Sporadic Occurrence of Recent Selective Sweeps from Standing Variation in Humans as Revealed by an Approximate Bayesian Computation Approach

Guillaume Laval, Etienne Patin, Pierre Boutillier, Lluis Quintana-Murci

Sporadic Occurrence of Recent Selective Sweeps from Standing Variation in Humans as Revealed by an Approximate Bayesian Computation Approach

Guillaume Laval[1]*, Etienne Patin[1], Pierre Boutillier[2] and Lluis Quintana-Murci[1,3]

[1]Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, CNRS, Paris 75015, France.

[2]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

[3]Human Genomics and Evolution, Collège de France, 75005 Paris, France

***Corresponding author**

E-mail: glaval@pasteur.fr

Abstract

During their dispersals over the last 100,000 years, modern humans have been exposed to a large variety of environments, resulting in genetic adaptation. While genome-wide scans for the footprints of positive Darwinian selection have increased knowledge of genes and functions potentially involved in human local adaptation, they have globally produced evidence of a limited contribution of selective sweeps in humans. Conversely, studies based on machine learning algorithms suggest that recent sweeps from standing variation are widespread in humans, an observation that has been recently questioned. Here, we sought to formally quantify the number of recent selective sweeps in humans, by leveraging approximate Bayesian computation and whole-genome sequence data. Our computer simulations revealed suitable ABC estimations, regardless of the frequency of the selected alleles at the onset of selection and the completion of sweeps. Under a model of recent selection from standing variation, we inferred that an average of 68 (from 56 to 79) and 140 (from 94 to 198) sweeps occurred over the last 100,000 years of human history, in African and Eurasian populations, respectively. The former estimation is compatible with human adaptation rates estimated since divergence with chimps, and reveal numbers of sweeps per generation per site in the range of values estimated in Drosophila. Our results confirm the rarity of selective sweeps in humans and show a low contribution of sweeps from standing variation to recent human adaptation.

INTRODUCTION

Evaluating the legacy of positive, Darwinian selection in humans has proved crucial for increasing our understanding of the genetic architecture of adaptive phenotypes (Fan et al. 2016; Jeong and Di Rienzo 2014; Vitti et al. 2013). Genome-wide scans of selection have identified large numbers of signals of selective sweeps (Maynard Smith and Haigh 1974; Pritchard et al. 2010; Stephan et al. 1992) but, in turn, have produced limited evidence of this mode of selection to recent human adaptation (Hernandez et al. 2011). Conversely, a study based on the 1000 Genomes (1000G) data (Auton et al. 2015) and a simulation-based machine learning classifier has suggested that recent sweeps from standing variation ('soft sweeps') have been pervasive (Schrider and Kern 2017). Yet, the majority of such sweeps have been proposed to result from mis-classified neutral regions (Harris et al. 2018). Schrider and Kern (2017) identified ~862 or ~18 sweeps on average per population using the default or a more stringent probability threshold of 0.9 (Harris et al. 2018), showing the strong dependency between the number of reported sweeps and the detection thresholds used (Akey 2009; Pavlidis and Alachiotis 2017; Teshima et al. 2006). This highlights the difficulty to genuinely assess the real number of sweeps occurred in humans when identifying them individually (Jensen 2009; Li and Stephan 2006). Li and Stephan (2006) advised against methods counting the number of detected selection events due to false positives and the relatively low power to detect weak selection. Alternatively, many studies, including this one, aim to directly estimate the number of sweeps due to beneficial mutations that have occurred during a given period of time (e.g., Jensen et al. 2008; Li and Stephan 2006).

Inferring the rates of beneficial mutations in various species has been an active area of research in evolutionary biology. The most convincing results come from studies in *Drosophila*, where various methods have been implemented to formally estimate the expected number of selected substitutions per nucleotide site per generation (λ). These estimates have

been calculated from divergence and/or polymorphism data, focusing on beneficial mutations that have fixed over a high number of generations (Andolfatto 2007; Eyre-Walker 2006; Jensen et al. 2008; Li and Stephan 2006; Macpherson et al. 2007; Sawyer and Hartl 1992; Sawyer et al. 2007; Smith and Eyre-Walker 2002). For example, Li and Stephan (2006) estimated λ considering beneficial mutations that have fixed over the last ~60,000 years in *D. melanogaster* populations (~900,000 generations, assuming 15 generations per year) (Barker 1962; Pool 2015). Here, we sought to assess the extent of recent positive selection considering beneficial mutations occurring at shorter evolutionary time scales, such as those occurring after the split of African and Eurasian populations (e.g., in humans ~60,000 years represent ~2,150 generations assuming a generation interval of 28 years) (Fenner 2005; Moorjani et al. 2016b). We considered both segregating and fixed beneficial mutations (incomplete and complete sweeps), as many candidate selection targets still segregate in humans (Fan et al. 2016; Jeong and Di Rienzo 2014; Vitti et al. 2013) with few numbers of fixed or nearly-fixed differences being observed between human populations (Coop et al. 2009; Pritchard et al. 2010). In this study, we leveraged genome-wide polymorphism data and neutrality statistics known to detect recent selection to assess the number of occurring sweeps (X) using an approximate Bayesian computation (ABC) method (Beaumont et al. 2002; Jensen et al. 2008).

The ABC summary statistics we used here have been shown to be efficient in capturing true signals of selection, as they detect genome-wide excesses of candidate selected SNPs (i.e., SNPs harboring extreme values for a given neutrality statistic) within and near genes relative to intergenic regions, e.g., coding SNPs or *cis*-acting eQTLs *vs* intergenic SNPs (Barreiro et al. 2008; Fagny et al. 2014; Frazer et al. 2007; Jin et al. 2012; Kudaravalli et al. 2009; Schmidt et al. 2019; Voight et al. 2006). Indeed, selective sweeps produce clusters of candidate SNPs in the vicinity of selection targets, whereas under neutrality candidate SNPs are more uniformly scattered (Voight et al. 2006). Specifically, we used the odds ratio for

selection (OR) (Kudaravalli et al. 2009), a statistic that depends on the ratio between the percentages of candidate SNPs identified within genic and intergenic regions (Barreiro et al. 2008; Fagny et al. 2014). Under neutrality, false positive candidate SNPs are expected in genic and intergenic regions at the same proportion (Barreiro et al. 2008; Kudaravalli et al. 2009) ($OR = 1$, no excess of candidate SNPs). Otherwise, in case of higher rates of candidate SNPs in genes due to selection ($OR > 1$), we assumed that OR correlates with *X* and provides suitable information to estimate *X*.

We applied the ABC method to several African, European and East Asian 1000G populations (Supplemental Material, Table S1), and explored various assumptions about the nature of the sweeps. To compare our results with the number of reported sweeps from standing variation (Schrider and Kern 2017), we simulated a single advantageous mutation per sweep region with a selection coefficient (*s*) that ranged from 0.001 to 0.05 and a specific time (*t*) that ranged from the present to 3,500 generations ago (or ~100,000 years ago). The frequency of the selected allele when the sweep begins (p_{start}) was similarly ranged from $1/2N$ to 0.2. We labeled the sweeps from a *de novo* mutation ($p_{start} = 1/2N$) and from standing variation ($p_{start} > 1/2N$) (Hermisson and Pennings 2005; Innan and Kim 2004; Orr and Betancourt 2001; Przeworski et al. 2005) SDN and SSV, respectively (Peter et al. 2012). Although these sweeps correspond to the hard and soft sweeps simulated in Schrider and Kern (2017), we avoided such a terminology (Hermisson and Pennings 2005) since the nature -- either hard or soft -- of detected sweeps is often ambiguous (Harris et al. 2018; Jensen 2014; Orr and Betancourt 2001).

MATERIALS AND METHODS

Overview

Using ABC, we sought to jointly estimate the number of occurring sweeps X, the average strength and the average age of selection, $S = 1/X \sum_1^X s_i$ and $T = 1/X \sum_1^X t_i$, in fifteen 1000G populations analyzed separately (five African, five European and five East Asian populations, see the section "1000 Genomes populations analyzed" below). We also considered three distinct categories of sweeps arbitrarily defined on the basis of the frequency of the selected allele at the onset of selection. Namely, we estimated with ABC X_1, X_2 and X_3, the number of sweeps with very low ($1/2N \leq p_{start} < 0.01$), low ($0.01 \leq p_{start} < 0.1$) and intermediate ($0.1 \leq p_{start} < 0.2$) initial frequencies respectively (in our model $X = X_1 + X_2 + X_3$). ABC point estimates (posterior means) and the 95% credible intervals (CI) boundaries are obtained from simulated whole-genome sequence (WGS) data best fitting with the 1000G empirical WGS data (see the section "ABC, acceptance rules and accuracy evaluation" below).

