Human Biology Open Access Pre-Prints

WSU Press

11-22-2017

Leveraging Multiple Populations across Time Helps Define Accurate Models of Human Evolution: A Reanalysis of the Lactase Persistence Adaptation

Chenling Xu Antelope
University of California, Berkeley

Davide Marnetto
University of Torino, Turin, Italy

Fergal Casey
University of California Merced

Emilia Huerta-Sanchez
University of California Merced, sanchez@ucmerced.edu

Leveraging Multiple Populations across Time Helps Define Accurate Models of Human Evolution: A Reanalysis of the Lactase Persistence Adaptation

Chenling Xu Antelope,[1,#] Davide Marnetto,[2,#] Fergal Casey,[3] and Emilia Huerta-Sanchez[3,*]

[1]Computational Biology, University of California, Berkeley, Berkeley, California.

[2]Department of Molecular Biotechnology and Health Sciences, University of Torino, Turin, Italy.

[3]School of Natural Sciences, University of California Merced, Merced, California.

[#]Contributed equally

E.H.S is supported by UC Merced start-up funds and an NSF-DEB 1557151 grant.

*Corresponding author: Emilia Huerta-Sanchez, Molecular and Cell Biology, School of Natural Sciences, University of California, Merced, 5200 Lake Rd., Merced, CA 95340. E-mail: ehuerta-sanchez@ucmerced.edu.

Using Modern and Ancient DNA to Characterize Human Adaptation

KEY WORDS: POSITIVE SELECTION, ADAPTATION, ANCIENT DNA

Abstract Access to a geographically diverse set of modern human samples from the present time and from ancient remains, combined with archaic hominin samples, provides an unprecedented level of resolution to study both human history and adaptation. The amount and quality of ancient human data continues to improve, and enables tracking the trajectory of genetic variation over time. These data have the potential to help us redefine or generate new hypotheses of how human evolution occurred, and revise previous conjectures. In this review, we argue that leveraging all these data will help us better detail adaptive histories in humans. As a case in point, we focus on one of the most celebrated examples of human adaptation: the evolution of lactase persistence. We briefly review this dietary adaptation, and argue that, effectively, the evolutionary history of lactase persistence is still not fully resolved, and propose that by leveraging data from multiple populations across time and space, we find evidence of a more nuanced history than just a simple selective sweep. We support our hypotheses with simulation results and make some cautionary notes regarding the use of haplotype-based summary statistics to estimate evolutionary parameters.

Understanding the evolutionary forces that human populations have been subject to in the past has been a central issue in evolutionary biology. In particular, determining how two principal mechanisms, genetic drift and natural selection, have shaped genetic variation remains a highly studied area of research. One of the difficulties in addressing this issue is that there are many forms of natural selection, and sometimes various types of selection lead to similar genetic patterns. For example, dips in genetic diversity along the genome have been associated with the effects of both negative and positive selection or with a neutral scenario in a region with low mutation rate (Otto, 2000). Furthermore, positive selection can occur through different mechanisms. For example, in humans, most studies have focused on two models: selection on a *de novo* mutation (SDM) (Otto, 2000, Bustamante et al. 2001, Nielsen 2005, Nielsen et al. 2007) and selection from standing genetic variation (SSV) (Przeworski, Coop, and Wall 2005, Fu and Akey 2013). In the former, a new mutation enters a population and it immediately confers an adaptive advantage. In the latter, a neutral mutation already pre-exists in a population at some frequency and only at a later time, due to an environmental change, becomes beneficial. More recently, another scenario has emerged as responsible for some adaptations in modern humans. In this scenario, adaptation is facilitated through the acquisition of mutations from archaic hominins, and has been termed adaptive introgression (AI) (Racimo et al. 2015, Hawks and Cochran, 2006, Xu et al. 2017). Notably, we can only confidently ascertain AI because we now have access to archaic human DNA, and without this resource, in the absence of making inferences of archaic sequence based on LD statistics (Plagnol and Wall, 2006) or frequencies (Hawks and Cochran, 2006), we would have likely ascribed particular examples of selection to either SDM or SSV. Therefore, as we compile more DNA from archaic hominins (e.g. Neanderthal (Green et al. 2010, Prüfer et al. 2014), Denisovan (Prüfer et al. 2014) (Reich et al.

2010, Meyer et al. 2012) and different modern human populations from the present and the past (Rasmussen et al. 2010, Rasmussen et al. 2011, Keller et al. 2012, Skoglund et al. 2012, Lazaridis et al. 2014, Fu et al. 2014, Seguin-Orlando et al. 2014, Schroeder et al. 2015), we have an unprecedented opportunity to study genetic variation across time and space axes. Joint analyses of these data have the potential to elucidate human evolution in a manner never before possible. Consequently, as we continue to acquire more data, we need to verify or update previous findings and better define models of human evolution.

Detecting Positive Selection in Present Day Human Populations. Determining the contributions of natural selection in the evolution of a population is one of the central questions in human evolutionary biology, as we seek the genetic basis underlying adaptations to the diverse challenging environments inhabited by humans. The HapMap (International HapMap Consortium, 2005) consortium produced one of the first large population genetic data sets from a sample of Africans, Europeans and Asians. The HapMap data was generated from genotyping arrays that probe allelic information for a specific set of known polymorphic positions in the genome. More recently, the 1000 Genomes project produced the whole genome sequences of more than 2500 individuals encompassing 26 populations from Africa, Europe, Asia and the Americas (1000 Genomes Project Consortium, 2012, 1000 Genomes Project Consortium, 2015). Using these data, genome-wide scans were performed to detect signatures of positive selection using summary statistics that exploit haplotype or frequency information. For example, haplotype-based statistics such as iHS (Voight et al. 2006), EHH (Sabeti et al. 2002), XP-EHH (Sabeti et al. 2007) and nSL (Ferrer-Admetlla et al. 2014) (Supplementary Table 1) are based on the observation that haplotype homozygosity should be greater around a positively selected locus than in a neutrally evolving locus. In parallel, statistics depending on allele frequencies such as the SFS (Bustamante et al. 2001, Nielsen 2005), F_{ST} (Reynolds, Weir, and Cockerham, 1983) and PBS (Yi et al. 2010) (Supplementary Table 2) can be applied to one, two and three populations respectively to identify regions under positive selection, and have been widely applied to both the HapMap and 1000 Genomes data sets (1000 Genomes Project Consortium, 2015, Sabeti et al. 2007). The SFS is the distribution of mutation frequencies present in a population, and in the presence of positive selection skews towards a larger proportion of high frequency variants. F_{ST} measures population differentiation from allele frequencies observed in a pair of populations, and local regions of large differentiation compared to the genome-wide F_{ST} value may be due to positive selection acting on genetic variation in one of the populations. The population branch statistic (PBS) calculates the allele frequency change at a given locus since its divergence from two other populations, one a closely related population and the other acting as a reference population. The statistic is functionally related to the F_{ST} statistic for population differentiation, but unlike F_{ST} between 2 populations, the reference population controls for local rates of drift.

These haplotype and frequency based approaches have led to the identification of several genomic regions under positive selection. Among the genes detected are those involved in lactase persistence (*LCT* (Voight et al. 2006, Sabeti et al. 2007, Bersaglieri et al. 2004, Enattah et al. 2002)), hair thickness (*EDAR* (Sabeti et al. 2007, Fujimoto et al. 2007, Kimura et al. 2009)), skin pigmentation (*SLC24A5* (Sabeti et al. 2007, Lamason et al. 2005)), and eye color (*HERC2/OCA2* (Sabeti et al. 2007, Sturm et al. 2008, Eiberg et al. 2008, Han et al. 2008)). Other genes are involved in the response to hypoxia (*EPAS1* (Yi et al. 2010, Simonson et al. 2010, Bigham et al. 2010, Peng et al. 2010, Xu et al. 2010, Wang et al. 2011, Beall et al. 2010), *EGLN1* (Yi et al. 2010, Simonson et al. 2010, Bigham et al. 2010, Peng et al. 2010, Xu et al. 2010, Wang et al. 2011), *BHLHE1* (Huerta-Sanchez et al. 2013)), immunity (*HLA-DQA1* (Lindo et al. 2016), *TLR* (Barreiro et al. 2009)) and metabolism (*TBX15* (Fumagalli et al. 2015, Racimo, Marnetto, and Huerta-Sánchez, 2017, Racimo et al. 2017), *FADS* (1000 Genomes Project Consortium, 2015, Fumagalli et al. 2015)). Overall it is likely these genes are subject to genetic changes in response to changing environmental conditions (Sabeti et al. 2007, Pickrell et al. 2009). Results from positive selection scans using haplotype based or frequency based statistics often identify the same loci because these summaries are sensitive to incomplete selective sweeps, i.e. when positively selected mutations are at high frequencies in the human population but have not yet reached fixation. Beneficial mutations rising in frequency due to sweeps leave a distinctive haplotype pattern: many almost identical haplotypes (low diversity) that segregate at high frequencies. However, as noted in Hernandez *et al.* (2011), the decreased level of genetic diversity (as measured by average pairwise difference, π) at nonsynonymous loci putatively under selected sweeps are generally not significantly lower than synonymous loci, and high population differentiation can alternatively be explained under a model of purifying selection, suggesting that the selective sweep model has in many cases been identifying non-neutral loci that were actually subject to purifying selection (Hernandez et al. 2011). The haplotype and frequency based statistics also have differential performance depending on the exact scenario of selection. Population differentiation statistics (F_{ST} and PBS) are more likely to pick up older sweeps, whereas homozygosity based statistics (EHH, iHS) fare better in detecting more recent sweeps (Sabeti et al. 2006). Detecting sweeps from standing variation is more difficult even though it is likely a common process in human evolution (Voight et al. 2006, Ferrer-Admetlla et al. 2014, Peter, Huerta-Sanchez, and Nielsen, 2012). Haplotype methods are prone to false positive calls of selection in regions where recombination rates are unusually low. Nevertheless, haplotype and frequency based summary statistics continue to be the mainstays for identifying positively selected loci in genomic sequence data.

