c0})$	0.535	0.532	0.503	0.591	0.497
χ_f^2	34.1	29.1	33.3	27.1	13.4
(c) (f)	(1.62) (3.7)	(1.90) (3.2)	(1.66) (3.6)	(2.04) (2.9)	(1.88) (3.2)
Combined P	$< 3 \times 10^{-7}$	$< 3 \times 10^{-6}$	$< 7 \times 10^{-7}$	$< 5 \times 10^{-6}$	0.005

The combined variable is given by $X^2 = \sum_{l=1}^{k}(-2 \ln P_l)$ where P_l is the probability that the l-th variable exceeds the observed value under the null hypothesis. The minimum value of $X^2 = 55.3$ corresponding to the three maximum P values of 10^{-4} (Table 2a) is used for the first four populations. $X^2/c = \chi_f^2$ is approximated by a chi-square variable with f degrees of freedom, provided that c and f are computed from the correlation coefficient $\rho = \rho(Y \cdot Z)$ between variables Y and Z. Brown (1975) fitted a quadratic function of ρ to covariance $Cov(-2\ln P_l, -2 \ln P_{l'})$, whereas Kost and McDermott (2002) fitted a third-order polynomial (Poole at el., 2016). Except for two cases between F_c and G_{c0}, ρ is as high as 0.6; yet, the combined method rejects the null hypothesis of neutrality at a level of significance of at most 0.005 (TSI).

Figure Legends

Figure 1. *Demographic models for five European populations.*

The effective population size (Y axis in log scale) is depicted against the past time (X axis in years). Four colored lines show population-specific demographic models with different exponential growth rates (α) (Table 1). The models assume that the size (500 years ago) was 600,000 for CEU, 30,000 for IBS, 20,000 for GBR and TSI, and 10,000 for FIN. All populations began to expand 45,000 years ago after experiencing a bottleneck phase of constant size 2000 since this ancestral Eurasian population migrated out of Africa 58,000 years ago. The ancestral African population size was 20,000.

Figure 2. *Unrooted NJ tree for a subsample of the T and C distinct haplotypes.*

The T/C sample size is 23/9 from CEU, 13/9 from IBS, 22/5 from GBR, 4/12 from TSI, 20/7 from FIN, and 11/15 from PJL. The LD region in each population is the same as that in the entire EUR sample (Satta et al., 2019). TA haplotypes at rs4988235 and rs182549 are marked by filled colored circles and CA haplotypes in TSI and PJL by two open inverse triangles. The tree was constructed from pairwise nucleotide differences without multiple-hit correction (Saitou and Nei, 1987). The inset is a cartoon of the haplotype evolution. The G to A mutation occurred >23,000 years ago, the C to T mutation emerged much later in an A lineage and swept through a population. The CA haplotype in TSI and PJL is a remnant of the A mutation that escaped from the selective sweep.

Figure 3. *Identification of a target site of selection*.

We searched for a target site from candidate SNPs surrounding *LCT* (X axis for genomic positions) by ranking their F_c and iSAFE values (Y axis). In F_c, candidate SNPs were all derived alleles whose frequencies are >45% in EUR and >54% in FIN. Note that, although derived alleles at low frequencies were excluded from candidates, they were used to compute F_c. The red points indicate the F_c or iSAFE value for the T allele at rs4988235. Needless to say, the ranking by F_c is in ascending order (Fujito et al., 2018), whereas that of iSAFE is in descending order (Akbari et al., 2018). (A) , (B) F_c scanning over the 2.5 MB genomic region in the EUR sample of size $n = 1006$ and the FIN sample of size $n = 189$; (C) iSAFE scanning over the 2.5 MB genomic region in the FIN sample.

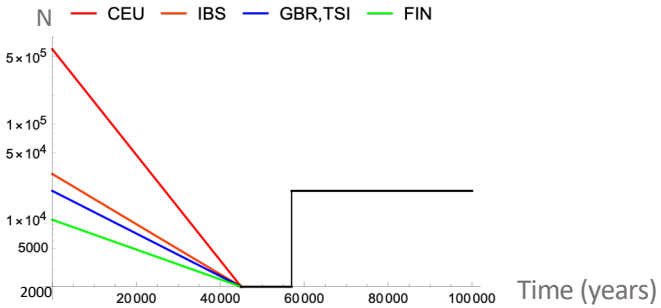
N
CEU
IBS
GBR,TSI
FIN
5 × 10^5
1 × 10^5
5 × 10^4
1 × 10^4
5000
2000
20 000
40 000
60 000
80 000
100 000
Time (years)

