Received: 10 June 2018 | Revised: 17 October 2018 | Accepted: 18 October 2018

DOI: 10.1111/1755-0998.12957

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

# McSwan: A joint site frequency spectrum method to detect and date selective sweeps across multiple population genomes

Rémi Tournebize[1] | Valérie Poncet[1] | Mattias Jakobsson[2,3] | Yves Vigouroux[1] | Stéphanie Manel[4]

[1]IRD, University of Montpellier, UMR DIADE BP 64501, Montpellier Cedex 5, France

[2]Department of Organismal Biology and SciLifeLab, Uppsala University, Uppsala, Sweden

[3]Centre for Anthropological Research, Department of Anthropology and Development Studies, University of Johannesburg, Auckland Park, South Africa

[4]EPHE, PSL Research University, CNRS, University of Montpellier, Montpellier SupAgro, IRD, INRA, UMR:5175 CEFE, Montpellier, France

**Correspondence**
Valérie Poncet, IRD, University of Montpellier, UMR DIADE BP 64501, Montpellier Cedex 5, France.
Email: valerie.poncet@ird.fr

**Funding information**
Forskningsrådet om Hälsa, Arbetsliv och Välfärd; the Knut and Alice Wallenberg Foundation; Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche

**Abstract**

Inferring the mode and tempo of natural selection helps further our understanding of adaptation to past environmental changes. Here, we introduce McSwan, a method to detect and date past and recent natural selection events in the case of a hard sweep. The method is based on the comparison of site frequency spectra obtained under various demographic models that include selection. McSwan demonstrated high power (high sensitivity and specificity) in capturing hard selective sweep events without requiring haplotype phasing. It performed slightly better than SweeD when the recent effective population size was low and the genomic region was small. We then applied our method to a European (CEU) and an African (LWK) human re-sequencing data set. Most hard sweeps were detected in the CEU population (96%). Moreover, hard sweeps in the African population were estimated to have occurred further back in time (mode: 43,625 years BP) compared to those of Europeans (mode: 24,850 years BP). Most of the estimated ages of hard sweeps in Europeans were associated with the Last Glacial Maximum and were enriched in immunity-associated genes.

KEYWORDS

age of selection, coalescent process, high-performance computing, human, selective sweep, site frequency spectrum

## 1 | INTRODUCTION

Positive selection drives an increase in beneficial traits, allowing species to adapt to varying climate or pathogenic pressures (Sabeti et al., 2006), for example. Determining the mode and tempo of positive selection can generate insight into events that have shaped the phenotypes and genotypes of current species and populations. Specifically, correlating the age of selection with known past environmental changes can improve our understanding of how adaptation occurs in response to environmental change (Ormond, Foll, Ewing, Pfeifer, & Jensen, 2016).

Several methods have been developed to: (a) detect selection in single or multiple populations (Duforet-Frebourg, Bazin, & Blum, 2014; Haasl & Payseur, 2016; Pavlidis, Živković, Stamatakis, & Alachiotis, 2013), (b) estimate the strength of selection (Vitalis, Gautier, Dawson, & Beaumont, 2014), and (c) distinguish whether the advantageous allele arose from novel mutations (hard sweeps) or from standing variation (soft sweeps; Peter, Huerta-Sanchez, & Nielsen, 2012). However, few methods estimate the age of selection by scanning individual genomes or focus on multiple populations (Supporting Information Table S1). Selection event dating methods differ with regard to the type of selection event they estimate, the nature of the selection they target (complete/incomplete, disruptive/directional) and the statistics they use (Chen, Hey, & Slatkin, 2015; Nakagome et al., 2016; Ormond et al., 2016; Peter et al., 2012; Przeworski, 2003; Smith, Coop, & Stephens, 2016; Supporting Information

Table S1). Recent approaches implementing hidden Markov models leverage haplotype lengths and the accumulation of mutations to estimate the age of beneficial alleles (Chen et al., 2015) or the time to the most recent common ancestor ($T_{\mathrm{MRCA}}$) that carried the beneficial allele (Smith et al., 2016). Yet, haplotype-based methods might be challenged by phasing quality (switch error rate, i.e., rate of phase misassignment between variants). They could also be less sensitive to ancient sweeps due to the rapid breakdown in linkage disequilibrium over time (Chen, Patterson, & Reich, 2010). Other approaches have implemented approximate Bayesian computations (ABC) to estimate the timing of selection based on various combinations of linkage disequilibrium statistics and statistics derived from the site frequency spectrum (Nakagome et al., 2016; Ormond et al., 2016; Peter et al., 2012; Przeworski, 2003). These methods allow estimation of the age of selection for already known candidate regions. Here, we developed the **M**ultiple-**C**ollision Coalescent **SW**eep **AN**alyzer (McSwan) method for whole-genome scan analysis, which can simultaneously identify hard selective sweeps over multiple populations while also estimating the age of the selective sweeps.

Hard selective sweeps (Smith & Haigh, 1974) constitute the selection-driven evolutionary process whereby a novel mutation increases in frequency in a population until it reaches fixation, while carrying along physically linked alleles. Consequently, the site frequency spectrum of a population haplotype that contains a positively selected mutation is expected to be skewed towards extreme allele frequencies (Fay & Wu, 2000; Nielsen et al., 2005). Our approach applies this property to classify locally observed site frequency spectra to a neutral or selective demographic model, and to predict the age of selection.

We applied the method to detect and date selective sweeps in two human populations: a population with north-western European ancestry (CEU) and Luhya individuals from Webuye, Kenya (LWK), and we found a strong correlation between the temporal distribution of hard sweeps in the ancestral CEU population and the environmental changes that occurred in Europe during the late glacial period —a pattern that was not found in the Luhya population.

## 2 | MATERIALS AND METHODS

### 2.1 | Rationale

The **M**ultiple-**C**ollision **C**oalescent **SW**eep **AN**alyzer (McSwan) is a genome scan-based approach, which simultaneously detects regions which likely experienced hard selective sweeps and estimates their age under a molecular clock assumption. McSwan implements an ensemble genome scan approach by iterating multiple genome scans over adjacent windows of various lengths and offsets (starting positions). For any scan window *i*, the encompassed biallelic polymorphic sites are summarized within a site frequency spectrum (SFS). The observed SFS is then classified to one of the neutral or selective models under the assumption that the given window *i* maps a haplotype which has not recombined internally. Considering a single population of *m* haploid individuals, SFS is defined as the vector ($\nu_1$, $\nu_2$, …, $\nu_{m-1}$), where $\nu_j$ is the number of biallelic SNPs for which exactly *j* gametes carry a mutation, divided by the total number of considered biallelic SNPs (normalized SFS). This mutation can be set as the derived allele if the ancestral/derived status is known, or else the allele found at minor frequency in the sampled population. In case of a multipopulation model, SFS is defined as the multidimensional joint distribution of polymorphic sites, that is, for *N* multiple populations of sizes $M_1$, $M_2$, …, $M_N$ haploid individuals, SFS is the vector $\left(\nu_{0,0,\ldots,1},\nu_{0,0,\ldots,2},\ldots,\nu_{0,1,\ldots,0},\ldots,\nu_{M_1,M_2,\ldots,M_m-1}\right)$ and the number of elements in SFS equals $-2+(M_1+1)\times(M_2+1)\times\ldots\times(M_N+1)$. In this population set, a minor mutation is that found at minor frequency throughout *all* the populations.

The site frequency spectrum expected in a sampled population which experienced a hard selective sweep is modelled using a modification of the time-continuous coalescent process: the coalescent with multiple collisions or $\Lambda$-coalescent (Pitman, 1999). McSwan implements a Kingman approximation of the $\Lambda$-coalescent assuming an infinite allele mutation model and a Wright–Fisher population model, thus facilitating integration with other Kingman coalescent simulators such as *MS* (Hudson, 2002) or *MSMS* (Ewing & Hermisson, 2010).

Practically, the McSwan approach comprises two steps (Figure 1).