Unlike discriminating between sweeps and neutral variation, these statistics have more difficulty differentiating between specific modes of positive selection, although some improvements have been made in the single population scenario (Peter, Huerta-Sanchez, and Nielsen 2012) and between populations (Key et al. 2014). The approach is to leverage multiple frequency and haplotype based summary statistics in an Approximate Bayesian Computation framework to distinguish between different modes of selection, but in some cases there is a loss of power (Peter, Huerta-Sanchez, and Nielsen 2012, Key et al. 2014). Spatial-temporal data could enhance performance in these cases. As an example, we earlier mentioned most studies have implicitly assumed that positive selection acts on *de novo* mutations (SDM) or on standing genetic variation (SSV) and summary statistics are tailored for these scenarios. However, evidence from ancient samples that predate the selection event and carry the subsequently selected allele will unambiguously indicate selection on standing variation. Likewise if a selected mutation discovered in one population is observed in a second spatially distinct population, and the time of the population split is before the estimated time of selection, we can conclude that selection happened on standing variation. Therefore both time and space may clarify between these two modes of selection.

Spatial-temporal data can go further and suggest alternative evolutionary mechanisms. For example, *EPAS1* (Yi et al. 2010), a gene involved in high altitude adaptation and identified using the PBS statistic, involves an evolutionary scenario that differs from SDM and SSV, because the beneficial variants originated in archaic hominins and modern humans acquired them through introgression referred to as adaptive introgression (Huerta-Sánchez et al. 2014, Huerta-

Sánchez 2015). Adaptive introgression leads to different patterns of genetic variation than either SSV or SDM. Under SDM, studies have shown that the expected pattern of variation leads to a reduction in effective population size around the selected region (Nielsen 2005), because the haplotype background in which the *de novo* mutation landed rises to high frequency. In contrast, under SSV, the neutral mutation may be segregating on multiple haplotype backgrounds before it becomes beneficial (Barrett and Schluter, 2008), leading to more than one beneficial haplotype present in the populations. If we consider a model of adaptive introgression, we also expect to have a single haplotype that increases in frequency in the population. However, the divergence between the selected haplotype and non-selected haplotypes will be much larger than for SSV or SDM models (Huerta-Sánchez et al. 2014). In a more recent analysis (Racimo, Marnetto, and Huerta-Sánchez 2017), we showed that both the number of uniquely archaic shared mutations and their allele frequency is powerful to detect AI, and these summaries are not sensitive to SSV or SDM suggesting that they could be used to distinguish between models. We also found that measures of linkage disequilibrium such as r^2 and D'have very little power to detect AI and high power to detect SDM, so these too could be used to distinguish between models (Racimo, Marnetto, and Huerta-Sánchez 2017).

Other examples of leveraging spatial genomic data to elucidate the mode of selection go beyond just population differentiation: clines of allele frequency can arise under geographically uniform and non-uniform selective pressures or even neutral scenarios (Novembre and Di Rienzo 2009). For some alleles this data can reveal the originating population, and correlations between allele frequency and environmental factors supply strong supplementary evidence for the selective pressures at play (Fumagalli et al. 2011). With the availability of serial samples collected over time, Wright-Fisher models or approximations thereof can be applied to explicitly define a likelihood function for a given allele trajectory (Malaspinas 2016). The likelihood function allows the estimation of crucial parameters: the age of the allele, the strength of selection and the population effective size. An extension to multilocus models allows for the inference of selection coefficients based on comparison to reference, putatively neutral, loci (Nishino 2013). These methods show that for more accurate parameter estimation, availability of samples distributed over time is generally preferable to having the same number of samples taken at a single contemporaneous time point, as intuitively the change in allele frequency over time is dramatically affected by the selection strength. In all these cases, the deviation from accepted models, either in model structure or model parameters, are often revealed when using complementary information from other time points or other populations, and therefore we suggest that incorporating those other data sources will better determine the mechanisms underlying human adaptations and will provide a more complete picture of our evolutionary history.

Integrating Ancient Data to Characterize Human Adaptation. Studying data from extant individuals has revealed several examples of human adaptation. The possibility of sequencing ancient samples, however, provides an unprecedented level of resolution to delineate a more detailed process of how positive selection acted on genetic variation through time. Additionally, we can also look for signatures of positive selection in the ancient populations and ask whether we also observe those in the population in the present. For example, one study sequenced the exomes of ancient and modern First Nations individuals of the Prince Rupert Harbour (PRH) region of British Columbia. This study found signals of positive selection in the immune-related *HLA-DQA1* gene in the ancient samples (Lindo et al. 2016). Natural selection likely acted on the genetic variation in this gene because as populations settled in the Americas, they adapted to local environments, including local pathogens. Notably, these signatures are absent in the present-day descendants of these populations. One plausible reason is that those variants became harmful when the environment changed with the introduction of new diseases, such as smallpox, during European colonization (Lindo et al. 2016). This work demonstrates the necessity of ancient DNA in deciphering past evolutionary events, and it also allows us to observe the genetic changes that occurred during the recent past and to correlate them with actual historical events.

Most studies, however, have utilized ancient DNA to provide new insights on positively selected loci identified in present-day human populations; to infer where the beneficial variants originated, when the adaptation occurred, and the magnitude of selection. As the geographic region with the highest number of sequenced ancient samples is west Eurasia (Slatkin and Racimo 2016), these data have already aided in obtaining additional details about European adaptations. For example, Allentoft *et al.* (2015) sequenced 101 ancient humans from across Eurasia from the Bronze age and compared them to populations from different time periods (Paleolithic, Mesolithic, and Neolithic) to study population migrations as well as the temporal dynamics of selected genetic variants. Among the variants are those associated with skin pigmentation (*SLC24A5, SLC45A2*), eye color (*HERC2-OCA2*) and lactase persistence (*LCT*). Interestingly, they found that while the Mesolithic hunter-gatherers carried an allele associated with blue eye color, the allele was absent in the populations from the Pontic Steppe. In subsistence related traits, they found that Bronze Age Europeans had the lactase persistence allele (rs4988235) at low frequencies. This is quite surprising because this allele is at high frequency (~70%) in Northern European populations today, suggesting a rapid increase in frequency in the very recent past. Previous estimates of the onset of selection from data collected from present-day Europeans varied widely (Bersaglieri et al. 2004, Peter, Huerta-Sanchez, and

Nielsen 2012, Burger et al. 2007, Itan et al. 2009 , Enattah et al. 2007); estimates range from the very recent (~3000y ago (Bersaglieri et al. 2004, Allentoft et al. 2015)) to older than 10000 years ago (Peter, Huerta-Sanchez, and Nielsen 2012). Interestingly, the ancient population with the largest allele frequency is the Bronze Age Yamnaya, leading to speculation that perhaps the allele originated in the steppe populations, and was introduced into Europe via migration.

Another study, Mathieson *et al.* (2015) (Mathieson et al. 2015), investigated data of 230 ancient individuals from West Eurasia from 6500 to 1000 BCE. They scanned for variants that deviated from the expected proportions of Mesolithic hunter-gatherer, early farmer and steppe pastoralist ancestries in present-day Europeans. Some of the genetic mutations that did not follow the expected proportions lie in genes involved in diet (*LCT, FADS1*), skin pigmentation (*SLC24A5, SLC45A2*), eye color (*HERC2-OCA2*) and immunity (e.g. *TLR1-TLR6-TLR10* gene cluster). Again, these two studies show how ancient DNA can inform us about human adaptation: they resolved previous highly variable estimates on the timing of selection and revealed the potential origin of the putative beneficial genetic variants.