Figure 1

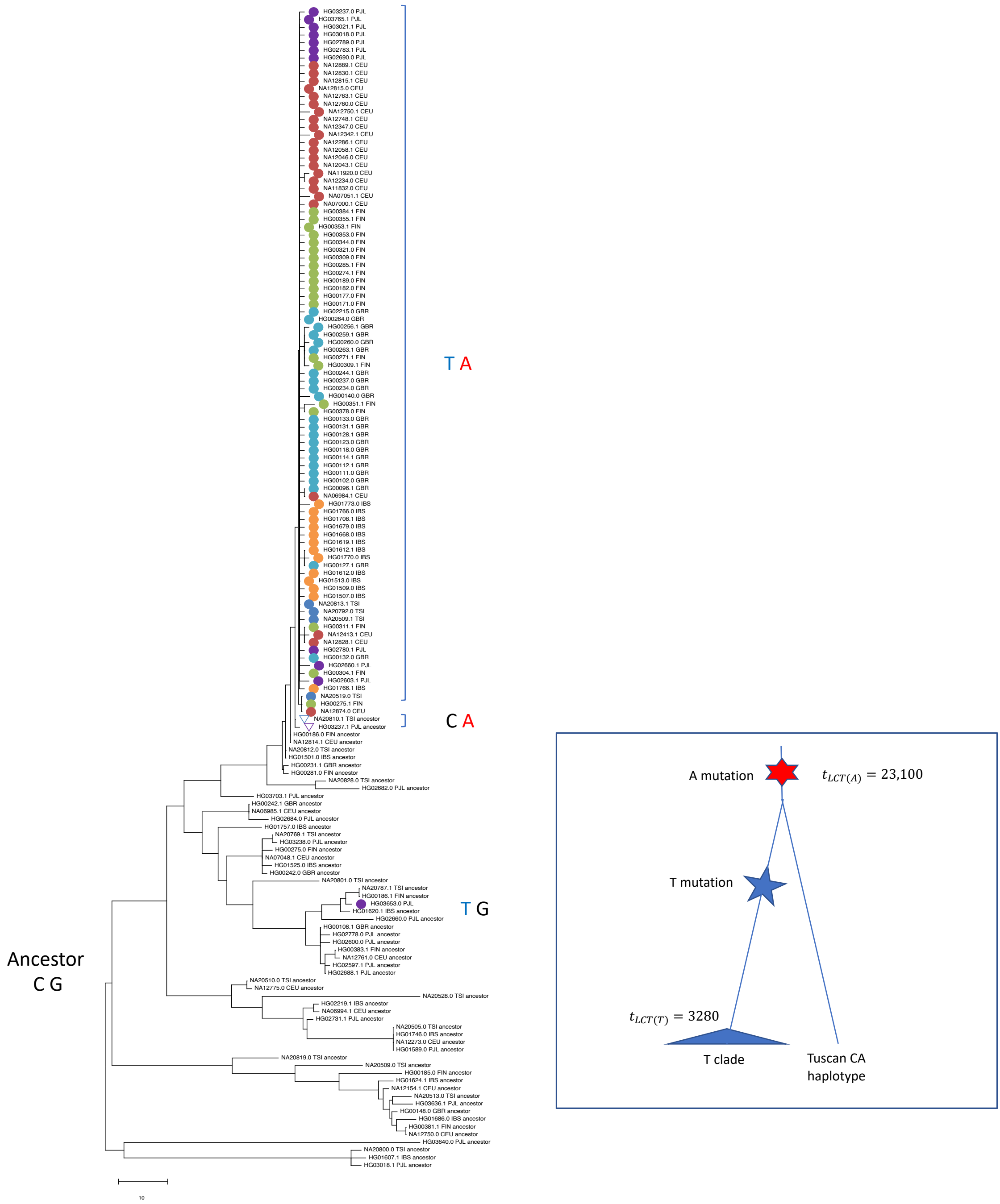
T A
C A
T G
Ancestor
C G
A mutation
$t_{LCT(A)} = 23{,}100$
T mutation
$t_{LCT(T)} = 3280$
T clade
Tuscan CA
haplotype
10