First, expected site frequency spectra (SFS) are simulated based on a user-defined neutral demographic model with and without population-specific selection. We define a set of $P+1$ models $[\Phi_o,(\Phi_p)_{\forall p\in P}]$ with $\Phi_o$ being the neutral model and $\Phi_{p\neq o}$ any population-specific selective model with a user-defined prior distribution of sweep ages. A set of *N* SFSs is simulated for each model. To identify the SFS signatures across the neutral and selective models, we trained a linear discriminant analysis (LDA) over all simulations from the neutral and selective models. LDA determines a set of linear combinations among SFS bins that maximize the variance between the neutral model and each of the selective models, thus allowing SFS classification according to the SFS model class. The most likely model can then be predicted for any SFS with the same dimensionality. The LDA robustness is increased by performing a calibration on a subset of orthogonal components extracted from a preliminary principal component analysis (PCA) over all simulations from the neutral and selective models. This PCA-LDA allows us to reduce overfitting in the LDA models (Yang & Yang, 2003). Note that LDA assumes equal variance for each SFS element between models which can be sometimes violated, leading to classification bias in favour of the model having the highest variance. This bias could ultimately be corrected using quadratic discriminant analyses (QDA) instead. For each independent selective model $\Phi_{p\neq o}$, the relationship between the sweep ages and SFSs is characterized using a partial least squares regression (PLSR, R package *pls*). PLSR is trained separately for each model. PLSR determines a set of linear combinations of SFS bins that explain the maximum variance in the sweep age vector. The best subset of features retained from the PLSR training can be automatically determined by McSwan.

Second, local empirical SFSs are computed along the genome from the real data set for each sliding window of size ***L*** base pairs (bp), using within-window SNPs (Figure 1b). Sliding windows are separated by $L/A$ bp, with $A$ denoting the number of offsets for the scan, thus allowing capture of the recombination and the selection signal variability. If a genomic region mapped by a given window is assigned to a selective model according to the LDA prediction (Figure 1b), the posterior classification probability (i.e., how likely this selective model could have generated the observed SFS) is stored for each SNP encompassed by the window, as well as their associated sweep age, as estimated using the PLSR model specific to the detected selective model. A genomic site is therefore potentially scanned by $A \times L$ windows. We then calculated a score for each SNP as the proportion of overlapping windows classified as being under selection (henceforth "selective windows"), divided by the number of all overlapping windows (Figure 1c). Sweep regions are inferred by merging contiguous fragments which were assigned to the same selective demographic model based on a score cut-off that can be estimated using cross-validation analyses (Figure 1c). The age of the whole inferred sweep region is estimated as the average of sweep ages estimated for the overlapping selective windows, weighted by their LDA posterior class probability. In this study, the "sweep age" refers to the time to the most recent common ancestor ($T_{MRCA}$) of lineages carrying the adaptive mutation, and this time will be shorter than or equal to the age of the adaptive mutation (see supplementary text 1).

## 2.2 | Inputs

McSwan is an R package (GitHub: https://github.com/sunyatin/McSwan). McSwan analysis inputs include the following: (a) a set of biallelic SNPs wrapped into a VCF file, and (b) a neutral demographic model given as an *MS*-formatted string comprising any possible demographic switch natively available in *MS* (Hudson, 2002). In its current implementation, McSwan discards SNPs having at least one missing genotype and applies to well-covered re-sequencing (e.g., >7×) data sets.

## 2.3 | Method validation

We assessed the accuracy of McSwan to detect selective sweeps and date sweep events using simulated genomic fragments (pseudo-observed data sets, PODs) under neutral and selective demographic

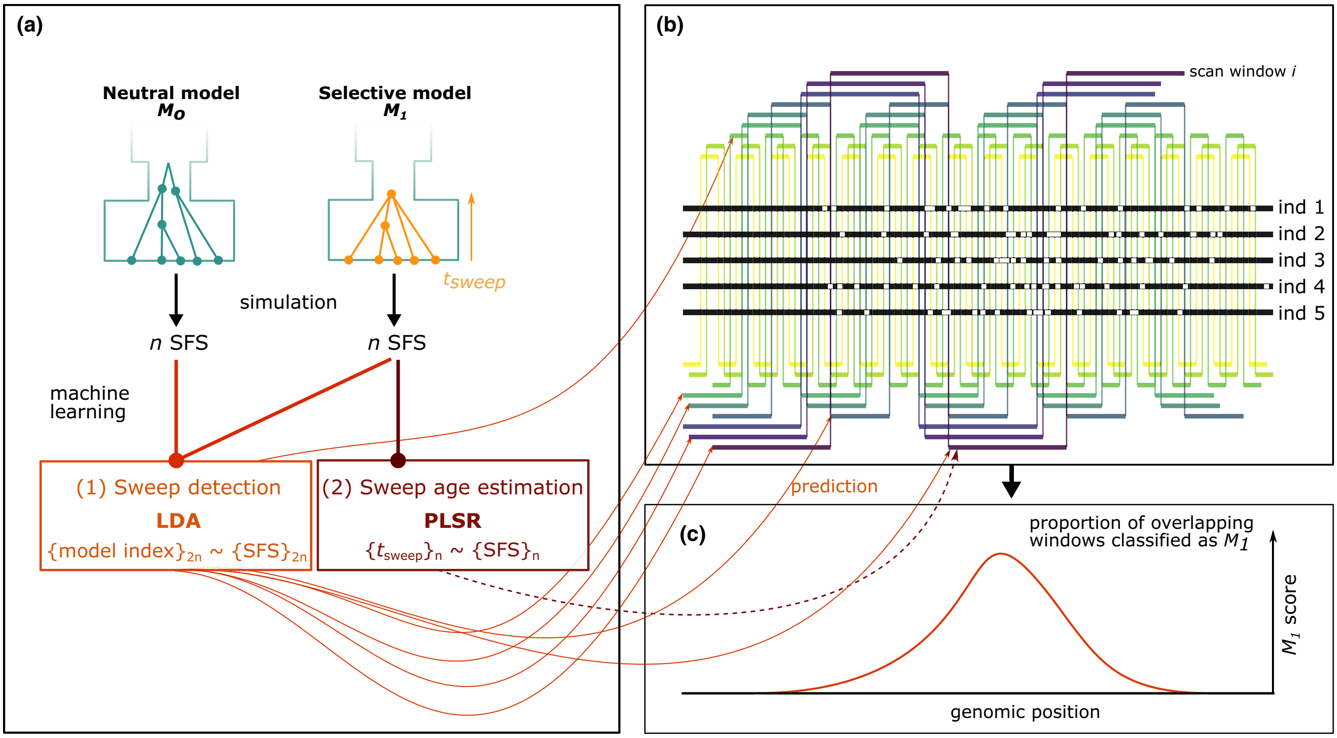

**FIGURE 1** Illustration of the genome scan approach. For illustration purposes, we consider a single population. (a) Generation of expected SFSs. A total of $n$ site frequency spectra (SFS) are simulated under both the neutral ($M_o$) and the selective ($M_1$) demographic models. For $M_1$, the age of the selection event ($t_{sweep}$) is randomly sampled in a prior distribution. A linear discriminant analysis (LDA) is then calibrated to classify the simulated SFSs to their demographic model of origin. For the selective model $M_1$, a partial least squares regression (PLSR) is fitted to predict the sweep age as a function of the linear combination of SFS elements. (b) Genome scan. Local SFSs are calculated for individual genomes (black segments) comprising polymorphic sites (white squares) using different window lengths (segments with yellow to purple tones). Window-specific SFSs are then classified as neutral or selective using the linear discriminant model trained on the simulated data set. If the predicted model is the selective model $M_1$, the sweep age is estimated using fitted partial least squares regression. The analysis is replicated for different window sizes and positions in the windows (offsets). (c) Analysis across window sizes and offsets. A score for each SNP is calculated as the proportion of overlapping selective windows over all overlapping windows. We then build the contiguous genomic fragments which likely underwent the same hard selective sweep in the past. The sweep age for the inferred sweep region is estimated by the weighted average of the sweep ages estimated for all overlapping selective windows [Colour figure can be viewed at wileyonlinelibrary.com]

models. We used the published demographic model for CEU and LWK populations, fitted with the published parameter point estimates (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013). A hundred independent data sets comprising 40 chromosomes of 1 Mb length were simulated. Simulations assumed either neutrality or selection in each population and were performed using *MSMS* software (Ewing & Hermisson, 2010). *MSMS* is more flexible than McSwan for simulating different selection scenarios. It is a time-continuous Kingman coalescent simulating selection by first tracing the frequency trajectory of a beneficial allele and sampling a random coalescent genealogy for neutral linked markers conditioned on the beneficial allele frequency. *MSMS* is faster than forward simulators like *SLiM* (Haller & Messer, 2017) which allows simulation of both complete and partial sweeps, and it is more flexible than other coalescent simulators like *mbs* which do not take the population structure into account (Teshima & Innan, 2009). We let the recombination rates range from $\mu/10$ to $10\mu$, where μ is the neutral substitution rate set at $2.5 \times 10^{-8}$/individual/generation. The ages of selection were randomly drawn from 1 to 2,000 generations BP. The sweep detection accuracy was estimated by calculating the specificity (i.e., true negative rate) and the sensitivity (i.e., true positive rate). The parameter inference accuracy was estimated using the mean-normalized root mean square error (NRMSE; Walther & Moore, 2005), defined as $\frac{1}{E(O)}\sqrt{E\left((O-E)^2\right)}$, where $O$ is the vector of simulated values and $E$ the paired vector of estimated values.

