Testing for ancient selection using cross-population allele frequency differentiation

Fernando Racimo[1,*]

1 Department of Integrative Biology, University of California, Berkeley, CA, USA

∗ E-mail: fernandoracimo@gmail.com

1 Abstract

A powerful way to detect selection in a population is by modeling local allele frequency changes in a particular region of the genome under scenarios of selection and neutrality, and finding which model is most compatible with the data. Chen et al. [1] developed a composite likelihood method called XP-CLR that uses an outgroup population to detect departures from neutrality which could be compatible with hard or soft sweeps, at linked sites near a beneficial allele. However, this method is most sensitive to recent selection and may miss selective events that happened a long time ago. To overcome this, we developed an extension of XP-CLR that jointly models the behavior of a selected allele in a three-population tree. Our method - called 3P-CLR - outperforms XP-CLR when testing for selection that occurred before two populations split from each other, and can distinguish between those events and events that occurred specifically in each of the populations after the split. We applied our new test to population genomic data from the 1000 Genomes Project, to search for selective sweeps that occurred before the split of Africans and Eurasians, but after their split from Neanderthals, and that could have presumably led to the fixation of modern-human-specific phenotypes. We also searched for sweep events that occurred in East Asians, Europeans and the ancestors of both populations, after their split from Africans.

2 Introduction

Genetic hitchhiking will distort allele frequency patterns at regions of the genome linked to a beneficial allele that is rising in frequency [2]. This is known as a selective sweep. If the sweep is restricted to a particular population and does not affect other closely related populations, one can detect such an event by looking for extreme patterns of localized population differentation, like high values of F_{st} at a specific locus [3]. This and other related statistics have in fact been used to scan the genomes of present-day humans from different populations, so as to detect signals of recent positive selection [4–7].

Once it became possible to sequence entire genomes of archaic humans (like Neanderthals) [8–10], researchers also began to search for selective sweeps that occurred in the ancestral population of all present-day humans. For example, ref. [8] searched for genomic regions with a depletion of derived alleles in a low-coverage Neanderthal genome, relative to what would be expected given the derived allele frequency in present-day humans. This is a pattern that would be consistent with a sweep in present-day humans. Later on, ref. [10] developed a hidden Markov model (HMM) that could identify regions where Neanderthals fall outside of all present-day human variation (also called "external regions"), and are therefore likely to have been affected by ancient sweeps in early modern humans. They applied their method to a high-coverage Neanderthal genome. Then, they ranked these regions by their genetic length, to find segments that were extremely long, and therefore highly compatible with a selective sweep. Finally, ref. [11] used summary statistics calculated in the neighborhood of sites that were ancestral in archaic humans but fixed derived in all or almost all present-day humans, to test if any of these sites could be compatible with a selective sweep model. While these methods harnessed different summaries of the patterns of differentiation left by sweeps, they did not attempt to explicitly model the process by which these patterns are generated over time.

Chen et al. [1] developed a method called XP-CLR, which is designed to test for selection in one population after its split from a second, outgroup, population t_{AB} generations ago. It does so by modeling the evolutionary trajectory of an allele under linked selection and under neutrality, and then comparing the likelihood of the data under each of the two models. The method detects local allele frequency differences that are compatible with the linked selection model [2], along windows of the genome.

XP-CLR is a powerful test for detecting selective events restricted to one population. However, it provides little information about when these events happened, as it models all sweeps as if they had immediately occurred in the present generation. Additionally, if one is interested in selective sweeps that took place before two populations a and b split from each other, one would have to run XP-CLR separately on each population, with a third outgroup population c that split from the ancestor of a and b t_{ABC} generations ago (with $t_{ABC} > t_{AB}$). Then, one would need to check that the signal of selection appears in both tests. This may miss important information about correlated allele frequency changes shared by a and b, but not by c, limiting the power to detect ancient events.

To overcome this, we developed an extension of XP-CLR that jointly models the behavior of an allele in all 3 populations, to detect selective events that occurred before or after the closest two populations split from each other. Below we briefly review the modeling framework of XP-CLR and describe our new test, which we call 3P-CLR. In the Results, we show this method outperforms XP-CLR when testing for selection that occurred before the split of two populations, and can distinguish between those events and events that occurred after the split, unlike XP-CLR. We then apply the method to population genomic data from the 1000 Genomes Project [12], to search for selective sweep events that occurred before the split of Africans and Eurasians, but after their split from Neanderthals. We also use it to search for selective sweeps that occurred in the Eurasian ancestral population, and to distinguish those from events that occurred specifically in East Asians or specifically in Europeans.

3 Methods

3.1 XP-CLR

First, we review the procedure used by XP-CLR to model the evolution of allele frequency changes of two populations a and b that split from each other t_{AB} generations ago (Figure 1.A). For neutral SNPs, Chen et al. [1] use an approximation to the Wright-Fisher diffusion dynamics [13]. Namely, the frequency of a SNP in a population a (p_A) in the present is treated as a random variable governed by a normal distribution with mean equal to the frequency in the ancestral population (β) and variance proportional to the drift time ω from the ancestral to the present population:

$$p_A|\beta \sim N(\beta, \omega\beta(1-\beta)) \tag{1}$$

where $\omega = t_{AB}/(2N_e)$ and N_e is the effective size of population A.

If a SNP is segregating in both populations - i.e. has not hit the boundaries of fixation or extinction - this process is time-reversible. Thus, one can model the frequency of the SNP in population a with a normal distribution having mean equal to the frequency in population b and variance proportional to the sum of the drift time (ω) between a and the ancestral population, and the drift time between b and the ancestral population (ψ):

$$p_A|p_B \sim N(p_B, (\omega+\psi)p_B(1-p_B)) \tag{2}$$

For SNPs that are linked to a beneficial allele that has undergone a sweep in population a only, Chen et al. [1] model the allele as evolving neutrally until the present and then apply a transformation to the normal distribution that depends on the distance to the selected allele r and the strength of selection s [14, 15]. Let $c = 1 - q_0^{r/2}$ where q_0 is the frequency of the beneficial allele in population A before the sweep begins. The frequency of a neutral allele is expected to increase from p to $1 - c + cp$ if the allele is linked to the beneficial allele, and this occurs with probability equal to the frequency of the neutral allele (p) before the sweep begins. Otherwise, the frequency of the neutral allele is expected to decrease from p to cp. This leads to the following transformation of the normal distribution:

$$f(p_A|p_B, r, s, \omega, \psi) = \frac{1}{\sqrt{2\pi}\sigma}\frac{p_A + c - 1}{c^2} e^{-\frac{(p_A+c-1-cp_B)^2}{2c^2\sigma^2}} I_{[1-c,1]}(p_A) + \frac{1}{\sqrt{2\pi}\sigma}\frac{c - p_A}{c^2} e^{-\frac{(p_A-cp_B)^2}{2c^2\sigma^2}} I_{[0,c]}(p_A) \tag{3}$$

where $\sigma^2 = (\omega + \psi)p_b(1 - p_b)$ and $I_{[x,y]}(z)$ is 1 on the interval $[x, y]$ and 0 otherwise.

For $s \to 0$ or $r >> s$, this distribution converges to the neutral case. Let $\mathbf{v}$ be the vector of all drift times that are relevant to the scenario we are studying. In this case, it will be equal to (ω, ψ) but in more complex cases below, it may include additional drift times. Let $\mathbf{r}$ be the vector of recombination fractions between the beneficial alleles and each of the SNPs within a window of arbitrary size. We can then calculate the product of likelihoods over all k SNPs in that window for either the neutral or the linked selection model, after binomial sampling of alleles from the population frequency and conditioning on the event that the allele is segregating in the population:

$$CL_{XP-CLR}(\mathbf{r}, \mathbf{v}, s) = \prod_{j=1}^{k} \frac{\int_0^1 f(p_A^j|p_B^j, \mathbf{v}, s, r^j)\binom{n}{m_j}(p_A^j)^{m_j}(1 - p_A^j)^{n-m_j} dp_A^j}{\int_0^1 f(p_A^j|p_B^j, \mathbf{v}, s, r^j) dp_A^j} \tag{4}$$

We note that the denominator in the above equation is not explicitly stated in ref. [1] for ease of notation, but appears in the published online implementation of the method. Because we are ignoring the correlation in frequencies produced by linkage, this is a composite likelihood [16, 17]. Finally, we obtain a composite likelihood ratio statistic S_{XP-CLR} of the hypothesis of linked selection over the hypothesis of neutrality:

$$S_{XP-CLR} = 2[sup_{\mathbf{r},\mathbf{v},s} log(CL_{XP-CLR}(\mathbf{r}, \mathbf{v}, s)) - sup_{\mathbf{v}} log(CL_{XP-CLR}(\mathbf{r}, \mathbf{v}, s = 0))] \tag{5}$$

For ease of computation, Chen et al. [1] assume that **r** is given (via a recombination map) and we will do so too. Furthermore, they empirically estimate **v** using F_2 statistics [18] calculated over the whole genome, and assume selection is not strong or frequent enough to affect their genome-wide values. Because we are interested in selection over long time scales, the new methods we will present below are optimally run using drift times calculated from population split times and effective population sizes estimated using model-based demographic inference methods, like $\partial a \partial i$ [19] or fastsimcoal2 [20].

3.2 3P-CLR

We are interested in the case where a selective event occurred more anciently than the split of two populations (a and b) from each other, but more recently than their split from a third population c (Figure 1.B). We begin by modeling p_A and p_B as evolving from an unknown common ancestral frequency β:

$$p_A|\beta,\omega \sim N(\beta,\omega\beta(1-\beta)) \tag{6}$$

$$p_B|\beta,\psi \sim N(\beta,\psi\beta(1-\beta)) \tag{7}$$

Let χ be the drift time separating the most recent common ancestor of a and b from the most recent common ancestor of a, b and c. Additionally, let ν be the drift time separating population c in the present from the most recent common ancestor of a, b and c. Given these parameters, we can treat β as an additional random variable that either evolves neutrally or is linked to a selected allele that swept immediately more anciently than the split of a and b. In both cases, the distribution of β will depend on the frequency of the allele in population c (p_C) in the present. In the neutral case:

$$f_{neut}(\beta|p_C,\nu,\chi) = N(p_C,(\nu+\chi)p_C(1-p_C)) \tag{8}$$

In the linked selection case:

$$f_{sel}(\beta|p_C,\nu,\chi,r,s) = \frac{1}{\sqrt{2\pi}\kappa}\frac{\beta+c-1}{c^2}e^{-\frac{(\beta+c-1-cp_C)^2}{2c^2\kappa^2}}I_{[1-c,1]}(\beta) + \frac{1}{\sqrt{2\pi}\kappa}\frac{c-\beta}{c^2}e^{-\frac{(\beta-cp_C)^2}{2c^2\kappa^2}}I_{[0,c]}(\beta) \tag{9}$$

where $\kappa^2 = (\nu+\chi)p_C(1-p_C)$

The frequencies in a and b given the frequency in c can be obtained by integrating β out. This leads to a density function that models selection in the ancestral population of a and b.

$$f(p_A, p_B|p_C, \mathbf{v}, r, s) = \int_0^1 f_{neut}(p_A|\beta, \omega) f_{neut}(p_B|\beta, \psi) f_{sel}(\beta|p_C, \nu, \chi, r, s) d\beta \tag{10}$$

Additionally, formula 10 can be modified to test for selection that occurred specifically in one of the terminal branches that lead to a or b (Figures 1.C and 1.D), rather than in the ancestral population of a and b. For example, the density of frequencies for a scenario of selection in the branch leading to a can be written as:

$$f(p_A, p_B|p_C, \mathbf{v}, r, s) = \int_0^1 f_{sel}(p_A|\beta, \omega, r, s) f_{neut}(p_B|\beta, \psi) f_{neut}(\beta|p_C, \nu, \chi) d\beta \tag{11}$$

We will henceforth refer to the version of 3P-CLR that is tailored to detect selection in the internal branch that is ancestral to a and b as 3P-CLR(Int). In turn, the versions of 3P-CLR that are designed to detect selection in each of the daughter populations will be designated as 3P-CLR(A) and 3P-CLR(B).

We can now calculate the probability density of specific allele frequencies in populations a and b, given that we observe m_C derived alleles in a sample of size n_C from population c:

$$f(p_A, p_B|m_C, \mathbf{v}, r, s) = \int_0^1 f(p_A, p_B|p_C, \mathbf{v}, r, s) f(p_C|m_C) dp_C \tag{12}$$

where B(x,y) is the Beta function and

$$f(p_C|m_C) = \frac{1}{B(m_C, n_C - m_C + 1)} p_C^{m_C - 1} (1 - p_C)^{n_C - m_C} \tag{13}$$

Conditioning on the event that the site is segregating in the population, we can then calculate the probability of observing m_A and m_B derived alleles in a sample of size n_A from population a and a sample of size n_B from population b, respectively, given that we observe m_C derived alleles in a sample of size n_C from population c, using binomial sampling:

$$P(m_A, m_B|m_C, \mathbf{v}, r, s) = \frac{\int_0^1 \int_0^1 P(m_A|p_A) P(m_B|p_B) f(p_A, p_B|m_C, \mathbf{v}, r, s) dp_A dp_B}{\int_0^1 \int_0^1 f(p_A, p_B|m_C, \mathbf{v}, r, s) dp_A dp_B} \tag{14}$$

where

$$P(m_A|p_A) = \binom{n_A}{m_A} p_A^{m_A} (1-p_A)^{n_A - m_A} \tag{15}$$

and

$$P(m_B|p_B) = \binom{n_B}{m_B} p_B^{m_B} (1-p_B)^{n_B - m_B} \tag{16}$$

This allows us to calculate a composite likelihood of the derived allele counts in a and b given the derived allele counts in c:

$$CL_{3P-CLR}(\mathbf{r}, \mathbf{v}, s) = \prod_{j=1}^{k} P(m_A^j, m_B^j | m_C^j, \mathbf{v}, r^j, s) \tag{17}$$

As before, we can use this composite likelihood to produce a composite likelihood ratio statistic that can be calculated over regions of the genome to test the hypothesis of linked selection centered on a particular locus against the hypothesis of neutrality. Due to computational costs in numerical integration, we skip the sampling step for population c (formula 13) in our implementation of 3P-CLR. In other words, we assume $p_C = m_C/n_C$, but this is also assumed in XP-CLR when computing its corresponding outgroup frequency. We implemented our method in a freely available C++ program that can be downloaded from here:

`https://github.com/ferracimo` [WILL POST IT AFTER PUBLICATION]

4 Results

4.1 Simulations

We generated simulations in SLiM [21] to test the performance of XP-CLR and 3P-CLR in a three-population scenario. We focused specifically on the performance of 3P-CLR(Int) in detecting ancient selective events that occurred in the ancestral branch of two sister populations. We assumed that the population history had been correctly estimated by the researcher (i.e. the drift parameters and population topology were known). First, we simulated scenarios in which a beneficial mutation arose in the ancestor of populations a and b, before their split from each other but after their split from c (Table 1). Although both XP-CLR and 3P-CLR are sensitive to partial or soft sweeps (as they do not rely on extended patterns of homozygosity [1]), we required the allele to have fixed before the split (at time t_{ab}) to ensure that the allele had not been lost before it, and also to ensure that the sweep was restricted to the internal branch of the tree. We fixed the effective size of all three populations at $N_e = 10,000$. Each simulation consisted in a 5 cM region and the beneficial mutation occurred in the center of this region. The mutation rate was set at $2.5 * 10^{-8}$ per generation and the recombination rate was set at 10^{-8} per generation.

To make a fair comparison to 3P-CLR(Int), and given that XP-CLR is a two-population test, we applied XP-CLR in two ways. First, we pretended population b was not sampled, and so the "test" panel consisted of individuals from a only, while the "outgroup" consisted of individuals from c. In the second implementation (which we call "XP-CLR-avg"), we used the same outgroup panel, but pretended that individuals from a and b were pooled into a single panel, and this pooled panel was the "test". The window size was set at 0.5 cM and the space between the center of each window was set at 600 SNPs. To speed up computation, and because we are largely interested in comparing the relative performance of the three tests under different scenarios, we used only 20 randomly chosen SNPs per window in all tests. We note, however, that the performance of all three tests can be improved by using more SNPs per window.