In each analyzed population, the *Xs* (X, X_1, X_2 and X_3), S and T were estimated using whole-genome sequence data simulated to reproduce the 1000G empirical WGS data. The simulated and the empirical 1000G WGS data are each summarized by a vector of K ORs, K being the number of neutrality statistic used. To compute these ORs, we considered as candidate SNPs of selection those with the most extreme values for neutrality statistics known to detect recent selection and previously found to be insensitive to background selection (BGS). The OR is based on the classic comparison between putatively neutral mutations and mutations potentially targeted or influenced by selection (Kimura 1977; Mcdonald and Kreitman 1991), labeled here ENVs (Evolutionary Neutral Variants) and PSVs (Possibly Selected Variants) respectively and defined below. Specifically, the OR computed for each WGS dataset and for each neutrality statistic separately is defined as

$$OR = \left[\frac{P(PSV|I(Candidate=1))}{P(ENV|I(Candidate=1))}\right]\left[\frac{P(ENV|I(Candidate=0))}{P(PSV|I(Candidate=0))}\right], \quad \text{(equation 1)}$$

with $I(Candidate = 1)$ being the indicator function equal to 1 if the SNP is a candidate SNP of selection or 0 otherwise (Kudaravalli et al. 2009). The candidate SNPs were determined using genome-wide thresholds defining the 1% most extreme values obtained in simulated or in empirical 1000G data. The OR is thus expected close to one under neutrality (no sweep on the genome, X=0). Otherwise, since selective sweeps produce clusters of candidate SNPs around targets of selection (Voight et al. 2006), an OR above one is an indicator of an excess of candidate SNPs in PSVs due to selection targeting beneficial mutations in this SNP class (X>0). The ENVs are used here as the neutral baseline to control for the rate of false positive candidate SNPs of selection (Barreiro et al. 2008; Fagny et al. 2014; Kudaravalli et al. 2009). Finally, as a sanity check, we verified that the selection signals detected by the neutrality statistics used correspond to well-known examples of selective sweeps. We performed, in each 1000G population, a classic selection scan based on the same neutrality statistics used to estimate X (see the section "Detecting genomic regions enriched in candidate SNPs of selection" below).

Odds ratio for selection in the simulated and 1000G populations

For each whole-genome sequence dataset, either simulated or empirical, and each of the K neutrality statistic used, the OR was computed using all SNPs genome-wide and the logistic region model set as follows (Kudaravalli et al. 2009):

$$Logit[I(PSV = 1)] = \beta_1 I(Candidate = 1) + \varepsilon \qquad \text{(equation 2)}$$

with $I(PSV = 1)$ being the indicator function equal to 1 for PSV SNPs or equal to 0 otherwise, $I(Candidate = 1)$ being the indicator function defined above, and β_1 being the coefficient of the logistic regression (the constant term β_0 was omitted). We used the 'glm' R package (family=binomial) to estimate β_1 and $exp(\beta_1)$ to compute the OR for the effect of selection according to equation (1). Indeed, if $\beta_1 > 0$, this implies an enrichment effect for

PSVs among SNPs with selection signals due to the enrichment of SNPs with selection signals in the PSV SNP class (Kudaravalli et al. 2009).

In addition, when analyzing the 1000G populations, we carefully investigated the sensitivity of our ABC method to various estimations of the ORs. To this end, we also used the logistic regression-based method proposed by Kudaravalli et al. (2009) to estimate the 1000G ORs while controlling for various covariates such as the coverage (or sequencing quality), a feature of the data which was not simulated in our ABC model. Moreover, while we performed our simulations using human recombination maps (see the section "Simulating WGSs") it is difficult to closely reproduce in genome-wide computer simulations the recombination pattern observed in humans. Recombination varies both along and across chromosomes in humans (Coop et al. 2008; Kong et al. 2010; Myers et al. 2005) and can alter the OR values. Due to the diminished hitchhiking effects of selection in high recombining regions, which result in reduced clusters of candidate SNPs, some selection signals can be penalized and ultimately contribute less to the ORs because of their genomic location only. The OR corrected for these covariates was computed using the logistic region model set as follows (Kudaravalli et al. 2009):

$$Logit[I(PSV = 1)] = \beta_1 I(Candidate = 1) + [\beta_2 Cov + \beta_3 Rec + \beta_4 NbSNP + \beta_5 Cov * Rec + \beta_6 Rec * NbSNP + \beta_7 NbSNP * Cov] + \varepsilon, \quad \text{(equation 3)}$$

with $I(PSV = 1)$ and $I(Candidate = 1)$ defined above, and Rec the mean recombination rate in cM/bp obtained from HapMap recombination maps, Cov the mean coverage and NbSNP the number of SNPs, all computed with 100 kb sliding windows centered on each SNP. Finally, the algorithm used to compute the various ORs was set as follows:

(i) for each WGS dataset either simulated or empirical

(ii) get the definition of PSVs ($PSV = 1$ or $PSV = 0$ otherwise) randomly determined when building the simulated WGS data or according to the 1000G VEP annotations for empirical WGS data (see below)

(iii) for each of the *K* neutrality statistics used

(iv) get the definition of candidate SNPs ($Candidate = 1$ or $Candidate = 0$ otherwise) determined when building the simulated or empirical WGSs

(v) make the logistic regression model without covariates, equation (2)

(vi) compute the uncorrected OR merging all chromosomes in the logistic model

$\mathrm{OR}_1 = exp(\hat{\beta}_1)$, $\hat{\beta}_1$ obtained from equation (2)

(vii) if empirical WGS data then

(viii) compute the uncorrected OR averaged across chromosomes

$\mathrm{OR}_2 = 1/C \sum_c \ exp(\hat{\beta}_{1,c})$, $\hat{\beta}_{1,c}$ obtained for the c^{th} chromosome from equation (2)

(ix) compute covariates

(x) make the logistic regression model with covariates, equation (3)

(xi) compute the corrected OR merging all chromosomes in the logistic model

$\mathrm{OR}_3 = exp(\hat{\beta}_1)$, $\hat{\beta}_1$ obtained from equation (3), which corresponds to that used by Kudaravalli et al. (2009)

(xii) compute the corrected OR averaged across chromosomes

$\mathrm{OR}_4 = 1/C \sum_c \ exp(\hat{\beta}_{1,c})$, $\hat{\beta}_{1,c}$ obtained for the c^{th} chromosome from equation (3)

back to (iii)

back to (i)

With this algorithm, the 10^5 simulated WGSs used to perform the ABC estimations in each 1000G population are summarized by a matrix of OR_1 with 10^5 rows and *K* columns. For each population, we systematically performed four rounds of ABC estimations using the four vectors of 1000G ORs of dimension ($1 \times K$) computed per population (OR_1, OR_2, OR_3 and

209 OR_4, see above). The matrix of simulated OR_1 and each vector of 1000G ORs were used
210 without any modifications to jointly estimate X, X_1, X_2, X_3, S and T with a standard ABC
211 method.
212
213 **Definitions of ENVs and PSVs**
214 The ENVs are Evolutionary Neutral Variants. In the simulations, they are neutral SNPs
215 simulated in absence of selection, while in real data, these are intergenic SNPs far from the
216 nearest gene and assumed to be unaffected by selection (Barreiro et al. 2008; Schmidt et al.
217 2019; Voight et al. 2006). Conversely, the PSVs are Possibly Selected Variants; they include
218 the potential targets of selection, that is mutations prone to alter individual phenotypes. In
219 addition, we also incorporated in PSVs the SNPs near the potential targets of selection in
220 order to capture the hitchhiking effects of linked selected variants on neutral polymorphisms.
221 Indeed, the extent of clusters of candidate SNPs around the targets of selection, which
222 depends on the intensity and on the age of selection, may provide valuable information for the
223 estimation of X, S and T, respectively. In simulations, the PSVs are thus the selected SNPs
224 and nearby SNPs simulated in the same genomic regions. In real data, the PSVs are
225 nonsynonymous, regulatory mutations and nearby SNPs, i.e., synonymous mutations, intronic
226 mutations, which corresponds to the genic SNPs classically used (Barreiro et al. 2008;
227 Schmidt et al. 2019; Voight et al. 2006). We also considered as PSVs the variants located
228 upstream/downstream of genes as well as any presumed regulatory sites located in distal
229 intergenic regions, e.g., SNPs located in transcription factor binding sites or in mature
230 miRNAs. In our study, we considered that all nonsynonymous or regulatory mutations,
231 denoted as ‘functional mutations’, are prone to alter phenotypes, as classically assumed in
232 studies assessing the impact of natural selection (Barreiro et al. 2008; Fagny et al. 2014).

In simulated WGS data, the simulated regions containing PSVs, namely the PSV regions, are randomly defined and contain X selective sweeps of various intensity, age and frequency of the selected allele at the onset of selection (the remaining mutations are ENVs, see the section "Simulating WGSs"). It must be emphasized that PSV regions in simulated WGS data do not necessarily include a selected variant. Indeed, in the real data, PSV regions contain functional mutations, but only a fraction of these regions exhibit selection signals and may thus contain functional selected mutations. The other PSV regions behave as neutral regions as they only contain functional mutations with no detectable advantageous effects on fitness.

For each analyzed population, the PSVs and ENVs are determined using the 1000G VEP (Ensembl Variant Effect Predictor) annotations. According to our definitions, all mutations in or near genes, as well as intergenic mutations annotated as "TF_binding_site_variant" or "mature_miRNA_variant" in the VEP files, were set to be PSVs, $I(PSV = 1)$. The rest of intergenic mutations annotated as "intergenic_variant" were set to be ENVs $I(PSV = 0)$ (Supplemental Material). With such filters, ~70% of 1000G SNPs were considered to be PSVs (the remaining 30% are ENVs). We thus systematically reproduced these proportions in the simulated WGS data. We also dealt with some particular situations: these include the selected sites that are unknown functional variants annotated as "intergenic_variant", small regulatory regions located far from other PSVs and regulatory variants located in edges of PSVs tracts. Positive selection targeting such SNPs can bias downward our estimations because the selection signals (clusters of candidate SNPs) may expand to ENVs. This will increase the proportion of candidate SNPs of selection in ENVs resulting in lower empirical ORs, which will be ultimately interpreted by the model as a lower number of sweeps. To minimize such estimation biases, we annotated as PSVs all SNPs with a genome-wide significant enrichment in candidate SNPs measured 100 kb around them, since such enrichments due to clusters of candidate SNPs are indicative of positive selection (Voight et

al. 2006). This step has only a marginal effect on the total number of SNPs classified as PSVs but it can inflate ORs. We thus reproduced this step in the simulated WGS data.