Ancient genomes from Africa were largely left aside until very recently – challenges with the rate of DNA deterioration and contamination, that were only recently overcome, had inhibited any extensive studies. Improved methods for DNA extraction, including from petrous bones, have now started to yield some exciting ancient data sets from this continent. Unpublished studies (Schlebusch et al. 2017) have proposed a deeper in-Africa branching, dating to 260 kya ago, signatures of back-to-Africa migrations of Levantine populations and evidence of a recent Bantu expansion into southeast Africa, previously populated by San related groups. The complex within-Africa demographics that is starting to emerge, enabled by ancient DNA sequencing, has the potential to be disruptive in the field, as it will completely upend our assumptions regarding Africans as a reference population for statistics of selection and introgression.

What Is the Evolutionary Scenario Underlying the LCT Region? Integrating ancient data has already provided new insights of how this human adaptation may have occurred; the increase in frequency of the beneficial allele in *LCT* was faster than most had estimated, and as mentioned earlier perhaps acquired from populations from the Pontic Steppe (Mathieson et al. 2015, Allentoft et al. 2015). It is worth noting, however, that these insights come from considering the frequency of the putative selected SNP (rs4988235) that carries the derived allele. In addition to SNP frequency, the sequences of variation (the haplotypes) near that SNP are also informative, as positive selection leaves a distinctive haplotype pattern. We argue that inspection of the genetic variation around the putative beneficial alleles may provide additional information regarding the adaptive history of the region, even in the case of a highly studied example such as *LCT*. For example, if we inspect the haplotype patterns of individuals of North-Western European descent (CEU population in 1000 genomes project, NW Europeans from now on), the *LCT* region exhibits a long haplotype at high frequency (Figure 1). One way to summarize the observed haplotype similarity in NW European individuals is by computing the haplotype shared tract length - a pairwise measure which is simply the number of shared base pairs from the site of interest until a mismatch is encountered - at this locus (see Methods). Therefore, a derived-derived shared tract length is generated from comparing two chromosomes both with the derived allele at the site of interest (rs4988235, pointed by blue arrow in Figure 1). On the other hand, derived-ancestral is the shared tract length generated from comparing two chromosomes, one with the derived allele and the other with the ancestral allele at the site of interest. If we measure all the pairwise shared tract lengths among the NW European haplotypes with the putative selected derived mutation (rs4988235), we find that most haplotypes with the derived allele are identical, see derived-derived histogram in Figure 2. This is consistent with previous findings that observed a longer than expected decay in haplotype homozygosity in regions of positive selection, and the basis for summary statistics for selection (Nielsen et al. 2007, Nielsen 2005).

The haplotype patterns in Europeans at this locus are often used as the canonical example of what happens under a selective sweep (Voight et al. 2006, Bersaglieri et al. 2004). However, comparisons with an Asian population (CHB population in the 1000 genomes project, Han from now on) show that the Han also carry a haplotype with striking similarity to the NW European population at intermediate frequencies even though the Han haplotype is missing the two putatively beneficial mutations (Figure 1). Indeed, if we compute the shared tract length between the Han haplotypes (denoted as controls), we observe a proportion of the Han haplotypes are quite similar to the putative selected haplotype in the NW Europeans (Figure 2, control-derived, peak at 100kb). This suggests that the haplotype background may have been an ancestral haplotype and after the Han-NW European split a mutation arose on that haplotype background in the NW European populations only. Previous haplotype studies of *LCT* have examined haplotype lengths and the relationship to the lactase persistence allele. Enattah *et al.* (2007) remarked on this observation of the similarity and frequency of selected and non-selected haplotypes (within a 30kb region), and hypothesize that a Central Asian haplotype (that does not have the lactase persistence allele) may in fact be the haplotype background of the current European selected haplotype. Bersaglieri *et al.* (2004) also summarize *LCT* haplotypes from 101 genotyped SNPs across a 3.2Mb region, noting that the parental core haplotype is present in Asians, but do not make the observation described here which challenges the underlying assumption that the extended high frequency haplotypes are solely due to a selective sweep on the putatively selected LP allele. Inspecting the patterns of genetic differentiation (as measured by F_{ST} between the NW European and Han populations) in a region of 1Mb (see Supplementary Figure 1) shows a region of size 200kb-400kb around the putative selected site that exhibits many variants with constant level of genetic differentiation (F_{ST}) around 0.3 that is maintained for at least +/- 100kb (200kb total length).

To understand if the haplotype similarity between populations is to be expected under our current demographic out-of-Africa model[73], we simulated chromosomes of length 1Mb under the reference demographic model, and applied selection on the European branch consistent with the estimated selection parameters on *LCT* from other studies (see methods). Under this model, we calculated shared haplotype lengths by the same method as we did with the sequenced data. The averages between ancestral-ancestral, derived-derived, ancestral-derived, and control-control lengths in the *LCT* data is similar to the values obtained by simulation under several parameter sets, especially with strong selection and low recombination rate. We did not conduct extensive optimization to find the best fit simulation parameters but rather derived them from reference papers (see methods). However, the control-derived (Han-NW European selected) and control-ancestral (Han-NW European not selected) haplotype sharing is much longer in real data compared to the simulated data under this demographic model (Figure 3). Therefore a standard model of demographic history coupled with strong selection parameters does not reflect the high level of shared similarity between the selected European haplotype and a subset of the Han haplotypes. The nature of the evolutionary scenario responsible for the patterns at this locus remains an open question. One possibility is the haplotype background, which pre-dated the arrival of the selected allele, may have already had some positive effect with respect to lactase persistence. Another alternative scenario may be that the descendants of the Pontic Steppe may have also independently introduced the haplotype into East Asia, which can be tested with ancient DNA by looking for genome-wide admixture from the Yamnaya. It would also be informative to assess the similarity of the ancient Yamnaya haplotype to the high frequency NW European haplotype (some modern central Asian populations have a divergent lactase persistence haplotype, as described in Enattah *et al.* 2007). The exact sequence of events that led to the rise of the selected haplotype in Europeans will therefore be better approximated when we have access to haplotype data of humans from the recent past, which may help clarify why some genomic features in the *LCT* gene are not easily explained by the current models explored here. Indeed, recent work by Segurel and Bon (2017) have highlighted the as yet unresolved history of the *LCT* gene. Based on comparison to ancient samples, they conclude that the selected European allele appears to only have grown to high frequency as recently as the Middle Ages. They also conclude that the selected allele is less likely derived from Yamnaya herders given the very low observed frequency in ancient Bronze Age samples from those populations.

Beyond determining what evolutionary process explains the data around the *LCT* gene region, these observations imply that we need to consider the patterns of genetic variation in more than one population and develop new summaries of the data that capture variation in multiple populations. Otherwise we may end up making spurious inferences regarding the model of positive selection and parameters of interest (e.g. the timing and the strength of selection). For example, often summaries of the data that measure the haplotype length are used to both distinguish between models of positive selection and to estimate parameters[25,54,44]. However, in the case of *LCT*, the haplotype length associated with the selection event has likely been overestimated when only considering the European population. Instead, we should be measuring the relative change in haplotype length with respect to the haplotype observed in the Han to obtain a more accurate estimate of how positive selection affected the length of the haplotype associated with the selected allele.

Discussion/Interpretations

Before genomics, the search for positive selection was only tested through candidate gene studies. Now, we can query whole genomes of hundreds of individuals to identify positively selected regions and build an understanding of the genetic signature of selection. Through this effort, we have discovered many candidate mutations that likely facilitated the necessary phenotypic changes underlying key adaptations in recent human evolution. However, the origin of these mutations, when they were introduced into a population, when they became advantageous and their biological mechanism of adaptation still largely remain unanswered questions. Considering possible evolutionary models enables us to address these questions, and by examining two specific evolutionary models (SDM and SSV) under current best estimates for demography and *LCT* selection parameters, we showed, although did not formally quantify, that the observed haplotype structure among the NW European and Han populations cannot be explained. Since most modeling of the *LCT* region (Bersaglieri et al. 2004, Peter, Huerta-Sanchez, and Nielsen 2012, Tishkoff et al. 2007) does not include comparisons of genetic variation between multiple populations nor ancient DNA in parameter estimation, it is plausible the estimates would need to be revised.

Many examples of adaptation were identified before whole-genome sequencing, and so we suggest re-inspecting those regions to determine the compatible evolutionary scenarios, which may validate or update previous conjectures. For example, selection on *EPAS1*was first discovered from exome sequencing data and genotyping array data (Yi et al. 2010, Simonson et al. 2010, Bigham et al. 2010, Peng et al. 2010, Xu et al. 2010, Wang et al. 2011, Beall et al. 2010). However it was only through full re-sequencing of *EPAS1* that we recognized a highly divergent haplotype that must have originated from archaic introgression, and was subsequently confirmed by comparing to DNA from an archaic Denisovan genome (Huerta-Sánchez et al. 2014). As we have highlighted, ancient DNA adds additional information that allows us to study populations in the past and the subsequent temporal changes. By filling in the details of the evolutionary scenarios underlying these adaptations we will be better able to explain humans' evolutionary past, and to contribute to the story of human evolution.