### 2.3.1 | Statistical inference parameters

A total of 10,000 unfolded joint SFSs were simulated per demographic model. LDA was trained using 168 orthogonal components. Genome scans were iteratively performed using $L = 20$ scan window lengths, uniformly sampled between $10^4$ and $2{\times}10^5$ bp, and $A = 30$ offsets. We considered only windows containing more than 10 SNPs. The selective or neutral status of windows was determined using two cut-offs for the score (i.e., for each SNP, the proportion of overlapping windows classified as selective divided by all overlapping windows) of 0.5 and 0.2 for the CEU and LWK populations, respectively (trade-off between specificity and sensitivity).

### 2.3.2 | Sensitivity analyses

We assessed the sensitivity of the detection performance and estimation accuracy to sample sizes ranging from 5 to 50 diploid individuals per population and selection strengths $s$ ranging from $5 \times 10^{-4}$ to 1. Genomic fragments were also simulated under a soft sweep from standing variation model with an initial frequency of the beneficial mutation, $f$, ranging from $10^{-3}$ to 0.5 in order to assess the potential analytical bias induced by standing variation.

## 2.4 | Biological application

We used 40 individuals from two human populations from the CEPH and HapMap-1 K collections: Utah residents with north-western European ancestry (CEU) and Luhya individuals in Webuye, Kenya (LWK). Genotypes from the CEU and LWK samples were called with 9.1 ± 2.7-fold and 7.1 ± 2.4-fold read depths on average. We retained only autosomal biallelic SNPs. Based on the ancestral allele annotation (The 1000 Genomes Project Consortium, 2015), polymorphisms were polarized using a custom Python script. Polymorphisms with ambiguous or missing ancestral annotations were discarded (representing a proportion of 12.2% SNPs). We first (re)validated the fit of the demographic model to the polymorphisms observed in our sampled CEU and LWK individuals by performing a goodness-of-fit test (R package *abc* (Csilléry, François, & Blum, 2012)). We also assessed the robustness of sweep detection and sweep age estimation to alternative demographic histories (Gazave et al., 2014; Keinan & Clark, 2012; Supporting Information Text 3.8).

### 2.4.1 | Enrichment analyses

We tested for excessive association of detected regions that had experienced a selective sweep with particular gene function categories, disease or gene cell expression. Enrichment tests were performed at the gene-wise level (i.e., counting every gene only once even if they overlapped with multiple candidate SNPs), for each population separately, using *Gowinda* 1.12 (Kofler & Schlotterer, 2012).

### 2.4.2 | Empirical result comparison

The sweeps detected by McSwan in the two human population data sets were compared with those obtained using two other recently published methods with the aim of detecting positive selection in multiple populations: (a) a composite-likelihood approach relying on cross-population differentiation metrics, XP-CLR (Chen et al., 2010), and (b) a machine learning method performing hierarchical boosting on multiple selection scores (Pybus et al., 2014).

### 2.4.3 | Environmental associations

We investigated putative associations between the temporal distribution of hard sweeps in the CEU population and the historical series of climatic and vegetation changes in Europe. A set of 14 time series of environmental variables was chosen to represent major features of the historical environment encountered by ancestral European populations, including mean temperature, variability, humidity, vegetation dynamics and composition (Sankararaman et al., 2014; Supporting Information Figure S2). Sweep events were counted by successive time bins of $\Delta t$ = 1,000 years length under the mutual independence assumption (Poisson point process), with a generation time set at 25 years. The regression of $S$ (i.e., the resulting time series of hard sweep counts) over the standardized environmental predictors was performed under a generalized linear model with Poisson-distributed errors. To simultaneously estimate regression coefficients and select the most important environmental features, a least absolute shrinkage and selection operator (LASSO) regularization was performed using the R package *glmnet*. To assess the impact of the sweep age estimation uncertainty on the putative climatic association, we performed a resampling approach by estimating the evidence ratio of the generalized linear model assuming the dependency of $S$ to the historical temperature series $T$, $\mathscr{M}_1 : [\log(E(S|T)) = a_0 + a_1 \cdot T]$, versus a model, where $S$ is a constant function of time, $\mathscr{M}_0 : [\log(E(S|T)) = a_0]$. Because the correlation between the sweep age density and the environmental conditions depends on the binning procedure and the generation time assumption, we further assessed the robustness of this association by using 50 equally spaced generation times (ranging from 20 to 30 years) and 50 equally spaced time bin sizes (100 to 2,500 years).

### 2.4.4 | Neanderthal ancestry

We investigated putative archaic Neanderthal introgression into the ancestral CEU population within our detected hard sweep regions using the published regions of Neanderthal ancestry in the CEU population (estimated from the 1000 Genomes Project data set and the 52× sequenced Altai Neanderthal genome (Prüfer et al., 2014; Sankararaman et al., 2014)). The significance of the odds ratios in favour of Neanderthal ancestry enrichment in the hard sweep regions was estimated using an exact Fisher's ratio test.

## 3 | RESULTS

### 3.1 | Method validation

The method was validated by simulating 300 data sets of 40 chromosomes sampled in two populations (20 per population). We used a previously published demographic model for these simulations, targeting the CEU and LWK populations (Figure 2a; Excoffier et al.,