Figure 2 shows receiver operating characteristic (ROC) curves comparing the sensitivity and specificity of 3P-CLR(Int), XP-CLR and XP-CLR-avg in the first six demographic scenarios described in Table 1. Each ROC curve was made from 100 simulations under selection (with $s = 0.1$ for the central mutation) and 100 simulations under neutrality (with $s = 0$ and no fixation required). In each simulation, 100 individuals were sampled from population a, 100 from population b and 10 from the outgroup population c. This emulates a situation in which only a few individuals have been sequenced from the outgroup, while large numbers of sequences are available from the tests (e.g. two populations of present-day humans). For each simulation, we took the maximum value at a region in the neighborhood of the central mutation (+/- 0.5 cM) and used those values to compute ROC curves under the two models.

When the split times are recent or moderately ancient (models A to D), 3P-CLR(Int) outperforms the two versions of XP-CLR. When the split times are very ancient (models E and F), none of the tests perform well. The root mean squared error (RMSE) of the genetic distance between the true selected site and the highest scored window is comparable across tests in all six scenarios (Figure S2). Finally, Figures S1 and S3 show the ROC curves and RMSE plots, respectively, for a case in which 100 individuals were sampled from all three populations (including the outgroup), with similar results.

Importantly, the usefulness of 3P-CLR(Int) resides not just in its performance at detecting selective sweeps in the ancestral population, but in its specific sensitivity to that particular type of events. Because the test relies on correlated allele frequency differences in both population a and population b (relative to the outgroup), selective sweeps that are specific to only one of the populations will not lead to high 3P-CLR(Int) scores. Figure 3 shows ROC curves in two scenarios in which a selective sweep event occurred only in population a (Models I and J in Table 1), using 100 sampled individuals from each of the 3 populations. Here, XP-CLR performs well, but 3P-CLR(Int) shows almost no sensitivity to the recent sweep, under reasonable specificity cutoffs. For example, in Model I, at a specificity of 95%, XP-CLR has 80% sensitivity, while at the same specificity, 3P-CLR(Int) only has 14% sensitivity. One can compare this to the same demographic scenario but with selection occurring in the ancestral population (Model C, Figure S1), where at 95% specificity, XP-CLR has 69% sensitivity, while 3P-CLR has 83% sensitivity.

4.2 Selection in Eurasians

We first applied 3P-CLR to modern human data from the 1000 Genomes Project [12]. We used the African-American recombination map [22] to convert physical distances into genetic distances. We focused on two populations (Europeans and East Asians), using Africans as the outgroup population (Figure S4.A). We randomly sampled 100 individuals from each population and obtained sample derived allele frequencies every 10 SNPs in the genome. We then calculated likelihood ratio statistics by a sliding window approach, where we sampled a "central SNP” once every 20 SNPs. The central SNP in each window was the candidate beneficial SNP for that window. We set the window size to 0.25 cM, and randomly sampled 100 SNPs from each window, centered around the candidate beneficial SNP. In each window, we calculated 3P-CLR to test for selection at three different branches of the population tree: the terminal branch leading to Europeans (3P-CLR Europe), the terminal branch leading to East Asians (3P-CLR East Asia) and the ancestral branch of Europeans and East Asians (3P-CLR Eurasia). Results are shown in Figure 4. For each scan, we selected the windows in the top 99.9% quantile of scores and merged them together if they were contiguous. Tables 2, 3 and 4 show the top hits for Europeans, East Asians and the ancestral Eurasian branch, respectively

We observe several genes that have been identified in previous selection scans. In the East Asian branch, one of the top hits is *EDAR*. This gene codes for a protein involved in hair thickness and incisor tooth morphology [23,24]. It has been repeatedly identified in earlier selections scans as having undergone a sweep in East Asians [25,26].

Furthermore, 3P-CLR allows us to narrow down on the specific time at which selection occurred in the history of particular populations. For example, ref. [1] performed a scan of the genomes of East Asians using XP-CLR with Africans as the outgroup, and identified a number of genes as being under selection [1]. 3P-CLR confirms this signal in several of these loci when looking specifically at the East Asian branch: *CYP26B1, EMX1, SPR, SFXN5, SLC30A9, PPARA, PKDREJ, GTSE1, TRMU, CELSR1, PINX1, XKR6, CD226, ACD, PARD6A, GFOD2, RANBP10, TSNAXIP1, CENPT, THAP11, NUTF2, CDH16, RRAD, FAM96B, CES2, CBFB, C16orf70, TRADD, FBXL8, HSF4, NOL3, EXOC3L1, E2F4, ELMO3, LRRC29, FHOD1, SLC9A5, PLEKHG4, LRRC36, ZDHHC1, HSD11B2, AtP6V0D1, AGRP, FAM65A, CTCF* and *RLTPR*. However, when applied to the ancestral Eurasian branch, 3P-CLR finds some genes that were previously found in the XP-CLR analysis of East Asians, but that are not among the top hits in 3P-CLR applied to the East Asian branch: *COMMD3, BMI1, SPAG6, CD226, SLC30A9,LONP2, SIAH1, ABCC11* and *ABCC12*. This suggests selection in these regions occurred earlier, i.e. before the European-East Asian split. Figure 5 shows a comparison between the 3P-CLR scores for the three branches in the region containing genes *BMI1* (a proto-oncogene [27]) and *SPAG6* (involved in sperm motility [28]). In that figure, the score within each window was standardized using its chromosome-wide mean and standard deviation, to make a fair comparison. One can observe that the signal of Eurasia-specific selection is evidently stronger than the other two signals.

Other selective events that 3P-CLR infers to have occurred in Eurasians include the region containing *HERC2* and *OCA2*, which are major determinants of eye color [29–31]. There is also evidence that these genes underwent selection more recently in the history of Europeans [32], which could suggest an extended period of selection - perhaps influenced by migrations between Asia and Europe - or repeated selective events at the same locus.

When running 3P-CLR to look for selection specific to Europe, we find that *TYRP1* and *MYO5A*, which play a role in human skin pigmentation [33–36], are among the top hits. Both of these genes have been previously found to be under strong selection in Europe [37], using a statistic called iHS, which measures extended patterns of homozygosity that are characteristic of selective sweeps. Interestingly, a change in the gene *TYRP1* has also been found to cause a blonde hair phenotype in Melanesians [38].

4.3 Selection in ancestral modern humans

We applied 3P-CLR to modern human data combined with recently sequenced archaic human data [10]. We sought to find selective events that occurred in modern humans after their spit from archaic groups. We used the combined Neanderthal and Denisovan high-coverage genomes [9,10] as the outgroup population, and, for our two test populations, we randomly sampled 100 Eurasian genomes and 100 African genomes from the 1000 Genomes data (Figure S4.B). We used previously estimated drift times as fixed parameters [10], and tested for selective events that occurred more anciently than the split of Africans and Eurasians, but more recently than the split from Neanderthals. We run 3P-CLR using 0.25 cM windows as above, but also verified that the density of scores was robust to the choice of window size and spacing (Figure S5). As before, we selected the top 99.9% windows and merged them together if they were contiguous. Table 5 and Figure S6 show the top hits. To find putative candidates for the beneficial variants in each region, we queried the catalogs of modern human-specific high-frequency or fixed derived changes that are ancestral in the Neanderthal and/or the Denisova genomes [10, 39].

We observe several genes that have been identified in previous scans that looked for selection in modern humans after their split from archaic groups [8, 10]: *SIPA1L1, ANAPC10, ABCE1, RASA1, CCNH, KCNJ3, HBP1, COG5, GPR22, DUS4L, BCAP29, CALDPS2, RNF133, RNF148, FAM172A, POU5F2, FGF7, RABGAP1, GPR21, STRBP, SMURF1, GABRA2, ALMS1, PVRL3, EHBP1, VPS54, OTX1, UGP2, HCN1, GTDC1, ZEB2, OIT3, USP54, MYOZ1* and *DPYD*. One of our strongest candidate genes among these is *ANAPC10*. This gene is a core subunit of the cyclosome, is involved in progression through the cell cycle [40], and may play a role in oocyte maturation and human T-lymphotropic virus infection (KEGG pathway [41]). *ANAPC10* is noteworthy because it was found to be significantly differentially expressed in humans compared to other great apes and macaques: it is up-regulated in the testes [42]. The gene also contains two intronic changes that are fixed derived in modern humans, ancestral in both Neanderthals and Denisovans and that have evidence for being highly disruptive, based on a composite score that combines conservation and regulatory data (PHRED-scaled C-scores $>$ 11 [10, 43]). The changes, however, appear not to lie in any obvious regulatory region [44, 45].

We also find *ADSL* among the list of candidates. This gene is known to contain a nonsynonymous change that is fixed in all present-day humans but homozygous ancestral in the Neanderthal genome, the Denisova genome and two Neanderthal exomes [39] (Figure 6.A). It was previously identified as lying in a region with strong support for positive selection in modern humans, using summary statistics implemented in an ABC method [11]. The gene is interesting because it is one of the members of the Human Phenotype ontology category "aggression / hyperactivity" which is enriched for nonsynonymous changes that occurred in the modern human lineage after the split from archaic humans [39, 46]. *ADSL* codes for adenylosuccinase, an enzyme involved in purine metabolism [47]. A deficiency of adenylosuccinase can lead to apraxia, speech deficits, delays in development and abnormal behavioral features, like hyperactivity and excessive laughter [48]. The nonsynonymous mutation (A429V) is in the C-terminal domain of the protein (Figure 6.B) and lies in a highly conserved position (primate PhastCons = 0.953; GERP score = 5.67 [43, 49, 50]). The ancestral amino acid is conserved across the tetrapod phylogeny, and the mutation is only three residues away from the most common causative SNP for severe adenylosuccinase deficiency [51–55]. The change has the highest probability of being disruptive to protein function, out of all the nonsynonymous modern-human-specific changes that lie in the top-scoring regions (C-score = 17.69). While *ADSL* is an interesting candidate and lies in the center of the inferred selected region (Figure 6.A), there are other genes in the region too, including *TNRC6B* and *MKL1*. *TNRC6B* may be involved in miRNA-guided gene silencing [56], while *MKL1* may play a role in smooth muscle differentiation [57], and has been associated with acute megakaryocytic leukemia [58].

RASA1 was also a top hit in a previous scan for selection [8], and was additionally inferred to have a high Bayes factor in favor of selection in ref. [11]. The gene codes for a protein involved in the control of cellular differentiation [59]. Human diseases associated with *RASA1* include basal cell carcinoma [60] and ateriovenous malformation [61, 62].

The $GABA_A$ gene cluster in chromosome 4p12 is also among the top regions. The genes within the putatively selected region code for three of the subunits of the $GABA_A$ receptor (*GABRA2, GABRA4, GABRB1*), which codes for a ligand-gated ion channel that plays a key role in synaptic inhibtion in the central nervous system (see review by ref. [63]). *GABRA2* is significantly associated with the risk of alcohol dependence in humans [64], perception of pain [65] and asthma [66]. In turn, GABRA4 is associated with autism risk [67, 68].

Two other candidate genes that may be involved in brain development are *FOXG1* and *CADPS2*. *FOXG1* was not identified in any of the previous selection scans, and codes for a protein called forkhead box G1, which plays an important role during brain development. Mutations in this gene have been associated with a slow-down in brain growth during childhood resulting in microcephaly, which in turn causes various intellectual disabilities [69, 70]. *CADPS2* was identified in [8] as a candidate for selection, and has been associated with autism [71]. The gene has been suggested to be specifically important in the evolution of all modern humans, as it was not found to be selected earlier in great apes or later in particular modern human populations [72].

Finally, we find a signal of selection in a region containing the gene *EHBP1* and *OTX1*. This region was identified in both of the two previous scans for modern human selection [8,10]. *EHBP1* codes for a protein involved in endocytic trafficking [73] and has been associated with prostate cancer [74]. *OTX1* is a homeobox family gene that may play a role in brain development [75]. Interestingly, *EHBP1* contains a single-nucleotide intronic change (chr2:63206488) that is almost fixed in all present-day humans and homozygous ancestral in Neanderthal and Denisova [10]. This change is also predicted to be highly disruptive (C-score = 13.1) and lies in a position that is extremely conserved across primates (PhastCons = 0.942), mammals (PhastCons = 1) and vertebrates (PhastCons = 1). The change is 18 bp away from the nearest splice site and overlaps a VISTA conserved enhancer region (element 1874) [76], which suggests a putative regulatory role for the change.

4.4 Modern human-specific high-frequency changes in GWAS catalog

We overlapped the genome-wide association studies (GWAS) database [77,78] with the list of fixed or high-frequency modern human-specific changes that are ancestral in archaic humans [10] and that are located within our top putatively selected regions in modern humans (Table 6). None of the resulting SNPs are completely fixed derived, because GWAS can only yield associations from sites that are segregating. Among these SNPs, the one with the highest probability of being disruptive (rs10003958, C-score = 16.58, Gerp score = 6.07) is located in a highly-conserved regulatory ("strong enhancer") region in the *RAB28* gene [44,45] (Primate PhastCons = 0.951), and is significantly associated with obesity [79] (Figure 7.A). Interestingly, the region containing *RAB28* is inferred to have been under positive selection in both the modern human and the Eurasian ancestral branches (Tables 4, 5). In line with this evidence, the derived allele of rs10003958 is absent in archaic humans, at very high frequencies in Eurasians ($> 94\%$), and only at moderately high frequencies in Africans (74%) (Figure 7.B).

We also find a highly disruptive SNP (rs10171434, C-score = 8.358) associated with urinary metabolites [80] and suicidal behavior in patients with mood disorders [81]. The SNP is located in an enhancer regulatory freature [44,45] located between genes *PELI1* and *VPS54*, in the same putatively selected region as genes *EHBP1* and *OTX1* (see above). Finally, there is a highly disruptive SNP (rs731108, C-score = 10.31) that is associated with renal cell carcinoma [82]. This SNP is also located in an enhancer regulatory feature [44,45], in an intron of *ZEB2*. In this last case, though, only the Neanderthal genome has the ancestral state, while the Denisova genome carries the modern human variant.

5 Discussion

We have developed a new method called 3P-CLR, which allows us to detect positive selection along the genome. The method is based on an earlier test (XP-CLR [1]) that uses linked allele frequency differences between two populations to detect population-specific selection. However, 3P-CLR can allow us to distinguish between selective events that occurred before and after the split of two populations. Our method also has some similiarities to an earlier method developed by [83], which used an F_{st}-like score to detect selection ancestral to two populations. In that case, though, the authors used summary statistics and did not explicitly model the process leading to allele frequency differentiation.

We used our method to confirm previously found candidate genes in particular human populations, like *EDAR*, *TYRP1* and *HERC2*, and find some novel candidates too (Tables 2, 3, 4). Additionally, we can infer that certain genes, which were previously known to have been under selection in East Asians (like *SPAG6*), are more likely to have undergone a sweep in the population ancestral to both Europeans and East Asians than in East Asians only.

We also used 3P-CLR to detect selective events that occurred in the ancestors of modern humans, after their split from Neanderthals and Denisovans (Table 5). These events could perhaps have led to the spread of phenotypes that set modern humans apart from other hominin groups. We find several intersting candidates, like *SIPA1L1, ADSL, RASA1, OTX1, EHBP1, FOXG1, RAB28* and *ANAPC10*, some of which were previously detected using other types of methods [8,10,11].

An advantage of differentiation-based tests like XP-CLR and 3P-CLR is that, unlike other patterns detected by tests of neutrality (like extended haplotype homozygostiy, [84]) that are exclusive to hard sweeps, the patterns that both XP-CLR and 3P-CLR are tailored to find are based on regional allele frequency differences between populations. These patterns can also be produced by soft sweeps from standing variation or by partial sweeps [1], and there is some evidence that the latter phenomena may have been more important than classic sweeps during human evolutionary history [85].