Simulating WGSs

Computer simulations were performed to reproduce the 1000G WGSs data, i.e., ~3Gb of DNA sequences for 100 unrelated individuals sampled per population. To do so, we used a demographic model previously inferred for the analyzed populations. To evaluate inter-population statistics, such as XP-EHH (Sabeti et al. 2007), we used a three-population demographic model calibrated to replicate the allele frequency spectrum, population structure and linkage disequilibrium levels in African (YRI), European (CEU) and East Asian (CHB) populations (Grossman et al. 2013; Grossman et al. 2010; Schaffner et al. 2005). We privileged this model because it has been used to detect selective sweeps in the YRI, CEU and CHB 1000G populations on the basis of simulation-based approaches (Grossman et al. 2013; Pybus et al. 2015). This model incorporates an African expansion, an Out-of-Africa exodus ~100 kya (3,500 generations) followed by a bottleneck, a split of Eurasians into European and Asian populations ~58kya and various migration rates between continents (~10^{-5} per haploid genome per generation). An interesting feature is the presence of a second bottleneck in each non-African population being four times stronger in Asia than in Europe (Pickrell et al. 2009). We systematically simulated triplets of African, European and East Asian populations. To simulate the *X* selective sweeps per WGS data in the focal population, we set $X = 0$ in the two other populations used as reference for the XP-EHH computation and simulated the non-equilibrium demography of each population using the three-population demographic model presented above.

We used SLiM (Haller and Messer 2017) to simulate 5Mb regions using human recombination rates drawn from HapMap recombination maps (Frazer et al. 2007). We

simulated 10^4 neutrally evolving DNA regions and $4x10^3$ DNA regions with selection for each sweep scenario envisaged (see the results section). In the latter case, we simulated selection in the focal population (Africa, Europe or East Asia) by inserting a single advantageous mutation in the middle of the region. At generation *t* being the onset of selection, the frequency of this mutation p_{start} was randomly drawn from the allele frequency spectrum at the generation *t*. In all the sweep scenarios investigated, we used initial frequency of the selected mutation p_{start}, intensity of selection *s* and age of selection *t* with the same ranges of values, i.e., respectively ranged from $1/2N$ to 0.2, from 0.001 to 0.05 and from present to 3,500 generations ago. We simulated long DNA regions to avoid premature truncations because selection signals can extent over mega bases for particularly strong sweeps, i.e., ~2Mb in the LCT region (various estimates of s for rs4988235 range from 0.025 to 0.069) (Chen et al. 2015; Peter et al. 2012; Tishkoff et al. 2007). Because SLiM is a forward-in-time simulator, the computation time depends on both the effective population size *N* and the number of generations considered. We thus rescaled effective population sizes and times according to N/λ and t/λ, with $\lambda = 10$, and used rescaled mutation, recombination and selection parameters, $\lambda\mu$, λr and λs (Haller and Messer 2017; Hoggart et al. 2007).

Lastly, simulated genomes were obtained by randomly drawing neutral simulated regions and regions simulated under the desired sweep scenario, some of which were considered to be PSVs, the rest being ENVs. The age *t* and the intensity *s* of selection, used to simulate the *X* sweeps per WGS data located in the PSV regions, were randomly drawn from various distributions depending on the sweep scenario investigated. We assigned to each SNP an indicator function $I(PSV)$ equal to 1 for the SNPs in PSV regions, which was subsequently used to compute the ORs (see above). In summary, the PSVs are simulated as tracts of SNPs with $I(PSV = 1)$ randomly spread over the genomes to reproduce the proportions of ENVs and PSVs observed in the 1000G populations, with numbers of simulated SNPs matching

those observed in 1000G populations. In the following, for each ABC estimation performed, we used 10^5 simulated WGS data, namely the ABC simulations, with X randomly drawn from a uniform prior distribution, $U(0, 300)$. A graphical illustration summarizing the different steps to simulate WGS data can be found in the Supplemental Material.

1000 Genomes populations analyzed

Analyses were performed on the 1000 Genomes Project phase 3 data, focusing on African, European and East Asian populations and excluding populations with diverse continental or admixed ancestry. We analyzed 1,511 individuals from five African, five European and five East Asian populations (85 to 113 individuals per population). Phased data (SHAPEIT2) (Delaneau et al. 2012), ancestral/derived states and VEP annotations were downloaded from the 1000G Project website.

Neutrality statistics

For each simulated and 1000G WGS, we computed six neutrality statistics. We used the haplotype-based neutrality statistics, iHS (Voight et al. 2006), DIND (Barreiro et al. 2009) and ΔiHH (Grossman et al. 2010), which compare the haplotypes carrying the derived and ancestral alleles. We computed two pairwise XP-EHHs (Sabeti et al. 2007), which compare the haplotypes carrying the derived allele between the focal population and two other populations of differing continental origin used as reference (for the 1000G WGS the two reference populations were chosen from YRI, CEU or CHB). We also used the Fay and Wu's *H* (F&W-H) (Fay and Wu 2000), which detects deviations from the neutral allele frequency spectrum in genomic regions. We used a sliding-windows approach for these computations (100 kb windows centered on each SNP) (Fagny et al. 2014). The sliding-windows began and ended 50kb from the edges of the 5Mb simulated regions, to prevent truncation in the 100 kb

sliding windows (a similar approach was applied to the 1000G chromosomes). As iHS, DIND, ΔiHH and XP-EHH are sensitive to the inferred ancestral/derived state, we computed these statistics only when the derived state was determined unambiguously and we normalized them by DAF bin (Fagny et al. 2014; Voight et al. 2006) (mutations grouped by DAF bin, from 0 to 1, in increments of 0.025). We applied classic filters by excluding minor alleles frequencies below 0.05 and we minimized the false-positive discovery by excluding SNPs with a DAF below 0.2, as the power to detect positive selection has been shown to be reduced at such low frequencies (Fagny et al. 2014; Voight et al. 2006). Finally, for each neutrality statistic, we defined the candidate SNPs of selection (top 1% of SNPs genome-wide). For iHS, DIND, ΔiHH and XP-EHHs, we considered extreme values indicative for selection targeting the derived alleles.

Because background selection may generate spurious positive selection signatures genome-wide (Coop et al. 2009; Hernandez et al. 2011; Pritchard et al. 2010), we excluded the F_{ST}, Tajima's D and other neutrality statistics previously found to be affected by BGS (Zeng et al. 2006). Indeed, the differences in allele frequencies between populations are expected to be exacerbated in regions affected by BGS, a pattern that can be confounded with positive selection (Coop et al. 2009; Pickrell et al. 2009; Pritchard et al. 2010). We only retained Fay and Wu's H and the haplotype-based statistics also previously found to be insensitive to BGS (Fagny et al. 2014; Zeng et al. 2006), e.g., the haplotype-based statistics use haplotypes carrying the ancestral alleles as internal controls that should be affected by BGS to a similar extent.

ABC, acceptance rules and accuracy evaluation

We used the 'abc' R package and the standard ABC method (Beaumont et al. 2002), in which posteriors are built from accepted simulated parameters subsequently adjusted by local linear

regression (method="Loclinear" in the 'abc' package). The accepted simulated parameters are those which provide the best fit with empirical data (the 'abc' parameter 'tol' was set to be equal to 0.005). We used the mean of the posterior distribution as point estimate and computed the 95% credible interval from these accepted parameters. To test the accuracy of our estimations, we compared estimated and simulated parameter values, $\widehat{\theta_i}$ and θ_i respectively, using classical accuracy indices: the relative error *rE* (i.e. difference between estimated and true values, expressed as a proportion of the true value, $rE = \left(\widehat{\theta_i} - \theta_i\right)/\theta_i$, $i = 1, \ldots, J$), the relative root of the mean square error, *r*RMSE (i.e. the root of the MSE expressed as a proportion of the true value), and the proportion of true values within the 90% credible interval of estimates, $90\%COV = \frac{1}{J}\sum_1^J 1(q_1 < \theta_i < q_2)$ where 1(*C*) is the indicative function (equal to 1 when *C* is true, 0 otherwise) and q_1 and $q_{2,}$ the corresponding percentiles of the posterior distributions.

ABC estimations obtained simulating BGS in PSVs or changing the PSVs annotation

To check the insensitivity of our estimation to background selection, we simulated widespread background selection in PSVs: we simulated DNA regions with a fraction of mutations targeted by purifying selection, determined following the coding region maps, in which 2/3 of mutations were considered to be nonsynonymous with selection coefficients randomly drawn from the Gamma distribution of fitness effects determined in Boyko et al. (2008). These new simulations are used as PSV regions in the ABC simulations and we re-estimated the parameters accordingly. In addition, for the reasons described above, the SNPs with a genome-wide significant enrichment in candidate SNPs 100 kb around them were annotated as PSVs to perform the ABC estimations. To check these estimations, we performed another round of estimations after excluding from the analysis all ENVs with such a significant enrichment (some selection signals due to selection targets located close to the

edges of PSVs were thus truncated while other selection signals that cover several genes and intergenic regions were diminished, Table S2C).

