Detecting Signatures of Positive Selection along Defined Branches of a Population Tree Using LSD

Pablo Librado*[,1,2] and Ludovic Orlando[1,2]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark

[2]Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, CNRS UMR 5288, Université de Toulouse, Université Paul Sabatier, Toulouse, France

*Corresponding author: E-mail: plibrado@snm.ku.dk.

Associate editor: Ryan Hernandez

Abstract

Identifying the genomic basis underlying local adaptation is paramount to evolutionary biology, and bears many applications in the fields of conservation biology, crop, and animal breeding, as well as personalized medicine. Although many approaches have been developed to detect signatures of positive selection within single populations and population pairs, the increasing wealth of high-throughput sequencing data requires improved methods capable of handling multiple, and ideally large number of, populations in a single analysis. In this study, we introduce LSD (levels of exclusively shared differences), a fast and flexible framework to perform genome-wide selection scans, along the internal and external branches of a given population tree. We use forward simulations to demonstrate that LSD can identify branches targeted by positive selection with remarkable sensitivity and specificity. We illustrate a range of potential applications by analyzing data from the 1000 Genomes Project and uncover a list of adaptive candidates accompanying the expansion of anatomically modern humans out of Africa and their spread to Europe.

Key words: **positive selection, population tree, ancient selection.**

Article

Introduction

Adaptation to local environments allowed life to spread across all ecosystems around the globe, ranging from deep-sea floors to high altitude mountain ranges (Yi et al. 2010; Fan et al. 2016; Sun et al. 2017). The evolutionary mechanism underlying adaptation is natural selection by which individuals carrying beneficial alleles show increased reproductive fitness in given environmental conditions. The variance in reproductive success across individuals results in abnormal gene genealogies, and ultimately distorts patterns of DNA variation at and around selected loci, leading to a local molecular footprint of positive selection. This footprint involves reduced levels of molecular diversity, extended homozygosity tracts, and increased differentiation from other populations. The development of statistical methods tailored to the detection of these signatures (Vitti et al. 2013) remains one of the most active research areas in evolutionary genomics (Hudson et al. 1987; McDonald and Kreitman 1991; Berg and Coop 2015; Field et al. 2016; Peyregne et al. 2017).

Amongst the wide range of methods available for detecting positive selection, Williamson et al. have designed a composite likelihood ratio (CLR) test that contrasts the fit of two competing models to patterns of DNA variation observed within a given genomic window (Williamson et al. 2007). The CLR test builds on Kim and Stephan (2002) (Kim and Stephan 2002) and is particularly relevant for detecting recent episodes of positive selection that led to the (almost) fixation of a novel variant emerging in a given population, the so-called recent hard sweeps. It has been implemented in a number of statistical packages, such as SweepFinder and SweeD (Pavlidis et al. 2013; DeGiorgio et al. 2016). Other methods, such as the long-range haplotype (LRH) and iHS tests, exploit genomic tracts of extended haplotype homozygosity (EHH) to detect haplotypes that have not been broken down by recombination and have reached intermediate population frequencies faster than expected under neutrality (Sabeti et al. 2002; Voight et al. 2006). EHH-derived methods are especially sensitive to ongoing sweeps, where the selected haplotype has not reached fixation. Other statistics, such as nsL (Ferrer-Admetlla et al. 2014) and H12 (Garud et al. 2015), provide complementary approaches capturing signatures of positive selection on standing variation, the so-called soft sweeps. In soft sweeps, the selected variant pre-existed in the population, possibly cosegregated in multiple genomic backgrounds, until it became adaptive following environmental changes.

Regardless of the nature of the underlying sweeps, populations submitted to different pressures develop lineage-specific adaptations to local environments. CLR and EHH methods were further extended to accommodate cross-population comparisons, through XP-CLR (Chen et al. 2010) and XP-EHH (Sabeti et al. 2007) tests. However, the first cross-population test for diversifying selection was developed by Cavalli-Sforza (1966) and Lewontin and Krakauer, in

the early 1970s, based on the F_{ST} index of genetic differentiation (LK test) (Lewontin and Krakauer 1973). The LK test established the basis for multiloci approaches, which assume that demography impacts the levels of genetic diversity at the scale of the whole genome, whereas adaptation only leaves local signatures of selection, which increases the levels of genetic differentiation around selected loci.

The LK test suffers from a number of well-characterized limitations (Nei and Maruyama 1975). These include a relatively high false-positive rate when only a reduced number of genetic markers are available, as the full F_{ST} distribution expected under neutrality cannot be then accurately estimated. The LK test was thus almost abandoned for decades, until molecular methodologies opened for a cost-effective characterization of large-scale SNP panels. This, however, did not mitigate another major limitation of the LK test, pertaining to the assumption of independent demes, which is violated if populations are related by complex demographic histories. This limitation was tackled through the development of novel statistical methods accounting for hierarchical population structures. For example, Bonhomme et al. presented FLK and hapFLK (Bonhomme et al. 2010; Fariello et al. 2013), two extensions of the LK test accommodating tree-like relationship between populations and dynamic changes in the effective size over time. FLK and hapFLK, however, cannot a priori identify which lineage(s) of the population tree were targeted by selection. These methods are thus uninformative regarding both the timing and environmental conditions underpinning adaptation.

Under the FLK framework, lineage(s) that experienced selection are identified a posteriori, by first building gene trees from specific genomic regions (hereafter referred to as local gene trees), and then searching for unusually long branches as a proxy for adaptation. In contrast, the locus-specific branch length (LSBL) (Shriver et al. 2004), as well as the derived population-branch statistics (PBS), consider a predefined trifurcating population tree, including two focal populations and an outgroup population. They then evaluate the fraction of the F_{ST} differentiation index that is exclusively attributable to each focal population (i.e., external branch) (Yi et al. 2010). Similar to PBS, 3P-CLR does not rely on the reconstruction of local gene trees, and is tailored to the detection of selective sweeps in the internal branch of a three-population tree (Racimo 2016).

In this study, we present LSD, a new cross-population framework that exploits the levels of exclusively shared differences existing between populations, in order to identify loci that underwent selection along the internal and external branches of a population tree. Using forward simulation, we demonstrate that LSD shows improved performance when compared with PBS under a wide range of conditions. Applying LSD to the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), we detect candidate loci that participated in the adaptive diversification of anatomically modern humans (AMH), following their dispersal out of Africa some ~50–100 thousand years ago, and their subsequent colonization of northern Europe some ~40–45 thousand years ago (Bae et al. 2017; Nielsen et al. 2017).

Results and Discussion

General Overview

Cross-population methods quantify shifts in allele frequencies as proxies for genomic signatures of positive selection. Only a limited number of cross-population methods, however, are currently devised for detecting lineage-specific adaptations, and their applicability is merely limited to three-population trees. PBS and LSBL, for example, leverage an outgroup population to calculate the Gromov product, a widely used concept in spatial geometry that reveals the allele frequency changes experienced by each population (fig. 1*A*).

Instead of relying on an outgroup, LSD calculates the Gromov product from the most recent common ancestor (MRCA) of two sister populations, namely A and B. This MRCA delineates the allele frequencies segregating within population ancestral to A and B (fig. 1*B*), and therefore represents a better initial point for tracking allele frequency changes, without the uncertainty provided by a phylogenetically distant outgroup. More importantly, working on population pairs opens for investigating any sort of internal or external branch along a bifurcating population tree (fig. 1*C*).

The Gromov product of two populations based on their MRCA could be calculated from ancestral allele frequencies, as detailed in the Materials and Methods section. Genome-wide data from ancestral populations are, however, generally not available. LSD overcomes this limitation by integrating three or more populations through local gene trees, which are rooted according to an external outgroup. Local gene trees are built from particular genomic regions (e.g., protein-coding regions or windows of predefined size), and assumed to have evolved within a given population history (fig. 1*C* and *D*).

As an illustration, assume that we have a five-population tree (populations A–E), and that we aim at detecting loci positively selected in the lineage leading to population A (fig. 1*C*). For each local gene tree, we first search for the subtree that includes all individuals from A, and all individuals from its sister population B (fig. 1*E*). From this MRCA, we then calculate the average genetic differences to the members of population A, provided these differences are not shared with individuals from population B (fig. 1*E*).

Beneficial mutations emerging within population A will tend to rapidly spread across that population, sweeping linked variants to moderate or high frequencies. Although often reducing the genetic diversity within population A, this process will also increase the number of genetic *differences* to population B. Looking backwards in time, if one allele underlies a selective advantage, carriers will rapidly coalesce to their MRCA. The branch leading from this MRCA, to the MRCA of populations A and B, will be thus longer than expected under neutrality, accumulating mutations that, on average, will be exclusively shared by a high fraction of individuals within population A (fig. 1*E*). Additionally, LSD quantifies whether genetic variants tend to be *exclusive* to one population only, since alleles common to both populations (e.g., following admixture or incomplete lineage sorting, ILS) are unlikely to underlie local adaptations.