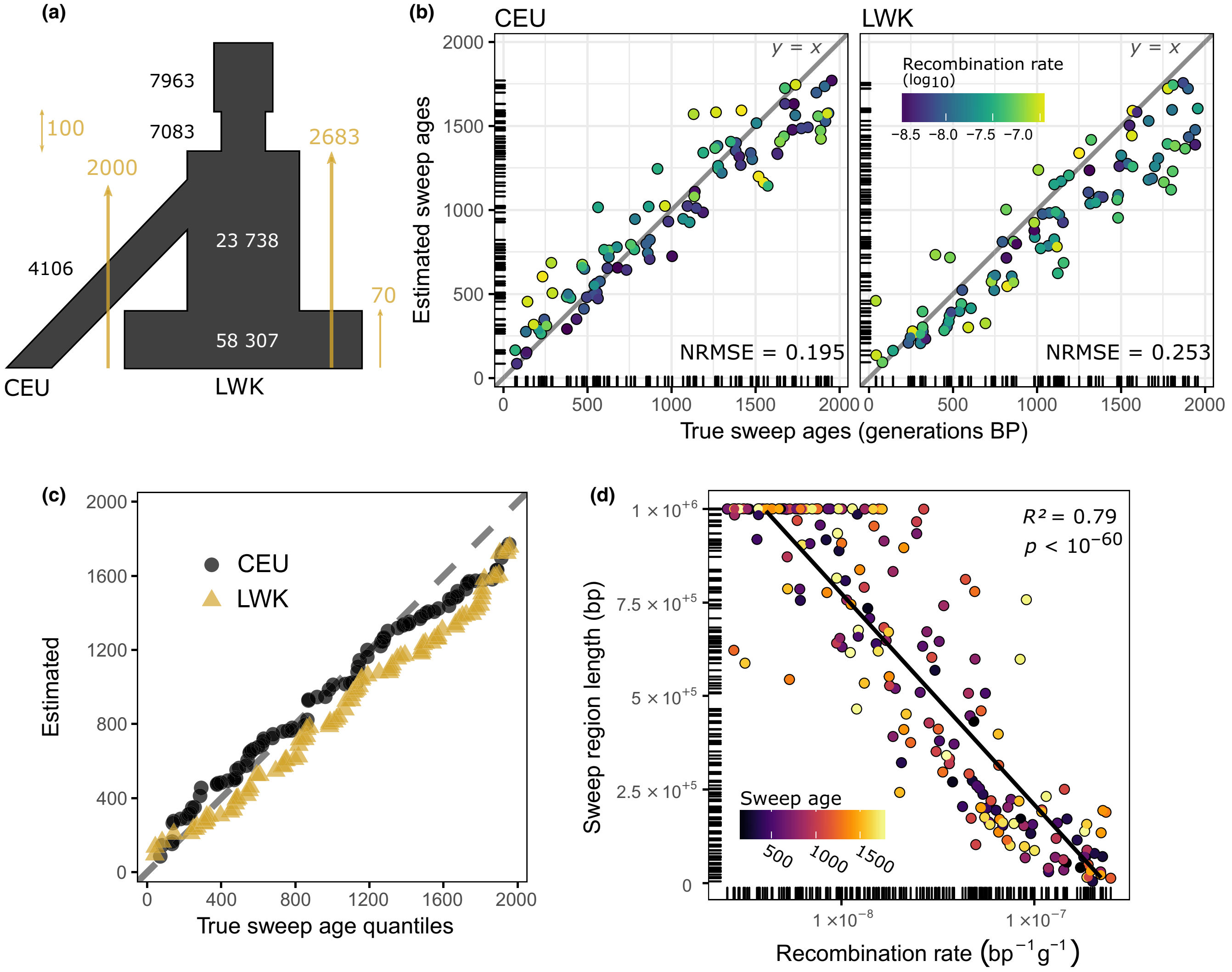

**FIGURE 2** Performance of McSwan evaluated from simulated data sets (codominant selection coefficient $s$ = 1). (a) Demographic model used to simulate pseudo-observed data sets on two populations (CEU and LWK, Excoffier et al. (2013), scenario A). (b) Estimated sweep ages as a function of the simulated sweep ages (for the true positive findings), for varying recombination rates ($r$). The bisector $y = x$ represents the ideal case where estimated values match the simulated data. (c) Q–Q plot of the estimated versus uniformly simulated sweep ages. (d) The length of the sweep regions is significantly and negatively correlated with the logarithm of the simulated recombination rate ($R^2$=0.79, $p < 10^{-60}$) [Colour figure can be viewed at wileyonlinelibrary.com]

2013). Based on these simulated data sets, we applied our methodology to detect hard sweeps and their ages. The method allowed us to accurately detect hard selective sweeps (sensitivity and specificity >0.95; false discovery rate FDR <0.05). More than 97% of all detected sweeps included the position of the truly adaptive mutation. We also found limited biases when estimating the sweep age (Figure 2b). The normalized root mean square error (NRMSE) was 0.19 and 0.25 for the CEU and LWK selective models, respectively. We found no major distortion in the posterior distribution of sweep ages compared to the simulated uniform prior for the time span considered (Figure 2c). Finally, the logarithm of the sweep region lengths was strongly and significantly correlated with the simulated recombination rate ($R^2$ = 0.79, $p < 10^{-60}$, Figure 2d). The sweep age estimation error was significantly and negatively correlated with the sweep length (itself related to the recombination rate; Pearson's $r$ = –0.23, $p$ = 0.004). The distance between the centre of the sweep region and the adaptive mutation position was positively correlated with the sweep length (Pearson's $r$ = 0.40, $p < 10^{-8}$). McSwan was found to perform slightly better than SweeD (Pavlidis et al., 2013) in selective sweep detection (Supporting Information Text 2.2).

## 3.2 | Accuracy analyses

We used simulated data sets and found that variations in (a) the sampling size, (b) the initial frequency $f$ of the beneficial mutation, and (c) the selection strength impacted the sweep detection and age inference performance (Figure 3 and Supporting Information Figures S3–S5). The method was tested for sampling sizes ranging from 5 to 50 diploid individuals per population and was found to perform well even with relatively few individuals (Supporting Information Figure S5), that is, the specificity (>0.90) and power (>0.85) were stable across the tested sampling sizes. Using more than 20 individuals did not significantly improve the sweep detection performance and the sweep age estimation accuracy, indicating that using higher sampling sizes would not be useful when studying genetically homogeneous populations. The specificity was high and stable (>0.90) for soft sweeps as long as the initial frequency $f$ remained low ($f < 0.01$; Supporting Information Figure S4). The detection power significantly decreased with $f$, tending towards 0 for increasing $f$. Our method was effective for strong selection coefficients, so the detection power fell below 0.2 for $s < 0.005$ but significantly increased (>0.4) when $s > 0.05$ (Supporting Information Figure S3). The sweep estimation accuracy significantly increased with the sampling size and selection strength. The accuracy decreased with increasing $f$ (NRMSE > 1 for $f > 0.05$), but the estimation precision was affected by the low number of detected sweeps (<2) which is essential for the NRMSE calculation.

## 3.3 | Detection of hard sweeps in the CEU and LWK populations

We investigated the adaptive dynamics in two human populations: Utah residents with north-western European ancestry (CEU) and Luhya individuals in Webuye, Kenya (LWK), from the 1000 Genomes Phase 3 public variant call set (The 1000 Genomes Project Consortium, 2015). These two populations were chosen because of their contrasting historical environments and phenotypes with known molecular determinisms, like lactase persistence (Enattah et al., 2002, 2008 ; Tishkoff et al., 2007). The demographic model published by Excoffier et al. (2013) closely fitted the genomic data set (goodness-of-fit test: $p$ = 1). We then empirically checked that the sweeps detected for the CEU population along chromosome 2 satisfactorily agreed with two other methods used to detect positive selection: XP-CLR (Chen et al., 2010) and a multistatistical hierarchical boosting approach (Pybus et al., 2014). We therefore performed a method comparison using Cohen's κ index (Supporting Information Figure S7 and Table S2).

## 3.4 | Distribution of hard sweep counts

Among the 172 hard sweep regions detected across the autosomes (Supporting Information Table S3), 96% were found in the CEU

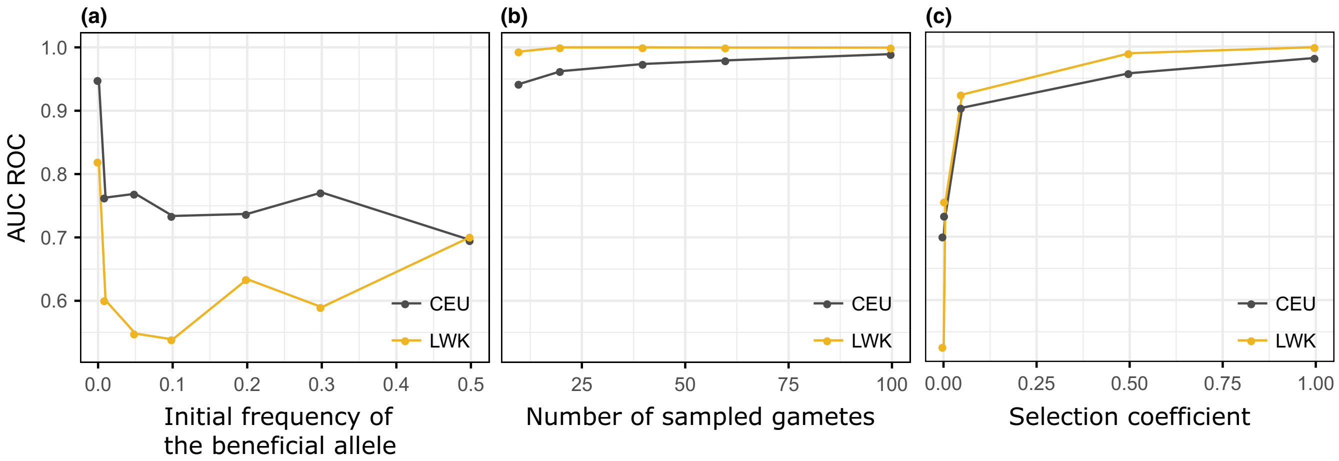

**FIGURE 3** Performance of McSwan evaluated from simulated data sets: (a) for varying initial frequencies of the beneficial allele, (b) for varying number of sampled gametes, and (c) for varying selection coefficients. The performance is evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve [Colour figure can be viewed at wileyonlinelibrary.com]