Methods

Simulation Using SLiM. We used SliM2 (Haller and Messer 2016) to simulate recent selection on standing variation in the NW European population, and no selection in the Han population. We implemented the (Gravel et al. 2011) model of human demography from the SliM2 manual. The recombination rate used to generate Figure 3 is 3e-9 events per basepair and the mutation rate is 2.36e-8 events per base pair. The demographic model is shown in Figure 4. We simulated two different selection scenarios: selection from *de novo* mutation, and selection from low frequency standing variation. In both cases, we simulate neutral evolution until generation 57800, then we pick one segregating site at low frequency (between 0.5% and 1%), or a new mutation present at exactly one copy in the NW European population, in the middle of a 1 Mb segment, and assign it a positive fitness value (0.1 in the *de novo* mutation case, or 0.06 in the standing variation case). We have chosen two different selection coefficients for each model because these values lead to a better match of the observed frequency of the putative selected allele in NW Europeans. We continue the simulation for another 200 generations, and after the end of simulation, we sample 198 chromosomes from the NW European population, and 204 from the Han population as these are the sample sizes in the observed data. The average final frequency of the adaptive allele in the NW population in the standing variation case is 0.70, with standard deviation of 0.09, and in the new mutation case it is 0.45, with standard deviation of 0.34 while the observed frequency is around 0.6 in European populations. We note that the selection strength coefficients we have matched to the observed data are consistent with previous inferences (see review by (Ségurel and Bon 2017) and (Enattah et al. 2007).

Our choice of recombination rate is derived from pedigree inferred recombination rates from the DECODE project. Our variant of interest rs4988235 is located at position 136608646 of chromosome 2, thus we take the average recombination for 1Mb around our variant from the DECODE recombination map, and have recombination rate of 0.6791328, 0.4306461 and 1.178304 cM/Mb for sex-averaged map, female map and male specific map respectively. 1cM/Mb correspond to 1e-8 recombinations per basepair per generation. Thus, 3e-9 (or 0.3 cM/Mb) is a lower bound for recombination rate, 1e-8 is close to the true estimate, and 2e-8 is an upper bound for true recombination rate in this region. 3e-9 gave haplotype lengths that are closest to the observed data thus we show the results using this parameter. We set the mutation rate to be 2.36e-8 since that is the rate implemented in the population demography model from Gravel *et al.* (2011). Even if this might not be the most accurate mutation rate, we think it is most appropriate to use the same parameter as the one used for our reference demographic model.

Although we only presented one parameter set (in Figure 3), we experimented with different selection start times (100 or 200 generations), selection coefficients (0.06, 0.1, and 0.15), starting frequencies (between 0.5% and 1%), and recombination rates (1e-8, 2e-8, 3e-9) in the case of selection on standing genetic variation. The results show a similar pattern and we chose to pick the parameter set that is closest to the observed data. For the case of selection on a *de novo* mutation, we only simulated under one parameter set (recombination rate=3e-9, 200 generations ago, and selection coefficient s=0.1). The reason is that forward simulations are slow, and we have to simulate the entire human history for each simulation, thus introducing *de novo* mutations is highly inefficient because the mutation is highly likely to be lost (each simulation takes about 5 mins and will be discarded if the allele is lost before reaching the present generation). For the results presented in Figure 3, there are 315 independent replicates in the de novo mutation scheme where 94 of them matches the observed frequency, and 437 independent replicates in the standing variation scheme where 386 of them matches the observed frequency (excluding the simulations where the variant was lost of fixed). Here, a frequency match means a frequency between 0.4 and 0.8.

Our simulation results do not match the haplotype similarity between populations in the observed data, and suggest that at least with the parameters used here, the two models of positive selection considered may not be the right model for the evolutionary history in this local region.

Calculating the Length of the Shared Track of Homozygosity.

The length of the shared track of homozygosity is one measure of haplotype length. We define the pairwise shared track of homozygosity length around a position *x* (from now on referred to as the track length around *x*) as the sum of the maximum number of base pairs to the left and right of position *x* until the two chromosomes differ by a base pair. This measure is negatively affected by the presence of rare variants, which depend on the number of chromosomes in the sample. Therefore, to make the simulation comparable to the real data, we filter out all SNPs under 5% frequency in both the simulation and the 1000 Genomes data. We calculate the track length by lexicographically sorting the chromosomes from the edge up to the base pair next to site *x*, and calculating the shared length between every adjacent pair. This is an efficient approach such that we can obtain all pairwise track lengths in a computational time that is proportional to N*(L+N), where N is the total number of sequences, and L is the length of the sequence (Durbin 2014). A naïve implementation would take a computational time that is proportional to N*N*L.

For the *LCT* locus, we first partitioned all sampled chromosomes into three types: the ones carrying the derived allele in NW Europeans, the ones carrying the ancestral allele in NW Europeans, and for control, the Han haplotypes (which all carry the ancestral allele). We then calculate the within-type shared track length and the between-type shared track length. We do similar calculations for the simulated haplotypes, except that since the derived allele frequency at the present time may be non-zero in the Han (either if the allele arose before the Han-European split or due to migration), the control is not all sampled Han haplotypes, but only those carrying the ancestral allele.

Haplotype Structure Plot. The Haplostrips (Marnetto and Huerta-Sánchez, 2017) software was used to produce the plot shown in Figure 1. This software displays each SNP within a predefined region as a column, and each row represents a phased haplotype. Derived alleles are represented as black spots and ancestral alleles are represented as white spots. Each haplotype is labeled with a color that corresponds to the population of its carrier individual. The haplotypes were first hierarchically clustered and the resulting dendrogram of haplotypes was reordered by decreasing similarity to the NW European consensus haplotype. Variant positions were filtered out if the minor allele had a population frequency smaller than 5% in NW Europeans and Han independently.

Received 24 February 2017; revision accepted for publication 4 September 2017.

Literature Cited

Allentoft, M. E., M. Sikora, K. G. Sjögren et al. 2015. Population genomics of bronze age Eurasia. *Nature* 522:167–172.

Barreiro, L. B., M. Ben-Ali, H. Quach et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5:e1000562.

Barrett, R. D., and D. Schluter. 2008. Adaptation from standing genetic variation. *Trends Ecol. Evol.* 23:38–44.

Beall, C. M., G. L. Cavalleri, L. Deng et al. 2010. Natural selection on EPAS1(HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *PNAS* 107:11,459–11,464.

Bersaglieri, T., P. C. Sabeti, N. Patterson et al. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74:1,111–1,120.

Bigham, A., M. Bauchet, D. Pinto et al. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* 6:e1001116.

Burger, J., M. Kirchner, B. Bramanti et al. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *PNAS* 104:3,736–3,741.

Bustamante, C. D., J. Wakeley, S. Sawyer et al. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159:1,779–1,788.

Durbin, R. 2014. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* 30:1,266–1,272.

Eiberg, H., J. Troelsen, M. Nielsen et al. 2008. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* 123:177–187.

Enattah, N. S., T. Sahi, E. Savilahti et al. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30:233.

Enattah, N. S., A. Trudeau, V. Pimenoff et al. 2007. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am. J. Hum. Genet.* 81:615–625.

Ferrer-Admetlla, A., M. Liang, T. Korneliussen et al. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* 31:1,275–1,291.

Fu, Q., H. Li, P. Moorjani et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445–449.

Fu, W., and J. M. Akey. 2013. Selection and adaptation in the human genome. *Annu. Rev. Genom. Hum. G.* 14:467–489.

Fujimoto, A., R. Kimura, J. Ohashi et al. 2007. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* 17:835–843.

Fumagalli, M., I. Moltke, N. Grarup et al. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349:1,343–1,347.

Fumagalli, M., M. Sironi, U. Pozzoli et al. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.

Gravel, S., B. M. Henn, R. N. Gutenkunst et al. 2011. Demographic history and rare allele sharing among human populations. *PNAS.* 108:11,983–11,988.

Green, R. E., J. Krause, A. W. Briggs et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.

Haller, B. C., and P. W. Messer. 2016. SLiM 2: Flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* 34:230–240.

Han, J., P. Kraft, H. Nan et al. 2008. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* 4:e1000074.

Hawks, J., and G. Cochran. 2006. Dynamics of adaptive introgression from archaic to modern humans. *PaleoAnthropology* 2006:101–115.