Another advantage of both XP-CLR and 3P-CLR is that they do not rely on an arbitrary division of genomic space. Unlike other methods which require the partition of the genome into small windows of fixed size, our composite likelihood ratios can theoretically be computed over windows that are as big as each chromosome, while only switching the central candidate site at each window. This is because the likelihood ratios use the genetic distance to the central SNP as input. SNPs that are very far away from the central SNP will not contribute much to the likelihood function of both the neutral and the selection models, while those that are close to it will. While we heuristically limit the window size in our implementation in the interest of speed, this can be arbitrarily adjusted by the user as needed. The use of genetic distance in the likelihood function also allows us to take advantage of the spatial distribution of SNPs as an additional source of information, rather than only relying on patterns of population differentiation restricted to tightly linked SNPs.

3P-CLR also has an advantage over HMM-based selection methods, like the one implemented in ref. [10]. The likelihood ratio scores obtained from 3P-CLR can provide an idea of how credible a selection model is for a particular region, relative to the rest of the genome. The HMM-based method previously used to scan for selection in modern humans [10] can only rank putatively selected regions by genetic distance, but cannot output a statistical measure that may indicate how likely each region is to have been selected in ancient times. In contrast, 3P-CLR provides a composite likelihood ratio score, which allows for a statistically rigorous way to compare the neutral model and a specific selection model (for example, recent or ancient selection). The score also gives an idea of how much fainter the signal of ancient selection in modern humans is, relative to recent selection specific to a particular present-day population. For example, the outliers from Figure 4 have much higher scores (relative to the rest of the genome) than the outliers from Figure S6. This may be due to both the difference in time scales in the two sets of tests and to the uncertainty that comes from estimating outgroup allele frequencies using only two archaic genomes. This pattern can also be observed in Figure S7, where the densities of the scores looking for patterns of ancient selection have much shorter tails than the densities of scores looking for patterns of recent selection.

Like XP-CLR, 3P-CLR is largely robust to the underlying population history, even when this is wrongly specified, as it relies on looking for extreme allele frequency differences that are restricted to a particular region. We have noticed, however, that these types of tests may not be robust to admixture events from the outgroup population used. For example, we observe that 3P-CLR finds evidence for selection in the region containing *HYAL2* (involved in the cellular response to ultraviolet radiation), when run in the East Asian branch. This makes sense, as a variant of this gene is known to have been pushed to high frequencies by selection specifically in East Asians. However, this variant likely came from Neanderthals via introgression [86, 87]. While we do not observe that the *HYAL2* region is a top hit in either the European or the Eurasian ancestral branches, we do observe it as a top hit in the modern human ancestral branch. This is puzzling, given that the selected haplotype should not have been introduced into modern humans until after the split of Africans and non-Africans. One explanation for the appearance of this region in both the East Asian and the modern human top hits is that the introgression event could perhaps confound the signal that 3P-CLR targets, as we are assuming no admixture in our demographic model. Another possibility is that the region has suffered multiple episodes of repeated selection. Incorporating admixture into the modeling procedure may help to disentangle this pattern better, but we leave this to a future work.

A further limitation of composite likelihood ratio tests is that the composite likelihood calculated for each model under comparison is obtained from a product of individual likelihoods at each site, and so it underestimates the correlation that exists between SNPs due to linkage effects [1, 16, 17, 88]. One way to mitigate this problem is by using corrective weights based on linkage disequilibrium (LD) statistics calculated on the outgroup population [1]. Our implementation of 3P-CLR allows the user to incorporate such weights, if appropriate LD statistics are available from the outgroup. However, in cases where these are unreliable, it may not be possible to fully correct for this (for example, when only a few unphased genomes are available, as in the case of the Neanderthal and Denisova genomes).

While 3P-CLR relies on integrating over the possible allele frequencies in the ancestors of populations a and b (formula 10), one could envision using ancient DNA to avoid this step. Thus, if enough genomes could be sampled from that ancestral population that existed in the past, one could use the sample frequency in the ancient set of genomes as a proxy for the ancestral population frequency. This may soon be possible, as several early modern human genomes have already been sequenced in recent years [89–91].

Though we have limited ourselves to a three-population model in this manuscript, it should be straightforward to expand our model to a larger number of populations, albeit with additional costs in terms of speed and memory. Our method relies on a similar framework to the demographic inference method implemented in TreeMix [92], which can estimate complex population trees that include migration events, using genome-wide data. With a more complex modeling framework, it may be possible to estimate the time and strength of selective events with better resolution, and to incorporate additional demographic forces, like continuous migration between populations or pulses of admixture.

Acknowledgments

We thank Montgomery Slatkin, Rasmus Nielsen, Joshua Schraiber, Nicolas Duforet-Frebourg, Emilia Huerta-Sánchez, Hua Chen, Nick Patterson, David Reich, Joachim Hermisson, Graham Coop and members of the Slatkin and Nielsen labs for helpful advice and discussions. This work was supported by NIH grant R01-GM40282 to Montgomery Slatkin.

References

1. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome research 20: 393–402.

2. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetical research 23: 23–35.

3. Lewontin R, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175–195.

4. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density snp map for signatures of natural selection. Genome research 12: 1805–1814.

5. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome research 15: 1468–1476.

6. Oleksyk TK, Zhao K, Francisco M, Gilbert DA, O'Brien SJ, et al. (2008) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. PLoS One 3: e1712.

7. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75–78.

8. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the neandertal genome. science 328: 710–722.

9. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic denisovan individual. Science 338: 222–226.

10. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. (2014) The complete genome sequence of a neanderthal from the altai mountains. Nature 505: 43–49.

11. Racimo F, Kuhlwilm M, Slatkin M (2014) A test for ancient selective sweeps and an application to candidate sites in modern humans. Molecular biology and evolution 31: 3344–3358.

12. Consortium GP, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

13. Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, et al. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64: 695–715.

14. Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theoretical population biology 66: 129–138.

15. Fay JC, Wu CI (2000) Hitchhiking under positive darwinian selection. Genetics 155: 1405–1413.

16. Lindsay BG (1988) Composite likelihood methods. Contemporary Mathematics 80: 221–39.

17. Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Statistica Sinica 21: 5–42.

18. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. (2012) Ancient admixture in human history. Genetics 192: 1065–1093.

19. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS genetics 5: e1000695.

20. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and snp data. PLoS genetics 9: e1003905.

21. Messer PW (2013) Slim: simulating evolution with selection and linkage. Genetics 194: 1037–1039.

22. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, et al. (2011) The landscape of recombination in african americans. Nature 476: 170–175.

23. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, et al. (2008) A scan for genetic determinants of human hair morphology: Edar is associated with asian hair thickness. Human Molecular Genetics 17: 835–843.

24. Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, et al. (2009) A common variation in edar is a genetic determinant of shovel-shaped incisors. The American Journal of Human Genetics 85: 528–535.

25. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

26. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327: 883–886.

27. Siddique HR, Saleem M (2012) Role of bmi1, a stem cell factor, in cancer recurrence and chemoresistance: preclinical and clinical evidences. Stem Cells 30: 372–378.

28. Sapiro R, Kostetskii I, Olds-Clarke P, Gerton GL, Radice GL, et al. (2002) Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6. Molecular and cellular biology 22: 6298–6305.

29. Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, 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. Human genetics 123: 177–187.

30. Han J, Kraft P, Nan H, Guo Q, Chen C, et al. (2008) A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS genetics 4: e1000074.

31. Branicki W, Brudnik U, Wojas-Pelc A (2009) Interactions between herc2, oca2 and mc1r may influence human pigmentation phenotype. Annals of human genetics 73: 160–170.

32. Mathieson I, Lazaridis I, Rohland N, Mallick S, Llamas B, et al. (2015) Eight thousand years of natural selection in europe. bioRxiv : 016477.

33. Pastural E, Ersoy F, Yalman N, Wulffraat N, Grillo E, et al. (2000) Two genes are responsible for griscelli syndrome at the same 15q21 locus. Genomics 63: 299–306.

34. Fukuda M, Kuroda T, Mikoshiba K (2002) Slac2-a/melanophilin, the missing link between rab27 and myosin va: implications of a tripartite protein complex for melanosome transport. The Journal of biological chemistry 277: 12432.

35. Halaban R, Moellmann G (1990) Murine and human b locus pigmentation genes encode a glycoprotein (gp75) with catalase activity. Proceedings of the National Academy of Sciences 87: 4809–4813.

36. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, et al. (2008) Two newly identified genetic determinants of pigmentation in europeans. Nature genetics 40: 835–837.

37. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS biology 4: e72.

38. Kenny EE, Timpson NJ, Sikora M, Yee MC, Moreno-Estrada A, et al. (2012) Melanesian blond hair is caused by an amino acid change in tyrp1. Science 336: 554–554.

39. Castellano S, Parra G, Sánchez-Quinto FA, Racimo F, Kuhlwilm M, et al. (2014) Patterns of coding variation in the complete exomes of three neandertals. Proceedings of the National Academy of Sciences 111: 6666–6671.

40. Pravtcheva DD, Wise TL (2001) Disruption of apc10/doc1 in three alleles of oligosyndactylism. Genomics 72: 78–87.

41. Kanehisa M, Goto S (2000) Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27–30.

42. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, et al. (2011) The evolution of gene expression levels in mammalian organs. Nature 478: 343–348.

43. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46: 310–315.

44. Consortium EP, et al. (2012) An integrated encyclopedia of dna elements in the human genome. Nature 489: 57–74.

45. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, et al. (2011) Encode whole-genome data in the ucsc genome browser: update 2012. Nucleic acids research : gkr1012.

46. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, et al. (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics 83: 610–615.

47. Van Keuren M, Hart I, Kao FT, Neve R, Bruns G, et al. (1987) A somatic cell hybrid with a single human chromosome 22 corrects the defect in the cho mutant (ade–i) lacking adenylosuccinase activity. Cytogenetic and Genome Research 44: 142–147.

48. Gitiaux C, Ceballos-Picot I, Marie S, Valayannopoulos V, Rio M, et al. (2009) Misleading behavioural phenotype with adenylosuccinate lyase deficiency. European Journal of Human Genetics 17: 133–136.

49. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research 15: 1034–1050.

50. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, et al. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nature methods 7: 250–251.

51. Kmoch S, Hartmannová H, Stibùrková B, Krijt J, Zikánová M, et al. (2000) Human adenylosuccinate lyase (adsl), cloning and characterization of full-length cdna and its isoform, gene structure and molecular basis for adsl deficiency in six patients. Human molecular genetics 9: 1501–1513.

52. Maaswinkel-Mooij P, Laan L, Onkenhout W, Brouwer O, Jaeken J, et al. (1997) Adenylosuccinase deficiency presenting with epilepsy in early infancy. Journal of inherited metabolic disease 20: 606–607.

53. Marie S, Cuppens H, Heuterspreute M, Jaspers M, Tola EZ, et al. (1999) Mutation analysis in adenylosuccinate lyase deficiency: Eight novel mutations in the re-evaluated full adsl coding sequence. Human mutation 13: 197–202.

54. Race V, Marie S, Vincent MF, Van den Berghe G (2000) Clinical, biochemical and molecular genetic correlations in adenylosuccinate lyase deficiency. Human molecular genetics 9: 2159–2165.

55. Edery P, Chabrier S, Ceballos-Picot I, Marie S, Vincent MF, et al. (2003) Intrafamilial variability in the phenotypic expression of adenylosuccinate lyase deficiency: a report on three patients. American Journal of Medical Genetics Part A 120: 185–190.

56. Meister G, Landthaler M, Peters L, Chen PY, Urlaub H, et al. (2005) Identification of novel argonaute-associated proteins. Current biology 15: 2149–2155.

57. Du KL, Chen M, Li J, Lepore JJ, Mericko P, et al. (2004) Megakaryoblastic leukemia factor-1 transduces cytoskeletal signals and induces smooth muscle cell differentiation from undifferentiated embryonic stem cells. Journal of Biological Chemistry 279: 17578–17586.

58. Mercher T, Busson-Le Coniat M, Monni R, Mauchauffé M, Khac FN, et al. (2001) Involvement of a human gene related to the drosophila spen gene in the recurrent t (1; 22) translocation of acute megakaryocytic leukemia. Proceedings of the National Academy of Sciences 98: 5776–5779.

59. Trahey M, Wong G, Halenbeck R, Rubinfeld B, Martin GA, et al. (1988) Molecular cloning of two types of gap complementary dna from human placenta. Science 242: 1697–1700.

60. Friedman E, Gejman PV, Martin GA, McCormick F (1993) Nonsense mutations in the c–terminal sh2 region of the gtpase activating protein (gap) gene in human tumours. Nature genetics 5: 242–247.

61. Eerola I, Boon LM, Mulliken JB, Burrows PE, Dompmartin A, et al. (2003) Capillary malformation–arteriovenous malformation, a new clinical and genetic disorder caused by rasa1 mutations. The American Journal of Human Genetics 73: 1240–1249.

62. Hershkovitz D, Bercovich D, Sprecher E, Lapidot M (2008) Rasa1 mutations may cause hereditary capillary malformations without arteriovenous malformations. British Journal of Dermatology 158: 1035–1040.

63. Whiting PJ, Bonnert TP, McKernan RM, Farrar S, Bourdelles BL, et al. (1999) Molecular and functional diversity of the expanding gaba-a receptor gene family. Annals of the New York Academy of Sciences 868: 645–653.

64. Edenberg HJ, Dick DM, Xuei X, Tian H, Almasy L, et al. (2004) Variations in gabra2, encoding the $\alpha 2$ subunit of the gaba a receptor, are associated with alcohol dependence and with brain oscillations. The American Journal of Human Genetics 74: 705–714.

65. Knabl J, Witschi R, Hösl K, Reinold H, Zeilhofer UB, et al. (2008) Reversal of pathological pain through specific spinal gabaa receptor subtypes. Nature 451: 330–334.

66. Xiang YY, Wang S, Liu M, Hirota JA, Li J, et al. (2007) A gabaergic system in airway epithelium is essential for mucus overproduction in asthma. Nature medicine 13: 862–867.

67. Ma D, Whitehead P, Menold M, Martin E, Ashley-Koch A, et al. (2005) Identification of significant association and gene-gene interaction of gaba receptor subunit genes in autism. The American Journal of Human Genetics 77: 377–388.

68. Collins AL, Ma D, Whitehead PL, Martin ER, Wright HH, et al. (2006) Investigation of autism and gaba receptor subunit genes in multiple ethnic groups. Neurogenetics 7: 167–174.

69. Ariani F, Hayek G, Rondinella D, Artuso R, Mencarelli MA, et al. (2008) Foxg1 is responsible for the congenital variant of rett syndrome. The American Journal of Human Genetics 83: 89–93.

70. Mencarelli M, Spanhol-Rosseto A, Artuso R, Rondinella D, De Filippis R, et al. (2010) Novel foxg1 mutations associated with the congenital variant of rett syndrome. Journal of medical genetics 47: 49–53.

71. Sadakata T, Furuichi T (2010) Ca 2+-dependent activator protein for secretion 2 and autistic-like phenotypes. Neuroscience research 67: 197–202.

72. Crisci JL, Wong A, Good JM, Jensen JD (2011) On characterizing adaptive events unique to modern humans. Genome biology and evolution 3: 791–798.

73. Guilherme A, Soriano NA, Furcinitti PS, Czech MP (2004) Role of ehd1 and ehbp1 in perinuclear sorting and insulin-regulated glut4 recycling in 3t3-l1 adipocytes. Journal of Biological Chemistry 279: 40062–40075.

74. Gudmundsson J, Sulem P, Rafnar T, Bergthorsson JT, Manolescu A, et al. (2008) Common sequence variants on 2p15 and xp11. 22 confer susceptibility to prostate cancer. Nature genetics 40: 281–283.

75. Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, et al. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature 425: 917–925.

76. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502.

77. Li MJ, Wang P, Liu X, Lim EL, Wang Z, et al. (2011) Gwasdb: a database for human genetic variants identified by genome-wide association studies. Nucleic acids research : gkr1182.

78. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The nhgri gwas catalog, a curated resource of snp-trait associations. Nucleic acids research 42: D1001–D1006.