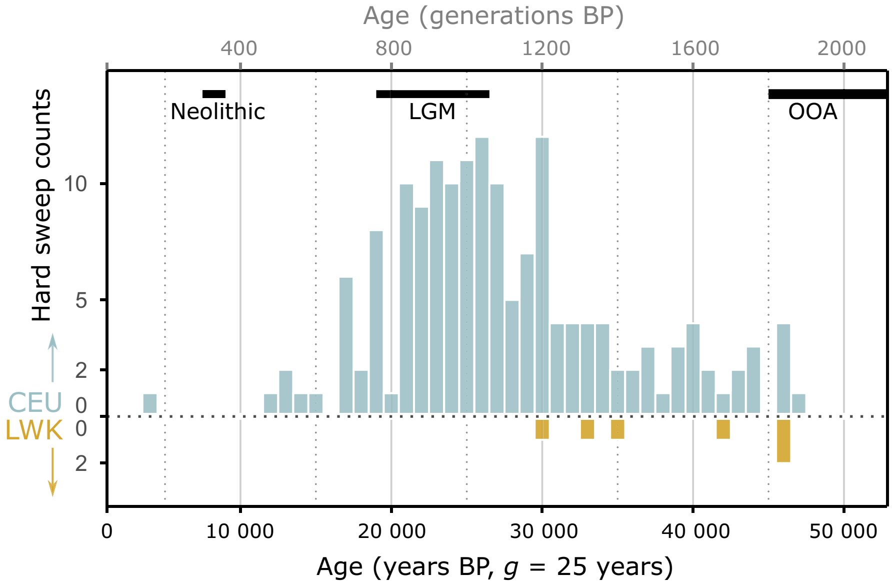

**FIGURE 4** Distribution of hard sweep counts in CEU (blue histogram) and LWK (yellow, mirrored scale) in the 0–50,000 years BP timeframe. Only hard sweeps posterior to the CEU-LWK divergence are represented. The calibration of sweep ages in years BP relies on a generation time of 25 years (Excoffier et al., 2013). Horizontal black segments represent climatic and cultural events assumed to have played a role in the recent evolution of human populations, with (right to left): (1) OOA: the "Out-of-Africa" period, ~51,000 years BP, 95% CI 45–69 ky BP (Gravel et al., 2011); (2) LGM: the Last Glacial Maximum period ~26,500–19,000 years BP (Clark et al., 2009); (3) Neolithic revolution in the Middle East and in North Africa, ~7,500–9,000 years BP (Tishkoff et al., 2007) [Colour figure can be viewed at wileyonlinelibrary.com]

population (Figure 4). The average hard sweep region length was over ~182 kb (median: 102 kb; 95% interquartile: 9–961 kb). There was no significant difference in the mean sweep length between the two populations (Mann–Whitney *U* test: W = 518, *p* = 0.87). The density in sweeps per autosome was significantly correlated with the chromosomal length ($R^2 = 0.71$; $p = 5{\times}10^{-7}$). The hard sweep age distribution mode was found to be significantly 1.7-fold older in LWK (43,625 years BP) than in CEU (24,850 years BP; Kolmogorov–Smirnov test: *D* = 0.699, *p* = 0.0025). Hard sweeps affecting the LWK ancestral populations occurred more likely following the Out-of-Africa (OOA) period which ended about 45,000 years BP. However, the estimation accuracy for these older LWK sweeps could be impacted by heterozygote identification due to the lower coverage depth (7 to 9×; Supporting Information Figure S6). The inference performed on randomly drawn haplotypes lowered the estimation accuracy of the detected LWK sweep ages around the Late Glacial Maximum, but had a minor impact on CEU sweep ages (Supporting Information Figure S6).

The hard sweep age distribution in CEU was centred on 14,000–28,000 years BP, with the highest hard sweep density noted around the Late Glacial Maximum (~24,500 years BP; Bos, Bohncke, Kasse, & Vandenberghe, 2001; Kasse, Huijzer, Krzyszkowski, Bohncke, & Coope, 1998). The distribution levelled off soon before the Holocene (~11,500 years BP) over the Neolithic revolution (~9,000 years BP) until present. A single sweep was detected after the Neolithic revolution (Figure 4).

As archaeological, historical and genetic evidence indicates that CEU populations experienced exponential growth since the Neolithic revolution, we assessed the robustness of our hard sweep detection and dating relative to alternative human demographic histories (Gazave et al., 2014; Keinan & Clark, 2012). Our inferences across the underlying demographic models were comparable and the hard sweep counts peaked at the LGM for all models (Supporting Information Text 4.3). More than 95% of the previously detected hard sweeps were also detected with the alternative demographic models.

### 3.5 | Characterization of selective sweeps: candidate genes and enriched cell types

The most recent sweep in the CEU population was detected at chromosome 2 (chr2q21.3):135.811–136.781 Mb. The sweep region fully overlapped with the lactase-coding gene (LCT, 136.545–136.594, GRCh37 reference) and four other protein-coding genes (DARS, MCM6, R3HDM1 and UBXN4; Supporting Information Figure S11). The mutation responsible for lactase persistence (rs4988235, chr2:136,608,646 (Tishkoff et al., 2007)) was found to have the 3.7% highest selection score among all the population-specific SNPs within the sweep boundaries (Supporting Information Figure S11). The age of this sweep was estimated at around 4,250 years BP (95% CI: [3,700–17,680]). Another hallmark of recent positive selection in human populations, associated with skin and eye pigmentation variation (SLC24A5 gene) (Beleza et al., 2013; Canfield et al., 2013; Nakagome et al., 2016; Smith et al., 2016) was detected in a sweep encompassing the (chr15q21.1):48.322–48.513 Mb region (Supporting Information Figure S12), which was dated at around 17,190 years BP [13,905–20,650].

We found that genes associated with the hard sweep regions in CEU were relatively more frequently expressed in immune cells (Table 1). The enrichment analyses suggested that genes included in the CEU hard sweep regions were more frequently expressed in eleven cell types (FDR-corrected $p < 0.05$). Among the eleven cell types, six were associated with the human immune system and three with the epidermis (Table 1). The gene ontology (GO) biological process related to the innate immune response (GO0045087) was not found to be enriched ($P_{\text{FDR}}$ = 0.387), but some genes were found significantly associated with processes like deamination, keratinization, protein and sugar metabolism, cellular cycle regulation and specific immunity processes like chemokine or cytokine activities (Table 1). Moreover, eight diseases were significantly enriched in selection signals and included lactose intolerance, rhesus blood group incompatibility, allergic hypersensitivity and hypothyroidism with goitre. No comparable enrichments were found for the LWK population except in genes associated with Parkinson disease (KANSL1, CRHR1, CRHR1-IT1, MAPT, MAPT-AS1 and SPPL2C, $P_{\text{FDR}} < 0.002$), thus calling for future functional analyses.

## 3.6 | Temporal correlations between environmental variables and hard selective sweeps

The temporal distribution of hard sweep counts in the CEU population was correlated with four environmental variables sampled every 1,000 years since 50,000 years BP (Supporting Information Figure S8). By decreasing order of importance, the hard sweep count was higher during periods of lower annual average temperature (standardized coefficient of regression = –0.366), higher snow depth (0.341) and in tundra/boreal forest-type environments (–0.324). Hard sweeps were denser during periods of higher short-term climatic instability (i.e., between-year mean annual temperature standard deviation, 0.133). The temperature change velocity, representing the direction and rate at which organisms must move to maintain constant climate (i.e., the ratio of spatial temperature gradient over temperature change rate (Loarie et al., 2009)), did not increase the hard sweep distribution prediction power (Supporting Information Figure S8). The evidence ratio (*ER*), which measures the relative penalized likelihood in favour of the association between the temporal distribution of CEU hard sweeps and the temperatures in Europe, was large regardless of the time bin length and the generation time used to analyse the association ($ER > 9 \times 10^9$; Supporting Information Figure S10). When accounting for the uncertainty in the hard sweep age estimates by resampling the sweep ages 999 times within their 95% confidence intervals, the association was still supported by a high evidence ratio ($ER > 2 \times 10^{12}$; Supporting Information Figure S9).

## 3.7 | Depletion in Neanderthal ancestry

When testing for potential Neanderthal ancestry enrichment along the detected CEU hard sweep regions, we instead found evidence of a significant depletion (13.3% of the sweeps overlapped Neanderthal haplotypes) relative to the random expectation at the genome scale (32.5%; Fisher's exact test: odds ratio OR = 0.319, $p = 1.3 \times 10^{-8}$).