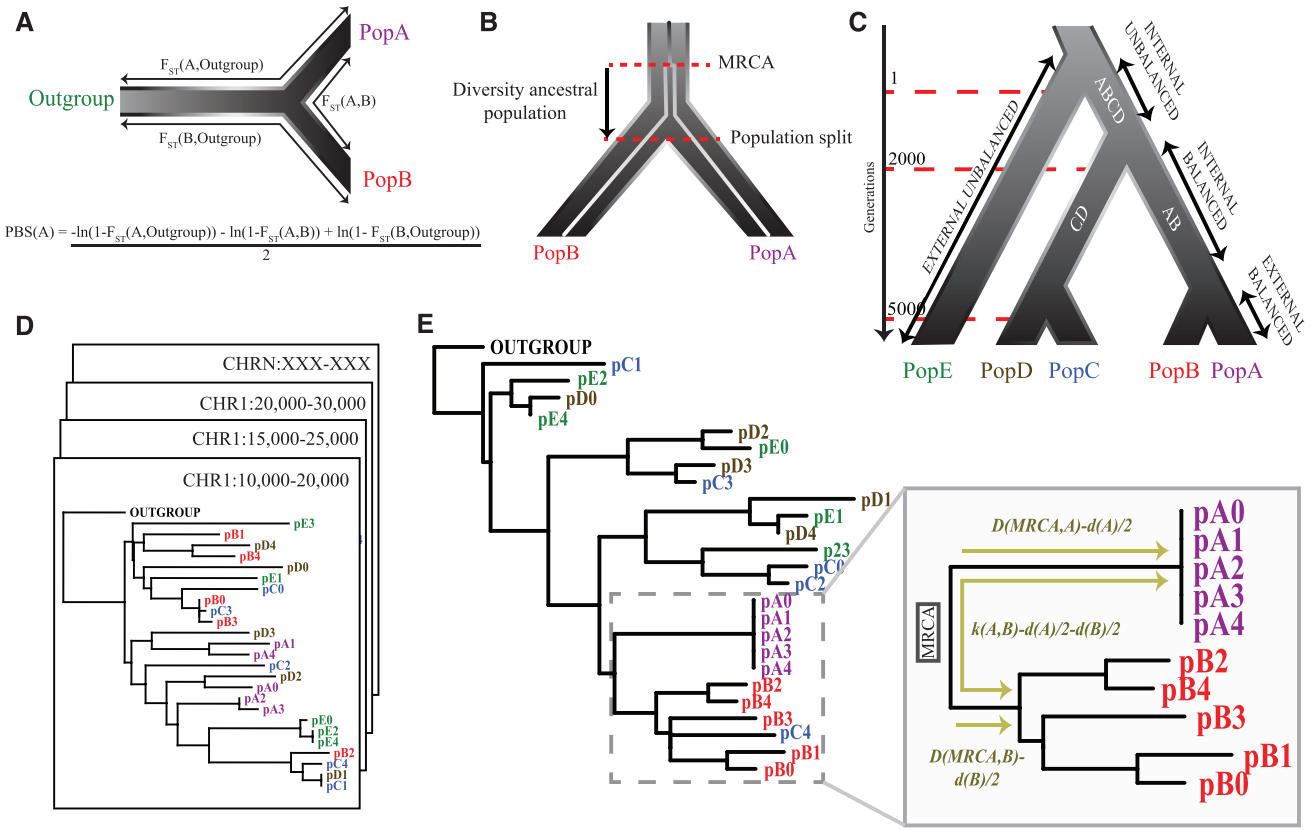

Fig. 1. Overview of methods for identifying lineage-specific adaptations. (*A*) PBS calculations for population A, applying the Gromov product on pairs of F_{ST} indexes. (*B*) The time spanned from the MRCA to the split of populations A and B determines the total branch length on which mutations could have been accumulated in the ancestral population, and thus its diversity. (*C*) Five population tree, with popA-popE representing external populations, AB and CD internal balanced branches, and ABCD an internal unbalanced branch. (*D*) Collection of gene trees, each summarizing local evolutionary forces operating on particular genomic regions, potentially including lineage-specific selection. (*E*) Local gene tree highlighting the footprint of selection on the individuals belonging to population A. The right panel provides a zoom into the population A undergoing selection at a particular locus, and its sister population B, altogether with the equations underlying the calculation of LSD(A). Note this local gene tree exemplifies strong selection signatures, characterized by low intrapopulation diversity and high interpopulation divergence. LSD can also be calculated from sister clades that are not reciprocally monophyletic.

LSD will be thus maximized for local gene trees where alleles from population A conform a monophyletic clade, harbor low levels of diversity, and are largely differentiated from alleles sampled for sister population B. Conversely, LSD will be minimized whenever alleles from population A exhibit high diversity, and are scattered across the same phylogenetic clade as individuals from population B.

Simulations

In order to evaluate the performance of LSD, we first simulated 10-kb-long DNA sequences evolving under a trifurcating population tree, with selection operating on the lineage leading to population A, as illustrated in figure 2*A* ($0 \leq s \leq 0.03$). Limiting such simulations to three populations enabled a comparison of the performance of LSD and PBS methods (see next section). From the simulated sequences, we sampled 10, 50, or 100 chromosomes from populations A and B. A single individual was sampled from the outgroup in order to root local gene trees. We then calculated the normalized LSD score (LSDnorm) for external branches A and B.

Results showed that LSDnorm(A) scaled with the selection coefficient, in line with the simulation of a selective advantage within this specific lineage (Spearman correlations; $P = 0.0068$, 0.0005, and 0.0005 for sample sizes 10, 50, and 100, respectively). In contrast, LSDnorm(B) remained steady (Spearman correlations; $P > 0.05$), as expected since population B was simulated to evolve under neutrality, solely subject to mutation and drift processes (fig. 2*A*). This suggests that LSDnorm can detect selection within the correct branch of the population tree with high sensitivity and specificity.

Simulations under scenarios of convergent evolution, where the same allele is independently selected in both populations A and B, reduced the sensitivity of LSD (fig. 2*B* and *C*). Nevertheless, selection was still detectable, as the LSDnorm increased with selection coefficients. In particular, simulations with selection operating only on population A, LSDnorm(A) reached values around 0.10 for $s = 0.03$ (fig. 2*A*). Assuming that the same beneficial mutation also appeared de novo in population B, the maximum magnitude of LSDnorm(A) was much lower, around 0.08 (fig. 2*B*). A similar result was

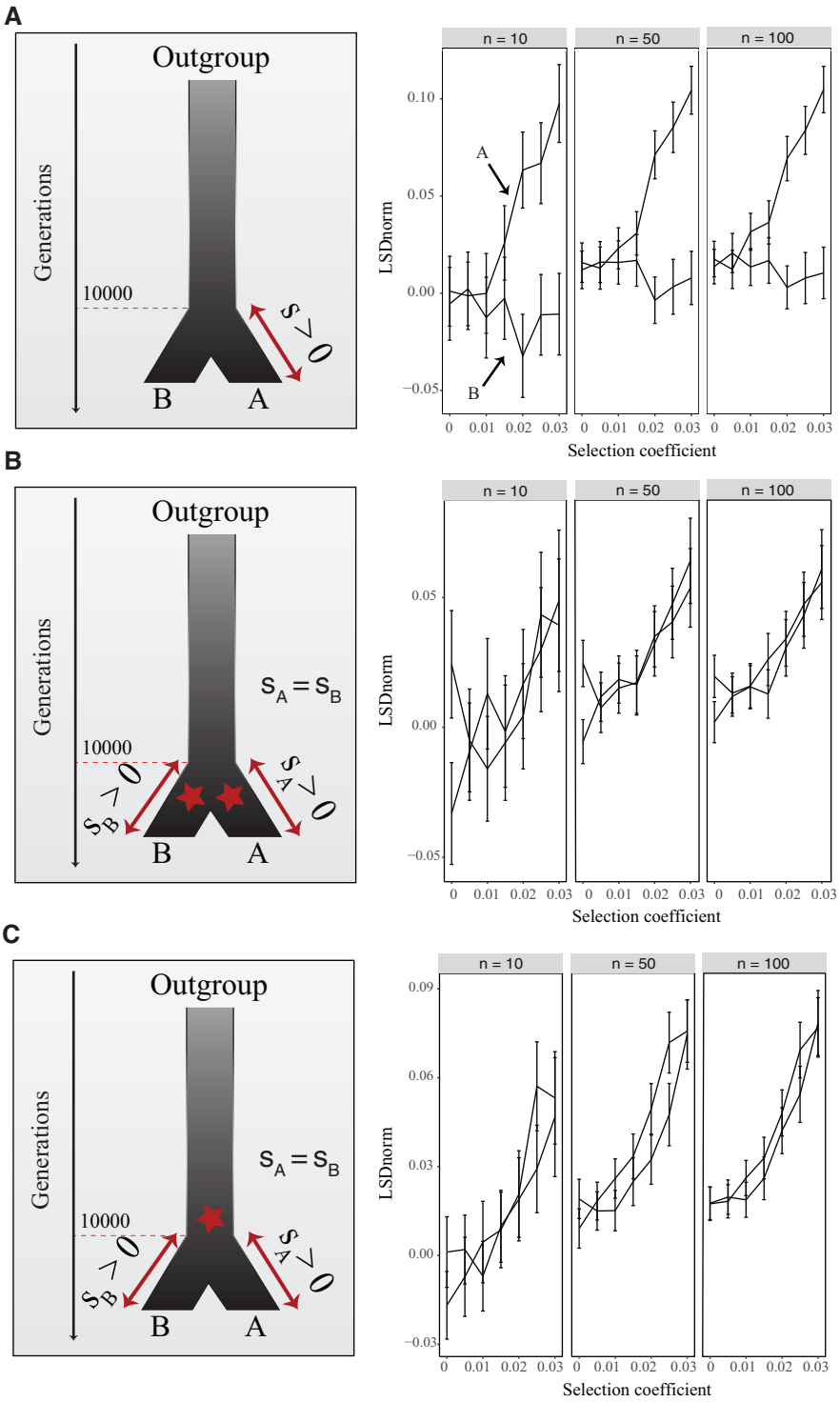

Fig. 2. Performance of LSDnorm in trifurcating population trees. Fifty chromosomes (25 diploid individuals) were sampled for populations A and B. The initial frequency and dominance coefficient of beneficial mutations were assumed to be 0.05 and 0.5, respectively. (*A*) The left panel shows the simulated scenario, where beneficial alleles were introduced in population A. The *x*–*y* scatterplot shows that only LSDnorm(A) scales with increasing selection coefficients. (*B*) The same adaptive mutations appeared, independently, in populations A and B. LSDnorm(A) reached lower values than in (*A*). (*C*) Convergent evolution from variants that were already segregating in the ancestral population, and became beneficial after the split of populations A and B.

obtained by simulating convergent evolution from alleles that were introduced in their ancestral population, but that only became beneficial after the split of populations A and B (fig. 2*C*). The relative reduction of LSDnorm values in cases of convergent evolution illustrates how changes independently experienced by populations A and B are partly interpreted as shared and erroneously attributed to the branch that leads to the outgroup population.