79. Paternoster L, Evans DM, Nohr EA, Holst C, Gaborieau V, et al. (2011) Genome-wide population-based association study of extremely overweight young adults–the goya study. PLoS One 6: e24303.

80. Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, et al. (2011) A genome-wide association study of metabolic traits in human urine. Nature genetics 43: 565–569.

81. Perlis RH, Huang J, Purcell S, Fava M, Rush AJ, et al. (2010) Genome-wide association study of suicide attempts in mood disorder patients. Genome 167.

82. Henrion M, Frampton M, Scelo G, Purdue M, Ye Y, et al. (2013) Common variation at 2q22. 3 (zeb2) influences the risk of renal cancer. Human molecular genetics 22: 825–831.

83. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, et al. (2012) Genomic variation in seven khoe-san groups reveals adaptation and complex african history. Science 338: 374–379.

84. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

85. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic selective sweeps were rare in recent human evolution. science 331: 920–924.

86. Ding Q, Hu Y, Xu S, Wang J, Jin L (2013) Neanderthal introgression at chromosome 3p21. 31 was under positive natural selection in east asians. Molecular biology and evolution : mst260.

87. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, et al. (2014) The genomic landscape of neanderthal ancestry in present-day humans. Nature 507: 354–357.

88. Pace L, Salvan A, Sartori N (2011) Adjusting composite likelihood ratio statistics. Statistica Sinica 21: 129.

89. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, et al. (2014) Genome sequence of a 45,000-year-old modern human from western siberia. Nature 514: 445–449.

90. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, et al. (2014) Genomic structure in europeans dating back at least 36,200 years. Science 346: 1113–1118.

91. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day europeans. Nature 513: 409–413.

92. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics 8: e1002967.

639 Tables

Table 1. Description of models tested. All times are in generations. Selection in the "ancestral population" refers to a selective sweep where the beneficial mutation and fixation occurred before the split time of the two most closely related populations. Selection in "daughter population A" refers to a selective sweep that occurred in one of the two most closely related populations (A), after their split from each other.

Model	Population where selection occurred	t_{AB}	t_{ABC}	t_M	s	N_e
A	Ancestral population	500	2,000	1,800	0.1	10,000
B	Ancestral population	1,000	4,000	2,500	0.1	10,000
C	Ancestral population	2,000	4,000	3,500	0.1	10,000
D	Ancestral population	3,000	8,000	5,000	0.1	10,000
E	Ancestral population	2,000	16,000	8,000	0.1	10,000
F	Ancestral population	4,000	16,000	8,000	0.1	10,000
I	Daughter population A	2,000	4,000	1,000	0.1	10,000
J	Daughter population A	3,000	8,000	2,000	0.1	10,000

Table 2. Top hits for 3P-CLR run on the European terminal branch, using Africans as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if they were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
17	19175100	18858600	19445800	173.751	SLC5A10,FAM83G,GRAP,GRAPL,EPN2,B9D1,MAPK7,MFAP4,RNF112,SLC47A1
10	74736500	74007800	75402200	161.092	DDIT4,DNAJB12,MICU1,MCU,OIT3,PLA2G12B,P4HA1,NUDT13,ECD,FAM149B1, DNAJC9,MRPS16,TTC18,ANXA7,MSS51,PPP3CB,USP54,MYOZ1
15	29241200	29210600	29338200	155.873	APBA2
12	113010000	111691000	113030000	145.07	BRAP,ACAD10,ALDH2,MAPKAPK5,TMEM116,ERP29,NAA25,TRAFD1,RPL6,PTPN11, RPH3A,CUX2,FAM109A,SH2B3,ATXN2
1	35623900	35380800	36584500	140.484	DLGAP3,ZMYM6NB,ZMYM6,ZMYM1,SFPQ,ZMYM4,KIAA0319L,NCDN, TFAP2E,PSMB2,C1orf216,CLSPN,AGO4,AGO1,AGO3,TEKT2,ADPRHL2,COL8A2
14	66765100	66471400	67923400	125.356	GPHN,FAM71D,MPP5,ATP6V1D,EIF2S1,PLEK2,TMEM229B
11	64581000	64217100	64588600	122.885	RASGRP2,PYGM,SF1,MAP4K2,MEN1,SLC22A11,SLC22A12,NRXN2
2	74507400	74365200	74970500	118.922	INO80B,WBP1,MOGS,MRPL53,CCDC142,TTC31,LBX2,PCGF1,TLX2,DQX1,AUP1, HTRA2,LOXL3,DOK1,M1AP,SEMA4F,BOLA3,MOB1A,MTHFD2,SLC4A5,DCTN1, WDR54,RTKN
15	72654000	72057800	73108700	115.575	THSD4,MYO9A,SENP8,GRAMD2,PKM,PARP6,CELF6,HEXA,TMEM202,ARIH1, GOLGA6B,BBS4,ADPGK
7	98882800	98717700	99369400	114.032	ZSCAN25,CYP3A5,CYP3A7,CYP3A4,SMURF1,KPNA7,ARPC1A,ARPC1B,PDAP1, BUD31,PTCD1,ATP5J2-PTCD1,CPSF4,ATP5J2,ZNF789,ZNF394,ZKSCAN5, FAM200A,ZNF655
15	45332100	45094600	45436600	109.189	C15orf43,SORD,DUOX2,DUOXA2,DUOXA1,DUOX1
7	81142700	81087600	81298600	107.683	-
10	31863100	31479100	31908500	105.903	ZEB1
18	66807000	66646700	66883000	104.605	CCDC102B
10	83601100	83597800	83761400	103.49	NRG3
4	167411000	167094000	167644000	102.882	-
15	35551700	35444900	35727400	102.53	DPH6
4	60872200	60814500	61356600	101.779	-
6	150686000	150637000	150738000	100.616	IYD
5	142116000	142074000	142194000	99.5731	FGF1,ARHGAP26
1	204823000	204680000	204872000	94.6291	NFASC
9	108572000	108412000	108755000	92.1087	TAL2,TMEM38B
2	104933000	104749000	105027000	91.4296	-
9	91155000	90913100	91201600	89.9796	SPIN1,NXNL2
9	12777200	12488900	12787600	89.8212	TYRP1,LURAP1L
15	52859500	52581800	52992200	88.5241	MYO5C,MYO5A,ARPP19,FAM214A
17	58512800	58075800	59174400	88.2992	HEATR6,CA4,USP32,C17orf64,APPBP2,PPM1D,BCAS3
11	75850500	75434700	75868100	87.6352	MOGAT2,DGAT2,UVRAG
21	21424100	21378700	21643900	87.5231	-
18	7330950	7259810	7374120	84.6369	-
6	76751700	76636000	77261200	84.398	IMPG1
22	25939100	25932200	26133300	84.3301	ADRBK2
15	48211200	48153900	48308500	83.8987	-
5	82679100	82488400	82790300	83.6618	XRCC4,VCAN
6	121627000	121082000	121788000	83.5179	TBC1D32,GJA1
20	53878400	53876100	54051800	82.7889	-
1	162116000	162002000	162228000	81.759	NOS1AP
8	18519100	18514800	18647300	81.1436	PSD3
3	97346000	96453200	97364600	80.9285	EPHA6
6	43624100	43419100	43688500	80.7957	DLK2,TJAP1,LRRC73,POLR1C,YIPF3,XPO5,POLH,GTPBP2,MAD2L1BP,RSPH9, MRPS18A
2	182591000	182360000	182839000	80.6618	ITGA4,CERKL,NEUROD1,SSFA2,PPP1R1C
7	78745700	78688500	78982700	80.4923	MAGI2
12	80298900	80117100	80435100	80.4656	PPP1R12A
7	137213000	137120000	137360000	79.5913	DGKI
2	216600000	216551000	216628000	79.4415	-
5	150045000	149992000	150386000	79.1225	SYNPO,MYOZ3,RBM22,DCTN4,SMIM3,IRGM,ZNF300
9	87890100	87820300	88099100	78.98	-
14	45251400	45194500	45849200	78.664	C14orf28,KLHL28,FAM179B,PRPF39,FKBP3,FANCM,MIS18BP1
6	138591000	138479000	138645000	78.5803	KIAA1244,PBOV1
11	129910000	129805000	130073000	78.4337	PRDM10,APLP2,ST14
10	93143500	93060300	93325000	78.2142	HECTD2
2	18352000	18284800	18517900	77.8332	KCNS3
2	194863000	194678000	196286000	77.7683	-
3	159281000	159244000	159477000	76.9608	IQCJ-SCHIP1
13	48977600	48726500	49291500	76.6641	ITM2B,RB1,LPAR6,RCBTB2,CYSLTR2

Table 3. Top hits for 3P-CLR run on the East Asian terminal branch, using Africans as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if they were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
5	117510000	117345000	117716000	249.186	-
3	58238900	58104900	58557500	221.826	FLNB,DNASE1L3,ABHD6,RPP14,PXK,PDHB,KCTD6,ACOX2,FAM107A
10	94874300	94840100	95720400	211.168	MYOF,CEP55,FFAR4,RBP4,PDE6C,FRA10AC1,LGI1,SLC35G1
2	72378700	72353700	73177300	210.631	CYP26B1,EXOC6B,SPR,EMX1,SFXN5
15	64166100	63692600	64339800	209.852	USP3,FBXL22,HERC1,DAPK2
4	42193900	41824100	42206800	207.344	TMEM33,DCAF4L1,SLC30A9,BEND4
11	25172400	25098500	25276200	200.866	LUZP2
1	234347000	234207000	234380000	180.97	SLC35F3
4	158638000	158481000	158740000	175.394	-
17	61536300	60912100	61549600	168.02	TANC2,CYB561
20	24793800	24570100	25037800	164.012	SYNDIG1,CST7,APMAP,ACSS1
4	86504400	86438300	86602900	160.869	ARHGAP24
10	56026900	55868800	56209000	157.583	PCDH15
1	75622900	75277800	76729300	155.376	LHX8,SLC44A5,ACADM,RABGGTB,MSH4,ASB17,ST6GALNAC3
7	112265000	112125000	112622000	147.988	LSMEM1,TMEM168,C7orf60
18	5299800	5203000	5314080	147.972	ZBTB14
4	135424000	134792000	135547000	146.919	-
7	1.09E+08	108741000	109226000	145.365	-
1	172931000	172670000	172950000	143.351	-
22	46760700	46594600	46831200	141.902	PPARA,CDPF1,PKDREJ,TTC38,GTSE1,TRMU,CELSR1
10	53363100	53226200	53440300	140.25	PRKG1
8	10836400	10467500	11126200	137.393	RP1L1,C8orf74,SOX7,PINX1,XKR6
3	102005000	101902000	102242000	135.116	ZPLD1
6	69974500	69524500	70359500	134.756	BAI3
2	26159900	25853900	26233500	133.01	KIF3C,DTNB
18	67572500	67533400	67877100	132.023	CD226,RTTN
3	104826000	104604000	104910000	130.642	-
2	17456900	17247000	17564600	126.069	-
12	93322200	92983200	93454700	125.576	C12orf74,PLEKHG7,EEA1
20	31604100	31304800	31614200	125.536	COMMD7,DNMT3B,MAPRE1,SUN5,BPIFB2
2	56096500	55929400	56198400	124.835	EFEMP1
9	107052000	106657000	107058000	124.094	SMC2
13	63542000	63261200	63971200	124.033	-
4	80074800	79878800	80250300	122.138	NAA11
2	109534000	108937000	109626000	121.83	LIMS1,RANBP2,CCDC138,EDAR,SULT1C4,GCC2
12	124021000	123925000	124275000	121.45	SNRNP35,RILPL1,TMED2,DDX55,EIF2B1,GTF2H3,TCTN2,ATP6V0A2,DNAH10
5	119814000	119666000	119870000	119.87	PRR16
16	67607200	66947800	68430200	119.175	ACD,PARD6A,ENKD1,C16orf86,GFOD2,RANBP10,TSNAXIP1,CENPT,THAP11, NUTF2,EDC4,NRN1L,PSKH1,CTRL,PSMB10,LCAT,SLC12A4,DPEP3,DPEP2,DUS2, DDX28,NFATC3,ESRP2,PLA2G15,SLC7A6,SLC7A6OS,PRMT7,SMPD3,CDH16,RRAD, FAM96B,CES2,CES3,CES4A,CBFB,C16orf70,B3GNT9,TRADD,FBXL8,HSF4,NOL3, KIAA0895L,EXOC3L1,E2F4,ELMO3,LRRC29,TMEM208,FHOD1,SLC9A5,PLEKHG4, KCTD19,LRRC36,TPPP3,ZDHHC1,HSD11B2,ATP6V0D1,AGRP,FAM65A,CTCF,RLTPR
10	97039700	96682100	97059000	118.353	CYP2C9,CYP2C8,C10orf129,PDLIM1
3	17197100	17188900	17897600	117.877	TBC1D5
4	28858500	28537900	28879000	116.991	-
12	55254800	55126200	55322100	113.527	MUCL1
12	103350000	103178000	103439000	112.494	PAH,ASCL1
5	10480500	10311500	10491100	110.472	MARCH6,ROPN1L
5	153541000	153053000	153736000	110.3	GRIA1,FAM114A2,MFAP3,GALNT10
3	37465800	36837400	37518800	109.917	TRANK1,EPM2AIP1,MLH1,LRRFIP2,GOLGA4,C3orf35,ITGA9
3	49165200	48396900	50415600	109.512	RBM5,SEMA3F,GNAT1,GNAI2,LSMEM2,IFRD2,HYAL3,NAT6,HYAL1,HYAL2,TUSC2, RASSF1,ZMYND10,NPRL2,CYB561D2,TMEM115,CACNA2D2,FBXW12,PLXNB1, CCDC51,TMA7,ATRIP,TREX1,SHISA5,PFKFB4,UCN2,COL7A1,UQCRC1,TMEM89, SLC26A6,CELSR3,NCKIPSD,IP6K2,PRKAR2A,SLC25A20,ARIH2OS,ARIH2,P4HTM, WDR6,DALRD3,NDUFAF3,IMPDH2,QRICH1,QARS,USP19,LAMB2,CCDC71,KLHDC8B, C3orf84,CCDC36,C3orf62,USP4,GPX1,RHOA,TCTA,AMT,NICN1,DAG1,BSN,APEH, MST1,RNF123,AMIGO3,GMPPB,IP6K1,CDHR4,FAM212A,UBA7,TRAIP,CAMKV, MST1R,MON1A,RBM6
17	48390900	48365600	48608400	109.226	XYLT2,MRPL27,EME1,LRRC59,ACSF2,CHAD,RSAD1,MYCBPAP