Detecting genomic regions enriched in candidate SNPs of selection

The enrichment of candidate SNPs in a given genomic region was determined by means of a combined selection score (CSS) computed using the same neutrality statistics used to estimate *X*. For each SNP and each of the K = 6 neutrality statistics used in the ABC estimations, we computed the proportion of candidate SNPs in a 100 kb window around the considered SNP. We next determined the empirical *P*-values (*P*) for these proportions and combined them into a single combined selection score using a Fisher score, $CSS = -2 \sum_1^6 \log(P_i)$ (Deschamps et al. 2016). The rationale behind such a composite approach (Deschamps et al. 2016; Grossman et al. 2013; Grossman et al. 2010) is that neutrality statistics are more strongly correlated for positively selected variants than for neutral sites (Grossman et al. 2010). Consequently, false positives may harbor extreme values for a few neutrality statistics only, whereas SNPs genuinely selected (or nearby SNPs) should harbor extreme values for several statistics together, a feature captured by the combined score. Finally, the genomic regions enriched in candidate SNPs were defined as consecutive SNPs with genome-wide significant CSS values, $P<0.01$ (gaps of maximal length of 100kb are allowed). In these enriched regions, the maximum CSS value was used as a proxy of the strength of selection signal; high CSS values are expected for SNPs targeted by strong selection.

For each population and each enriched genomic region, we computed an overlap score calculated as the number of populations for which the same region was identified. Overlap scores were calculated either within or between continents (upper limits of 5 and 10 for within- and between-continent overlap scores respectively, because we analyzed three continents of five populations each). Within continents, the overlap scores range from 1 (the

considered region was identified in a single population) to 5 (the considered genomic region was identified in all populations of the same continental origin). Between continents, the overlap scores range from 0 (the considered region was never identified in another continent) to 10 (the considered region was identified in all populations of differing continental origin).

Data availability

Genome-wide data and SNPs annotations can be downloaded from the 1000 Genome website. The software used to compute the neutrality statistics can be downloaded from https://github.com/h-e-g/selink. The scripts used to perform the main analysis can be downloaded from https://github.com/h-e-g/ABCnumSweeps. Supplemental material, including Figures S1 to S26, Tables S1 to S4 (Tables S2-S4 are supplied as Excel files) and Files S1 and S2 can be found with this article online.

RESULTS

Odds Ratios for selection capture the number of occurring sweeps

We estimated the number of selective sweeps using ABC and summary statistics that measure the enrichment of selection signals in PSVs regions, relative to ENV regions; a graphical illustration summarizing the rationale behind our approach is shown in Figures 1 and S1. Because ABC inferences are based on summary statistics that exhibit a monotonic relationship with the parameter to estimate, we first present the relationships between the odds ratio for selection (OR) and the number of sweeps in African, European and East Asian simulated populations. We explored two different scenarios (Figure 2, Supplemental Material, Figures S2 and S3): in the first, we simulated *X* incomplete sweeps per whole-genome sequence (WGS) dataset (Materials and Methods) by randomly drawing *s* and *t* from flat distributions, $U(0.001, 0.05)$ and $U(0, 100)$ kya, and excluding complete sweeps using a

rejection algorithm (Figure 2). The underlying distributions of *s* and also *t* are therefore enriched in low values (Figure 2A), since (all other things being equal) complete sweeps tend to be stronger or older than incomplete sweeps. In a second, more realistic scenario, we simulated *X* sweeps per WGS data while keeping the complete sweeps (Figure S3). The age of selection was randomly drawn from a flat distribution, $U(0, 100)\ kya$, whereas we aimed to reproduce the excess of mutations with low or moderate effect on fitness (Boyko et al. 2008) with a distribution enriched in low values (a mixture between a Gamma distribution with 60% of $s < 0.01$ and a L-shape distribution with 90% of $s < 0.01$, Figure S3A).

For each scenario, we simulated WGS data with fixed numbers of sweeps, $X = [0, 50, 100, 150]$, using human recombination maps and a demographic model previously inferred for these populations (Materials and Methods). Our simulations clearly show that the OR well captures the enrichment of candidate SNPs in PSVs owing to the action of selection (Figures 2B and S3B). PSV regions with more than 1% of candidate SNPs in the vicinity of the selection targets contribute to the greater proportions of candidate SNPs in PSVs relative to ENVs ($OR = 1$ for $X = 0$ *vs* $OR > 1$ otherwise). As assumed, the six ORs increase with the number of simulated sweeps, resulting in a monotonic relationship between *X* and the OR over the parameter space investigated. In addition, the OR values depend on the frequencies of alleles at the onset of selection (Figure S2); SSV typically have weaker effects on linked sites (Pritchard et al. 2010; Przeworski et al. 2005), which reduces the numbers of candidate SNPs around selection targets and ultimately the OR. However, very similar ORs were obtained for SDN and SSV with p_{start} lower than 0.01. This is expected since such sweeps are difficult to distinguish due to similar signatures when p_{start} tends to $1/2N$ (Ferrer-Admetlla et al. 2014; Jensen 2014; Peter et al. 2012). In light of this, we focused on the estimation of the numbers of sweeps corresponding to these two sweep models merged ($p_{start} < 1\%$, X_1).

Odds Ratios for selection are moderately sensitive to demographic assumptions

The ORs values follow the power to detect selection as previously assessed for iHS and XP-EHH (highest, intermediate and lowest power in the African, European and East Asian demography respectively) (Pickrell et al. 2009). The lower European and East Asian ORs relative to those in Africans (Figures 2B and S3B), which reflect lower enrichments of candidate SNPs in PSVs, are in agreement with a reduced power to detect selection in bottlenecked relative to expanding populations (Fagny et al. 2014; Grossman et al. 2013; Gunther and Schmid 2011; Huff et al. 2010; Pickrell et al. 2009). However, little differences in the behavior of ORs were observed across the populations simulated with contrasting demographic histories (i.e., African expansion *vs* Eurasian bottleneck). Whereas the East Asian bottleneck is stronger than that in Europe, the East Asian and European simulated ORs are virtually identical (e.g., 2.3 *vs* 2.4 in average for DIND with $X = 100$, Figure 2B). These observations suggest that the ABC estimations of X will be moderately affected by potential demographic misspecifications.

Accuracy of the ABC estimations

To assess the accuracy of the estimations in the two sweep scenarios explored, we treated the simulations shown in Figures 2 (incomplete sweeps) and S3 (complete and incomplete sweeps) as pseudo-empirical data for which the parameter values are known. We found virtually unbiased estimations of X (Figure 3). An in-depth analysis (Figures S4-S12) showed a similar accuracy for the estimations of X_1, X_2 and X_3 under the two assumed scenarios (Figures S4 and S8). When simulating complete sweeps, the neutrality statistics with low power to detect sweeps near to fixation, iHS or DIND, remained informative as they captured the enrichment of candidate SNPs around beneficial mutations at fixation (Figure S3). This illustrates the importance of incorporating in PSVs the SNPs in the vicinity of selection

targets to leverage information provided by the hitchhiking effects of selected variants. Our computer simulations showed that incorporating nearby SNPs in PSVs, which encompass 70% of the simulated genome, did not dilute the true adaptive signals. Using such proportions of simulated PSVs allow to define large 1000G PSVs regions by incorporating all the functional mutations and neighboring variants to capture a large fraction of the existing selective sweeps, including selected regulatory mutations located up/down-stream genes (see the next section). However, because of largely overlapping posterior distributions, our method is not able to distinguish between estimations obtained when the real number of sweeps is low (real $X < 10$) and those obtained under neutrality (real $X = 0$). Finally, we noticed a diminished accuracy of *S* and *T*, e.g., underestimations and overestimations of *S* and *T* respectively when the selection is strong and recent in average (Figures S5 and S9), together with large CIs that sometimes exceed the range of simulated priors (Figures S7 and S11).

The estimations of *X* based on the combined six ORs performed reasonably well. They are equal -- in expectation -- to the numbers of occurring sweeps (Figure 3), with CIs that consistently predict the range of uncertainty within which the true parameters are expected (Figures S6 and S10), as shown by the 90% *COV* indicated in Figures S4 and S8. Thus, our estimations of *X* appeared to be neither systematically biased upward by neutral genomic regions with spurious signatures of selection nor systematically biased downward by sweep regions with few candidate SNPs (i.e., false negatives). As expected, false positives do not contribute to ORs, owing to the use of ENVs as an internal neutral control. In the absence of selection ($X = 0$ in simulated pseudo-empirical data), the estimations of *X* are close to 0 as ORs are close to one. Under selection ($X > 0$ in simulated pseudo-empirical data), the expected proportion of false negatives is accounted by the model thanks to the simulation of parameters that normally drive the statistical power to detect sweeps, such as the age of selection, the intensity of selection or the demography. Altogether, our method accurately

estimates *X* from OR values; yet, the estimations were obtained under a known demography whereas model-based methods are expected to be affected by partially unknown demography. We thus evaluated the effect of incorrect demographic assumptions on the estimations of *X* in 1000G populations.