Benchmarking LSDnorm and PBS

Overlapping confidence intervals might hamper the ability to discriminate loci evolving under selection or neutrality, especially if selection is weak. In order to assess the sensitivity and specificity of LSDnorm, we compared LSDnorm values estimated from selective and neutral scenarios. Briefly, we randomly sampled a LSDnorm value from a scenario assuming $s = 0.03$, to which we subtracted another LSDnorm value sampled from neutral simulations. This difference controls for the underlying LSDnorm baseline under neutrality. In order to obtain an empirical distribution that fully captures the stochastic nature of the evolutionary process, we repeated this sampling procedure 10,000 times with replacement.

This sampling procedure was applied to two sets of local gene trees, first using phased chromosomes (LSDnorm.phased), then unphased data (LSDnorm.unphased). The same procedure was applied to estimate the accuracy of the PBS statistics for comparison. Since PBS relies on allele frequency changes, measured by pairwise F_{ST} indexes, we sampled 10, 50, or 100 chromosomes also from the outgroup population.

With recessive beneficial alleles, or with $s < 0.02$, both LSDnorm and PBS statistics performed poorly, mainly because the impact of selection on intra- and interpopulation diversity was limited (see supplementary fig. S1*B–D*, Supplementary Material online). In scenarios where alleles with $s = 0.03$ were introduced, LSDnorm and PBS substantially increased, relative to the neutral expectations, in line with their increasing power to pinpoint selected loci (fig. 3). More specifically, such increments were found to be generally higher for LSDnorm.phased, followed by LSDnorm.unphased, and then PBS (fig. 3 and see supplementary table S1, Supplementary Material online). Interestingly, LSDnorm.phased and LSDnorm.unphased performed similarly with small sample sizes, revealing the latter as a cost-effective alternative if sampling and/or sequencing capabilities are scarce, a restriction often experienced in ancient DNA and conservation studies (fig. 3).

Confidence intervals were found to be larger for LSDnorm than PBS estimates, for several reasons. LSDnorm can range from $-\infty$ to 1, in contrast to PBS, which often has a minimum boundary around zero. Phylogenetic inference of local gene trees can also contribute to this elevated LSDnorm variance, as discussed below. Finally, in the current experimental design, PBS was allowed to use a substantially higher number of chromosomes from the outgroup population, in comparison to LSDnorm. Given the simulated conditions, including large effective population sizes and short evolutionary distances, substantial amounts of incomplete lineage sorting are expected between the outgroup and focal populations. This implies that many local gene trees might be incorrectly rooted, provided that the chromosome sampled from the outgroup population is not truly external to populations A and B. In order to accurately root local gene trees, LSD users are advised to rely on a more distant outgroup, such as chimpanzees for humans or even a more distant species. Assuming that such distant outgroup is available, which is often the case, LSDnorm could still incorporate multiple samples from additional populations, including the one that was here simulated to serve as outgroup. As reflected by sampling more chromosomes from populations A and B, larger sampling from additional populations would likely decrease confidence intervals (fig. 3). Since the calculation of the PBS statistics relies on more outgroup samples than that of LSDnorm in this experimental setup, our benchmarking scheme was disadvantageous for LSDnorm and thus conservative.

The increased accuracy of LSDnorm, in comparison to the PBS statistics, was found regardless of the initial frequency of the beneficial allele (iAF) and its dominance coefficient (see supplementary material fig. S1, Supplementary Material online). The only exception corresponded to selective episodes starting from iAFs $= 0.05$, provided that $n > 10$. In these scenarios, PBS outperformed LSDnorm.unphased but not LSDnorm.phased (fig. 3). For iAF $= 0.10$, this trend vanished, and the accuracy of LSDnorm.unphased was greater than that of PBS (see supplementary fig. S2*A*, Supplementary Material online). Taken together, PBS might outperform LSDnorm.unphased in intermediate situations, where high iAFs promote multiple haplotype backgrounds to moderate frequencies, increasing intrapopulation levels of diversity (e.g., soft sweeps). In such cases, LSDnorm.phased and especially LSDnorm.unphased can become negative (see supplementary fig. S2*B*, Supplementary Material online).

Exploiting unphased data through genotype distances assumes individuals with the same genotype are identical, as no allele frequency changes (drift) occurred after their divergence. This assumption does not hold if individuals showing the same genotypes actually carry different haplotype pairs, which could reveal recombination events, implying that both individuals are not identical, but separated at least by a few generations of divergence since their MRCA (tMRCA). Genotype distances provide thus an underestimation of tMRCAs, especially for small sample sizes (see supplementary fig. S2*C*, Supplementary Material online). This downward bias in tMRCA detection reduced the accuracy of LSDnorm.unphased, in comparison with LSDnorm.phased.

The Wright–Fisher model predicts that the variance in allele frequency increases with the number of generations, by $1 - e^{-t/N}$, where t is the number of generations and N the population size (Tataru et al. 2017). The variance of F_{ST} values will then increment with longer distances to the outgroup (i.e., $F_{ST}(A, \text{outgroup})$ and $F_{ST}(B, \text{outgroup})$). Wider confidence intervals imply larger overlap between scenarios simulated under selection and neutrality, which reduces PBS ability to discriminate adaptive loci. Furthermore, the F_{ST} index contrasts between-population divergence to within-population diversity. The latter is often calculated by averaging over the nucleotide diversities of both populations (Hudson et al. 1992). Averaging dilutes, however, the footprint of evolutionary processes that affect a single population, including lineage-specific adaptations. By using the population immediately ancestral to A and B, instead of an outgroup (fig. 1*A* and *B*), LSDnorm overcomes both limitations. Indeed, inferring the allele frequencies segregating within the ancestral population enables to directly track lineage-specific

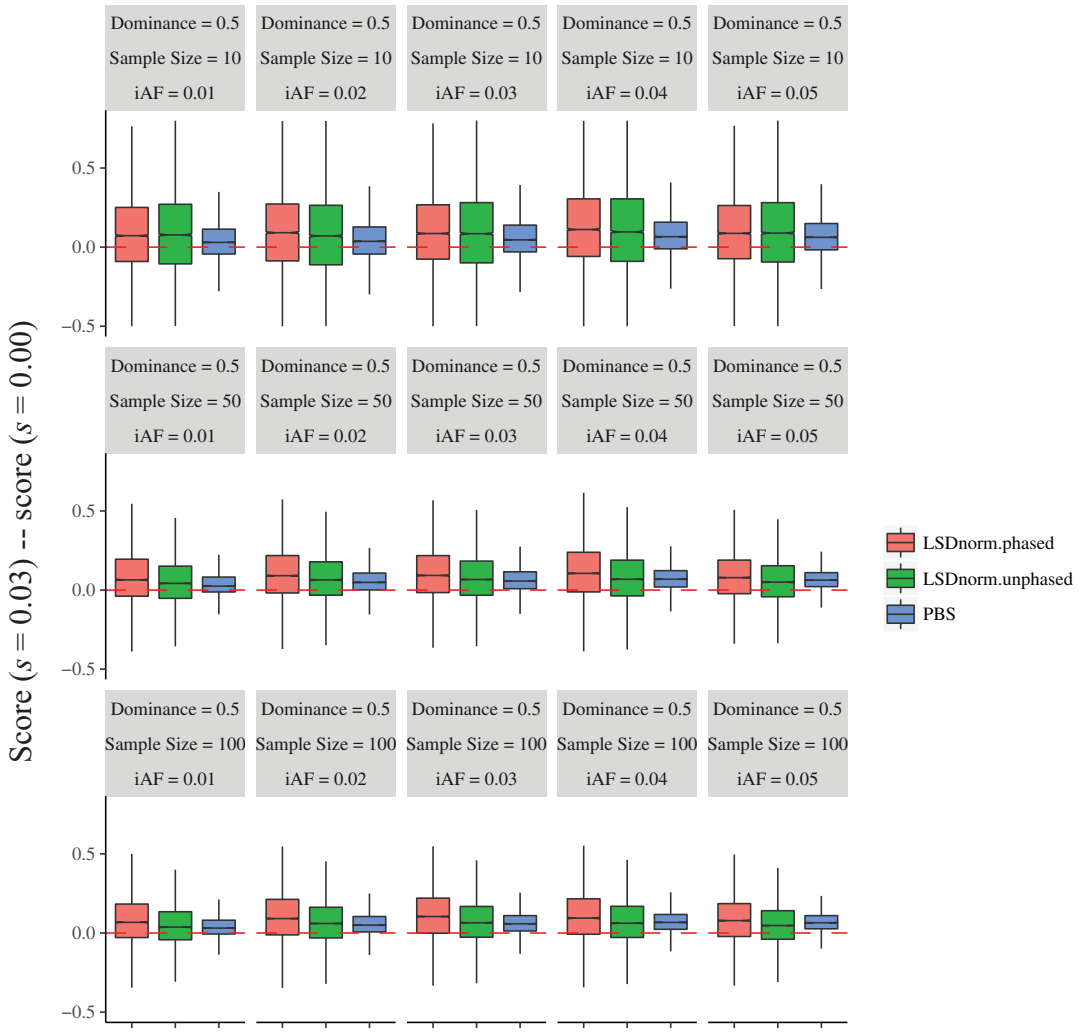

Fig. 3. Comparing LSDnorm.phased, LSDnorm.unphased, and PBS, across a wide range of initial allele frequencies (iAFs) and sample sizes. Intermediate dominance was assumed. The *y*-axis measures the difference between the scores estimated at $s = 0.03$ and $s = 0.00$, as proxies for the discriminatory power of LSDnorm.phased, LSDnorm.unphased, and PBS. LSDnorm.phased was generally associated with increased accuracy, compared with LSDnorm.unphased, and this, in turn, to PBS. Wider confidence intervals for LSDnorm than for PBS were partly due to different sample sizes from the outgroup population.

changes, without the uncertainty associated with a more distant outgroup.