Table 4. Top hits for 3P-CLR run on the Eurasian ancestral branch, using Africans as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if they were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
17	58658700	58117400	59309500	541.05	HEATR6,CA4,USP32,C17orf64,APPBP2,PPM1D,BCAS3
10	22705100	22428900	22798800	535.104	EBLN1,COMMD3,COMMD3-BMI1,BMI1,SPAG6
17	62870600	62655400	63068200	511.926	SMURF2,LRRC37A3,GNA13
18	67572500	67533400	67881500	477.032	CD226,RTTN
2	22420000	22187700	22469200	461.001	-
1	230018000	229910000	230132000	444.349	-
7	99227800	98717700	99374100	435.299	ZSCAN25,CYP3A5,CYP3A7,CYP3A4,SMURF1,KPNA7,ARPC1A,ARPC1B,PDAP1, BUD31,PTCD1,ATP5J2-PTCD1,CPSF4,ATP5J2,ZNF789,ZNF394,ZKSCAN5, FAM200A,ZNF655
20	54054100	53877600	54056600	425.515	-
4	41834200	41823000	42195900	421.548	TMEM33,DCAF4L1,SLC30A9,BEND4
17	61536300	60942800	61549600	406.862	TANC2,CYB561
1	25592800	25517600	25869600	404.258	SYF2,C1orf63,RHD,TMEM50A,RHCE,TMEM57
10	93143500	93060300	93325000	398.359	HECTD2
9	90946300	90908100	91200000	397.639	SPIN1,NXNL2
3	97346000	96453200	97364600	395.762	EPHA6
6	3149410	3073260	3204820	395.04	RIPK1,BPHL,TUBB2A
6	10644100	10578900	10784300	389.704	GCNT2,C6orf52,PAK1IP1,TMEM14C,TMEM14B,SYCP2L,MAK
5	121498000	121486000	121640000	385.618	ZNF474
10	31863100	31479100	31908500	383.622	ZEB1
1	64483200	64340800	64538400	381.831	ROR1
10	66018600	65795200	66311900	380.584	-
4	13424000	13143500	13535500	378.487	RAB28
15	65012200	64308500	65208800	376.663	DAPK2,FAM96A,SNX1,SNX22,PPIB,CSNK1G1,KIAA0101,TRIP4,ZNF609,OAZ2, RBPMS2,PIF1,PLEKHO2,ANKDD1A
4	33576600	33301000	33643000	373.953	-
3	188751000	188647000	188859000	371.551	TPRG1
4	177625000	177608000	177889000	370.287	VEGFC
16	61316300	61123100	61456600	369.203	-
2	73545700	73488400	74117400	368.52	FBXO41,EGR4,ALMS1,NAT8,TPRKB,DUSP11,C2orf78,STAMBP
13	49170200	48726500	49293400	366.868	ITM2B,RB1,LPAR6,RCBTB2,CYSLTR2
11	19609000	19591300	19731200	366.495	NAV2
12	111447000	111331000	111655000	365.613	CCDC63,MYL2,CUX2
7	142800000	142639000	143022000	364.893	OR9A2,OR6V1,PIP,TAS2R39,TAS2R40,GSTK1,TMEM139,CASP2,CLCN1,KEL
16	47937600	33585200	48471700	362.804	SHCBP1,VPS35,ORC6,MYLK3,C16orf87,GPT2,DNAJA2,NETO2,ITFG1,PHKB, ABCC12,ABCC11,LONP2,SIAH1
20	35293200	35003300	35596300	362.528	DLGAP4,MYL9,TGIF2,TGIF2-C20orf24,C20orf24,SLA2,NDRG3,DSN1,SOGA1, TLDC2,SAMHD1
7	30270500	30178800	30471600	361.081	MTURN,ZNRF2,NOD1
15	28362400	28295700	28630200	360.613	OCA2,HERC2
8	30625000	30515900	30891400	360.176	GSR,PPP2CB,TEX15,PURG,WRN
5	159211000	159139000	159271000	359.48	-
14	90449500	90301800	90531400	359.187	EFCAB11,TDP1,KCNK13
11	39699100	39604100	39937900	357.303	-
5	11741300	11640500	11850200	356.6	CTNND2
17	27227600	26844500	27341600	354.11	RPL23A,TLCD1,NEK8,TRAF4,FAM222B,ERAL1,FLOT2,DHRS13,PHF12,FOXN1, UNC119,PIPOX,PIGS,ALDOC,SEZ6,SPAG5,KIAA0100,SDF2,SUPT6H,PROCA1,RAB34

Table 5. Top hits for 3P-CLR run on the ancestral branch to Eurasians and Africans, using archaic humans as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if they were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
21	34916200	34737300	35222100	852.869	IFNGR2,TMEM50B,DNAJC28,GART,SON,DONSON,CRYZL1,ITSN1
17	5z6595700	56373200	57404800	832.783	BZRAP1,SUPT4H1,RNF43,HSF5,MTMR4,SEPT4,C17orf47,TEX14,RAD51C,PPM1E,TRIM37,SKA2,PRR11,SMG8,GDPD1
12	79919700	79756800	80109400	827.892	SYT1,PAWR
14	29635300	29222200	29696100	823.504	FOXG1
14	71790600	71658900	72283600	822.057	SIPA1L1
12	116589000	116366000	116760000	814.828	MED13L
2	37989300	37917400	38021500	813.181	CDC42EP3
3	36941700	36836900	37517500	805.571	TRANK1,EPM2AIP1,MLH1,LRRFIP2,GOLGA4,C3orf35,ITGA9
4	146155000	145355000	146222000	803.332	HHIP,ANAPC10,ABCE1,OTUD4
5	86911000	86463700	87101400	802.795	RASA1,CCNH
2	156468000	155639000	156767000	800.114	KCNJ3
1	213498000	213145000	213561000	798.967	VASH2,ANGEL2,RPS6KC1
7	107229000	106619000	107308000	787.005	PRKAR2B,HBP1,COG5,GPR22,DUS4L,BCAP29,SLC26A4
17	61237300	60906000	61544500	784.538	TANC2,CYB561
7	121700000	121620000	122369000	784.297	PTPRZ1,AASS,FEZF1,CADPS2,RNF133,RNF148
21	36769600	36689700	36842100	775.26	RUNX1
5	93214400	92677500	93645500	773.258	NR2F1,FAM172A,POU5F2,KIAA0825
15	49269100	49247500	50036600	771.33	SECISBP2L,COPS2,GALK2,FAM227B,FGF7,DTWD1,SHC4
13	96782600	96180700	97420500	767.375	CLDN10,DZIP1,DNAJC3,UGGT2,HS6ST3
4	13346700	13140300	13533100	762.438	RAB28
1	176411000	175890000	176437000	762.017	RFWD2,PAPPA2
3	50576000	50177100	51929300	760.176	SEMA3F,GNAT1,GNAI2,LSMEM2,IFRD2,HYAL3,NAT6,HYAL1,HYAL2,TUSC2,RASSF1,ZMYND10,NPRL2,CYB561D2,TMEM115,CACNA2D2,C3orf18,HEMK1,CISH,MAPKAPK3,DOCK3,MANF,RBM15B,RAD54L2,TEX264,GRM2,IQCF6,IQCF3,IQCF2,IQCF5,IQCF1
9	125562000	125505000	126059000	755.867	ZBTB26,RABGAP1,GPR21,STRBP,OR1L6,OR5C1,PDCL,OR1K1,RC3H2,ZBTB6
7	99167000	98722100	99375300	754.442	CYP3A5,ZSCAN25,CYP3A7,CYP3A4,SMURF1,KPNA7,ARPC1A,ARPC1B,PDAP1,BUD31,PTCD1,ATP5J2-PTCD1,CPSF4,ATP5J2,ZNF789,ZNF394,ZKSCAN5,FAM200A,ZNF655
4	46634200	46361400	47004100	753.008	GABRA2,COX7B2,GABRA4,GABRB1
22	40521100	40350900	41080200	750.394	GRAP2,FAM83F,TNRC6B,ADSL,SGSM3,MKL1,MCHR1
13	44935700	44887400	45237200	750.165	SERP2,TSC22D1
2	73506800	73482800	74054300	747.479	FBXO41,EGR4,ALMS1,NAT8,TPRKB,DUSP11,C2orf78
5	89579400	89008500	89654700	746.591	-
19	19313800	19100800	19788800	746.184	SUGP2,ARMC6,SLC25A42,TMEM161A,MEF2BNB-MEF2B,MEF2B,MEF2BNB,RFXANK,NR2C2AP,NCAN,HAPLN4,TM6SF2,SUGP1,MAU2,GATAD2A,TSSK6,NDUFA13,YJEFN3,CILP2,PBX4,LPAR2,GMIP,ATP13A1,ZNF101
3	110709000	110513000	110932000	746.168	PVRL3
10	119809000	119698000	119995000	745.865	RAB11FIP2
2	63889800	62767900	64394800	745.126	EHBP1,OTX1,WDPCP,MDH1,UGP2,VPS54,PELI1
5	50148300	44575900	50411900	742.91	MRPS30,HCN1,EMB,PARP8
1	66833100	66772600	66952600	741.972	PDE4B
2	145109000	144689000	145219000	740.09	GTDC1,ZEB2
11	65212900	65129200	65379700	738.826	KCNK7,MAP3K11,SLC25A45,FRMD8,SCYL1,LTBP3,SSSCA1,FAM89B,EHBP1L1
4	22923300	22826000	23196800	737.716	-
14	76101700	76054700	76450300	737.296	FLVCR2,TTLL5,C14orf1,IFT43,TGFB3
2	201156000	200636000	201355000	736.589	C2orf69,TYW5,C2orf47,SPATS2L,KCTD18
10	74730800	74007800	75399900	735.921	DDIT4,DNAJB12,MICU1,MCU,OIT3,PLA2G12B,P4HA1,NUDT13,ECD,FAM149B1,DNAJC9,MRPS16,TTC18,ANXA7,MSS51,PPP3CB,USP54,MYOZ1
9	88868400	88695200	89027700	733.578	GOLM1,C9orf153,ISCA1,ZCCHC6
1	98272100	97949300	98565500	731.566	DPYD
8	79211300	78656900	79554100	729.158	PKIA
12	97223900	96828100	97424100	728.739	NEDD1
11	111391000	111309000	111981000	728.464	POU2AF1,BTG4,C11orf88,LAYN,SIK2,PPP2R1B,ALG9,FDXACB1,C11orf1,CRYAB,HSPB2-C11orf52,C11orf52,DIXDC1,DLAT,PIH1D2,C11orf57,TIMM8B,SDHD
3	147223000	146941000	147360000	726.503	ZIC4,ZIC1
15	75631100	75458200	76012400	724.9	C15orf39,GOLGA6C,GOLGA6D,COMMD4,NEIL1,MAN2C1,SIN3A,PTPN9,SNUPN,IMP3,SNX33,CSPG4
1	21083000	21012100	21629100	720.302	KIF17,SH2D5,HP1BP3,EIF4G3,ECE1
12	111254000	110248000	111324000	719.87	ANKRD13A,C12orf76,IFT81,ATP2A2,ANAPC7,ARPC3,GPN3,FAM216A,VPS29,RAD9B,PPTC7,TCTN1,HVCN1,PPP1CC,CCDC63,TRPV4,GLTP,TCHP,GIT2

Table 6. Overlap between GWAS catalog and catalog of modern human-specific high-frequency changes in the top modern human selected regions. Chr = chromosome. Pos = position (hg19). ID = SNP rs ID. Hum = Present-day human major allele. Anc = Human-Chimpanzee ancestor allele. Arch = Archaic human allele states (Altai Neanderthal, Denisova) where H=human-like allele and A=ancestral allele. Freq = present-day human derived frequency. Cons = consequence. C = C-score. PubMed = PubMed article ID for GWAS study.

Chr	Pos	ID	Hum	Anc	Arch	Freq	Gene	Cons	C	GWAS trait	PubMed
2	64279606	rs10171434	C	T	A/A,A/A	0.92	NA	regulatory	8.358	Suicide attempts in bipolar disorder	21041247
2	64279606	rs10171434	C	T	A/A,A/A	0.92	NA	regulatory	8.358	Urinary metabolites	21572414
2	144783214	rs16823411	T	C	A/A,A/A	0.93	GTDC1	intron	4.112	Body mass index	21701565
2	144783214	rs16823411	T	C	A/A,A/A	0.93	GTDC1	intron	4.112	Body mass index	21701565
2	145213638	rs731108	G	C	A/A,H/H	0.92	ZEB2	regulatory	10.31	Renal cell carcinoma	23184150
2	156506516	rs4407211	C	T	A/A,A/A	0.92	NA	intergenic	1.348	Alcohol consumption	23953852
3	51142359	rs4286453	T	C	A/A,A/A	0.91	DOCK3	intron	4.96	Multiple complex diseases	17554300
3	51824167	rs6796373	G	C	A/A,A/A	0.94	NA	intergenic	1.381	Response to taxane treatment (placlitaxel)	23006423
3	147200492	rs9876193	G	A	H/H,A/A	0.95	ZIC1	intron,nc	6.856	Type 2 diabetes	17463246
4	13325741	rs2867467	G	C	A/A,A/A	0.91	NA	intergenic	0.476	Obesity (extreme)	21935397
4	13328373	rs6842438	T	C	A/A,A/A	0.92	NA	intergenic	5.241	Obesity (extreme)	21935397
4	13330095	rs10019897	C	T	A/A,A/A	0.92	NA	upstream	1.472	Multiple complex diseases	17554300
4	13330095	rs10019897	C	T	A/A,A/A	0.92	NA	upstream	1.472	Obesity (extreme)	21935397
4	13333413	rs9996364	A	G	A/A,A/A	0.92	HSP90AB2P	upstream	5.865	Obesity (extreme)	21935397
4	13338465	rs11945340	C	T	A/A,A/A	0.92	HSP90AB2P	non coding exon	12.04	Obesity (extreme)	21935397
4	13340249	rs6839621	T	C	A/A,A/A	0.92	HSP90AB2P	non coding exon	0.074	Obesity (extreme)	21935397
4	13346602	rs11930614	C	T	A/A,A/A	0.92	NA	intergenic	0.587	Obesity (extreme)	21935397
4	13350973	rs10021881	T	C	A/A,A/A	0.92	NA	regulatory	3.032	Obesity (extreme)	21935397
4	13356393	rs16888596	G	A	A/A,A/A	0.94	NA	intergenic	2.344	Obesity (extreme)	21935397
4	13357274	rs11732938	A	G	A/A,A/A	0.94	NA	intergenic	15.45	Obesity (extreme)	21935397
4	13360622	rs11947529	T	A	A/A,A/A	0.93	RAB28	downstream	4.356	Obesity (extreme)	21935397
4	13363958	rs12331157	A	G	A/A,A/A	0.97	RAB28	intron	1.3	Obesity (extreme)	21935397
4	13363974	rs12332023	C	T	A/A,A/A	0.97	RAB28	intron	0.75	Obesity (extreme)	21935397
4	13366481	rs7673680	C	T	A/A,A/A	0.93	RAB28	downstream	4.16	Obesity (extreme)	21935397
4	13370308	rs10003958	T	C	A/A,A/A	0.93	RAB28	regulatory	16.58	Obesity (extreme)	21935397
4	13373583	rs9999851	C	T	A/A,A/A	0.97	RAB28	intron	1.305	Obesity (extreme)	21935397
4	13374462	rs9291610	G	A	A/A,A/A	0.93	RAB28	intron	3.264	Obesity (extreme)	21935397
4	13393897	rs9998914	A	T	A/A,A/A	0.96	RAB28	intron	0.414	Obesity (extreme)	21935397
4	13403855	rs11943295	G	A	A/A,A/A	0.94	RAB28	intron	1.702	Multiple complex diseases	17554300
4	13403855	rs11943295	G	A	A/A,A/A	0.94	RAB28	intron	1.702	Obesity (extreme)	21935397
4	13403998	rs11943330	G	A	A/A,A/A	0.93	RAB28	intron	3.295	Obesity (extreme)	21935397
4	13404130	rs7677336	G	T	A/A,A/A	0.94	RAB28	intron	0.752	Obesity (extreme)	21935397
4	13404717	rs7673732	A	C	A/A,A/A	0.93	RAB28	intron	0.702	Obesity (extreme)	21935397
4	13440031	rs11737264	C	G	A/A,A/A	0.93	RAB28	intron	1.159	Obesity (extreme)	21935397
4	13440271	rs11737360	C	T	A/A,A/A	0.94	RAB28	intron	2.745	Obesity (extreme)	21935397
4	13449532	rs16888654	A	C	A/A,A/A	0.94	RAB28	intron	0.46	Obesity (extreme)	21935397
4	13452022	rs16888661	C	A	A/A,A/A	0.91	RAB28	intron	5.359	Obesity (extreme)	21935397
4	13463991	rs11933841	T	C	A/A,A/A	0.93	RAB28	intron	4.193	Obesity (extreme)	21935397
4	13465710	rs11947665	T	A	A/A,A/A	0.93	RAB28	intron	4.41	Obesity (extreme)	21935397
4	23095293	rs6825402	C	T	A/A,A/A	0.96	NA	intergenic	2.599	Multiple complex diseases	17554300
5	45393261	rs6874279	G	A	A/A,A/A	0.93	HCN1	intron	1.47	Alcohol dependence	20201924
5	45393261	rs6874279	G	A	A/A,A/A	0.93	HCN1	intron	1.47	Alcoholism	pha002891
5	89540468	rs2935504	C	T	A/A,A/A	0.97	RP11-61G23.1	non coding exon	4.52	Multiple complex diseases	17554300
7	106720932	rs12154324	G	A	A/A,A/A	0.93	NA	regulatory	5.411	Multiple complex diseases	17554300
13	44978167	rs9525954	C	A	A/A,A/A	0.95	RP11-269C23.3	intron	2.731	Type 2 diabetes	17463246
13	45034814	rs9533862	G	C	A/A,A/A	0.93	FILIP1LP1	intron	2.026	Suicide attempts in bipolar disorder	21041247
13	45055091	rs17065868	T	C	A/A,A/A	0.92	FILIP1LP1	intron	3.214	Antineutrophil cytoplasmic antibody-associated vasculitis	22808956