Figure 2

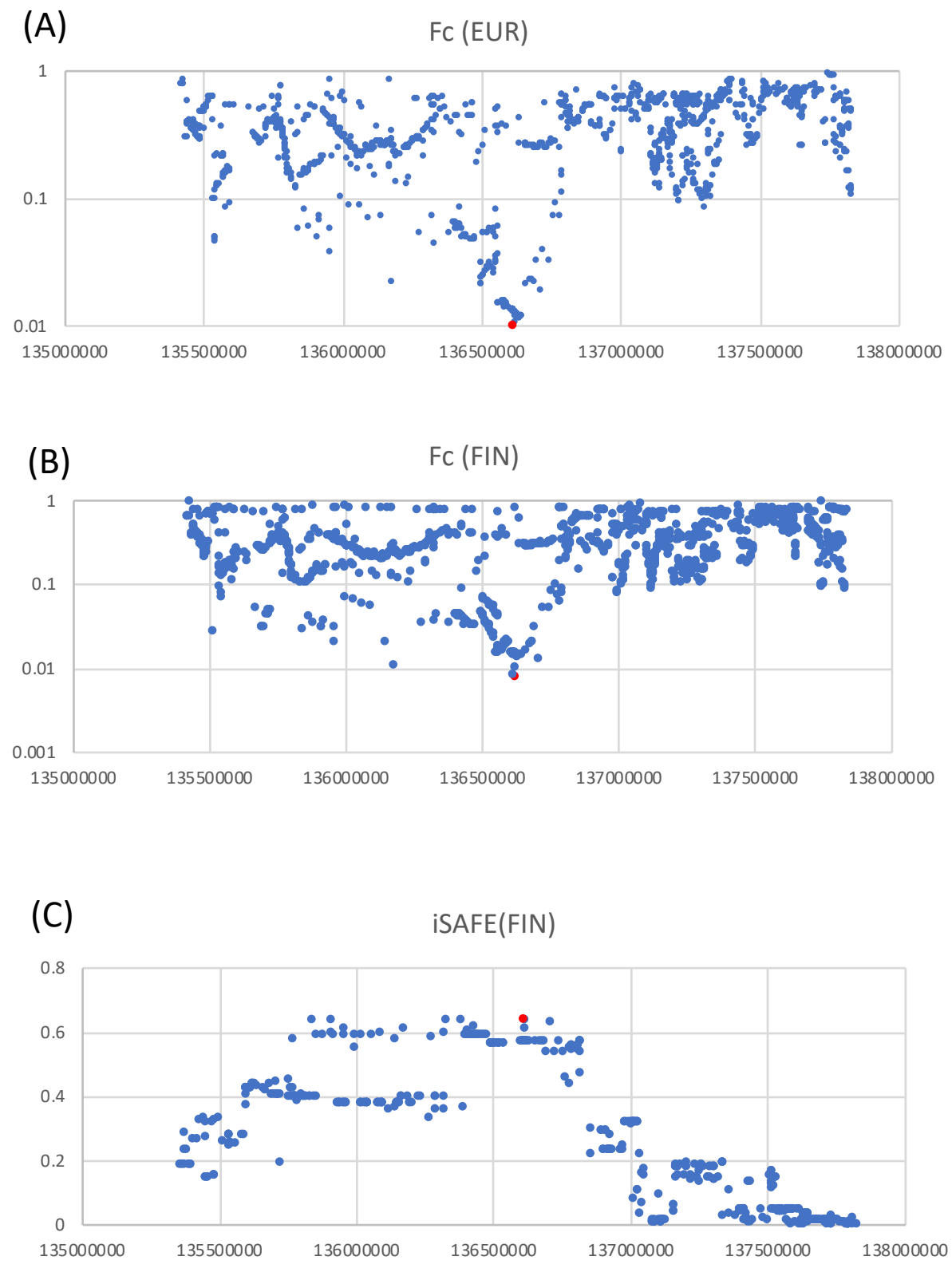
(A)
Fc (EUR)
1
0.1
0.01
135000000
135500000
136000000
136500000
137000000
137500000
138000000
(B)
Fc (FIN)
1
0.1
0.01
0.001
135000000
135500000
136000000
136500000
137000000
137500000
138000000
(C)
iSAFE(FIN)
0.8
0.6
0.4
0.2
0
135000000
135500000
136000000
136500000
137000000
137500000
138000000

Figure 3