# 4 | DISCUSSION

Few methods so far have sought to detect the age of selection (Nakagome et al., 2016; Ormond et al., 2016; Peter et al., 2012; Przeworski, 2003; Smith et al., 2016). Most of the existing methods concern restricted inferences for a single population or are conditioned by previous knowledge of genomic regions affected by selective sweeps (Supporting Information Table S1). Here, we introduced the **M**ultiple-**C**ollision **C**oalescent **SW**eep **AN**alyzer (McSwan), which leverages on biallelic SNP data sets (VCF file) to: (a) scan hard sweeps across genomes and subsequently, and (b) infer the times to the most recent common ancestors that carried the adaptive loci at

**TABLE 1** Significantly enriched cell types in candidate genes

| Significantly enriched cell types | Number of candidates genes[a] | *p*-Value[b] | FDR[c]-corrected *p*-value | Associated with the immune system |
|---|---|---|---|---|
| Promyelocyte | 19 | 0.0000050000 | 0.00014 | Yes |
| T lymphocyte | 93 | 0.0000500000 | 0.0007725 | Yes |
| Alveolar macrophage | 30 | 0.0013950000 | 0.01236 | Yes |
| Dendritic cell | 22 | 0.0022550000 | 0.01236 | Yes |
| Natural killer cell | 72 | 0.0023450000 | 0.01236 | Yes |
| Keratinocyte | 67 | 0.0023500000 | 0.01236 | |
| Squamous cell | 91 | 0.0034100000 | 0.01501 | |
| Neuroepithelium | 46 | 0.0036900000 | 0.01501 | |
| Muscle cell | 46 | 0.0058500000 | 0.02084889 | |
| Leucocyte | 80 | 0.0078300000 | 0.0255375 | Yes |
| Epithelium | 178 | 0.0137600000 | 0.04530227 | |

[a]Candidate genes are defined as genes overlapping hard sweep regions. [b]Uncorrected *P*-values of the enrichment tests. [c]False discovery rate.

the origin of the selective sweeps. The method implements an approximation of the coalescent with multiple collisions (Durrett & Schweinsberg, 2005; Pitman, 1999) to model selective histories. It then scans multiple genomes using an original combination of linear discriminant analysis and partial least squares regression. McSwan is based on joint site frequency spectra to detect sweeps and estimate their ages. Other approaches have shown high performance in detecting selection at a whole-genome scale based on haplotype-derived statistics, for example, extended haplotype homozygosity (EHH; Sabeti et al., 2002), the integrated haplotype score (iHS; Voight, Kudaravalli, Wen, & Pritchard, 2006) and between-population counterparts like the XP-EHH (Sabeti et al., 2007), as recently implemented in the *rehh* package (Gautier & Vitalis, 2012; Gautier, Klassmann, & Vitalis, 2017). Yet, so far none of these approaches have been able to simultaneously detect and date sweeps. They also require haplotype data sets that might be hard to generate for nonmodel organisms. McSwan has demonstrated high power (sensitivity and specificity above 95%; FDR <5%) to recover hard selective sweep events before 50,000 years BP, without requiring haplotype phasing. Regarding the sweep detection performance, McSwan was found to perform slightly better than SweeD, a popular likelihood-based approach, particularly when the recent effective population size was low and the genomic region to scan was limited in size (Supporting Information Text 4.2). McSwan can handle several populations in a joint analysis. However, since the computation load scales exponentially with the number of populations, we advise decreasing the sample size per population when increasing the number of populations since McSwan performs well with relatively few individuals per population.

Applying the method to detect and compare the time frames of the putative hard sweeps that have occurred in the European (CEU) and Luhya (LWK) populations since their split, we detected ~28-fold fewer signals of hard selective sweeps in the genomes of the LWK population (4%, 6 sweeps) relative to CEU (96%, 166 sweeps). This difference between the CEU and LWK populations could possibly be explained by the difficulty in detecting older selection in LWK populations (supplementary text 2.1.1). The higher LWK effective size might have increased the likelihood of recombination events, thus reducing haplotype lengths (Voight et al., 2006) and potentially lowering the sensitivity in detecting selection relative to CEU (Figure 2d). Yet, previous studies (based on haplotype-derived statistics and distortion from neutral SFS, respectively) also indicated lower selection imprints in African populations, including LWK (Carlson et al., 2005; Liu et al., 2013). Biologically, this unbalanced signal may be caused by different predominant modes of selection between the two populations, leading, for instance, to various degrees of prevalence in complete versus incomplete selective sweeps. The historically higher effective size and genetic diversity in African populations (Excoffier et al., 2013) could maintain a reservoir of standing variants, some of which might be adaptive in certain environments, thus increasing the probability of soft selective sweeps. In addition to different modes of selection, this discrepancy could stem from spatiotemporally varying selective pressures.

The temporal dynamics of the hard sweeps in the CEU population peaked around the Last Glacial Maximum (~24,500 years BP) and levelled off after ~11,000 years BP (Figure 4). We cannot completely rule out the possibility that this enrichment of sweep signals during the Last Glacial Maximum could be an artefact of the dating method due to undetected soft sweep biases. However, we think that it is less likely that this enrichment could be explained by statistical artefacts (Supporting Information Text 2.1.1). Moreover, although McSwan is mostly sensitive to complete sweeps, we cannot rule out the possibility of a sweep age overestimation in the case of incomplete sweeps. The four most significant environmental variables associated with the CEU hard sweep dynamics indicated that climate cooling and high short-term temperature instability were associated with several adaptations linked with novel variants in ancestral palaeolithic Europeans (Figure 5). These candidate adaptive mutations were significantly associated with genes involved in the immune system, suggesting possible adaptation to climate-mediated co-evolution with the European pathogen environment that ancestral modern humans were faced with during their north-westward expansion (Deschamps et al., 2016; Hancock et al., 2011). It has often been assumed that human dispersal out of Africa (~45,000–69,000 years BP; Gravel et al., 2011) and the agricultural transition (~9,000–7,500 in the Middle East and North Africa; Tishkoff et al., 2007) were two periods in human history associated with major shifts in selective pressure. Our results suggest that the LGM may have played a major role in shaping the hard sweep landscape in ancestral European populations, but further research is necessary to confirm the robustness of the results.

Although previous studies revealed several signatures of positive selection in CEU during the Neolithic (Nakagome et al., 2016; Smith et al., 2016), we found few hard sweeps dated during this period. A large region overlapping the lactase gene (LCT) and the upstream regulatory MCM6 gene was detected here as having experienced a strong selective sweep during this period. The LCT gene codes for the lactase enzyme which breaks down lactose, a sugar found in dairy milk. A mutation on a regulatory region of this gene, 13,910 bp upstream of LCT (Bersaglieri et al., 2004; Enattah et al., 2002), confers the ability to digest lactose into adulthood. This gene is the strongest target of recent positive selection in European populations (Bersaglieri et al., 2004; Smith & Haigh, 1974; Smith et al., 2016). Adoption of a diary-centred diet is hypothesized to have driven this selection. Our sweep age estimation ($T_{\mathrm{MRCA}}$ = 4,250 years BP) is in line with the findings of recent studies based on ancient DNA which did not detect any LCT persistence variants until the Bronze Age (Haber, Mezzavilla, Xue, & Tyler-Smith, 2016). Our estimation is also in line with archaeological evidence of dairy farming in Europe after ~9,000 years BP (Gerbault et al., 2011). As in other studies, we estimated a large $T_{\mathrm{MRCA}}$ confidence interval (3,700–17,680 years BP) that substantially overlapped previous confidence interval estimates based on modern-day genetic data sets (Table 2), although two previous estimates showed a lower bound prior to the Bronze Age (Coelho et al., 2005; Smith et al., 2016). As expected for selection on de novo mutations (Hermisson & Pennings, 2017), our $T_{\mathrm{MRCA}}$ point estimate was slightly earlier than the age of the adaptive

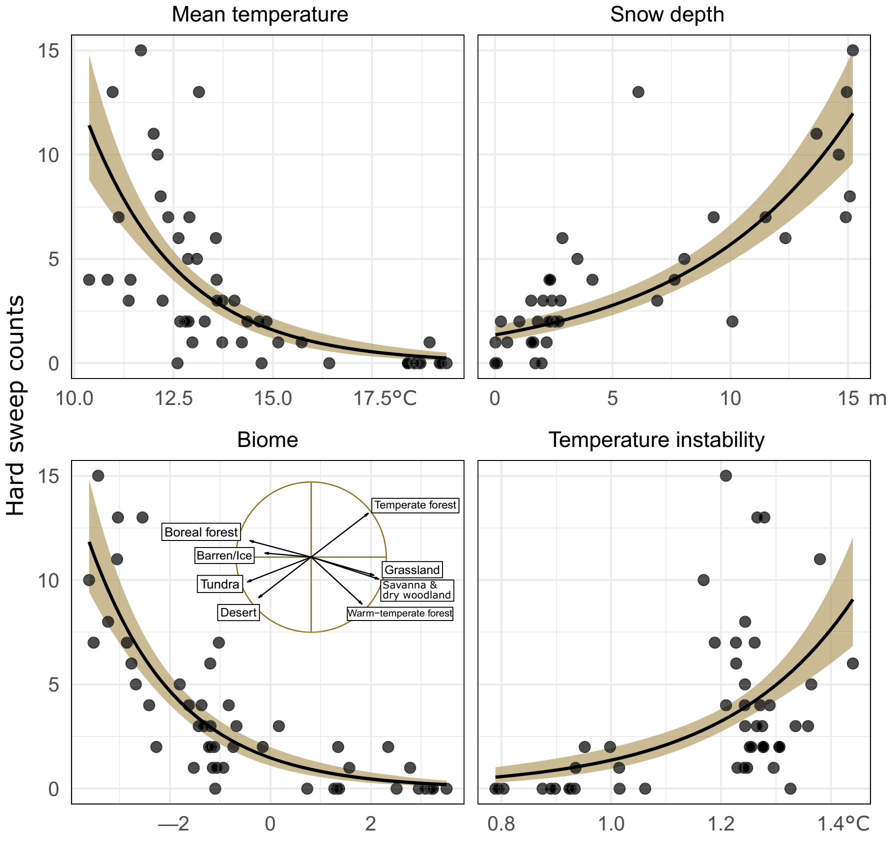

**FIGURE 5** Plots of the hard sweep counts in relation to the four most explanatory environmental variables, as identified using a LASSO Poisson regression. In black, the fitted univariate Poisson regression curves. The biome variable is measured as a continuous gradient from tundra/boreal forest to grasslands/temperate forest (projection on the first axis of a principal component analysis, the correlation circle is overlaid) [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 2** Published estimates of selection ages for LCT and SLC24A5 adaptive mutations in the CEU population