In order to identify MRCAs, LSD leverages local gene trees that jointly incorporate multiple populations. This approach comes with two potential limitations. First, inferring local gene trees raises computational costs, albeit running times remained suitable for rapidly scanning whole genome panels (see supplementary fig. S2*D*, Supplementary Material online). Second, the accuracy of gene tree inference is limited if the amount of genetic changes is small (e.g., within short time scales). However, we note that LSD does not rely on the exact topology of local gene trees, but mainly uses branch lengths as estimates of genetic distances between pairs of phylogenetic nodes. In fact, in the simulations above, accuracy was found to be higher for LSDnorm than for PBS, despite LSDnorm was based on suboptimal local gene trees, reconstructed using the Neighbor-Joining (NJ) algorithm. Alternative approaches might deserve future consideration, including UPGMA-based local gene trees. Assuming strict molecular clock, which is presumed at short times scales, this method directly infers midpoint-rooted trees. Depending on the amount of data, and its computational feasibility, more sophisticated methods for phylogenetic inference, such a maximum likelihood algorithms (Stamatakis 2014) or Bayesian approaches integrating over the marginal probability of local gene trees (Yang and Rannala 1997; Ronquist and Huelsenbeck 2003) might be embedded to the calculation of LSDnorm, likely improving its performance.

Inferring LSD from local gene trees conversely offers multiple advantages. First, incorporating multiple populations enables partitioning the past in evolutionary time periods delimited by the internal and external branches of the population tree. This allows for the detection of positive selection along the different branches of a population tree. Furthermore, integrating substitution models, such as the generalized time reversible (GTR), during gene tree inference is straightforward, improving predictions especially when comparing distant populations. Finally, local gene trees can

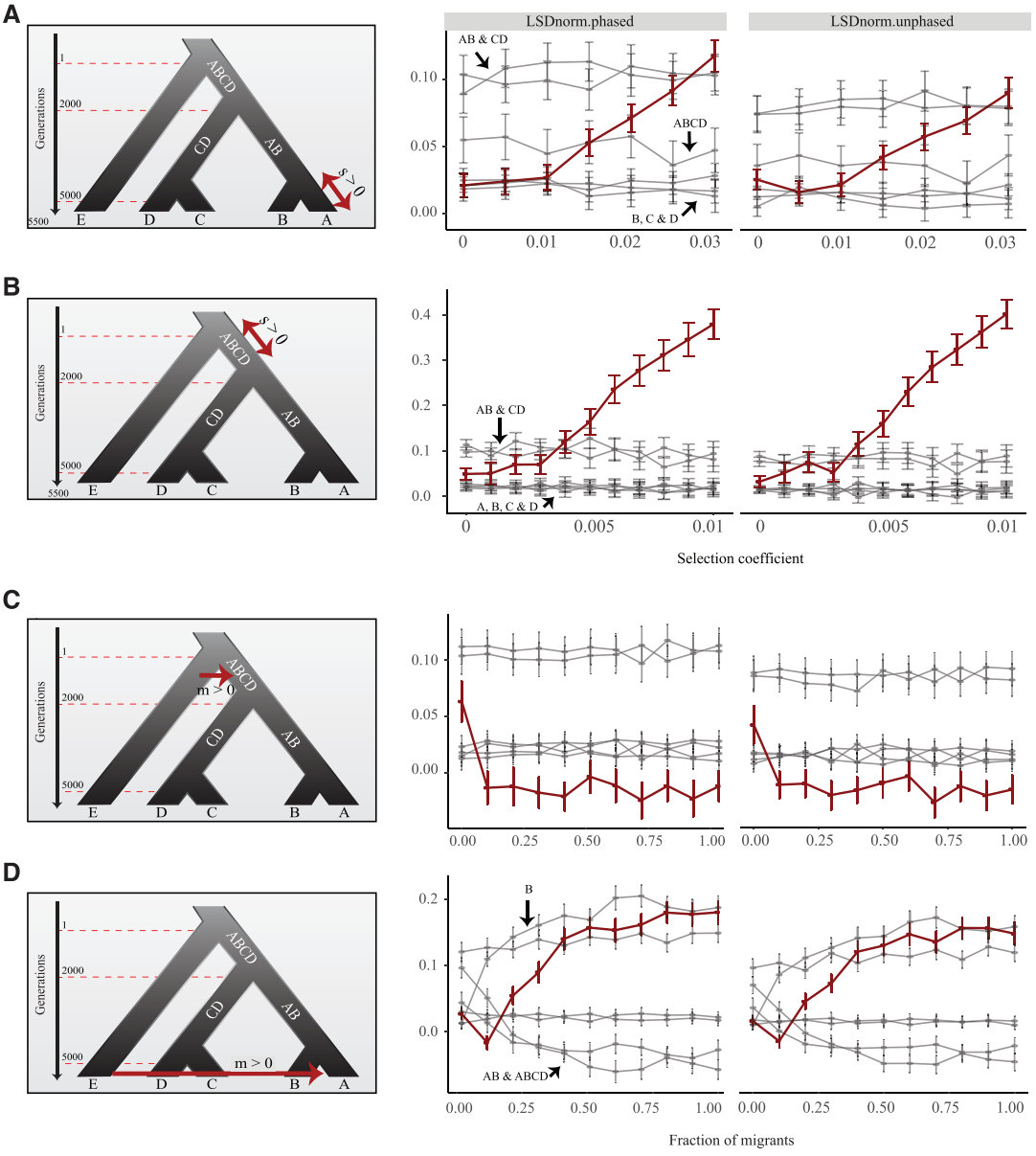

Fig. 4. Performance of LSDnorm in five-population trees. Fifty chromosomes (25 diploid individuals) were sampled for all populations. The initial frequency and dominance coefficient of beneficial mutations were assumed to be 0.05 and 0.5, respectively. (*A*) The left panel shows the simulated scenario, where beneficial alleles are introduced in the unbalanced lineage ABCD. The *x*–*y* scatterplot shows that only LSDnorm(ABCD) scales with increasing selection coefficients. (*B*) Beneficial alleles, associated with a range of selection coefficients, are introduced in population A, increasing LSDnorm(A) scores. (*C*) The unbalanced lineage ABCD experiences introgression from its sister lineage, leading to population E. As the proportion of migrants increases towards one, LSDnorm(ABCD) approaches to zero, which is indicative of the absence of genetic differentiation from its sister lineage. (*D*) Population A experiences introgression from the distant population population E. LSDnorm(ABCD) values can eventually become negative.

be easily built from multiallelic data, including genetic markers other than nucleotide polymorphisms, such as copy-number variants (CNVs) or small insertion and deletions, which opens for possible investigation of the adaptive role of such genetic variants.

Extending LSD to More than Three Populations

In contrast to PBS, LSD is not limited to trifurcating population trees, and is thus applicable to a wider range of scenarios. Suppose now that samples from two additional populations, namely C and D, are available, as well as a more distant outgroup (fig. 4A). We follow the same simulation framework as the previous section, where selection takes place within population A ($0 \leq s \leq 0.03$) during 500 generations. At $s = 0$, differences in the LSDnorm baseline solely reflect population divergence. For $s > 0$, LSDnorm(A) scaled with increasing selection coefficients (fig. 4A), following a trend similar to that found for trifurcating trees (fig. 2A). We note that LSDnorm did not propagate the selection footprint to other branches, which demonstrates the specificity of the method (fig. 4A).

We next introduced alleles conferring selective advantage in the population ancestral to populations A, B, C, and D

(ABCD), right after the split from population E. If such alleles were inherited by descending populations, they were then simulated to be no longer advantageous, and to evolve under neutrality. This ensured that the selective episode was confined within population ABCD. In such scenarios, we found that LSDnorm accurately identified selection within the expected branch (fig. 4*B*). Since LSDnorm depends on the duration of selection, in the form $(1+s)^t$ (eqs. 12 and 13), selection on the long ABCD branch resulted in high LSDnorm(ABCD) scores (fig. 4*B*). Given that the average time for sweep completion is $2\ln(2Ns)/s$ (Hermisson and Pennings 2005), many sweeps simulated for only 500 generations were expected to be still ongoing (incomplete), producing a moderate impact on DNA variation, and hence, on LSDnorm values (fig. 4*A*), in comparison to situations where selection was simulated within population (ABCD) (fig. 4*B*).

As the signal left by a sweep is transient, the exact timing of the selective episode within each lineage could potentially affect LSDnorm. After a sweep, novel neutral mutations slowly restore the levels of diversity (π), by a maximum rate of $\Delta\pi = 2\mu(1 - 1/N)^t$. Using the same parameters as in the above simulations, completely recovering from $\pi = 0.0015$ to 0.0020 would require over 185,000 generations. The time elapsed since the end of the sweep (here, 3,000 generations at best) is, therefore, expected to have negligible impact on LSDnorm, provided the population remains stable during this period, and does not experience a massive demographic collapse or migration flux.