640 Figures

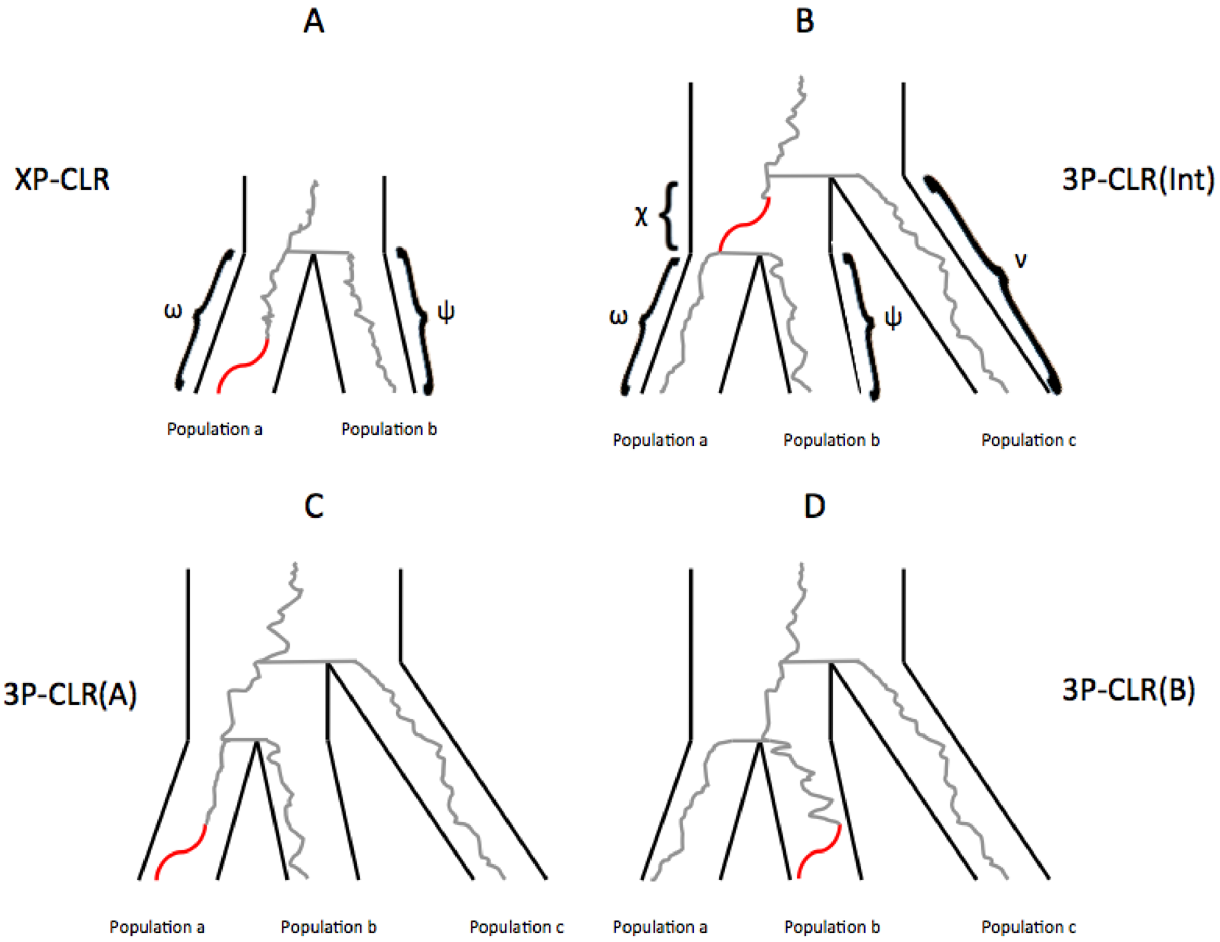

Figure 1. Schematic tree of selective sweeps detected by XP-CLR and 3P-CLR. While XP-CLR can only use two populations (an outgroup and a test) to detect selection (panel A), 3P-CLR can detect selection in the ancestral branch of two populations (3P-CLR(Int), panel B) or on the branches specific to each population (3P-CLR(A) and 3P-CLR(B), panels C and D, respectively). The greek letters denote the known drift times for each branch of the population tree.

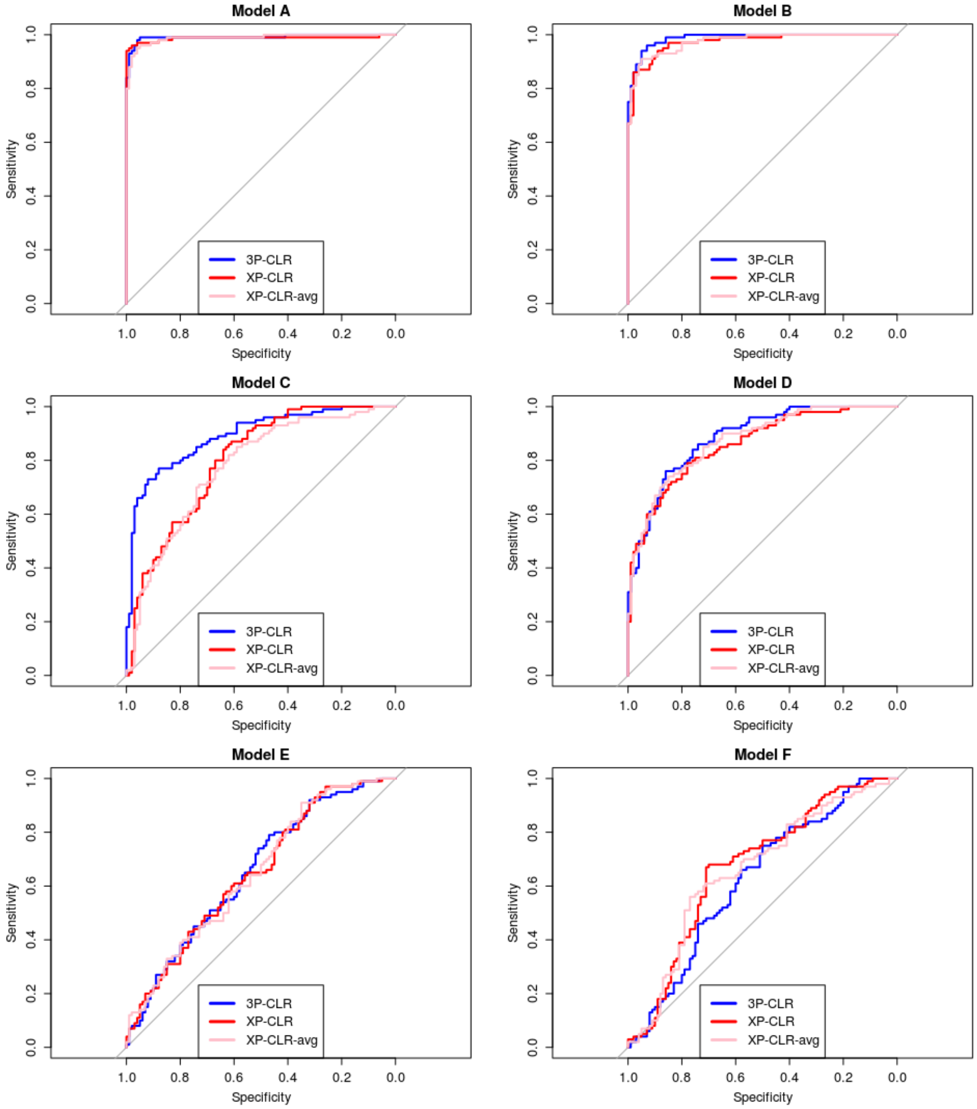

Figure 2. ROC curves for performance of 3P-CLR(Int) and two variants of XP-CLR in detecting selective sweeps that occurred before the split of two populations a and b, under different demographic models. In this case, the outgroup panel from population c contained 10 haploid genomes. The two sister population panels (from a and b) have 100 haploid genomes each.

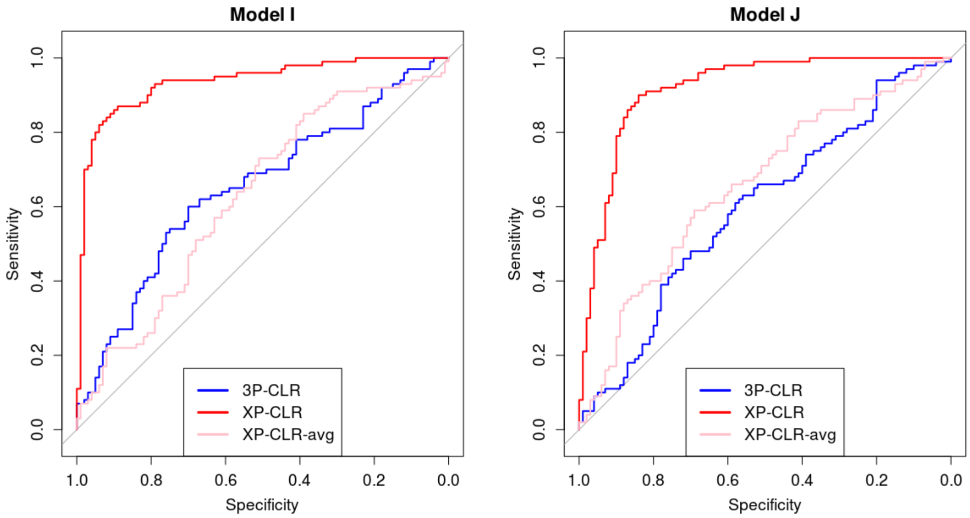

Figure 3. 3P-CLR(Int) is tailored to detect selective events that happened before the split t_{ab}, so it is largely insensitive to sweeps that occurred after the split. ROC curves show performance of 3P-CLR(Int) and two variants of XP-CLR for models where selection occurred in population a after its split from b.

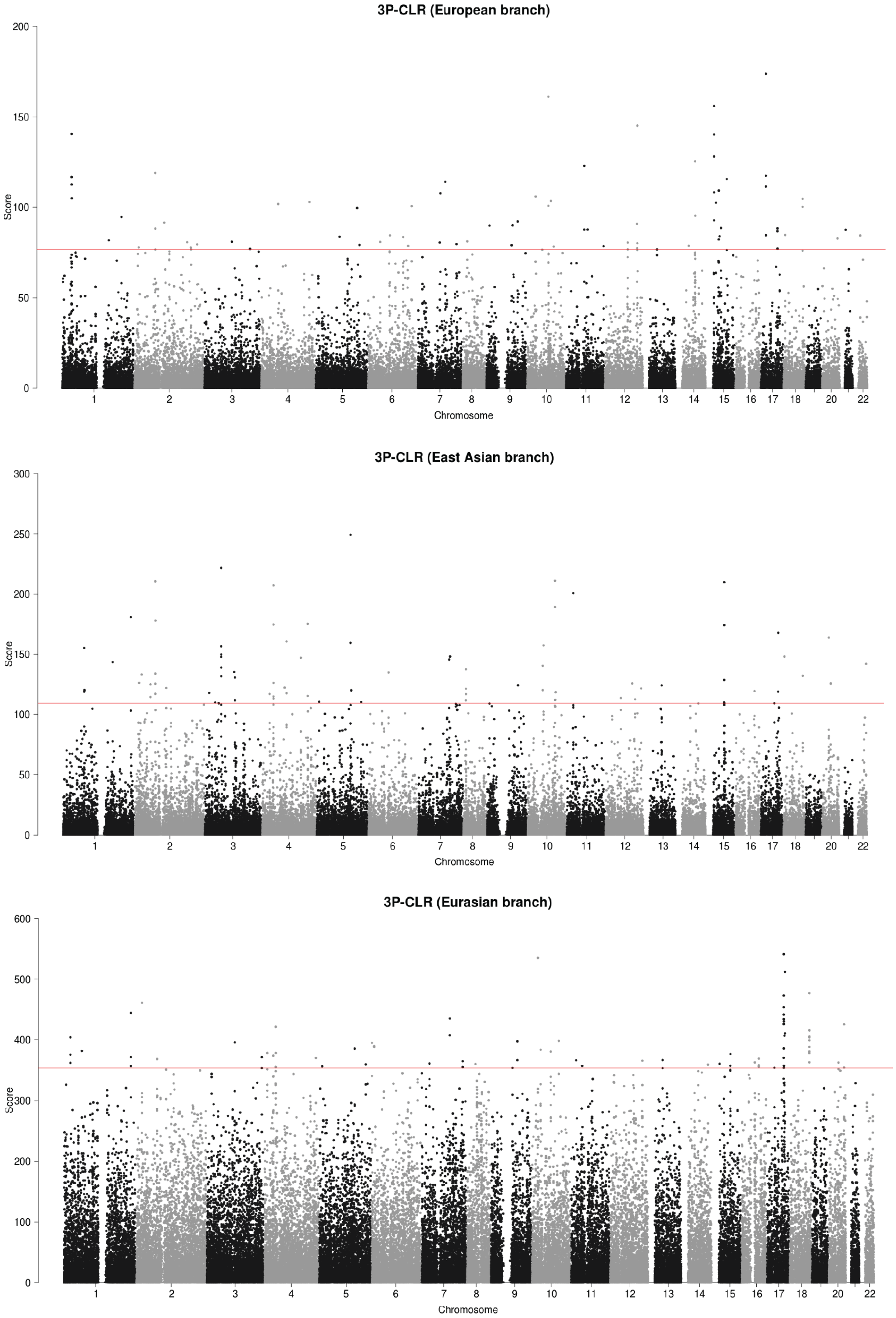

Figure 4. 3P-CLR scan of Europeans (upper panel), East Asians (middle panel) and the ancestral population to Europeans and East Asians (lower panel), using Africans as the outgroup in all 3 cases. The red line denotes the 99.9% quantile cutoff.

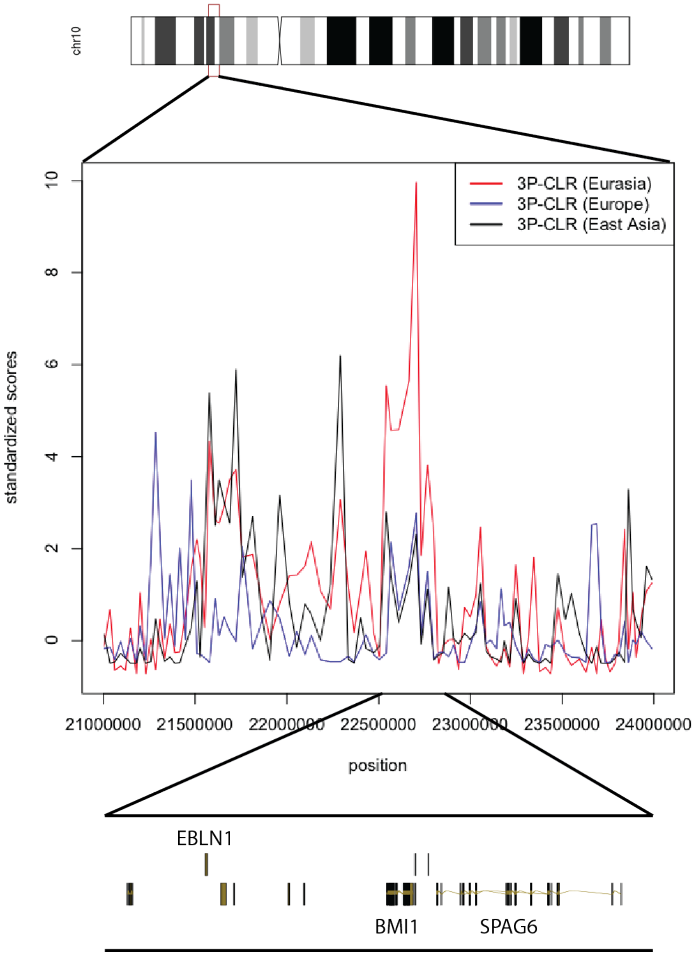

Figure 5. 3P-CLR scan of Europeans (blue), East Asians (black) and the ancestral Eurasian population (red) reveals the region containing genes SPAG6 and BMI1 to be candidates for selection in the ancestral population. To make a fair comparison, all 3P-CLR scores were standardized by substracting the chromosome-wide mean from each window and dividing the resulting score by the chromosome-wide standard deviation. The image was built using the GenomeGraphs package in Bioconductor.