Hernandez, R. D., J. L. Kelley, E. Elyashiv et al. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.

Huerta-Sánchez, E., and F. P. Casey 2015. Archaic inheritance: Supporting high-altitude life in Tibet. *J. Appl. Physiol.* 119:1,129–1,134.

Huerta-Sánchez, E., M. DeGiorgio, L. Pagani et al. 2013. Genetic signatures reveal high-altitude adaptation in a set of Ethiopian populations. *Mol. Biol. Evol.* 30:1,877–1,888.

Huerta-Sánchez, E., X. Jin, Z. Bianba et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512:194–197.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1,299.

Itan, Y., A. Powell, M. A. Beaumont et al. 2009. The origins of lactase persistence in Europe. *PLoS Comput. Biol.* 5:e1000491.

Keller, A., A. Graefen, M. Ball et al. 2012. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* 3:698.

Key, F. M., B. Peter, M. Y. Dennis et al. 2014. Selection on a variant associated with improved viral clearance drives local, adaptive pseudogenization of interferon lambda 4 (IFNL4). *PLoS Genet.* 10:e1004681.

Kimura, R., T. Yamaguchi, M. Takeda et al. 2009. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am. J. Hum. Genet.* 85:528–535.

Lamason, R. L., M. P. Mohideen, J. R. Mest et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310:1,782–1,786.

Lazaridis, I., N. Patterson, A. Mittnik et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513:409.

Lindo, J., E. Huerta-Sánchez, S. Nakagome et al. 2016. A time transect of exomes from a Native American population before and after European contact. *Nat. Commun.* 7:13,175.

Malaspinas, A. S. 2016. Methods to characterize selective sweeps using time serial samples: An ancient DNA perspective. *Mol. Ecol.* 25:24–41.

Marnetto, D., and E. Huerta-Sánchez. 2017. Haplostrips: Revealing population structure through haplotype visualization. *Methods Ecol. Evol.*

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

Meyer, M., M. Kircher, M. T. Gansauge et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222–226.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* 39:197–218.

Nielsen, R., I. Hellmann, M. Hubisz et al. 2007. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* 8:857.

Nishino, J. 2013. Detecting selection using time-series data of allele frequencies with multiple independent reference loci. *G3 (Bethesda)* 3:2,151–2,161.

Novembre, J., and A. Di Rienzo. 2009. Spatial patterns of variation due to natural selection in humans. *Nat. Rev. Genet.* 10:745.

Otto, S. P. 2000. Detecting the form of selection from DNA sequence data. *Trends Genet.* 16:526–529.

Peng, Y., Z. Yang, H. Zhang et al. 2010. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol. Biol. Evol.* 28:1,075–1,081.

Peter, B. M., E. Huerta-Sánchez, and R. Nielsen. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.* 8:e1003011.

Pickrell, J. K., G. Coop, J. Novembre et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.

Plagnol, V., and J. D. Wall. 2006. Possible ancestral structure in human populations. *PLoS Genet.* 2:e105.

Prüfer, K., F. Racimo, N. Patterson et al. 2014. The complete genome sequence of a Neandertal from the Altai Mountains. *Nature* 505:43.

Przeworski, M., G. Coop, and J. D. Wall. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59:2,312–2,323.

Racimo, F., D. Gokhman, M. Fumagalli et al. 2017. Archaic adaptive introgression in TBX15/WARS2. *Mol. Biol. Evol.* 34:509–524.

Racimo, F., D. Marnetto, and E. Huerta-Sánchez. 2017. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* 34:296–317.

Racimo, F., S. Sankararaman, R. Nielsen et al. 2015. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* 16:359.

Rasmussen, M., X. Guo, Y. Wang et al. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334:94–98.

Rasmussen, M., Y. Li, S. Lindgreen et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463:757.

Reich, D., R. E. Green, M. Kircher et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1,053.

Reynolds, J., B. S. Weir, and C. Cockerham. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105:767–779.

Sabeti, P. C., D. E. Reich, J. M. Higgins et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832.

Sabeti, P. C., S. F. Schaffner, B. Fry et al. 2006. Positive natural selection in the human lineage. *Science* 312:1,614–1,620.

Sabeti, P. C., P. Varilly, B. Fry et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913.

Schlebusch, C. M., H. Malmström, T. Günther et al. 2017. Ancient genomes from southern Africa pushes modern human divergence beyond 260,000 years ago. *bioRxiv* 145409.

Schroeder, H., M. C. Ávila-Arcos, A. S. Malaspinas et al. 2015. Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *PNAS* 112:3,669–3,673.

Seguin-Orlando, A., T. S. Korneliussen, M. Sikora et al. 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346:1,113–1,118.

Ségurel, L., and C. Bon. 2017. On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.* 18:297–319.

Simonson, T. S., Y. Yang, C. D. Huff et al. 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science* 329:72–75.

Skoglund, P., H. Malmström, M. Raghavan et al. 2012. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* 336:466–469.

Slatkin, M., and F. Racimo. 2016. Ancient DNA and human history. *PNAS* 13:6,380–6,387.

Sturm, R. A., D. L. Duffy, Z. Z. Zhao et al. 2008. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet.* 82:424–431.

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56.

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68.

Tishkoff, S. A., F. A. Reed, A. Ranciaro et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39:31.

Voight, B. F., S. Kudaravalli, X. Wen et al. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wang, B., Y. B. Zhang, F. Zhang et al. 2011. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PloS One* 6:e17002.

Xu, S., S. Li, Y. Yang et al. 2010. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol. Biol. Evol.* 28:1,003–1,011.

Xu, D., P. Pavlidis, R. O. Taskent et al. 2017. Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. *Mol. Biol. Evol.* 34:2,704–2,715.

Yi, X., Y. Liang, E. Huerta-Sánchez et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.

Table 1. Haplotype based methods to detect positive selection

Method	Formula	Explanation	Main Information	Ref
EHH	$EHH(X_i) = \sum_{h\in C(X_i)} \frac{\frac{n_h}{2}}{\frac{n}{2}}$ $EHH_c(X_i) = \sum_{h\in H_c(X_i)} \frac{\frac{n_h}{2}}{\frac{n}{2}}$	Extended Haplotype Homozygosity: The EHH score is calculated given a locus of interest and the i^{th} marker upstream or downstream x_i. h is all the observed types (unique combinations of SNPs) from SNP x_0 to x_i. n_h is the number of haplotypes of type h, and n is the total number of haplotypes. EHH_c is very similar to EHH, but is only calculated for samples carrying the core haplotype c.	LD and Haplotype	1
iHS	$iHH_c = \sum_{i=1}^{\lvert D\rvert} \frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}x_i) + \sum_{i=1}^{\lvert U\rvert} \frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i)$ $unstandardized\ iHS = ln(\frac{iHH_1}{iHH_0})$ $standardized\ iHs = \frac{ln(\frac{iHH_1}{iHH_0}) - E_p[ln(\frac{iHH_1}{iHH_0})]}{SD_p[ln\frac{iHH_1}{iHH_0}]}$	Integrated Haplotype Score: iHs score calculates the decay of EHH for the ancestral and derived haplotype extending from a query site. D is a set of marker downstream from the marker of interest, and U is a set of markers upstream from the marker of interest. $g(x_{i-1}, x_i)$ is the genetic distance between x_{i-1} and x_i	LD and haplotype	2
XP-EHH	$XP - EHH = \frac{ln(\frac{IHH_A}{iHH_B}) - ln(\frac{iHH_A}{iHH_B})}{SD[ln(\frac{iHH_A}{iHH_B})]}$	XP-EHH is a cross population EHH score that looks at the ratio of the iHH score in two different populations.	LD and Haplotype, cross population	3

nSL	$SL_c = \sum_{i=1}^{\lvert D \rvert} \frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i) \sum_{i=1}^{\lvert U \rvert} \frac{1}{2}(EHH_c(x_{i-1}) + EHH_c(x_i))g(x_{i-1}, x_i)$ $nSL = \frac{ln(\frac{SL_1}{SL_0}) - E_p[ln \frac{SL_1}{SL_0}]}{SD_p[ln \frac{SL_1}{SL_0}]}$	nSL is very similar to the iHS score, and the only difference is that $g(x_{i-1}, x_i)$correspond to the distance in base pairs, rather than the recombination distance, so $g(x_{i-1}, x_i) = 1$.	LD and Haplotype	4

Table 2. Frequency based methods for detecting positive selection.