We finally assessed the impact of introgression from population E into population ABCD. On the limit of high migration rates, introgression from the branch leading to the sister population E effectively counteracted genetic divergence, and the expected LSDnorm(ABCD) value should have approached zero, as predicted by equation (11) following the minimization of the $(p_{CD} - p_E)$ and $(p_{AB} - p_E)$ terms. Nevertheless, LSDnorm values were found to be slightly negative, likely reflecting the downward bias in tMRCA estimation (see supplementary fig. S2C, Supplementary Material online).

Beyond convergent evolution, only another scenario caused LSDnorm to vary in several branches of the population tree. This scenario involved introgression from the branch leading to population E into a population descending from ABCD, such as population A (fig. 4*D*). The influx of derived alleles into the incipient population A contributed to further differentiate population A from B, elevating LSDnorm(A). As exogenous alleles introgressing into population A deepened the tMRCA of populations A and B, LSDnorm(B) also showed an upward trend. Reciprocally, introgression of derived alleles depleted the incipient population A from ancestral variants, defined as p_A. As expected from equation (11), LSDnorm(AB) declined with lower p_A values, eventually becoming negative. The impact of this introgression was propagated to ABCD, indicating that consistent negative LSDnorm values could help, in the future, to unravel introgression events from distant populations.

Empirical Data Set

Adaptations in Modern Europeans

The peopling of Europe has been a major research focus, and is characterized by initial admixture with Neanderthals (Green et al. 2010; Fu et al. 2014; Seguin-Orlando et al. 2014; Fu et al. 2015; Sikora et al. 2017), demographic changes, and multiple waves of population replacement (Lazaridis et al. 2014; Haak et al. 2015). Multiple studies have also started to decipher the genetic basis of human adaptations to their new environments, reporting selection signatures for alleles most notably associated with skin color and eye pigmentation (Marciniak and Perry 2017; Nielsen et al. 2017). In the following section, we used LSDnorm to scan for genomic adaptations in modern Europeans, as represented by the CEU population (fig. 5*A*), in order to assess whether LSD can retrieve selection targets, already identified in previous selection scans based on other statistical methods.

The distribution of LSDnorm values within 10-kb genomic windows was bell shaped, and centered at 0.032 (fig. 5*B*). That the mode of the distribution is positive indicates that the CEU population harbors exclusively shared variants in the majority of genomic regions. We also identified an excess of regions showing the maximum score of one. These windows can correspond to loci driven to fixation by positive selection, but could also result from neutral processes. For instance, allelic surfing at the front wave of successive range expansions throughout Europe can lead neutral alleles to fixation and can spread these variants over vast geographic regions, thereby mimicking the signature of local adaptation (Edmonds et al. 2004; Hofer et al. 2009). Haplotypes that increased their frequency due to allelic surfing are relatively old (surfing occurs at the initial stage of colonization) and are thus more likely to have been shortened by recombination. Conditioning our analyses on long-enough regions (e.g., showing at least three consecutive 10-kb windows ranking amongst the top 1000 LSDnorm scores) is expected to minimize the impact of allelic surfing. Following this strategy, we delineated 72 LSDnorm outlier regions, in which the exons of 31 genes coding for proteins could be found (see supplementary table S2, Supplementary Material online). Of note, five such genes overlapped with selection candidates previously detected using 3P-CLR (*GMLC1*, *TEKT2*, *CLSPN*, *ADPRHL2* and *PSMB2*; Racimo 2016). LSDnorm also recapitulates other well-known examples of adaptation, such as at *MYO5A* and *SLC45A2* (Voight et al. 2006; Sabeti et al. 2007; Barreiro et al. 2008), which underpin adaptation for lighter skin color. We also confirmed the overrepresentation of some functional categories described previously, such as melanin biosynthesis (Fisher exact test; adjusted *P*-value < 0.0189), albinism (0.0217), and pigmentation disorders (0.0251) (see supplementary table S3, Supplementary Material online). Together with the *RP1* and *TTC21B* genes, *MYO5A* and *SLC45A2* contributed to the significant enrichment of two functional categories, namely

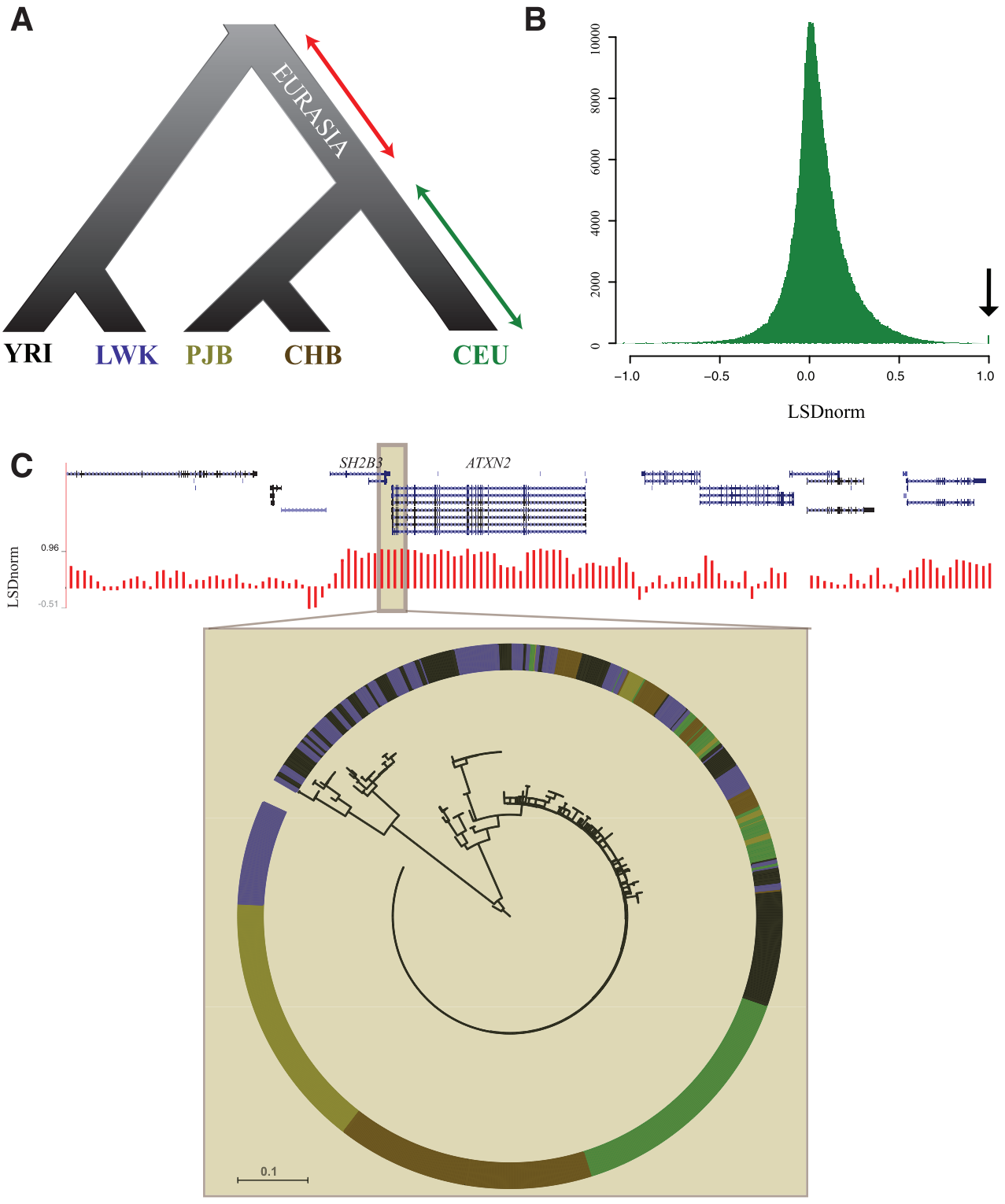

Fig. 5. LSDnorm applied to the 1000 human genome project (phase III). (*A*) Human population tree assumed to investigate selection along the lineage ancestral to Europeans (CEU) and Asians (PJB and CHB) (red arrow), and the European population (green arrow). (*B*) Histogram showing the empirical LSDnorm distribution for the CEU population. The black arrow highlights the excess of 10-kb windows with a maximum score of one. (*C*) LSDnorm scores along the *SH2B3-ATXN2* locus, altogether with the local gene tree at chr12: 111,440,001–111,450,000, which shows widely spread derived alleles amongst CEU (green), PJL (dark yellow) and CHB (dark brown) populations.

the photoreceptor outer segment (0.0163) and strabismus (0.0227). Loss-of-function *RP1* mutations, for example, are associated with retinitis pigmentosa, a disease characterized by limited night vision (Blanton et al. 1991; Pierce et al. 1999), and thus potentially related to reduced sunlight exposure at northern latitudes.

The selective pressures underlying the remaining selection candidates are more speculative. The *PIK3CB* gene,

participating in insulin resistance (Clément et al. 2009), has been recently found as a selection candidate in Europe in an independent study (Vatsiou et al. 2016). Selection at this gene might support the "carnivory connection" hypothesis, which posits that recent shift toward starch-rich diets shaped the levels of insulin resistance (Colagiuri and Brand Miller 2002). Likewise, the *LMLN* gene encodes for leishmanolysin, a glycoprotein acting against *Leishmania* protozoans. Such pathogens can cause a range of diseases in humans and dogs and are often transmitted by blood-sucking insects that are nearly absent in North European latitudes (Pace 2014).