Low numbers of selective sweeps in the 1000G populations

We applied our method to each African, European and East Asian 1000G population analyzed separately. To capture the selection signals as exhaustively as possible, the PSVs defined using the VEP annotation of the 1000G data (Materials and Methods, Figure S13) correspond to large genomic regions as variants near functional mutations must be included in PSVs to account for the clustering of candidate SNPs around putative selection targets. We used the ABC simulations with the same proportion of simulated PSVs (Figure 3) and 1000G ORs corrected for genomic variation in coverage, mutation and recombination rates (Kudaravalli et al. 2009) (Materials and Methods, Figure S14), which were found to be compatible with the simulated ORs used to perform the estimations (Figure S15). While the incomplete sweep scenario is not applicable to real populations since it ignores complete sweeps, the results obtained under this scenario are reported for comparison (Tables 1 and 2, Figures 4 and 5). We noticed a stronger influence of this assumption in Eurasia (Table 1), mainly in East Asia, likely because sweeps can reach fixation faster due to lower effective size (large proportions of the fixed differences between human populations are due to fixation events outside of Africa, mainly in East Asia) (Coop et al. 2009).

In all cases, we found lower numbers of sweeps than previously detected (Pybus et al. 2015; Schrider and Kern 2017), i.e., 116 sweeps in average across all populations (Table 1) with 198 [150-240] sweeps at most in the CHS population (Table S2A) (see also Figure 4, Table S2B and Figures S16 and S17). This result was found to be robust to the inclusion of

background selection in simulated PSVs and to the mode of computation of the 1000G ORs (Materials and Methods), whether they are corrected for various covariates (Table 1) or not (Table S2C). In Africa, we found that an average of ~70 sweeps occurred over ~100,000 years account for the genome-wide selection signal, or probably a few less since older sweeps are ignored. Such sweeps are not expected to have inflated much the 1000G ORs computed from neutrality statistics known to have low power to detect old selection events in humans (Grossman et al. 2013; Voight et al. 2006).

Beside the estimations of *X*, we obtained average strengths of selection close to 0.01 and average ages of selection compatible with expected ages of selection, e.g., recent ages when considering incomplete sweeps only (Table 2, Tables S2A and S2B). However, the CIs obtained for *S* and *T* covered the whole range of the simulated priors (Figures S7 and S11), preventing precise inferences for these parameters. In addition, we mapped the genomic regions significantly enriched in candidate SNPs by means of a classic genome-wide scan of selection performed with the same neutrality statistics used to estimate *X* (Materials and Methods). Reassuringly, the regions that most contribute to the ORs capture many examples of previously reported selected regions (Table S3), including iconic cases such as *TLR5* in Africa, *LCT* in Northern Europe and *EDAR* in East Asia (Fan et al. 2016; Jeong and Di Rienzo 2014; Vitti et al. 2013) (Figures S18 and S19, File S1 for a short description of the genomic regions identified).

We found a low overlap of selection signals between continents (Materials and Methods), confirming previous observations (Voight et al. 2006), while the most enriched regions tend to be highly shared across the 1000G populations from the same continent (Figure S20, Tables S3A and S3B), in agreement with recent ages of selection estimated under the incomplete sweep scenario. Indeed, with the neutrality statistics used herein, incomplete sweeps, younger by definition and thus more continent-specific than complete sweeps that

may have occurred before the split of Eurasian populations, result in the highest enrichment in candidate SNPs and consequently the lowest empirical P-values in our selection scan. We also tested the sensitivity of our ABC estimations to the use of neutral reference populations for the XP-EHH statistics. We thus performed another round of ABC estimations after removing the XP-EHH ORs, which can be impacted by shared selective sweeps. We found that estimates of *X* were unchanged (Figure S21), with similar average ages of selection, e.g., 36.0 [23.5-45.8] kya under the incomplete sweep scenario, suggesting that our estimations are not sensitive to the use of neutral outgroups.

Consistent detection of low numbers of selective sweeps in humans

To test the discrepancies observed between our results and the high numbers of sweeps previously reported (Schrider and Kern 2017), we used our highest estimations of *X*, i.e., those obtained with the 1000G ORs averaged across chromosomes (Table 1). First, estimations performed using different combinations of neutrality statistics, or modifying several technical steps, were consistently found low (Figure S21, Table S2C). For example, excluding the most correlated neutrality statistics and their corresponding ORs does not affect our ABC estimations (Figures S21 and S22). We next checked the influence of the demographic model assumed (Figures S23 and S24). We adopted a strategy based on stress tests to evaluate the differences in numbers of estimated sweeps when modifying the demographic assumptions, e.g., replacing an expansion with a bottleneck or increasing the bottleneck intensity. We performed this by swapping empirical and simulated ORs from differing continental regions, e.g., estimations performed using the 1000G African ORs with a "wrong" East Asian simulated demography and inversely. In doing so, we obtained similar numbers of sweeps (Tables S2D and S2E).

582 The graphical explanations given in Figure S23 illustrate how the new estimations of *X*
583 may differ. For example, we re-analyzed the European populations under an East Asian
584 demographic model, and found higher numbers of sweeps than initially estimated, i.e., 141
585 [95-187] *vs* 115 [68-160] sweeps on average (Table S2E *vs* Table 1). These higher estimations
586 of *X* are due to lower simulated ORs in East Asia compared to Europe (Figure 2B, Figure
587 S3B). The estimations are increased because the East Asian ORs simulated with higher *X*
588 values provided a better fit with the European 1000G ORs (see the $X_{Europe}^{(\mathrm{ASIdemo})}$ in the Figure
589 S23). The low discrepancies obtained were confirmed using pseudo-empirical datasets
590 simulated with bottlenecks four times stronger (or weaker) than in the ABC simulations used
591 for the estimations (Figure S24). Overall, we did not obtain point estimates that largely
592 exceeded the estimations shown in Tables 1 and S2. We acknowledge that new ABC
593 estimations based on different demographic models may differ. Nevertheless, our results
594 suggest that they should be of the same order of magnitude.
595 Lastly, we performed estimations of *X* in the studied 1000G populations using various
596 distributions of *s*. In all cases, we found low numbers of selective sweeps, although we
597 noticed a moderate influence of the shape of the distributions of *s* on the estimations of *X*
598 (Table S2F). As expected, the estimations of *X* are inversely correlated with the intensity of
599 selection (e.g., two sweeps of high intensity should result in ORs similar to those obtained
600 with more sweeps of lower intensity). For example, we estimated 70 [35-106] and 126 [90-
601 162] sweeps in average assuming a flat (~20% of $s < 0.01$ with an average equal to ~0.025)
602 or a L-shape distribution of *s* (~90% of $s < 0.01$ with an average equal to ~0.007). Under
603 these two extreme scenarios, the posterior distributions obtained are largely non-overlapping
604 but the point estimates are of the same order of magnitude (Table S2F). Note that we did not
605 consider this flat distribution to set the composite distribution of *s* used to perform our main
606 estimations (Tables 1 and 2, Figures 4 and 5), because assuming similar proportions of

sweeps with high and low intensity is an unrealistic assumption. To exclude that our estimations are due to an underestimation of X when X is large, we verified the ability of our method to infer high numbers of selection events when they are present in the data using pseudo-empirical WGS simulated with high numbers of sweeps. We used priors extended up to 1,000 sweeps per population ($X \sim U(0, 1000)$) and found unbiased estimations of X (Figure S25). We then re-estimated X in each 1000G population, and found similar estimations (Table S2G), supporting further our findings of low numbers of sweeps occurred.

A trend towards higher numbers of sweeps occurred in non-African populations

We found more sweep signals in Eurasian populations (Table 1), an observation that has been attributed to a greater extent of genetic adaptation outside Africa due to drastic changes of environmental pressures (Coop et al. 2009; Granka et al. 2012; Pybus et al. 2015; Schrider and Kern 2017). For example, we detected a recent European-specific selection signal likely due to the late Pleistocene warming at the end of the last ice age in Northern hemisphere (Cooper et al. 2015), a cold climatic period which could favor both light skin pigmentation and increased sensitivity to UV-induced melanoma (Key et al. 2016; Lopez et al. 2014). Regulatory mutations, protective against melanoma as they downregulate the *PLEK2* gene in skin cells exposed to the sun (GTEx database) (Ardlie et al. 2015), exhibit genome-wide significant signals of positive selection in Europe only (Table S3, Figure S26, and File S2). However, we cannot exclude that the higher X estimated in Eurasia, particularly in East Asia (see the non-overlapping CIs in Table 1), are due to the demographic model used (our stress tests and simulations also indicated that the estimations of X in East Asia could be lower under models with weaker bottlenecks, Table S2E, Figure S24). A more formal comparison would require a perfectly known demography or the inclusion of alternative demographic models.

With respect to the numbers of sweeps as a function of the initial frequency of selected alleles, X_2 and X_3 ($0.01 \leq p_{start} < 0.2$) were found in excess compared to X_1 ($X_1 < X_2 + X_3$, Figure 5), particularly in Eurasian populations. This observation may reflect the loss of selected alleles at very low frequency (X_1) during the first generations of selection, which is exacerbated in bottlenecked populations relative to expanding populations. However, in all cases our results confirm sweeps from standing variation as main drivers of adaptation (Schrider and Kern 2017), since sweeps from *de novo* mutations are contained in the X_1 category (numbers of SDN ranged from 0 to X_1). In practice, the neutrality statistics used to identify sweeps detect large frequency changes (Ferrer-Admetlla et al. 2014; Schrider and Kern 2017) whereas they are not sensitive to moderate changes in allele frequencies due to polygenic adaptation (Field et al. 2016; Pritchard and Di Rienzo 2010; Pritchard et al. 2010; Stephan 2016); some major loci controlling complex traits with extreme phenotypes recently selected (Perry et al. 2014) may resemble incomplete sweeps.