| Gene | Approach | Time | Estimation[a] | Reference |
|---|---|---|---|---|
| LCT | Haplotype | TMRCA | 2,188–20,650 | Bersaglieri et al. (2004) |
| LCT[b] | (Seixas et al., 2001; Stumpf & Goldstein, 2001) | TMRCA | 8,125–23,640 | Coelho et al. (2005) |
| LCT | Haplotype | Age of mutation | 4,612 [3,948–5,312] | Chen et al. (2015) |
| LCT | Haplotype | TMRCA | 7,646–9,225 | Smith et al. (2016) |
| LCT | ABC | Age of mutation | 3,466–18,191 | Tishkoff et al. (2007) |
| LCT | ABC | Age of mutation | 1,500–64,900 | Peter et al. (2012) |
| LCT | ABC | Age of mutation | 4,458–8,905 | Nakagome et al. (2016) |
| LCT | LDA-PLSR | TMRCA | 4,245 [3,700–17,680] | This study |
| SLC24A5 | ABC | Age of mutation | 16,700 [5,200–34,225] (dominant model) or 10,150 [925–48,450] (additive model) | Beleza et al. (2013) |
| SLC24A5 | ABC | Age of mutation | 30,000–40,000 | Nakagome et al. (2016) |
| SLC24A5 | LDA-PLSR | TMRCA | 17,190 [13,905–20,650] | This study |

[a] In years BP, recalibrated using a mutation rate $\mu$ = 2.5 × $10^{-8}$ and a generation time $g$ = 25 years. [b] For a Finnish population taken from Enattah et al. (2002).

mutation inferred by Chen et al. (2015) using a hidden Markov model (4,612 years BP, $g$ = 25 years) and by Nakagome et al. (2016) using an ABC approach (4,458–8,905 years BP).

Another hallmark of more ancient positive selection is found at the SLC24A5 gene (also known as NCKX5), which codes for a cation exchanger protein involved in skin pigmentation (Lamason et al., 2005). The derived allele of this gene (rs1426654) is fixed in European populations and is the main determinant of lighter skin. This phenotype is considered as being adaptive due to the low ultraviolet exposure at high latitudes and UV involvement in vitamin D synthesis (Canfield et al., 2013). The same specific region was identified in several studies (Beleza et al., 2013; Lamason et al., 2005; Liu et al., 2013; Table 2). Our method identified a 191-kb-long region encompassing the SLC24A5 gene, with a $T_{MRCA}$ estimate of 17,190 years BP [13,905–20,650], suggesting that the ancestral individual which carried the derived allele likely lived during the Late Pleniglacial (15,000–24,000 years BP), before the Neolithic transition. A different study targeting this gene estimated a comparable time frame for the age of the adaptive mutation, although the point estimates were slightly earlier than our $T_{MRCA}$ estimate (Beleza et al., 2013; Table 2). Investigations of the variant in prehistoric humans show that the derived mutation was likely rare among Mesolithic hunter-gatherers (Olalde et al., 2014) but increased in frequency in Europe during the Early Neolithic following the immigration of populations originating from Anatolia (Gunther & Jakobsson, 2016; Kılınç et al., 2016; Mathieson, Lazaridis, & Rohland, 2015). We estimated the $T_{MRCA}$ based on genetic data from the CEU population, whose ancestry is known to be drawn from (at least) three different main ancestral sources (Allentoft et al., 2015; Haak et al., 2015; Lazaridis et al., 2014; Skoglund et al., 2012, 2014 ), including early Neolithic farmers. Although part of the selective sweep period likely occurred in Europe within the last 10,000 years, our sweep age estimate goes back into the ancestral groups contributing to the CEU population. Mathieson et al. (2015) noted that the SLC24A5-derived variant was already very common in early Anatolian farmers living around 8,000 years BP, and our sweep estimate captured this early time period.

Multiple studies have emphasized the role of adaptive introgression from Neanderthal or Denisovan archaic species for specific human adaptation (Abi-Rached et al. (2011), Deschamps et al. (2016) and Racimo, Sankararaman, Nielsen, and Huerta-Sanchez (2015) for a review). However, we detected a depletion of Neanderthal ancestry in the hard sweep regions, in agreement with similar findings for functionally rich regions (Sankararaman et al., 2014).

We have introduced a new method to simultaneously detect and infer the age of hard selective sweeps. This method performed well in comparison with other methods. In the future, approaches to improve detection of soft sweeps due to recurrent mutations could benefit from the characterization of more general coalescents, like the $\Xi$-coalescent which allows simultaneous merging of ancestral lineage blocks (Blath, Cronjager, Eldon, & Hammer, 2016). Considering the empirical analyses, McSwan was successful in identifying the hard sweep dynamics in two human populations. Overall, simultaneously determining the location of hard selective sweeps along the genome and their age using a genome scan approach appears to be a powerful tool for clarifying the spatiotemporal dynamics of selective sweeps in recombining organisms. Our approach may also be very useful for nonmodel organisms where a reference genome has been sequenced and for which little is currently known about the distribution of hard selective sweeps.

## ACKNOWLEDGEMENTS

The authors warmly thank François Munoz, Miguel Navascués, Olivier François, Ophélie Ronce and Concetta Burgarella for useful discussions and suggestions. RT was funded by a PhD grant from the French Ministry of Higher Education and Research. MJ was funded by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

## AUTHOR CONTRIBUTIONS

The method was designed and developed by RT. The biological study was designed by all authors. The analyses were performed by RT. The first draft with contributions from all authors was written by RT.

## DATA ACCESSIBILITY

The R package, with data examples and a tutorial, can be downloaded at: https://github.com/sunyatin/McSwan.

## ORCID

*Valérie Poncet* http://orcid.org/0000-0002-1099-2846
*Yves Vigouroux* http://orcid.org/0000-0002-8361-6040
*Stéphanie Manel* http://orcid.org/0000-0001-8902-6052

## REFERENCES

Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., … Parham, P. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, *334*, 89–94. https://doi.org/10.1126/science.1209202

Allentoft, M. E., Sikora, M., Sjögren, K. G., Rasmussen, S., Rasmussen, M., Stenderup, J., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*, 167. https://doi.org/10.1038/nature14507

Beleza, S., Santos, A. M., McEvoy, B., Alves, I., Martinho, C., Cameron, E., … Rocha, J. (2013). The timing of pigmentation lightening in Europeans. *Molecular Biology and Evolution*, *30*, 24–35. https://doi.org/10.1093/molbev/mss207

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., … Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, *74*, 1111–1120. https://doi.org/10.1086/421051

Blath, J., Cronjager, M. C., Eldon, B., & Hammer, M. (2016). The site-frequency spectrum associated with Xi-coalescents. *Theoretical Population Biology*, *110*, 36–50.