Despite representing one of strongest signals of recent adaptation within European populations (Bersaglieri et al. 2004), the locus responsible for lactase persistence (LP) was not among the selection candidates detected by LSDnorm. Yet, the locus is embedded within a >2 Mb tract that shows LSDnorm values ranking amongst the top 2% of the genome-wide distribution (see supplementary fig. S3, Supplementary Material online). Beyond Europeans, LP is present in other populations, including the Punjabi (Abbas et al. 1983), often due to evolutionary convergence at the genetic level. For example, the −13910C/T (rs4988235) SNP, responsible for LP in Europe, also contributes to LP in Central Asia (Heyer et al. 2011). As illustrated in figure 2*B* and *C*, such cases of convergent evolution limit the sensitivity of LSDnorm to detect lineage-specific signatures of selection.

Adaptations in the Human Lineage That Left Africa

We next investigated phenotypic adaptations that accompanied another major transition in human history, namely the expansion out of Africa. This expansion started possibly as early as ∼100 kya ago (Grün et al. 2005, but see also Hershkovitz et al. 2018), when a reduced group of AMH left Africa for the first time, expanded throughout the Old World, and were exposed to novel environmental conditions and pathogens (fig. 4*A*).

Applying the same significance threshold as above, to the branch ancestral to Punjabi (PJB), Han Chinese (CHB), and European (CEU) populations (labeled as Eurasia in fig. 4*A*), we identified 41 genomic regions overlapping with the exons of only 28 genes coding for proteins (see supplementary table S4, Supplementary Material online). This relatively limited number of selection candidates was enriched for functional categories such as for cardiovascular alterations (Fisher exact test; adjusted *P*-value < 0.0327), movement disorders such as Parkinson (0.0399), celiac disease (0.0399), regulation of actin cytoskeleton (<0.0489), and disorders of the penis (0.0327) (see supplementary table S5, Supplementary Material online). These functional categories mirror traits commonly associated with populations of African ancestry, including high hypertension prevalence (adjusted *P*-value = 0.0327) (Wong et al. 2002; Agyemang and Bhopal 2003). We, however, note that the statistical significance of these categories is driven by a few pleiotropic genes, most often as combinations of *SH2B3*, *ATXN2*, and *SEPT4*.

The *SH2B3–ATXN2* locus, for example, is associated with a vast range of autoimmune disorders, including diabetes type I, celiac disease, or systolic blood pressure (Orrù et al. 2013; Auburger et al. 2014; Kullo et al. 2014). Interestingly, this locus has been previously reported as a selection candidate, mostly in Europe (Zhang et al. 2013; Brinkworth and Barreiro 2014; Mathieson et al. 2015). Inspection of local gene trees overlapping *SH2B3–ATXN2* confirmed that some derived variants are widely spread in Europe (CEU), but also amongst Han Chinese (CHB) and Punjabi populations (PJB), a pattern compatible with selection acting prior to the Europeans–Asians split (fig. 4*C*).

The *SEPT4* gene encodes a cytoskeletal regulator that, like *ATXN2*, has been associated with the Parkinson disease (Ihara et al. 2007). Using 3P-CLR with archaic hominins as outgroup, *SEPT4* was also recently reported as a selection candidate, however, in the branch leading to AMH (Racimo 2016). Local gene trees from the *SEPT4* region reveal that all Eurasians (CEU, PJB, and CHB) share an almost identical haplotype, whereas evident substructure emerges within Yoruba (YRI), and within Luhya (LWK) African populations (see supplementary fig. S4, Supplementary Material online). Although compatible with an incomplete sweep in the branch ancestral to AMH, this pattern of population divergence also supports positive selection associated with the out-of-Africa expansion.

Incongruences between LSDnorm and 3P-CLR analyses can result from distinct significance cut-offs, and/or differences in the analyzed data sets. Alternatively, as 3P-CLR is limited to two populations and an outgroup, the study relying on this method investigated the branch ancestral to AMH by pooling individuals from multiple European and Asian groups, as a single homogeneous proxy for the Eurasian population. The presence of population structure within Eurasians, however, violates the assumptions underlying binomial allele sampling, by preventing alleles to spread over the Eurasian metapopulation. This reduces the power to detect true selective episodes, unless hierarchical structure within Eurasia is explicitly accounted for. This cannot be done in current implementations of the 3P-CLR test but can easily be carried out using LSDnorm, which is not limited to three populations.

Conclusions

We present a new framework, based on the LSD between predefined phylogenetic clades to accurately identify loci undergoing selection across a population tree. Through simulations and the analysis of real genome-scale data sets, we demonstrate the sensibility and specificity of our method in assigning the episode of positive selection to the correct branch of the population tree. We also identify population histories inevitably misleading when characterizing diversifying selection, including convergence and gene flow. Future work exploiting phased data to detect recombination breakpoints and subsequently partition the genome into independently evolving regions can increase the resolution to precisely identify the selected haplotypes.

Materials and Methods

LSD in Practice

Assume A and B are sister populations that represent external lineages in the population tree (fig. 1C). LSD can be computed as

$$\mathrm{LSD}(A) = \frac{\left[D(MRCA,\ A) - \frac{d(A)}{2}\right] + \left[k(A,\ B) - \frac{d(A)}{2} - \frac{d(B)}{2}\right] - \left[D(MRCA,\ B) - \frac{d(B)}{2}\right]}{2} = \frac{D(MRCA,\ A) + k(A,\ B) - D(MRCA,\ B) - d(A)}{2}, \quad (1)$$

where $D(MRCA, A)$ represents the average distance from the MRCA of A and B to each individual belonging to population A, as estimated from the branch lengths in the local gene tree. Distance units are typically expressed in substitutions per site. $d(A)$ corresponds to the average pairwise distance between individuals belonging to population A, and $k(A, B)$ is the average pairwise distance between all pairs of individuals consisting of one individual from each population.

Figure 1E provides a visual interpretation of the mathematical terms, and illustrates that $D(MRCA, A) - d(A)/2$ accounts for the amount of differences shared between individuals belonging to population A. The fraction of these mutations that are exclusive to population A is quantified by $k(A, B) - d(A)/2 - d(B)/2$. Allele sharing between populations A and B, due to ILS or gene flow, will reduce $k(A, B) - d(A)/2 - d(B)/2$, thereby reducing LSD(A). The last operation consists in subtracting the amount of genetic differences accumulated along the branch leading to population B (fig. 1E), as provided by the term $D(MRCA, B) - d(B)/2$. This ensures that selective episodes occurring in population B are not attributed to LSD(A), making LSD independent for each branch in the population tree. We can thus write LSD(B) as

$$\mathrm{LSD}(B) = \frac{D(MRCA,\ B) + k(A,\ B) - D(MRCA,\ A) - d(B)}{2}. \quad (2)$$

We classify branches in the population tree as unbalanced or balanced, depending on whether their sister clade is an external or an internal lineage. Branches labeled as A and B are, for example, both external and balanced (fig. 1C). Branch E, instead, is external but unbalanced, because its sister lineage is the internal branch ABCD. Internal branches can be similarly classified as balanced, such as ABCD, or unbalanced, such as AB. The LSD calculation slightly varies according to this, as

$$\begin{cases} LSD(E) = \dfrac{D(MRCA,\ E) + \frac{k(CD,\ E) + k(AB,\ E)}{2} - \frac{D(MRCA,\ CD) + D(MRCA,\ AB)}{2} - d(E)}{2} \\ LSD(AB) = \dfrac{\frac{D(MRCA,\ A) + D(MRCA,\ B)}{2} + \frac{k(A,\ C) + k(A,\ D) + k(B,\ C) + k(B,\ D)}{4} - \frac{D(MRCA,\ C) + D(MRCA,\ D)}{2} - k(A,\ B)}{2} \\ LSD(ABCD) = \dfrac{\frac{D(MRCA,\ CD) + D(MRCA,\ AB)}{2} + \frac{k(CD,\ E) + k(AB,\ E)}{2} - D(MRCA,\ E) - k(CD,\ AB)}{2} \end{cases}. \quad (3)$$

In equation (3), the metapopulation CD represents the merger of individuals from C and D, and AB the merger of individuals from A and B. As our method can handle both unbalanced and balanced lineages, it is generalizable to any population tree of any given size, and is not limited to the population tree provided as an illustration in figure 1C.

Normalized LSD Values Provide Proxies for Selection Coefficients

The previous section presented the fundamental operations underpinning the LSD calculations, given predefined (groups of) population(s) in a given population tree. In the following, we demonstrate analytically that LSD captures the impact of natural selection on DNA variation. We first transform branch lengths, measured in numbers of differences per site, in allele frequency changes, according to

$$\begin{cases} k(A,B) = p_A(1 - p_B) + p_B(1 - p_A) \\ D(MRCA,A) = p_A(1 - 1p_{MRCA}) + p_{MRCA}(1 - p_A) \\ d(A) = \dfrac{2np_A(1-p_A)}{n-1} \approx 2p_A(1-p_A) \end{cases}, \quad (4)$$

where p_A, p_B, and p_{MRCA} stand for the allele frequencies in populations A, B, and their immediate parental population. The approximation in the last equality assumes large sample sizes, so that $n/(n-1) \approx 1$. Replacing equation (4) into equation (1), we have

$$\mathrm{LSD}(A) = \frac{[p_A(1-p_{MRCA})+p_{MRCA}(1-p_A)]+[p_A(1-p_B)+p_B(1-p_A)]+[p_{MRCA}(1-p_B)+p_B(1-p_{MRCA})]-2p_A(1-p_A)}{2}$$
$$= {p_A}^2 - p_A p_{MRCA} - p_A p_B + p_B p_{MRCA} = (p_A - p_{MRCA})(p_A - p_B) \quad (5)$$