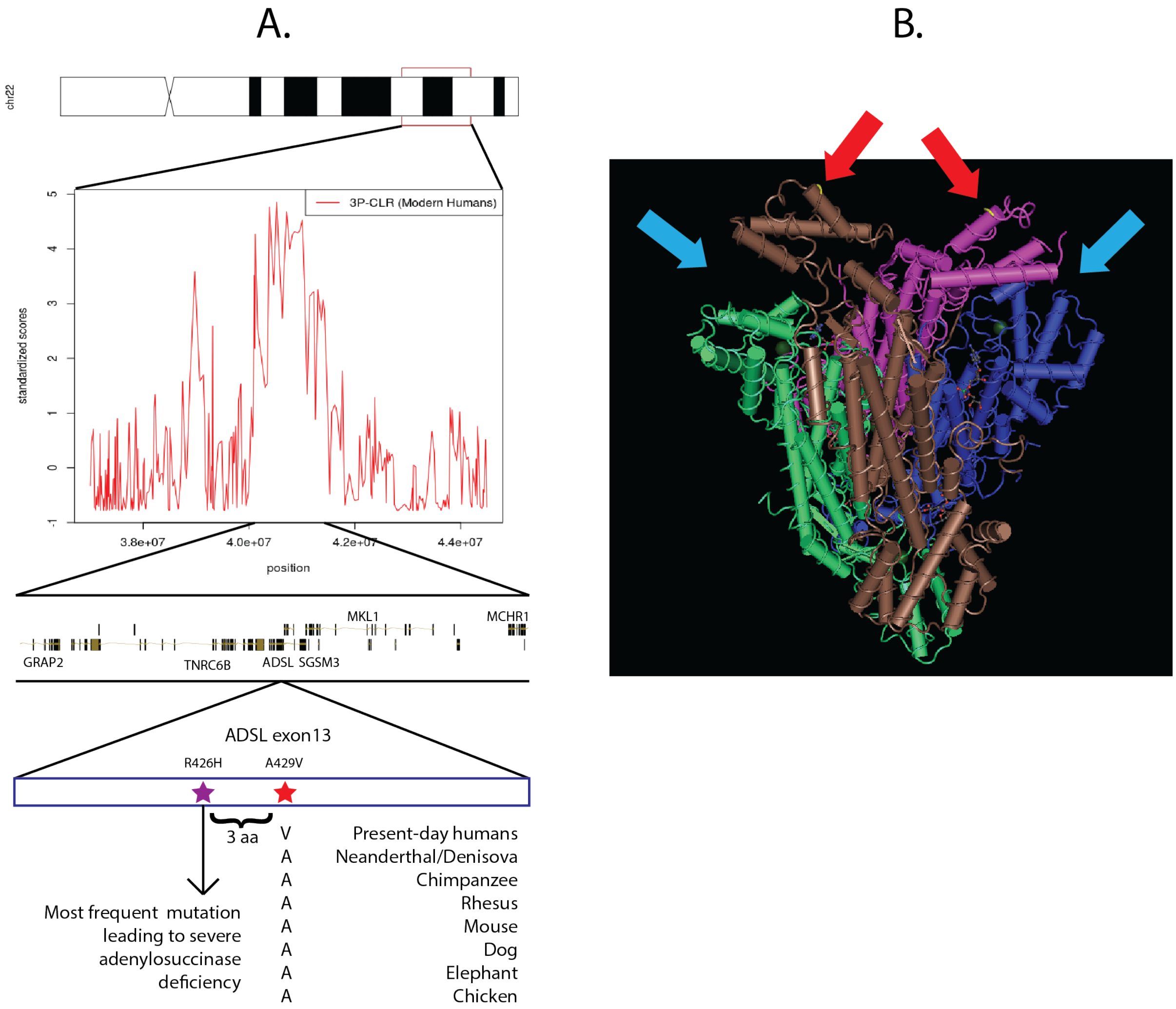

Figure 6. ADSL is a candidate for selection in the modern human lineage, after the split from Neanderthal and Denisova. A) One of the top-scoring regions when running 3P-CLR on the modern human lineage contains genes TNRC6B, ADSL, MKL1, MCHR1, SGSM3 and GRAP2. The most disruptive nonsynonymous modern-human-specific change in the entire list of top regions is in an exon of ADSL and is fixed derived in all present-day humans but ancestral in archaic humans. It is highly conserved accross tetrapods and lies only 3 residues away from the most common mutation leading to severe adenylosuccinase deficiency. B) The gene codes for a tetrameric protein. The mutation is in the C-terminal domain of each tetramer (red arrows), which are near the active sites (light blue arrows). Scores in panel A were standardized using the chromosome-wide mean and standard deviation. Vertebrate alignments were obtained from the UCSC genome browser (Vertebrate Multiz Alignment and Conservation track) and the image was built using the GenomeGraphs package in Bioconductor and Cn3D.

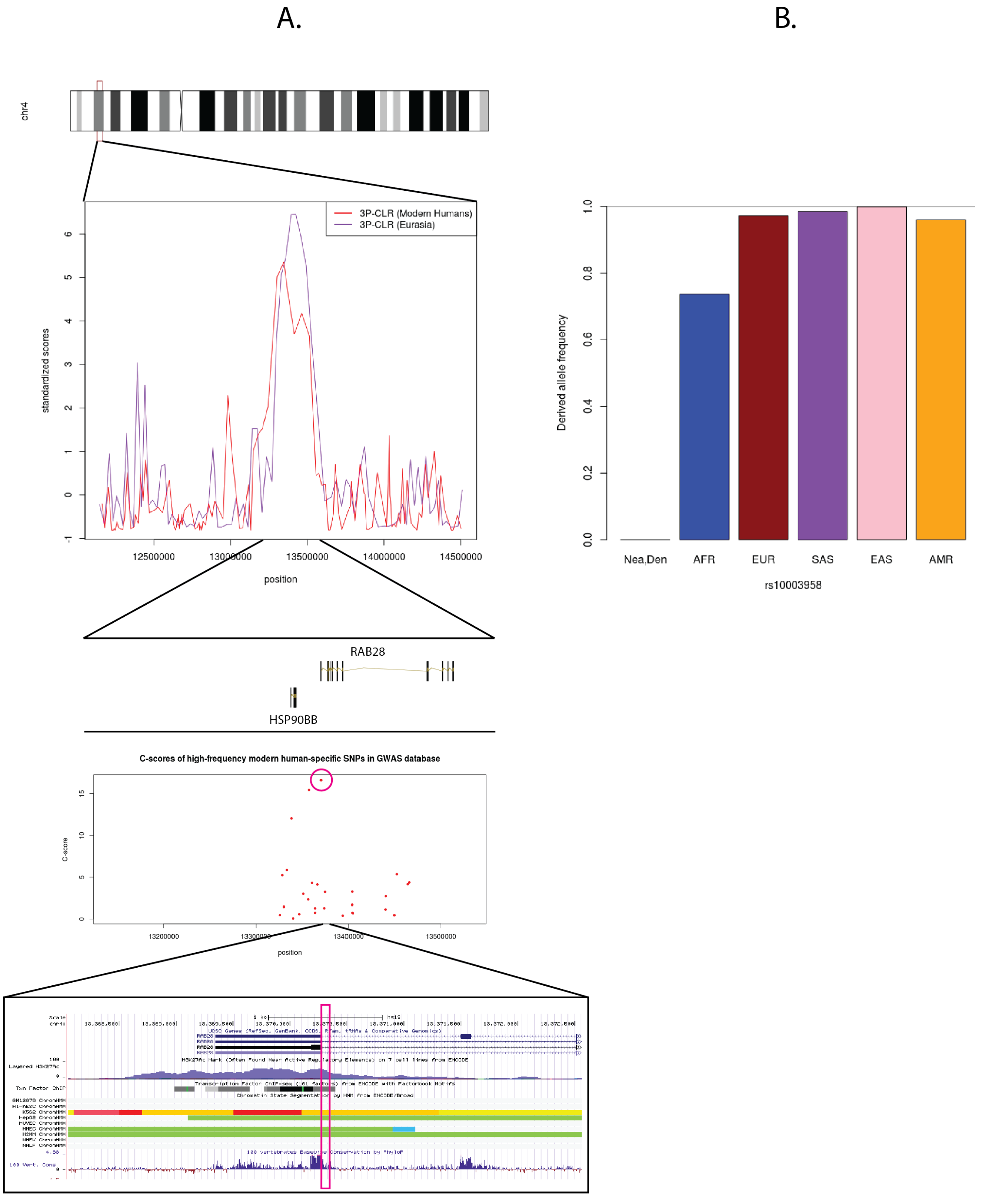

Figure 7. RAB28 is a candidate for selection in both the Eurasian and the modern human ancestral lineages. A) The gene lies in the middle of a 3P-CLR peak for both ancestral populations. The putatively selected region also contains several SNPs that are significantly associated with obesity and that are high-frequency derived in present-day humans (> 93%) but ancestral in archaic humans (red dots). The SNP with the highest C-score among these (rs10003958, pink circle) lies in a highly conserved strong enhancer region adjacent to the last exon of the gene. Color code for ChromHMM segmentation regions in UCSC genome browser: red = promoter, orange = strong enhancer, yellow = weak enhancer, green = weak transcription, blue = insulator. The image was built using the GenomeGraphs package in Bioconductor and the UCSC Genome Browser. B) Derived allele frequencies of SNP rs10003958 in the Denisova and Neanderthal genomes, and in different 1000 Genomes continental populations. AFR = Africans. AMR = Native Americans. SAS = South Asians. EUR = Europeans. EAS = East Asians.

641 Supplementary Figures

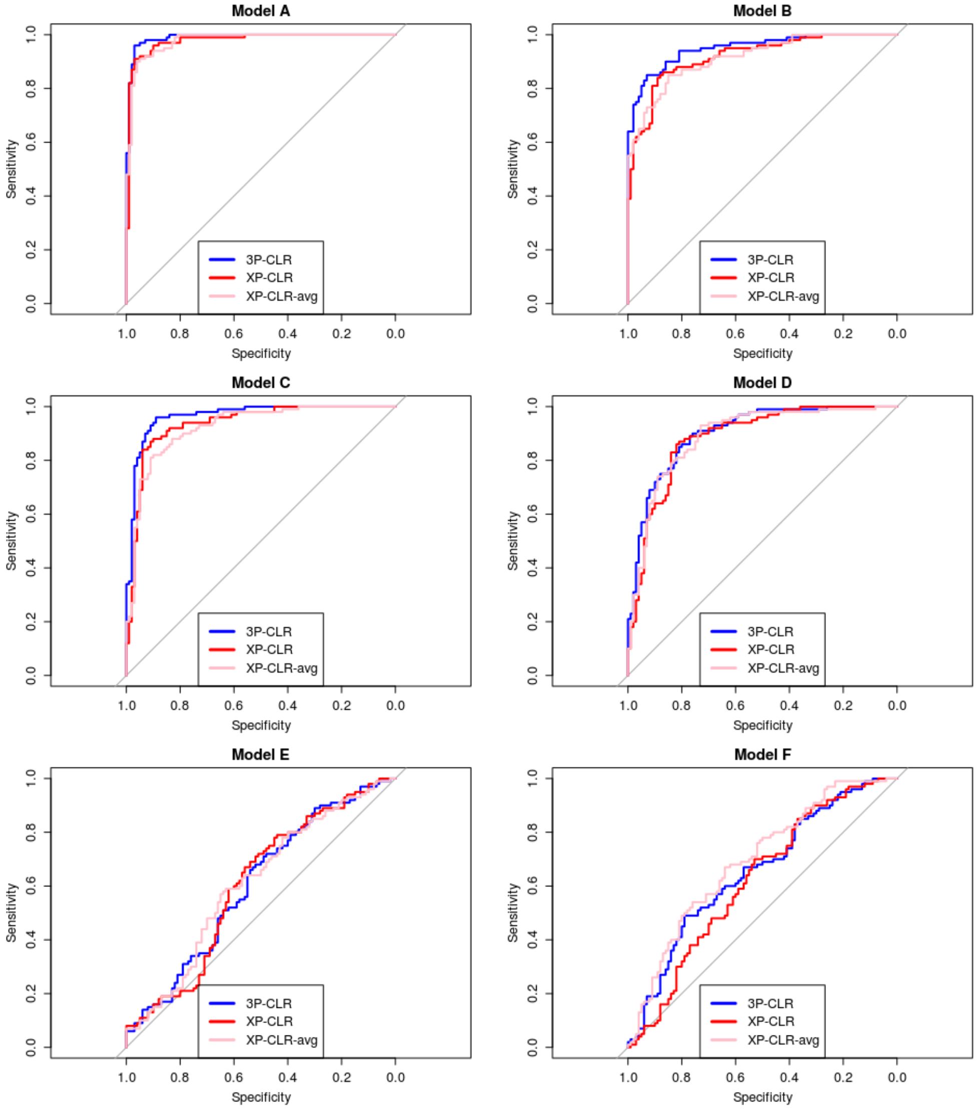

Figure S1. ROC curves for performance of 3P-CLR(Int) and two variants of XP-CLR in detecting selective sweeps that occurred before the split of two populations a and b, under different demographic models. In this case, the outgroup panel from population c contained 100 haploid genomes. The two sister population panels (from a and b) have 100 haploid genomes each.

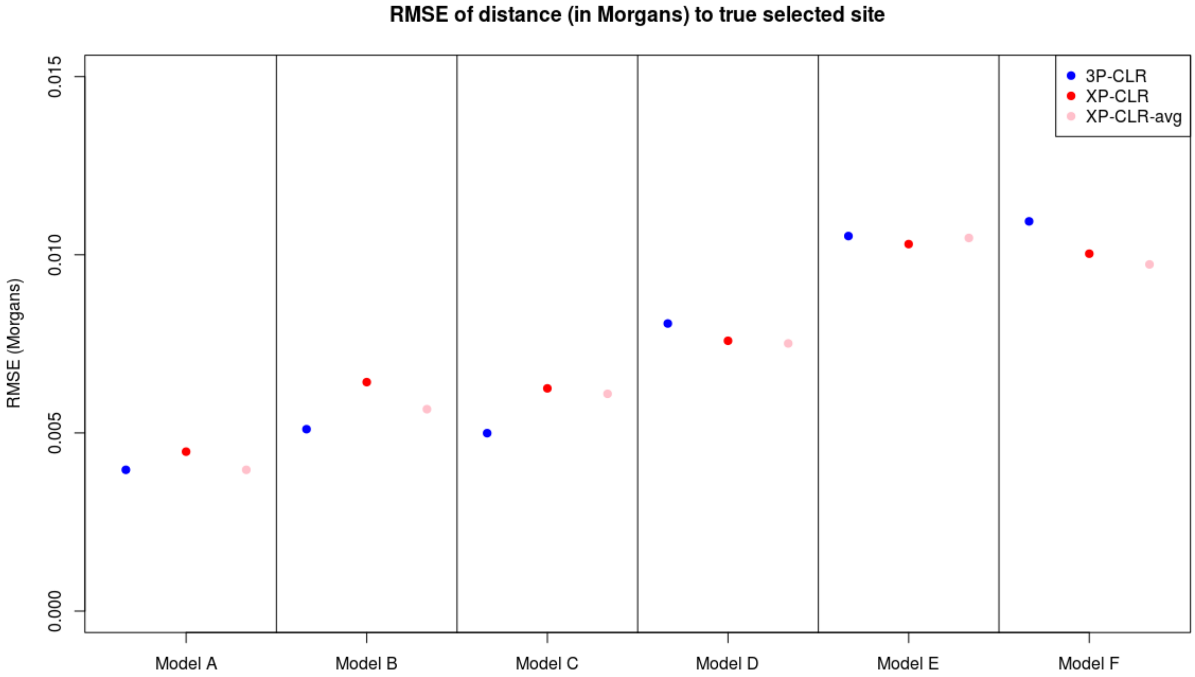

Figure S2. Root-mean squared error for the location of the sweep inferred by 3P-CLR(Int) and two variants of XP-CLR under different demographic scenarios. In this case, the outgroup panel from population c contained 10 haploid genomes and the two sister population panels (from a and b) have 100 haploid genomes each.