DISCUSSION

Our study shows that ABC approaches can be helpful to formally assess the extent and nature of recent positive selection in humans, and provide valuable information about the proportion of true sweeps identified by selection scans. A result emerging from this study is the low contribution of sweeps to recent human adaptation, including sweeps from standing variation, which supports the rarity of recent selective sweeps in humans (Coop et al. 2009; Harris et al. 2018; Hernandez et al. 2011; Pritchard et al. 2010). In our analyses, ~70% of the genome was considered as potentially influenced by selection and that all significant signals of positive selection found in intergenic regions were accounted for suggests that we capture a large fraction of the existing selective sweeps. Yet, our estimations are far lower than numbers of recent sweeps individually detected using classic selection scans; they correspond to ~35% of

657 the numbers of putatively selected regions identified (Table S3C). They also correspond to
658 ~30% and ~10% of sweeps identified using machine learning algorithms based on simulations
659 of the same demographic model used herein (Pybus et al. 2015) and on similar models of
660 selection from standing variation (Schrider and Kern 2017), respectively. Such discrepancies
661 are in agreement with studies in *Drosophila* that have shown that the numbers of sweeps
662 identified using scans of selection are roughly an order of magnitude greater than would be
663 predicted, likely because of high false positive rates (Jensen 2009; Teshima et al. 2006).
664 To evaluate the validity of our estimations, we compared estimations of adaptation rates
665 obtained in human and *Drosophila*. We first compared with α, the fraction of nonsynonymous
666 mutations driven to fixation by positive selection (Messer and Petrov 2013), estimated in
667 1000G African populations using ABC and 29,925 nonsynonymous fixed differences with
668 chimps (Uricchio et al. 2019). With the number of sweeps estimated in Africa (Table 1)
669 translated in number of sweeps per generation, we predicted α in the human lineage assuming
670 divergence times between human and chimp of ~7.9Mya (Moorjani et al. 2016a) (Tables 3
671 and S4A). These predictions assume constant numbers of beneficial mutations per generation
672 whereas these numbers did vary over time due to changes in environmental conditions or in
673 effective sizes (Hawks et al. 2007). Our predicted values of α, which can be higher given that
674 we also considered selection targeting regulatory sites, are quite similar to those previously
675 estimated by Uricchio et al. (2019) and Zhen et al. (2021) (Table 3). This indicates that our
676 results are compatible with previous estimates obtained at broader evolutionary scales and
677 using different methods and data.
678 We next compared our results with the numbers of beneficial mutations per generation per
679 nucleotide site (λ) estimated in *Drosophila* under a recurrent hitchhiking model (Kaplan et al.
680 1989; Przeworski 2002; Wiehe and Stephan 1993) with sweeps occurring at random locations
681 in the genome (Andolfatto 2007; Jensen et al. 2008; Li and Stephan 2006; Macpherson et al.

2007). The estimations of λ previously obtained in *Drosophila* largely vary across different studies (Table 4). Using the per generation numbers of sweeps obtained with our approach, which can predict λ in humans (Tables 4 and S4B), we found λ values similar or even higher than estimated in *Drosophila*. A recent analysis suggests that the species with a smaller population size and greater complexity (i.e., humans) may have stronger and/or more abundant new beneficial mutations than other species with much larger population sizes (i.e., mice and D. melanogaster) (Zhen et al. 2021), suggesting that the complexity of the organisms affects adaptation rates. Otherwise, we found lower λ values when the intensity of selection estimated in *Drosophila* is much lower than that estimated in this study (Table 4). For example, the λ estimated by Andolfatto (2007) is approximately two orders of magnitude higher than ours, which roughly corresponds to the expected differences in *N* values between *Drosophila* and humans ($N{\sim}10^6$ vs $N{\sim}10^4$). Empirical evidence in great apes also suggests that the numbers of experienced sweeps per unit of time may increase with effective population sizes (Nam et al. 2017). Beside our predicted λ, our results suggest a lower impact of beneficial mutations on the human genome-wide diversity compared to *Drosophila*, as indicated by lower $2N\lambda S$ values (Table S4B) (the reduction of heterozygosity below the neutral level as a function of the recombination distance to selected sites is essentially determined by this compound parameter) (Wiehe and Stephan 1993). Applying our methodology to *Drosophila* data may help to understand if lower $2N\lambda S$ are due to lower λ or lower $2NS$ driving lower rates of detectable beneficial mutations.

From a methodological standpoint, our ABC framework appears to provide unbiased estimations of *X* when the distribution on *s* is known. Although here we leveraged widely used neutrality statistics, the OR can be computed for any other kind of statistics, keeping in mind that background selection should be simulated (Johri et al. 2020) when using statistics sensitive to this selection regime. Our study presents however some limitations. We noticed a

dependency on the *s* distributions used, which suggests that a careful evaluation of their influence on the estimation of *X* is needed. We also observed a lack of accuracy in the estimation of *S* or *T*, indicating that other summary statistics that better describe the genomic extent of selection signals should be incorporated in selection analysis. Another important point that should be considered is the calibration of the period of selection studied, which depends on the sensitivity of the neutrality statistics to the age of selection simulated. For example, Li and Stephan (2006) used statistics with low power to detect selection events older than 60,000 years in *Drosophila*. Here, we used neutrality statistics with a low power to detect sweeps that are older than 100,000 years in humans (Grossman et al. 2013; Voight et al. 2006). Such sweeps, which are ignored by the model, marginally contribute to empirical data (i.e., the 1000G ORs) and should have thus a low influence on the final estimations (Li and Stephan 2006). Because simulating whole chromosomes with selection is barely tractable, we concatenated shorter regions together and avoided the computation of neutrality statistics across the junctions to ensure that they were computed over genomic regions simulated according to human recombination rates. Also, our ABC estimates were obtained using the common assumption that the demography is known, which can affect the validity of model-based estimations. The minor differences in point estimates observed within continents are likely due to specific local demographic histories not accounted by the model, e.g., some lower *X* in the Finnish population are potentially due to strong bottlenecks (Bulik-Sullivan et al. 2015) not accounted by the CEU demography. However, our stress tests indicate that the low numbers of sweeps estimated herein are likely not due to demographic misspecifications, although more efforts are needed to refine the estimations. Future work simulating, in reasonable computation times, whole chromosomes with beneficial mutations in genic regions and incorporating recent population admixture as revealed by ancient genomes (Skoglund and Mathieson 2018) into ABC should provide more refined estimations of the true numbers of

732 selective sweeps occurred in humans. Furthermore, while our results suggest that our method
733 is not sensitive to the use of neutral outgroups, the impact of this assumption needs to be
734 formally evaluated using more complex scenarios of shared selection with adaptive gene flow
735 (Laso-Jadart et al. 2017; Patin et al. 2017; Refoyo-Martinez et al. 2019).
736 To conclude, our study provides novel evidence in support of the paucity of selective
737 sweeps in humans, with posteriors distributions of X that consistently excluded a number of
738 sweeps higher than ~250 in all investigated scenarios. The observed enrichments of candidate
739 SNPs in and near genes were found to be lower than those obtained with computer
740 simulations assuming high numbers of sweeps (Figure S25A-C). Furthermore, the α
741 adaptation rates predicted from our estimations were found compatible with those estimated
742 using different sources of data and methods. Our estimated numbers of recent sweeps in
743 humans are helpful to better understand the outcomes of classic scans of selection, keeping in
744 mind that genomic regions evolving under complex selection regimes can only be detected
745 with powerful methods such as those recently implemented (Schrider and Kern 2017). The
746 methodology proposed in this study also provides an alternative to methods implemented in
747 *Drosophila*, and can also be applied to great apes (Nam et al. 2017; Prado-Martinez et al.
748 2013; Schmidt et al. 2019) or other mammals (Freedman et al. 2016; Ihle et al. 2006; Librado
749 et al. 2015; Roux et al. 2015; Stella et al. 2010). Adapting these approaches to study more
750 complex selection regimes, such as polygenic adaptation, should provide a broader, and more
751 precise, picture of the recent adaptive history of humans and other species.
752
753 **Acknowledgments**
754 We thank the IT infrastructure of Institut Pasteur (Paris) for the management of computational
755 resources. We also thank two anonymous reviewers for their fruitful comments and
756 suggestions which greatly improved the quality of our manuscript. This work was supported

by the Agence Nationale de la Recherche (ANR) grant "DEMOCHIPS" ANR-12-BSV7-0012. The Human Evolutionary Genetics laboratory is supported by the Institut Pasteur, the Collège de France, the CNRS, the Fondation Allianz-Institut de France, and the French Government's Investissement d'Avenir program, Laboratoires d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (ANR-10-LABX-62-IBEID) and 'Milieu Intérieur' (ANR-10-LABX-69-01).

Literature Cited

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res. 19(5):711-722.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the drosophila melanogaster genome. Genome Res. 17(12):1755-1762.

Ardlie KG, DeLuca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M et al. 2015. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. Science. 348(6235):648-660.

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature. 526(7571):68-74.

Barker JSF. 1962. The estimation of generation interval in experimental populations of drosophila. Genet Res. 3(3):388-404.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B et al. 2009. Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. PLoS Genet. 5(7):e1000562.