The amount of *differences* accumulated since the ancestral population is quantified by $(p_A - p_{MRCA})$, whereas $(p_A - p_B)$ evaluates whether these differences are *exclusive* of population A or not. The combination of both terms represents then the Levels of exclusively Shared Differences (LSD). It is noteworthy that this development reveals a particular configuration of the so-called f_3 statistics (Patterson et al. 2012), in the form $f_3(A; MRCA, B)$, which depends on p_{MRCA}. Negative LSD(*A*) values can only be explained by alleles introgressed into A, whereas positive values reflect derived changes accumulated in population A. Applying the transformation shown in equation (4) to equation (3), we can now write LSD(*E*), LSD(*AB*), and LSD(*ABCD*) as

$$\begin{cases} LSD(E) = (p_E - p_{MRCA})\left(p_E - \frac{p_{AB}+p_{CD}}{2}\right) \\ LSD(AB) = \frac{(p_B - \frac{P_C+P_D}{2})(p_A - p_{MRCA}) + (p_B - p_{MRCA})(p_A - \frac{P_C+P_D}{2})}{2} \\ LSD(ABCD) = \frac{(p_{CD}-p_E)(p_{AB}-p_{MRCA}) + (p_{CD}-p_{MRCA})(p_{AB}-p_E)}{2} \end{cases} \quad (6)$$

where p_{AB} is the allele frequency in the metapopulation AB, which considers all individuals from A and B jointly. Allele frequency changes can be visualized as the overlap between paths connecting the split nodes in a population tree (Patterson et al. 2012). For LSD(*ABCD*), this corresponds to the overlap between the paths connecting populations CD and E (CD→E) on the one hand, and AB→MRCA on the other hand, averaged with the overlap between CD→MRCA and AB→E. This clearly coincides with the lineage representing the ancestral population labeled as ABCD (fig. 1C). This can be extended to any other lineage, provided that the correct equation is applied depending on whether it is external or internal, and balanced or unbalanced.

We have shown that LSD can be expressed in terms of allele frequency changes, captured by terms that ultimately depend on p_{MRCA}. Isolating p_{MRCA} in equation (5) allows us to calculate the allele frequency in the parental node of population A

$$p_{MRCA}(A) = \frac{LSD(A) - {p_A}^2 + p_A p_B}{p_B - p_A}. \quad (7)$$

Likewise,

$$\begin{cases} p_{MRCA}(E) = \frac{LSD(E) - {p_E}^2 + p_E\left[\frac{p_{AB}+p_{CD}}{2}\right]}{\frac{p_{AB}+p_{CD}}{2} - p_A} \\ p_{MRCA}(AB) = \frac{2LSD(AB) - 2p_A p_B + \left[\frac{P_C+P_D}{2}\right](p_A+p_B)}{p_C+p_D-p_A-p_B} \\ p_{MRCA}(ABCD) = \frac{2LSD(ABCD) - 2p_{CD}p_{AB} + p_E(p_{CD}+p_{AB})}{2p_E - p_{CD} - p_{AB}} \end{cases} \quad (8)$$

Assuming strong selection and large population sizes, beneficial alleles arising at low frequencies (e.g., $p_{MRCA} = 1/N$) experience small fluctuations in their frequency right after their emergence, usually escaping accidental extinction, and following a quasi-deterministic allele trajectory. Such quasi-deterministic trajectory is also applicable for weak selection on standing variation ($p_{MRCA} \gg 1/N$) (Smith and Haigh 1974; Kaplan et al. 1988, 1989). In either case, the expected frequency of an allele associated with an increase of *s* in fitness, after *t* generations is given by

$$p_t = \frac{p_{MRCA}(1+s)^t}{p_{MRCA}(1+s)^t + (1-p_{MRCA})}. \quad (9)$$

In the simplest case, if selection acts on a single population, with no introgression, the expected values for equations (5) and (6) simplify to $(p_t - p_{MRCA})^2$. Replacing p_t by equation (9) reveals the relation between LSD and *t*, p_{MRCA} and *s*

$$\mathrm{LSD} = \left[\frac{p_{MRCA}(1-p_{MRCA})\left[(1+s)^t - 1\right]}{p_{MRCA}\left[(1+s)^t - 1\right] + 1}\right]^2. \quad (10)$$

This function draws a sigmoid curve, with its maximum boundary depending on the levels of standing variation, as defined by $\max(\mathrm{LSD}) = (1 - p_{MRCA})^2$. Since variations in the upper limit preclude comparisons along the genome, we normalize LSD following an approach similar to the calculation of F_{ST} indices. More specifically, and following our notation for LSD(*A*), F_{ST} can be defined as $F_{ST} = 1 - (d(A) + d(B))/2k(A, B)$ (Hudson et al. 1992). In contrast to the F_{ST} index, based on a population pair, LSD focalizes on a single population. Consequently, LSD is not normalized by $k(A, B)$, but by the

total variability exclusively accumulated in population A (i.e., $D(MRCA, A) + k(A, B) - D(MRCA, B)$)

$$\mathrm{LSDnorm}(A) = 1 - \frac{d(A)}{D(MRCA,A) + k(A,B) - D(MRCA,B)} = 1 - \frac{2p_A(1-p_A)}{2(p_A - p_{MRCA})(p_A - p_B) + 2p_A(1-p_A)}$$
$$= 1 - \frac{p_A(1-p_A)}{(p_A - p_{MRCA})(p_A - p_B) + p_A(1-p_A)}$$

$$\begin{cases} \mathrm{LSDnorm}(E) = 1 - \dfrac{p_E(1-p_E)}{(p_E - p_{MRCA})\left(p_E - \dfrac{p_{AB}+p_{CD}}{2}\right) + p_E(1-p_E)} \\ \mathrm{LSDnorm}(AB) = 1 - \dfrac{p_A(1-p_B) + p_B(1-p_A)}{(p_B - \frac{P_C+P_D}{2})(p_A - p_{MRCA}) + (p_B - p_{MRCA})(p_A - \frac{P_C+P_D}{2}) + p_A(1-p_B) + p_B(1-p_A)} \\ \mathrm{LSDnorm}(ABCD) = 1 - \dfrac{p_{AB}(1-p_{CD}) + p_{CD}(1-p_{AB})}{(p_{CD} - p_E)(p_{AB} - p_{MRCA}) + (p_{CD} - p_{MRCA})(p_{AB} - p_E) + p_{AB}(1-p_{CD}) + p_{CD}(1-p_{AB})} \end{cases}. \quad (11)$$

Following equation (10), the relation between LSDnorm and t, p_{MRCA} and s is given by

$$\mathrm{LSDnorm} = 1 - \frac{(1+s)^t}{p_{MRCA}(1-p_{MRCA})[(1+s)^t - 1]^2 + (1+s)^t} = \frac{p_{MRCA}(1-p_{MRCA})[(1+s)^t - 1]^2}{p_{MRCA}(1-p_{MRCA})[(1+s)^t - 1]^2 + (1+s)^t}. \quad (12)$$

For moderate to large $(1+s)^t$ values, so that $[(1+s)^t - 1]^2 \approx (1+s)^{2t}$, this simplifies to the following function

$$\mathrm{LSDnorm} \approx \frac{p_{MRCA}(1-p_{MRCA})(1+s)^t}{p_{MRCA}(1-p_{MRCA})(1+s)^t + 1}, \quad (13)$$

where the initial variability is distorted by natural selection at a $(1+s)$ rate each generation. With increasing $(1+s)^t$, LSDnorm asymptotically reaches a plateau with a maximum boundary of one, corresponding to hard sweep scenarios. Negative selection ($s < 0$) can also yield a similar allele trajectory, just because if an allele starts to be heavily purged from the genetic pool, the other variant segregating in the population will symmetrically increase in frequency, equally resulting in a diversification from the sister population.

Forward Simulations

To evaluate the accuracy of the method, we carried out forward simulations using SLiM 2.1 under a range of population histories (Messer 2013). We simulated 10-kb DNA sequences evolving under the symmetric five-population tree provided in figure 1C. We fixed the effective population sizes at $N = 20{,}000$, as well as mutation and recombination rates of 2.36×10^{-8} mutations and 1×10^{-8} recombination events per generation per site. In order to simulate different types of sweeps, we introduced beneficial alleles at different initial frequencies (iAF), ranging from 1% to 5%. Their dominance coefficient was simulated to be recessive, dominant, or intermediate. According to these parameters, we let beneficial alleles to evolve conferring an advantage s, ranging from 0.001 to 0.01 into the ancient population ABCD (i.e., ancestral to populations A to D), or from $s = 0.005$ to 0.03 into the external branch leading to population A. Additionally, we evaluated how the impact of an increasing fraction of migrants per generation, m, from a distant population E into any of the two branches subject to selection, namely A or ABCD.

From the simulated scenarios, we sampled 10, 50, and 100 chromosomes for each population, corresponding to 5, 25, and 50 diploid individuals, respectively. Two approaches were applied for local gene tree inference. The first approach consisted in constructing NJ trees from phased chromosomes, with FastTree 2 (Price et al. 2010). The second pertained to an estimation of genotype distances, with a script implemented in-house to quantify allele frequency changes between pairs of diploid individuals, that is, null distances were assigned to sites with the same genotype, distances of one to sites with alternative homozygous genotypes, and half otherwise. Pairwise genotype distances were then used as input to reconstruct a NJ tree with FastME 2.1.4 (Lefort et al. 2015). Local gene trees based on genotype distances, and thus on unphased information, were only used in the section comparing LSDnorm to the PBS statistics. The remaining simulated scenarios are based on local gene trees build from phased chromosomes.