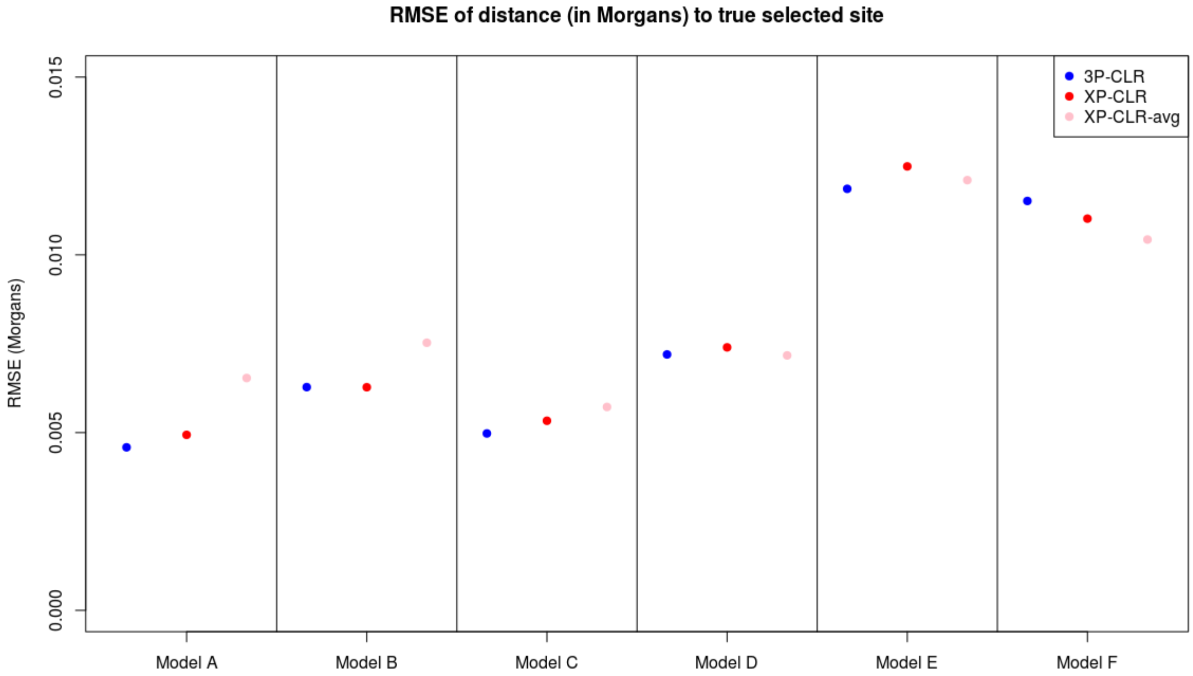

Figure S3. Root-mean squared error for the location of the sweep inferred by 3P-CLR(Int) and two variants of XP-CLR under different demographic scenarios. In this case, the outgroup panel from population c contained 100 haploid genomes and the two sister population panels (from a and b) have 100 haploid genomes each.

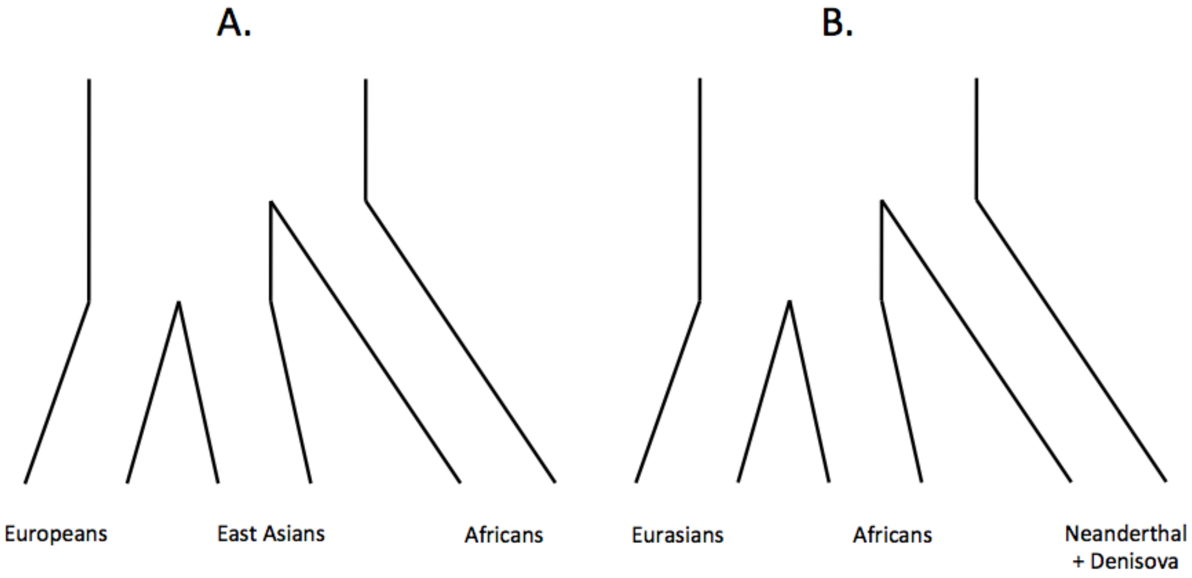

Figure S4. A. Three-population tree separating Europeans, East Asians and Africans. B. Three-population tree separating Eurasians, Africans and archaic humans (Neanderthal+Denisova).

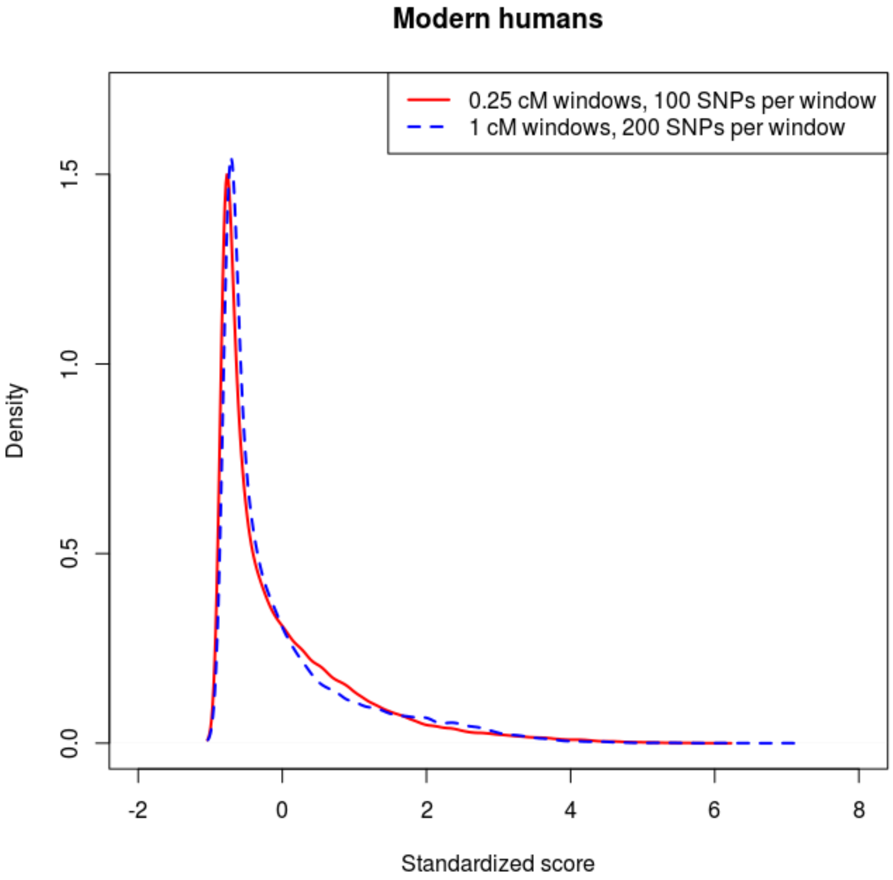

Figure S5. Comparison of 3P-CLR on the modern human ancestral branch under different window sizes and central SNP spacing. The red density is the density of standardized scores for 3P-CLR run using 0.25 cM windows, 100 SNPs per window and a spacing of 20 SNPs between each central SNP. The blue dashed density is the density of standardized scores for 3P-CLR run using 1 cM windows, 200 SNPs per window and a spacing of 80 SNPs between each central SNP.

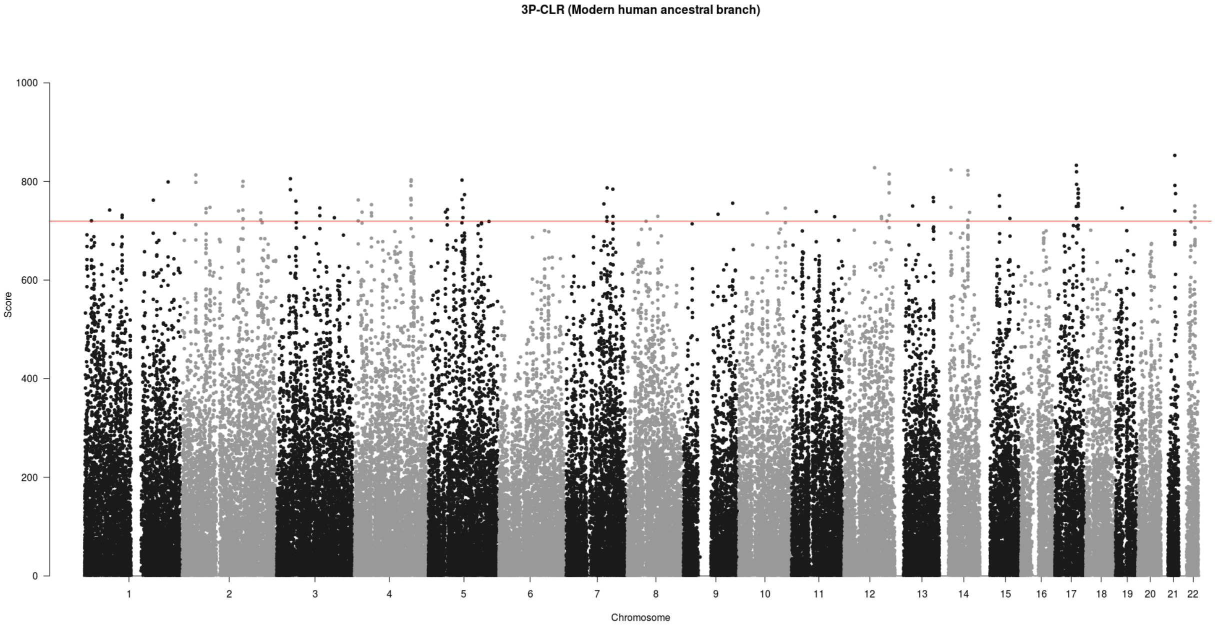

Figure S6. 3P-CLR scan of the ancestral branch to Africans and Eurasians, using the Denisovan and Neanderthal genomes as the outgroup. The red line denotes the 99.9% quantile cutoff.

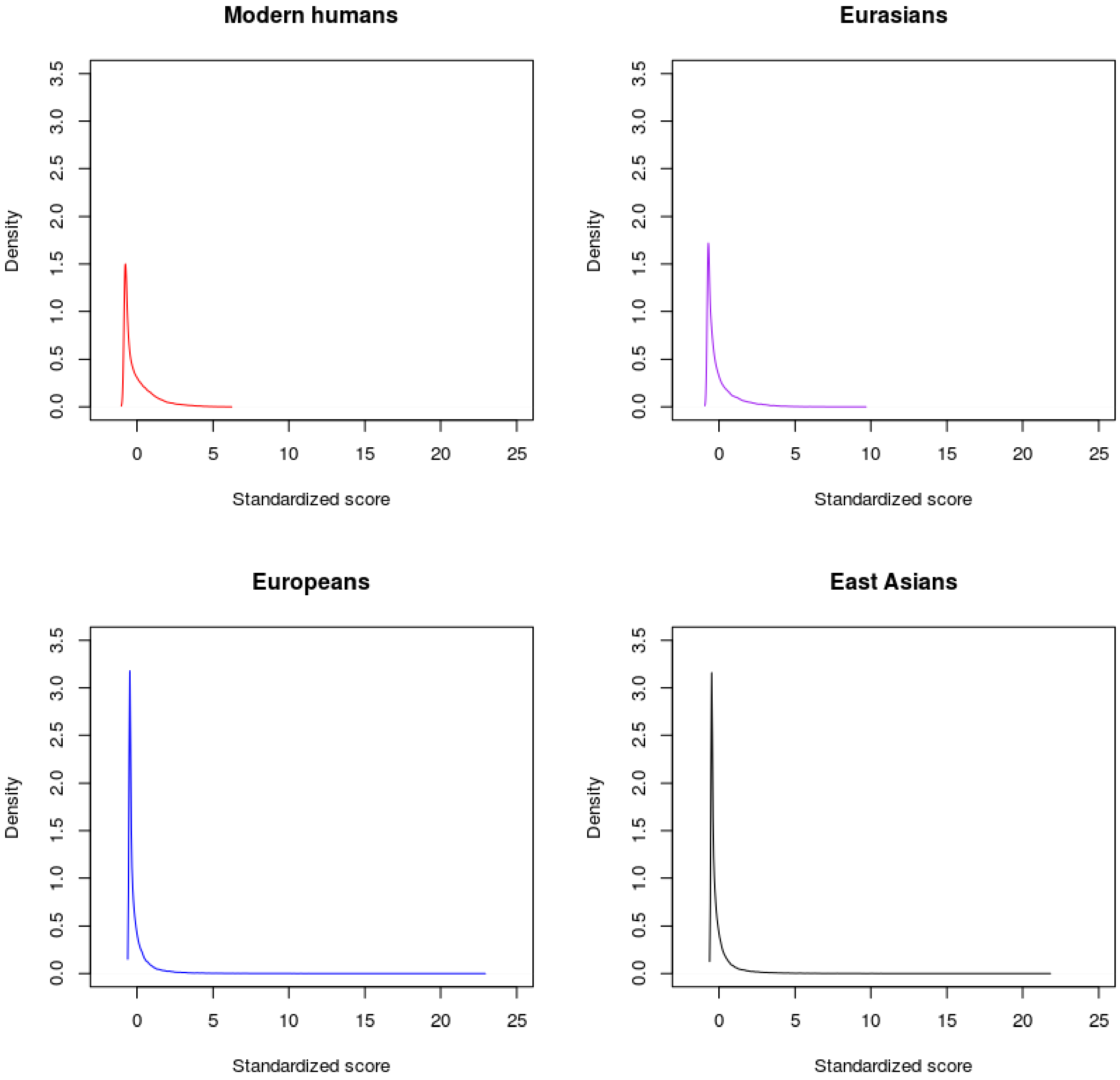

Figure S7. Genome-wide densities of each of the 3P-CLR scores described in this work. The distributions of scores testing for recent selection (Europeans and East Asians) have much longer tails than the distributions of scores testing for more ancient selection (Modern Humans and Eurasians). All scores were standardized using their genome-wide means and standard deviations.