781 Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has
782 driven population differentiation in modern humans. Nat Genet. 40(3):340-345.
783 Beaumont MA, Zhang W, Balding DJ. 2002. Approximate bayesian computation in
784 population genetics. Genetics. 162(4):2025-2035.
785 Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE,
786 Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al. 2008. Assessing the
787 evolutionary impact of amino acid mutations in the human genome. PLoS Genet.
788 4(5):e1000083.
789 Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL,
790 Neale BM. 2015. Ld score regression distinguishes confounding from polygenicity in
791 genome-wide association studies. Nat Genet. 47(3):291-295.
792 Chen H, Hey J, Slatkin M. 2015. A hidden markov model for investigating recent positive
793 selection through haplotype structure. Theor Popul Biol. 99:18-30.
794 Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza
795 LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation.
796 PLoS Genet. 5(6):e1000500.
797 Coop G, Wen XQ, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of
798 crossovers reveals extensive variation in fine-scale recombination patterns among
799 humans. Science. 319(5868):1395-1398.
800 Cooper A, Turney C, Hughen KA, Brook BW, McDonald HG, Bradshaw CJ. 2015.
801 Paleoecology. Abrupt warming events drove late pleistocene holarctic megafaunal
802 turnover. Science. 349(6248):602-606.
803 Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands
804 of genomes. Nat Methods. 9(2):179-181.

805 Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, Patin E, Quintana-Murci L.
806 2016. Genomic signatures of selective pressures and introgression from archaic
807 hominins at human innate immunity genes. Am J Hum Genet. 98(1):5-21.
808 Eyre-Walker A. 2006. The genomic rate of adaptive evolution. Trends Ecol Evol. 21(10):569-
809 575.
810 Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the
811 occurrence of classic selective sweeps in humans using whole-genome sequencing
812 data sets. Mol Biol Evol. 31(7):1850-1868.
813 Fan S, Hansen ME, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of
814 recent human adaptation. Science. 354(6308):54-59.
815 Fay JC, Wu CI. 2000. Hitchhiking under positive darwinian selection. Genetics. 155(3):1405-
816 1413.
817 Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in
818 genetics-based population divergence studies. Am J Phys Anthropol. 128(2):415-423.
819 Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft
820 or hard selective sweeps using haplotype structure. Mol Biol Evol. 31(5):1275-1291.
821 Field Y, Boyle EA, Telis N, Gao ZY, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel
822 P, McCarthy MI et al. 2016. Detection of human adaptation during the past 2000
823 years. Science. 354(6313):760-764.
824 Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau
825 A, Hardenbol P, Leal SM et al. 2007. A second generation human haplotype map of
826 over 3.1 million snps. Nature. 449(7164):851-861.
827 Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, Silva
828 PM, Galaverni M, Fan Z, Marx P et al. 2016. Demographically-based evaluation of
829 genomic regions under selection in domestic dogs. PLoS Genet. 12(3):e1005851.

Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. 2012. Limited evidence for classic selective sweeps in african populations. Genetics. 192(3):1049-1064.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH et al. 2013. Identifying recent adaptations in large-scale genomic data. Cell. 152(4):703-713.

Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science. 327(5967):883-886.

Gunther T, Schmid KJ. 2011. Improved haplotype-based detection of ongoing selective sweeps towards an application in arabidopsis thaliana. BMC Res Notes. 4:232.

Haller BC, Messer PW. 2017. Slim 2: Flexible, interactive forward genetic simulations. Mol Biol Evol. 34(1):230-240.

Harris RB, Sackman A, Jensen JD. 2018. On the unfounded enthusiasm for soft selective sweeps ii: Examining recent evidence from humans, flies, and viruses. PLoS Genet. 14(12):e1007859.

Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proc Natl Acad Sci USA. 104(52):20753-20758.

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

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. Science. 331(6019):920-924.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. Genetics. 177(3):1725-1731.

Huff CD, Harpending HC, Rogers AR. 2010. Detecting positive selection from genome scans of linkage disequilibrium. BMC Genomics. 11:8.

Ihle S, Ravaoarimanana I, Thomas M, Tautz D. 2006. An analysis of signatures of selective sweeps in natural populations of the house mouse. Mol Biol Evol. 23(4):790-797.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci USA. 101(29):10667-10672.

Jensen JD. 2009. On reconciling single and recurrent hitchhiking models. Genome Biol Evol. 1:320-324.

Jensen JD. 2014. On the unfounded enthusiasm for soft selective sweeps. Nat Commun. 5:5281.

Jensen JD, Thornton KR, Andolfatto P. 2008. An approximate bayesian estimator suggests strong, recurrent selective sweeps in drosophila. PLoS Genet. 4(9):e1000198.

Jeong C, Di Rienzo A. 2014. Adaptations to local environments in modern human populations. Curr Opin Genet Dev. 29:1-8.

Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. 2012. Genome-wide detection of natural selection in african americans pre- and post-admixture. Genome Res. 22(3):519-527.

Johri P, Charlesworth B, Jensen JD. 2020. Towards an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics. 215(1):173-192.

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

Key FM, Fu Q, Romagne F, Lachmann M, Andres AM. 2016. Human adaptation and population differentiation in the light of ancient genomes. Nat Commun. 7:10775.

879 Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of
880 molecular evolution. Nature. 267(5608):275-276.
881 Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters
882 GB, Jonasdottir A, Gylfason A, Kristinsson KT et al. 2010. Fine-scale recombination
883 rate differences between sexes, populations and individuals. Nature. 467(7319):1099-
884 1103.
885 Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2009. Gene
886 expression levels are a target of recent natural selection in the human genome. Mol
887 Biol Evol. 26(3):649-658.
888 Laso-Jadart R, Harmant C, Quach H, Zidane N, Tyler-Smith C, Mehdi Q, Ayub Q, Quintana-
889 Murci L, Patin E. 2017. The genetic legacy of the indian ocean slave trade: Recent
890 admixture and post-admixture selection in the makranis of pakistan. Am J Hum Genet.
891 101(6):977-984.
892 Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in
893 drosophila. PLoS Genet. 2(10):e166.
894 Librado P, Sarkissian CD, Ermini L, Schubert M, Jonsson H, Albrechtsen A, Fumagalli M,
895 Yang MA, Gambo C, Seguin-Orlando A et al. 2015. Tracking the origins of yakutian
896 horses and the genetic basis for their fast adaptation to subarctic environments. Proc
897 Natl Acad Sci USA. 112(50):E6889-E6897.
898 Lopez S, Garcia O, Yurrebaso I, Flores C, Acosta-Herrera M, Chen H, Gardeazabal J,
899 Careaga JM, Boyano MD, Sanchez A et al. 2014. The interplay between natural
900 selection and susceptibility to melanoma on allele 374f of slc45a2 gene in a south
901 european population. PLoS One. 9(8):e104367.

902 Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence
903 between nonsynonymous divergence and neutral polymorphism reveals extensive
904 adaptation in drosophila. Genetics. 177(4):2083-2099.
905 Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res.
906 23(1):23-35.
907 Mcdonald JH, Kreitman M. 1991. Adaptive protein evolution at the adh locus in drosophila.
908 Nature. 351(6328):652-654.
909 Messer PW, Petrov DA. 2013. Frequent adaptation and the mcdonald-kreitman test. Proc Natl
910 Acad Sci USA. 110(21):8615-8620.
911 Moorjani P, Amorim CE, Arndt PF, Przeworski M. 2016a. Variation in the molecular clock of
912 primates. Proc Natl Acad Sci USA. 113(38):10607-10612.
913 Moorjani P, Sankararaman S, Fu QM, Przeworski M, Patterson N, Reich D. 2016b. A genetic
914 method for dating ancient genomes provides a direct estimate of human generation
915 interval in the last 45,000 years. Proc Natl Acad Sci USA. 113(20):5652-5657.
916 Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of
917 recombination rates and hotspots across the human genome. Science. 310(5746):321-
918 324.
919 Nam K, Munch K, Mailund T, Nater A, Greminger MP, Krutzen M, Marques-Bonet T,
920 Schierup MH. 2017. Evidence that the rate of strong selective sweeps increases with
921 population size in the great apes. Proc Natl Acad Sci USA. 114(7):1613-1618.
922 Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic
923 variation. Genetics. 157(2):875-884.
924 Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH,
925 Barreiro LB, Froment A et al. 2017. Dispersals and genetic adaptation of bantu-
926 speaking populations in africa and north america. Science. 356(6337):543-546.

927 Pavlidis P, Alachiotis N. 2017. A survey of methods and tools to detect recent and strong
928 positive selection. J Biol Res-Thessalon. 24:7.
929 Perry GH, Foll M, Grenier JC, Patin E, Nedelec Y, Pacis A, Barakatt M, Gravel S, Zhou X,
930 Nsobya SL et al. 2014. Adaptive, convergent origins of the pygmy phenotype in
931 african rainforest hunter-gatherers. Proc Natl Acad Sci USA. 111(35):E3596-3603.
932 Peter BM, Huerta-Sanchez E, Nielsen R. 2012. Distinguishing between selective sweeps from
933 standing variation and from a de novo mutation. PLoS Genet. 8(10): e1003011.
934 Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS,
935 Myers RM, Feldman MW et al. 2009. Signals of recent positive selection in a
936 worldwide sample of human populations. Genome Res. 19(5):826-837.
937 Pool JE. 2015. The mosaic ancestry of the drosophila genetic reference panel and the d.
938 Melanogaster reference genome reveals a network of epistatic fitness interactions. Mol
939 Biol Evol. 32(12):3236-3251.
940 Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR,
941 Woerner AE, O'Connor TD, Santpere G et al. 2013. Great ape genetic diversity and
942 population history. Nature. 499(7459):471-475.
943 Pritchard JK, Di Rienzo A. 2010. Adaptation - not by sweeps alone. Nat Rev Genet.
944 11(10):665-667.
945 Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: Hard sweeps,
946 soft sweeps, and polygenic adaptation. Curr Biol. 20(4):R208-215.
947 Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics.
948 160(3):1179-1189.
949 Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing
950 genetic variation. Evolution. 59(11):2312-2323.