In order to calculate the PBS statistics, we followed the equation in figure 1A. Pairs of F_{ST} values were estimated with mstatspop v.0.1beta (20171109), an efficient C implementation to calculate statistics of variability, and that is freely available at https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html.

Application to Empirical Data

We applied LSD to data from the 1000 Human Genome Project (phase 3) (The 1000 Genomes Project Consortium 2015), restricting the analysis to 30 individuals from each of the following populations: CEU (northern and western European), CHB (Han Chinese), PJL (Punjabi), LWK (Luhya), and YRI (Yoruba). These comprise a total of 150 individuals spread across Africa, Europe, south and east Asia.

As the available VCF files encode haplotype phase information for the 150 individuals (ftp://ftp.1000genomes.ebi.ac.uk/Vol1/ftp/release/20130502/), we transformed their genotypes into 300 haploid fasta sequences. This fasta file was further split into 10-kb genomic windows, with a 5-kb stepwise overlap. Each fasta file was used as input to build a separate local gene tree, with FastTree 2 (with default parameters plus the GTR option) (Price et al. 2010). For each local gene tree, we calculated LSDnorm in two lineages, namely the external lineage leading to CEU, and the internal lineage including the ancestral population that left Africa (hereafter, referred to as Eurasians). Genes overlapping with outlier LSDnorm regions were further analyzed for functional enrichment, using WebGestalt (Zhang et al. 2005).

Implementation

The LSD framework is implemented in C, and available for download under the GPL license at https://bitbucket.org/plibrado/LSD. The program requires two input files, one including a collection of local gene trees (potentially multifurcating), and a tabulated file summarizing the topology of the population tree. As it leverages pairwise genetic distances, its computational time scales quadratically with the number of terminal lineages. In a standard CPU processor, LSD processes thousands of genealogies in minutes.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank the anonymous reviewer, Dr Joshua Schraiber and Dr Nathan Wales for their helpful comments on the manuscript. This work was supported by the Danish Council for Independent Research, Natural Sciences (Grant 4002-00152B); the Danish National Research Foundation (Grant DNRF94); Initiative d'Excellence Chaires d'attractivité, Université de Toulouse (OURASI); the Villum Fonden miGENEPI research project; and the European Research Council (ERC-CoG-2015-681605).

Author Contributions

P.L. and L.O. designed the study; P.L. performed all analyses, with input from L.O.; P.L. coded the LSD method; P.L and L.O. wrote the text.

References

Abbas H, Ahmad M, Ahmad M. 1983. Persistence of high intestinal lactase activity in Pakistan. *Hum Genet*. 64(3):277–278.

Agyemang C, Bhopal R. 2003. Is the blood pressure of people from African origin adults in the UK higher or lower than that in European origin white people? A review of cross-sectional data. *J Hum Hypertens*. 17(8):523–534.

Auburger G, Gispert S, Lahut S, Ömür Ö, Damrath E, Heck M, Başak N. 2014. 12q24 locus association with type 1 diabetes: *SH2B3* or *ATXN2*? *World J Diabetes* 5(3):316–327.

Bae CJ, Douka K, Petraglia MD, Bae CJ, Petraglia MD. 2017. On the origin of modern humans: Asian perspectives. *Science* 358(6368): eaai9067.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet*. 40(3):340–345.

Berg JJ, Coop G. 2015. A coalescent model for a sweep of a unique standing variant. *Genetics* 201(2):707–725.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 74(6):1111–1120.

Blanton SH, Heckenlively JR, Cottingham AW, Friedman J, Sadler LA, Wagner M, Friedman LH, Daiger SP. 1991. Linkage mapping of autosomal dominant retinitis pigmentosa (RP1) to the pericentric region of human chromosome 8. *Genomics* 11(4):857–869.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M. 2010. Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics* 186(1):241–262.

Brinkworth JF, Barreiro LB. 2014. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Curr Opin Immunol*. 31:66–78.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res*. 20(3):393–402.

Clément K, Le Stunff C, Meirhaeghe A, Dechartres A, Ferrieres J, Basdevant A, Boitard C, Amouyel P, Bougnères P. 2009. In obese and non-obese adults, the cis-regulatory rs361072 promoter variant of *PIK3CB* is associated with insulin resistance not with type 2 diabetes. *Mol Genet Metab*. 96(3):129–132.

Colagiuri S, Brand Miller J. 2002. The "carnivore connection"–evolutionary aspects of insulin resistance. *Eur J Clin Nutr*. 56(S1):S30–S35.

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. 2016. SweepFinder2: Increased sensitivity, robustness and flexibility. *Bioinformatics* 32(12):1895–1897.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A*. 101(4):975–979.

Fan S, Hansen MEB, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of recent human adaptation. *Science* 354(6308):54–59.

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193(3):929–941.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol*. 31(5):1275–1291.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. 2016. Detection of human adaptation during the past 2000 years. *Science* 354(6313):760–764.

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al. 2015. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524(7564):216.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, Filippo C. d, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet*. 11(2):e1005004.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979):710–722.

Grün R, Stringer C, McDermott F, Nathan R, Porat N, Robertson S, Taylor L, Mortimer G, Eggins S, McCulloch M. 2005. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *J Hum Evol*. 49(3):316–334.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555):207–211.

Hermisson J, Pennings PS. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* 169(4):2335–2352.

Hershkovitz I, Weber GW, Quam R, Duval M, Grün R, Kinsley L, Ayalon A, Bar-Matthews M, Valladas H, Mercier N, et al. 2018. The earliest modern humans outside Africa. *Science* 359(6374):456–459.

Heyer E, Brazier L, Ségurel L, Hegay T, Austerlitz F, Quintana-Murci L, Georges M, Pasquet P, Veuille M. 2011. Lactase persistence in central Asia: Phenotype, genotype, and evolution. *Hum Biol*. 83(3):379–392.

Hofer T, Ray N, Wegmann D, Excoffier L. 2009. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet*. 73(1):95–108.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116(1):153–159.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132(2):583–589.

Ihara M, Yamasaki N, Hagiwara A, Tanigaki A, Kitano A, Hikawa R, Tomimoto H, Noda M, Takanashi M, Mori H, et al. 2007. Sept4, a component of presynaptic scaffold and Lewy bodies, is required for the suppression of alpha-synuclein neurotoxicity. *Neuron* 53(4):519–533.

Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. *Genetics* 120(3):819–829.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. *Genetics* 123(4):887–899.

Kim Y, Stephan W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160(2):765–777.

Kullo IJ, Shameer K, Jouni H, Lesnick TG, Pathak J, Chute CG, de Andrade M. 2014. The *ATXN2-SH2B3* locus is associated with peripheral arterial disease: An electronic medical record-based genome-wide association study. *Front Genet*. 5:166.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513(7518):409.

Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol*. 32:2798–2800.

Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74(1):175–195.

Marciniak S, Perry GH. 2017. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet*. 18:659–674.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528(7583):499–503.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature* 351(6328):652–654.

Messer PW. 2013. SLiM: Simulating evolution with selection and linkage. *Genetics* **194**(4):1037–1039.

Nei M, Maruyama T. 1975. Lewontin-Krakauer test for neutral genes. *Genetics* 80(2):395.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. *Nature* 541(7637):302–310.

Orrù V, Steri M, Sole G, Sidore C, Virdis F, Dei M, Lai S, Zoledziewska M, Busonero F, Mulas A, et al. 2013. Genetic variants regulating immune cell levels in health and disease. *Cell* 155(1):242–256.

Pace D. 2014. Leishmaniasis. *J. Infect* 69(1 Suppl):S10–S18.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192(3):1065–1093.

Pavlidis P, Živković D, Stamatakis A, Alachiotis N. 2013. SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol*. 30(9):2224–2234.

Peyrégne S, Boyle MJ, Dannemann M, Prüefer K. 2017. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res*. 27:1563–1572.

Pierce EA, Quinn T, Meehan T, McGee TL, Berson EL, Dryja TP. 1999. Mutations in a gene encoding a new oxygen-regulated photoreceptor protein cause dominant retinitis pigmentosa. *Nat Genet*. 22(3):248–254.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

Racimo F. 2016. Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* 202(2):733–750.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinforma Oxf Engl*. 19(12):1572–1574.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419(6909):832–837.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164):913–918.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, et al. 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346(6213):1113–1118.

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum Genomics* 1(4):274.

Sikora M, Seguin-Orlando A, Sousa VC, Albrechtsen A, Korneliussen T, Ko A, Rasmussen S, Dupanloup I, Nigst PR, Bosch MD, et al. 2017. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* 358:659–662.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res*. 23(1):23–35.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, Lan Y, Fields CJ, Hui JHL, Zhang W, et al. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat Ecol Evol*. 1(5):121.

Tataru P, Simonsen M, Bataillon T, Hobolth A. 2017. Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data. *Syst. Biol*. 66:e30–e46.

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

Vatsiou AI, Bazin E, Gaggiotti OE. 2016. Changes in selective pressures associated with human population expansion may explain metabolic and immune related pathways enriched for signatures of positive selection. *BMC Genomics* 17:504.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet*. 47:97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. 4(3):e72.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet*. 3(6):e90.

Wong MD, Shapiro MF, Boscardin WJ, Ettner SL. 2002. Contribution of major diseases to disparities in mortality. *N Engl J Med*. 347(20):1585–1592.

Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol Biol Evol*. 14(7):717–724.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of fifty human exomes reveals adaptation to high altitude. *Science* 329:75–78.

Zhang B, Kirov S, Snoddy J. 2005. WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*. 33(web server issue):W741–W748.

Zhang G, Muglia LJ, Chakraborty R, Akey JM, Williams SM. 2013. Signatures of natural selection on genetic variants affecting complex human traits. *Appl. Transl. Genomics* 2:78–94.