Method	Formula	Explanation	Main Information	Ref
π	$\pi = \frac{1}{\frac{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i$	$\pi$is the mean pairwise sequence differences among a sample of n haplotypes. $\xi_i$is the unfolded site frequency spectrum.	Allele Frequency	5
FST	$F_{ST} = \frac{H_T - H_S}{H_T}$ $H_s = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{I} \sum_{i=1}^{I} p_{ij}^2$ $H_T = \frac{1}{I} \sum_{i=1}^{I} (\frac{1}{J} \sum_{i=1}^{I} p_{ij})^2$	$H_T$is the mean pairwise sequence differences in the total population and $H_s$is the same statistic in a subpopulation. It can be used to detect population structure, and local outliers of high values of F_{ST}suggests selection specific to a subpopulation.	Allele Frequency, cross population	6
PBS	$T^{A,B} = -log(1 - F_{ST}^{A,B})$ $PBS_A = \frac{T^{A,B} + T^{A,O} - T^{B,O}}{2}$	$T^{A,B}$is a transformed F_{ST}measure, where the total population is the combined population of	Allele Frequency, cross population	7

1. Sabeti, Pardis C., et al. "Detecting recent positive selection in the human genome from haplotype structure." Nature 419.6909 (2002): 832-837.
2. Voight, Benjamin F., et al. "A map of recent positive selection in the human genome." *PLoS Biol* 4.3 (2006): e72.
3. Sabeti, Pardis C., et al. "Genome-wide detection and characterization of positive selection in human populations." *Nature* 449.7164 (2007): 913-918.
4. Ferrer-Admetlla, Anna, et al. "On detecting incomplete soft or hard selective sweeps using haplotype structure." *Molecular biology and evolution* 31.5 (2014): 1275-1291.
5. Nei, Masatoshi. *Molecular evolutionary genetics*. Columbia university press, 1987.
6. Nei, Masatoshi. "Analysis of gene diversity in subdivided populations." *Proceedings of the National Academy of Sciences* 70.12 (1973): 3321-3323.
7. Yi, Xin, et al. "Sequencing of 50 human exomes reveals adaptation to high altitude." *science* 329.5987 (2010): 75-78.

Figure 1. Rows are haplotypes and columns are variable sites. Black denotes derived alleles and white denotes ancestral alleles. The blue arrow points to the site with the derived allele present at high frequency in NW European haplotypes and absent in Han haplotypes. Data is from 1000 Genomes Phase 3. This is a 88 kb region encompassing *LCT* and *MCM6* in their entirety, and the sample sizes are 103 Han individuals and 99 NW European individuals. For more details see Methods section "Haplotype Structure Plot."

Figure 2. The six haplotypes sharing comparisons are shown in six histograms. The x-axis shows the one-sided shared length with pairwise comparisons starting at the selected locus until 500kb to the left, or right of the locus. Note that the length of sharing could be longer than 500kb. The y-axis indicates the proportion of pairs of haplotypes in a 10kb bin of length sharing. Control is a haplotype in the Han population (that do not have the ancestral allele in the putatively selected position), ancestral is the NW European haplotype with ancestral allele in the selected position, and derived is the NW European haplotype with the selected allele.

Figure 3. These box plots show the distribution of average length of sharing between haplotypes of the same, or different allelic states, from the same, or different populations (NW European or Han). The y-axis indicates the average length of sharing of the comparison in one replicate of simulation. There are in total two selection schemes: the locus determining the allelic state is either selected or neutral in the European population, and is always neutral in other populations. The result from the selection simulation is plotted in blue, and the result from the neutral simulation is in red. The real data values for the *LCT* haplotype length sharing of each comparison is shown in red diamond, showing how extreme the average control-derived shared haplotype length is for the real data. The recombination rate is low to reflect the rate at the *LCT* region: 3e-9. The top panel shows the result when selection is acting on a *de novo* mutation with

s=0.1 and the bottom panel shows the result when selection is acting on a low frequency standing variation SNP with s=0.06 (see methods section for more details of how the simulations were run).

Figure 4. Demographic model of Gravel *et al* (2012)[73]. The brown, red and blue colors represent African, Asian and European populations, respectively. The rectangles show constant population size, and triangles show the exponential growth of the Asian and European populations. The red numbers are the effective population sizes. The black numbers are the times of each event, and blue numbers are the times between each event in years. Note that the size of the shapes is not proportional to the time or population size. There is low level of migration between all three populations that is shown in dashed line arrows.

Supplementary Figure 1. F_{ST} between NW European and Han populations for all SNPs in the region around rs4988235 (+/- 500 kb, position of SNP denoted by black vertical line). Each SNP is a dot/point with X coordinate given by the position on chromosome 2 and Y representing the F_{ST} computed according to Weir and Cockerham, 1984. Blue SNPs have a minor alllele fequency higher than 0.05 while grey ones have minor allele frequency <=0.05. Diamonds (in dark blue or black) represent the 101 SNPs included by Bersaglieri et al. 2004.

Supplementary Figure 2. The haplotype length of the ancestral-ancestral, derived-derived, ancestral-derived, control-control, control-derived and control-ancestral comparisons with varying selection scheme and recombination rate. Each column has the same recombination rate and increases going from left to right (3e-9,1e-8, 2e-8) and each row has the same selection coefficient (0.06 starting 200 generations ago, 0.1 starting 200 generations ago, 0.1 starting 1 generation ago). All the simulations started from standing variation. Please note that the y-axis scale is different among the plots.

Figure 1.

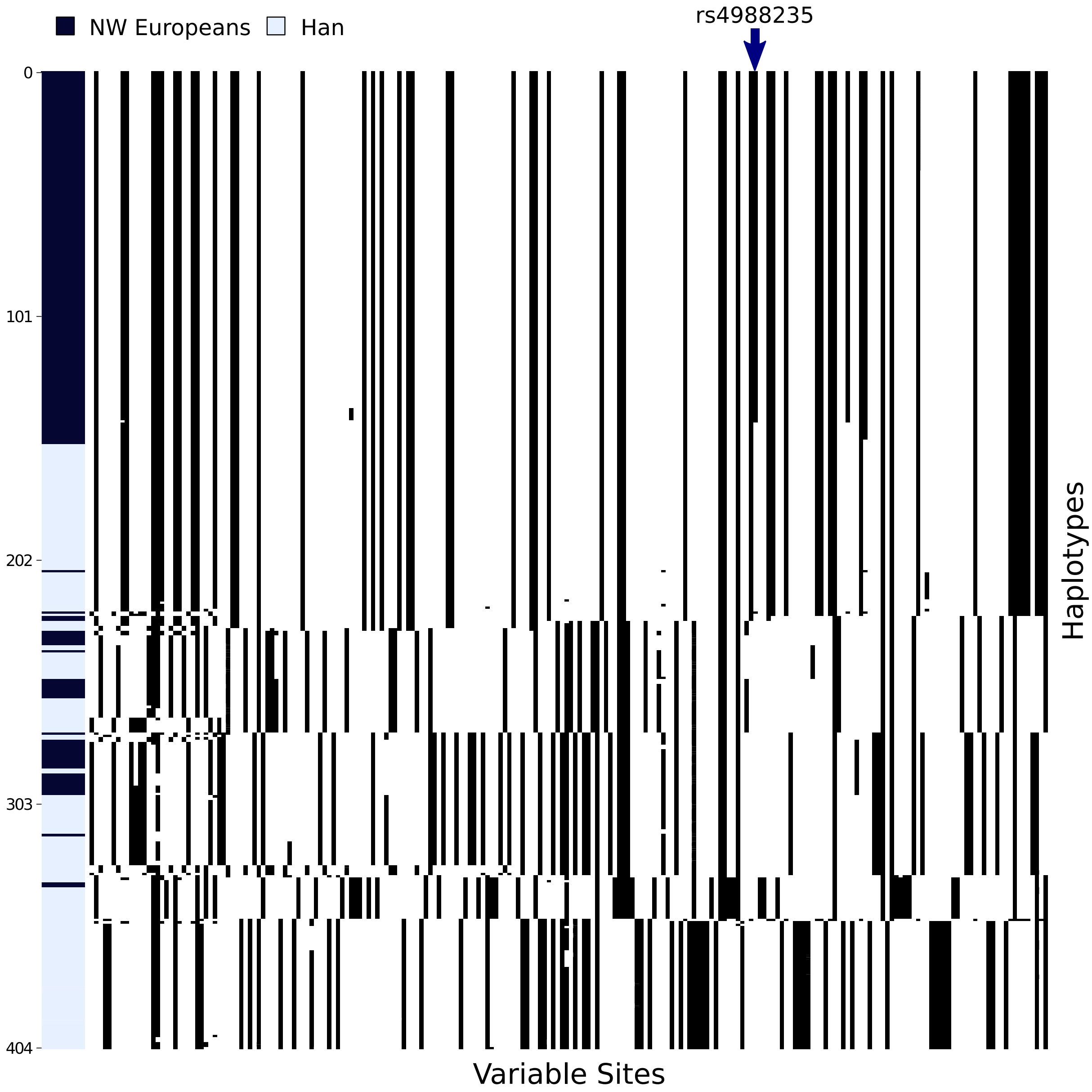

■ NW Europeans
□ Han
rs4988235
0
101
202
303
404
Variable Sites
Haplotypes

Figure 2.