Bos, J. A. A., Bohncke, S. J. P., Kasse, C., & Vandenberghe, J. (2001). Vegetation and climate during the Weichselian Early Glacial and Pleniglacial in the Niederlausitz, eastern Germany — macrofossil and pollen evidence. *Journal of Quaternary Science*, *16*, 269–289. https://doi.org/10.1002/jqs.606

Canfield, V. A., Berg, A., Peckins, S., Wentzel, S. M., Ang, K. C., Oppenheimer, S., & Cheng, K. C. (2013). Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3. Genes—genomes—genetics*, *3*, 2059–2067. https://doi.org/10.1534/g3.113.007484

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., & Nickerson, D. A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, *15*, 1553–1565. https://doi.org/10.1101/gr.4326505

Chen, H., Hey, J., & Slatkin, M. (2015). A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical Population Biology*, *99*, 18–30. https://doi.org/10.1016/j.tpb.2014.11.001

Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research*, *20*, 393–402. https://doi.org/10.1101/gr.100545.109

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., … McCabe, A. M. (2009). The last glacial maximum. *Science*, *325*, 710–714. https://doi.org/10.1126/science.1172873

Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A. I., Seixas, S., Destro-Bisol, G., & Rocha, J. (2005). Microsatellite variation and evolution of human lactase persistence. *Human Genetics*, 117. https://doi.org/10.1007/s00439-005-1322-z

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*, 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.-L., … Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *The American Journal of Human Genetics*, *98*, 5–21. https://doi.org/10.1016/j.ajhg.2015.11.014

Duforet-Frebourg, N., Bazin, E., & Blum, M. G. B. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, *31*, 2483–2495. https://doi.org/10.1093/molbev/msu182

Durrett, R., & Schweinsberg, J. (2005). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, *115*, 1628–1657. https://doi.org/10.1016/j.spa.2005.04.009

Enattah, N. S., Jensen, T. G. K., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., … Peltonen, L. (2008). Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *American Journal of Human Genetics*, *82*, 57–72. https://doi.org/10.1016/j.ajhg.2007.09.012

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., & Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, *30*, 233–237. https://doi.org/10.1038/ng826

Ewing, G., & Hermisson, J. (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, *26*, 2064–2065. https://doi.org/10.1093/bioinformatics/btq322

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*, e1003905. https://doi.org/10.1371/journal.pgen.1003905

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*, 1405–1413.

Gautier, M., Klassmann, A., & Vitalis, R. (2017). rehh 2.0: A reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, *17*, 78–90.

Gautier, M., & Vitalis, R. (2012). rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*, 1176–1177. https://doi.org/10.1093/bioinformatics/bts115

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., … Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 757–762. https://doi.org/10.1073/pnas.1310398110

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74. https://doi.org/10.1038/nature15393

Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., … Thomas, M. G. (2011). Evolution of lactase persistence: An example of human niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*, 863–877. https://doi.org/10.1098/rstb.2010.0268

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., … Bustamante, C. D. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 11983–11988. https://doi.org/10.1073/pnas.1019276108

Gunther, T., & Jakobsson, M. (2016). Genes mirror migrations and cultures in prehistoric Europe-a population genomic perspective. *Current Opinion in Genetics & Development*, *41*, 115–123. https://doi.org/10.1016/j.gde.2016.09.004

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., … Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, *522*, 207–211. https://doi.org/10.1038/nature14317

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*, 5–23. https://doi.org/10.1111/mec.13339

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

Haller, B. C., & Messer, P. W. (2017). SLiM 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, *34*, 230–240. https://doi.org/10.1093/molbev/msw211

Hancock, A. M., Brachi, B., Faure, N., Horton, M. W., Jarymowycz, L. B., Sperone, F. G., … Bergelson, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, *334*, 83–86. https://doi.org/10.1126/science.1209244

Hermisson, J., & Pennings, P. S. (2017). Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution*, *8*, 700–716. https://doi.org/10.1111/2041-210X.12808

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*, 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Kasse, C., Huijzer, A. S., Krzyszkowski, D., Bohncke, S. J. P., & Coope, G. R. (1998). Weichselian Late Pleniglacial and Late-glacial depositional environments, Coleoptera and periglacial climatic records from central Poland (Bełchatów). *Journal of Quaternary Science*, *13*, 455–469. https://doi.org/10.1002/(SICI)1099-1417(1998090)13:5<455:AID-JQS398>3.0.CO;2-T

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, *336*, 740–743. https://doi.org/10.1126/science.1217283

Kılınç, G. M., Omrak, A., Özer, F., Günther, T., Büyükkarakaya, A. M., Bıçakçı, E., … Götherström, A. (2016). The demographic development of the first farmers in anatolia. *Current Biology*, *26*, 2659–2666. https://doi.org/10.1016/j.cub.2016.07.057

Kofler, R., & Schlotterer, C. (2012). Gowinda: Unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, *28*, 2084–2085. https://doi.org/10.1093/bioinformatics/bts315

Lamason, R. L., Mohideen, M. A., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., … O'donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC,, (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, *310*, 1782–1786. https://doi.org/10.1126/science.1116238

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*, 409–413. https://doi.org/10.1038/nature13673

Liu, X., Ong, R. T., Pillai, E. N., Elzein, A. M., Small, K. S., Clark, T. G., … Teo, Y. Y. (2013). Detecting and characterizing genomic signatures of positive selection in global populations. *The American Journal of Human Genetics*, *92*, 866–881. https://doi.org/10.1016/j.ajhg.2013.04.021

Loarie, S. R., Duffy, P. B., Hamilton, H., Asner, G. P., Field, C. B., & Ackerly, D. D. (2009). The velocity of climate change. *Nature*, *462*, 1052–1055. https://doi.org/10.1038/nature08649

Mathieson, I., Lazaridis, I., Rohland, N., et al. (2015). Eight thousand years of natural selection in Europe. *bioRxiv*.

Nakagome, S., Alkorta-Aranburu, G., Amato, R., Howie, B., Peter, B. M., Hudson, R. R., & Di Rienzo, A. (2016). Estimating the ages of selection signals from different epochs in human history. *Molecular Biology and Evolution*, *33*, 657–669. https://doi.org/10.1093/molbev/msv256

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, *15*, 1566–1575. https://doi.org/10.1101/gr.4252305

Olalde, I., Allentoft, M. E., Sánchez-Quinto, F., Santpere, G., Chiang, C. W., DeGiorgio, M., … Lalueza-Fox, C. (2014). Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, *507*, 225. https://doi.org/10.1038/nature12960

Ormond, L., Foll, M., Ewing, G. B., Pfeifer, S. P., & Jensen, J. D. (2016). Inferring the age of a fixed beneficial allele. *Molecular Ecology*, *25*, 157–169. https://doi.org/10.1111/mec.13478

Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, *30*(9), 2224–2234. https://doi.org/10.1093/molbev/mst112

Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLOS Genetics*, *8*, e1003011. https://doi.org/10.1371/journal.pgen.1003011

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab*, *27*, 1870–1902. https://doi.org/10.1214/aop/1022874819

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., … Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*, 43–49. https://doi.org/10.1038/nature12886

Przeworski, M. (2003). Estimating the time since the fixation of a beneficial allele. *Genetics*, *164*, 1667–1676.

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., … Engelken, J. (2014). 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research*, *42*, D903–D909.

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sanchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*, 359–371. https://doi.org/10.1038/nrg3936

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., … Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*, 832–837. https://doi.org/10.1038/nature01140

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., … Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, *312*, 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., … Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*, 913–918. https://doi.org/10.1038/nature06250

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., … Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*, 354–357. https://doi.org/10.1038/nature12961

Seixas, S., Garcia, O., Trovoada, M. J., Santos, M. T., Amorim, A., & Rocha, J. (2001). Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: Insights into the natural history of the α1-antitrypsin polymorphism. *Human Genetics*, *108*, 20–30. https://doi.org/10.1007/s004390000434

Skoglund, P., Malmström, H., Omrak, A., Raghavan, M., Valdiosera, C., Günther, T., … Jakobsson, M. (2014). Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers. *Science*, *344*, 747–750. https://doi.org/10.1126/science.1253448

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., … Jakobsson, M. (2012). Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science*, *336*, 466–469. https://doi.org/10.1126/science.1216304

Smith, J., Coop, G., & Stephens, M. (2016). Estimating time to the common ancestor for a beneficial allele. *bioRxiv*.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*, 23–35. https://doi.org/10.1017/S0016672300014634

Stumpf, M. P. H., & Goldstein, D. B. (2001). Genealogical and evolutionary inference with the human Y chromosome. *Science*, *291*, 1738–1742. https://doi.org/10.1126/science.291.5509.1738

Teshima, K. M., & Innan, H. (2009). mbs: Modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics*, *10*, 166. https://doi.org/10.1186/1471-2105-10-166

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., … Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *NatureGenetics*, *39*, 31–40. https://doi.org/10.1038/ng1946

Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, *196*, 799–817. https://doi.org/10.1534/genetics.113.152991

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*, e72.

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*, 815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

Yang, J., & Yang, J.-Y. (2003). Why can LDA be performed in PCA transformed space? *Pattern Recognition*, *36*, 563–566. https://doi.org/10.1016/S0031-3203(02)00048-1

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Tournebize R, Poncet V, Jakobsson M, Vigouroux Y, Manel S. McSwan: A joint site frequency spectrum method to detect and date selective sweeps across multiple population genomes. *Mol Ecol Resour*. 2019;19:283–295. https://doi.org/10.1111/1755-0998.12957