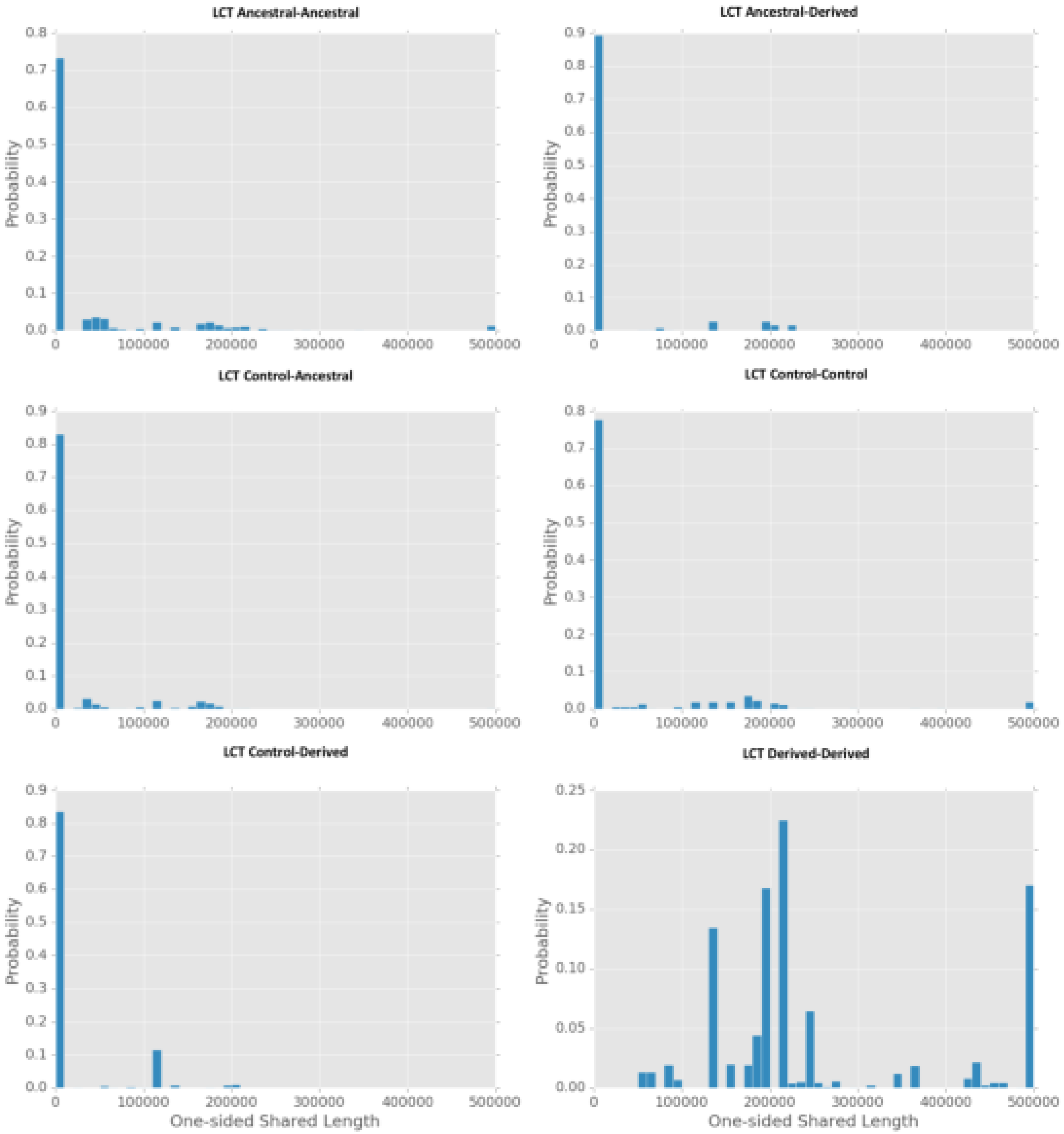

LCT Ancestral-Ancestral
LCT Ancestral-Derived
LCT Control-Ancestral
LCT Control-Control
LCT Control-Derived
LCT Derived-Derived
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.00
0.05
0.10
0.15
0.20
0.25
0
100000
200000
300000
400000
500000
One-sided Shared Length

Figure 3.

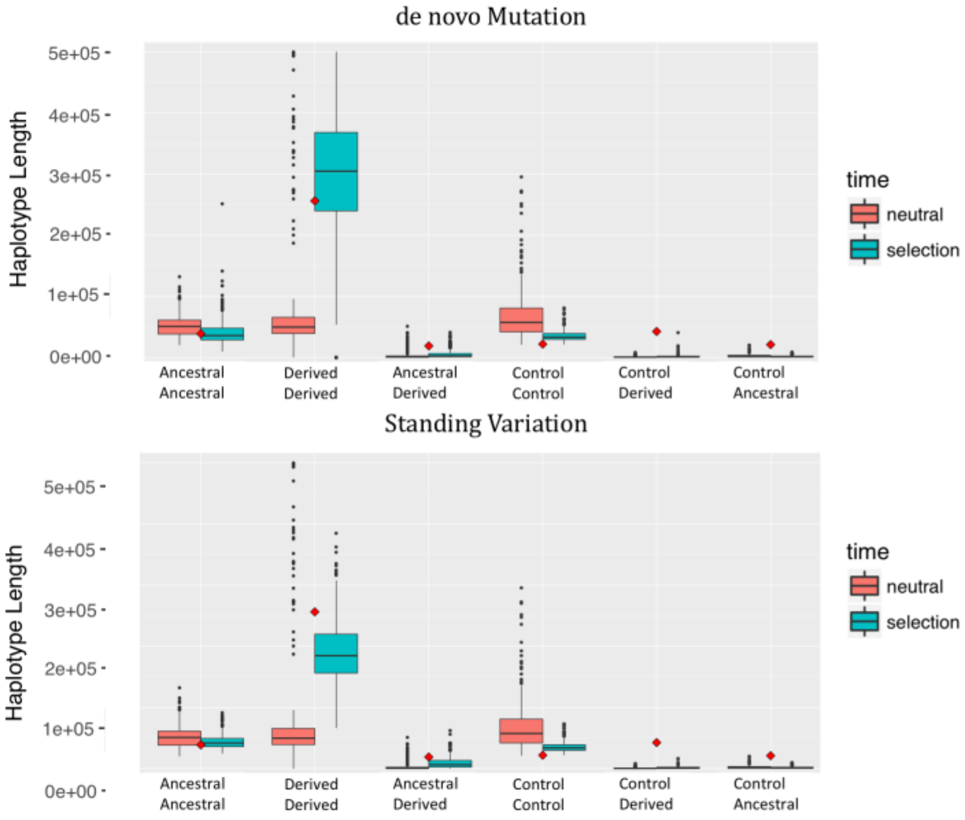
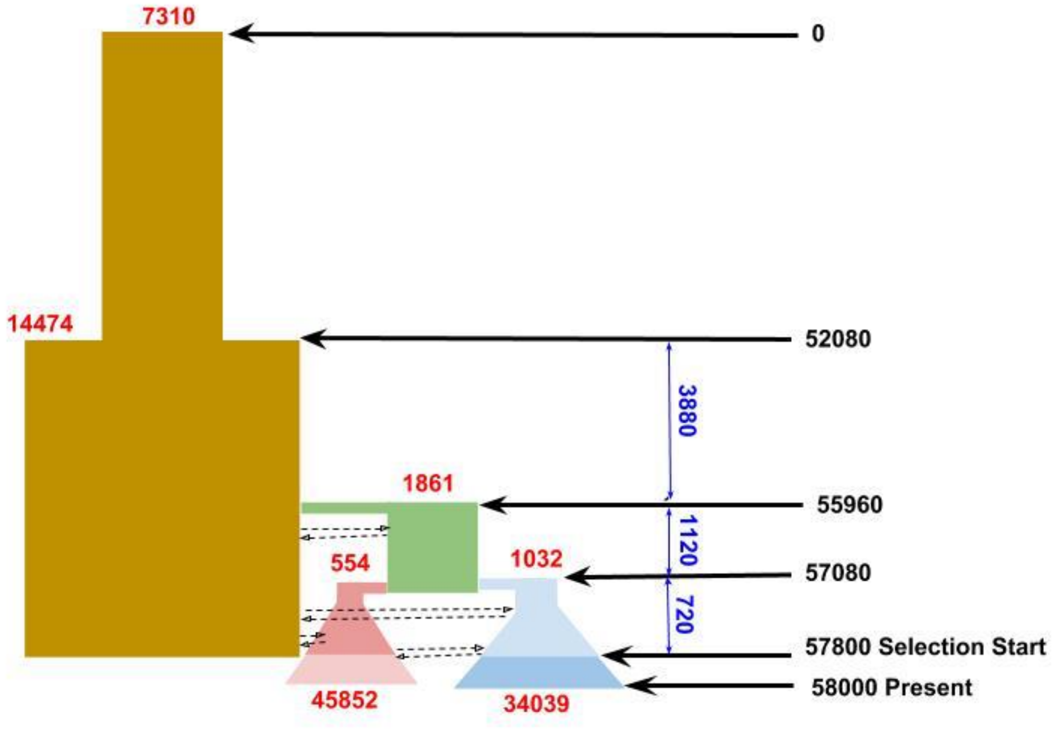
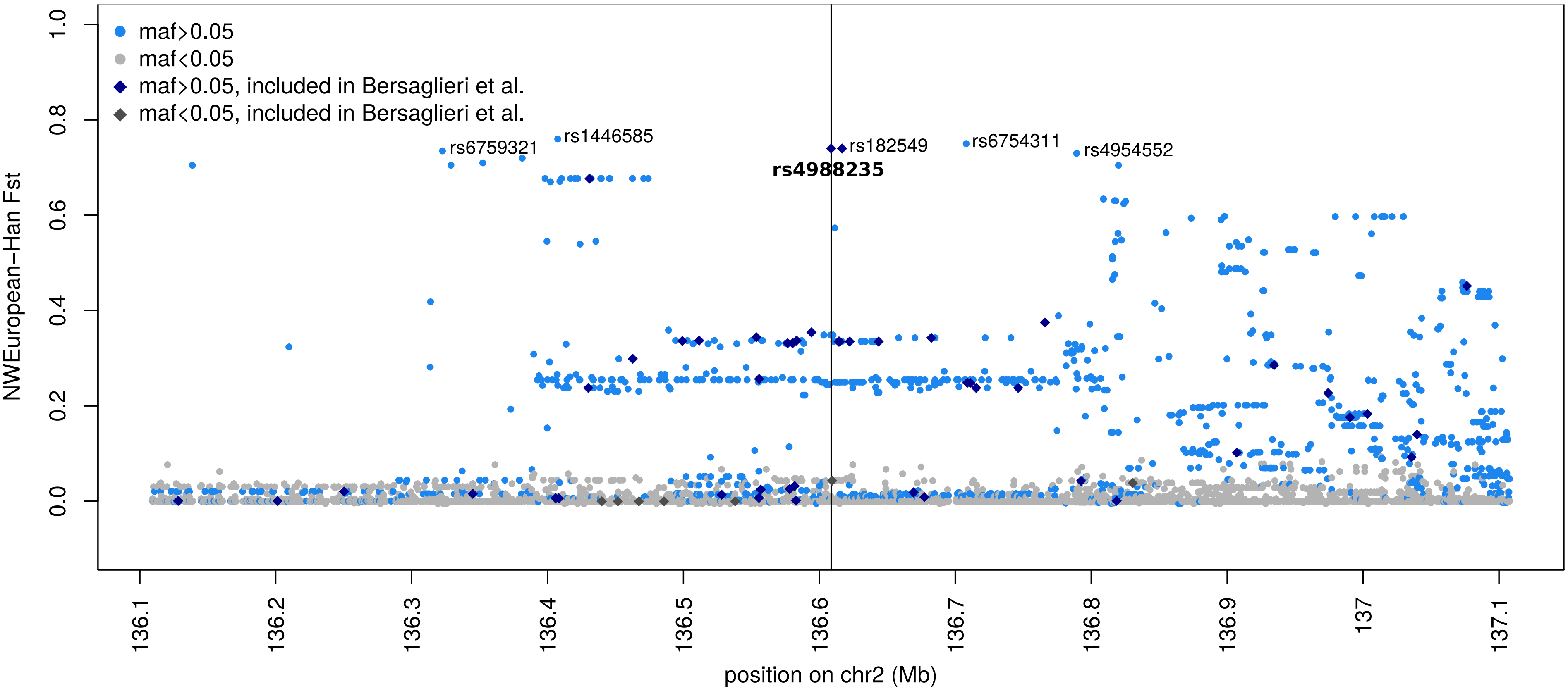

de novo Mutation
Haplotype Length
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
Ancestral Ancestral
Derived Derived
Ancestral Derived
Control Control
Control Derived
Control Ancestral
time
neutral
selection
Standing Variation
Haplotype Length
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
Ancestral Ancestral
Derived Derived
Ancestral Derived
Control Control
Control Derived
Control Ancestral
time
neutral
selection

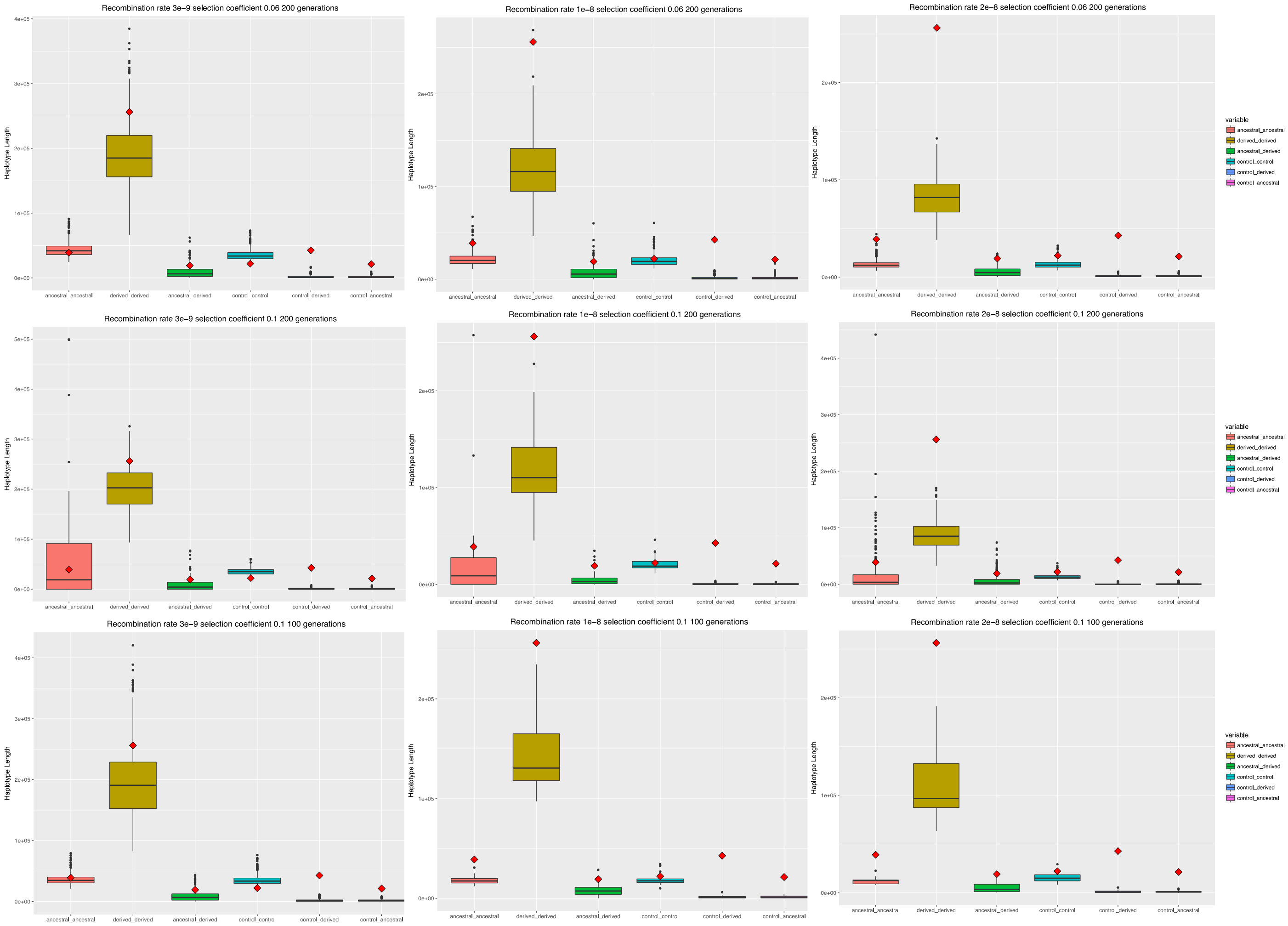
Figure 4.


7310
0
14474
52080
3880
1861
55960
1120
554
1032
57080
720
57800 Selection Start
58000 Present
45852
34039

Supplementary Figure 1.



Supplementary Figure 2.



Pre-print version. Visit http://digitalcommons.wayne.edu/humbiol/ after publication to acquire the final version.