951 Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, Engelken J. 2015.
952 Hierarchical boosting: A machine-learning framework to detect and classify hard
953 selective sweeps in human populations. Bioinformatics. 31(24):3946-3952.
954 Refoyo-Martinez A, da Fonseca RR, Halldorsdottir K, Arnason E, Mailund T, Racimo F.
955 2019. Identifying loci under positive selection in complex population histories.
956 Genome Res. 29(9):1506-1520.
957 Roux PF, Boitard S, Blum Y, Parks B, Montagner A, Mouisel E, Djari A, Esquerre D, Desert
958 C, Boutin M et al. 2015. Combined qtl and selective sweep mappings with coding snp
959 annotation and cis-eqtl analysis revealed park2 and jag2 as new candidate genes for
960 adiposity regulation. G3-Genes Genomes Genetics. 5(4):517-529.
961 Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH,
962 McCarroll SA, Gaudet R et al. 2007. Genome-wide detection and characterization of
963 positive selection in human populations. Nature. 449(7164):913-918.
964 Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. Genetics.
965 132(4):1161-1176.
966 Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among
967 nearly neutral amino acid replacements in drosophila. Proc Natl Acad Sci USA.
968 104(16):6504-6510.
969 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a
970 coalescent simulation of human genome sequence variation. Genome Res.
971 15(11):1576-1583.
972 Schmidt JM, de Manuel M, Marques-Bonet T, Castellano S, Andres AM. 2019. The impact of
973 genetic adaptation on chimpanzee subspecies differentiation. PLoS Genet.
974 15(11):e1008485.

975 Schrider DR, Kern AD. 2017. Soft sweeps are the dominant mode of adaptation in the human
976 genome. Mol Biol Evol. 34(8):1863-1877.
977 Skoglund P, Mathieson I. 2018. Ancient genomics of modern humans: The first decade. Annu
978 Rev Genomics Hum Genet. 19:381-404.
979 Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in drosophila. Nature.
980 415(6875):1022-1024.
981 Stella A, Ajmone-Marsan P, Lazzari B, Boettcher P. 2010. Identification of selection
982 signatures in cattle breeds selected for dairy production. Genetics. 185(4):1451-1461.
983 Stephan W. 2016. Signatures of positive selection: From selective sweeps at individual loci to
984 subtle allele frequency changes in polygenic adaptation. Mol Ecol. 25(1):79-88.
985 Stephan W, Wiehe THE, Lenz MW. 1992. The effect of strongly selected substitutions on
986 neutral polymorphism: Analytical results based on diffusion theory. Theor Popul Biol.
987 41:237–254.
988 Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for
989 selective sweeps? Genome Res. 16(6):702-712.
990 Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K,
991 Mortensen HM, Hirbo JB, Osman M et al. 2007. Convergent adaptation of human
992 lactase persistence in africa and europe. Nat Genet. 39(1):31-40.
993 Uricchio LH, Petrov DA, Enard D. 2019. Exploiting selection at linked sites to infer the rate
994 and strength of adaptation. Nat Ecol Evol. 3(6):977-984.
995 Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu
996 Rev Genet. 47:97-120.
997 Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in
998 the human genome. PLoS Biol. 4(3):e72.

Wiehe TH, Stephan W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from drosophila melanogaster. Mol Biol Evol. 10(4):842-854.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 174(3):1431-1439.

Zhen Y, Huber CD, Davies RW, Lohmueller KE. 2021. Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and drosophila melanogaster. Genome Res. 31(1):110-120.

Legends

Figure 1. The clustering of candidate SNPs of selection in PSVs

(**A**) A sweep in a PSV region (yellow) with the target of selection indicated in red and neutral nearby candidate SNPs indicated in black for SNPs unaffected by selection and in blue for SNPs driven by selection (ENV regions are indicated in gray). (**B**) Enrichment of candidate SNPs around the simulated targets of selection in 5Mb genomic regions simulated using the African demographic model and the recombination and selection parameters used in this study. The blue line in bold shows the proportions of candidate SNPs observed in non-overlapping 100Kb regions averaged across 1,000 simulations. The blue surface shows the minimum and maximum values obtained. The same statistics obtained using 5Mb regions simulated in absence of selection are indicated in black bold and dashed lines respectively. (**C**) An example of $X = 5$ selective sweeps (each targeting the SNPs indicated in red) in a whole-genome sequence (WGS) dataset, illustrating the expected enrichment of candidate SNPs in PSVs.

Figure 2. Relationships between the simulated ORs and *X* assuming incomplete sweeps only

ORs obtained with fixed numbers of incomplete sweeps per simulated WGS data in African, European and East Asian populations (1,000 simulated WGSs in each case). The frequency at the onset of selection varies from $1/2N$ to 0.2 (SDN and SSV merged together). (**A**) Distributions of *s* (selection coefficient per sweep, left hand side) and *t* (age of selection per sweep, right hand side) obtained after the exclusion of complete sweeps in Africa (yellow), in Europe (blue) and in East Asia (green). (**B**) Simulated ORs for Fay & Wu's *H* (F&W-H), iHS, DIND, ΔiHH and the two pairwise XP-EHHs.

Figure 3. Accuracy of the ABC estimations of *X* in African, European and East Asian simulated populations

Accuracy of the estimations of *X* under the incomplete sweep scenario (**A**) and under the sweep scenario (**B**), assessed by means of simulated WGSs used as pseudo-empirical data, each containing 0, 50, 100 or 150 selective sweeps (200 pseudo-empirical data in each situation). Each boxplot display the 200 point estimates obtained. The horizontal red lines indicate the true values of *X*. (**A**) For both pseudo-empirical data and ABC simulations, the selective sweeps are incomplete only. The pseudo-empirical data are the simulated WGSs shown in Figure 2B. The distributions of *s* and *t* used to simulate the WGS data are shown in Figure 2A. (**B**) For both pseudo-empirical data and ABC simulations, the selective sweeps can be either complete or incomplete. The pseudo-empirical data are the simulated WGSs shown in Figure S3B. The distribution of *t* used to simulate the WGS data is a uniform distribution whereas we used the distribution of *s* enriched in low values shown in Figure S3A.

Figure 4. Posterior distributions of *X* in African, European and East Asian 1000G populations.

Posterior distributions of *X* obtained for each African (**A**), European (**B**) and East Asian (**C**) 1000G population analyzed separately (the distributions are shifted for visibility, the population names are ranked in order of appearance in the plots, the YRI, CEU and CHB populations are indicated in bold). These distributions correspond to the estimations shown in Table 1 obtained using the 1000G ORs averaged across chromosomes. The left hand side panels show the posterior distributions obtained under the incomplete sweep scenario. The ABC simulations used to perform these estimations are those used in Figure 3A (ABC simulations with incomplete selective sweeps only). The right hand side panels show the

posterior distributions obtained under the sweep scenario. The ABC simulations used to perform these estimations are those used in Figure 3B (ABC simulations with selective sweeps that can be either complete or incomplete).

Figure 5. Numbers of selective sweeps as a function of allele frequencies at the onset of selection

Point estimates of X_1, X_2 and X_3, i.e., the numbers of selective sweeps with very low ($1/2N \leq p_{start} < 0.01$), low ($0.01 \leq p_{start} < 0.1$) and intermediate ($0.1 \leq p_{start} < 0.2$) initial frequencies of the selected alleles. The estimations were obtained using the 1000G ORs averaged across chromosomes and the ABC simulations used in Figure 4. (**A**) Point estimates obtained under the incomplete sweep scenario. (**B**) Point estimates obtained under the sweep scenario. The vertical bars show the 95% CIs edges averaged in a given continent (the estimations obtained in each 1000G population can be found in Table S2A).

Table 1. Estimations of the number of selective sweeps.

	Incomplete sweeps		Sweeps	
Continent[a]	**X**	**X (BGS)**[b]	**X**	**X (BGS)**[b]
Africa[c]	62 [36-91]	71 [43-102]	68 [46-91]	63 [41-86]
Europe[c]	71 [35-111]	74 [36-111]	115 [68-160]	137 [89-180]
East Asia[c]	88 [50-127]	104 [66-143]	165 [119-211]	158 [109-206]
All[c]	74 [41-110]	83 [48-119]	116 [78-154]	119 [80-157]
Africa[d]	46 [21-73]	59 [34-87]	58 [37-81]	53 [32-75]
Europe[d]	56 [26-90]	46 [17-79]	84 [40-130]	103 [57-148]
East Asia[d]	60 [29-93]	73 [41-108]	128 [83-173]	124 [80-170]
All[d]	54 [25-85]	59 [31-91]	90 [53-128]	93 [56-131]

[a]Point estimates and 95% CIs edges averaged across populations of the same continental origin (95% CIs are indicated in squared brackets). [b]BGS stands for background selection simulated in PSVs. [c]Estimations obtained using the 1000G ORs averaged across chromosomes. [d]Estimations obtained using the 1000G ORs computed merging chromosomes.

Table 2. Estimations of the average strength and average age of selection.

	Incomplete sweeps		Sweeps	
Continent[a]	**S**	**T (Kya)**	**S**	**T (Kya)**
Africa[b]	0.010 [0.005-0.018]	43.9 [31.4-52.4]	0.013 [0.007-0.022]	53.6 [48.3-58.0]
Europe[b]	0.017 [0.010-0.029]	28.2 [15.4-38.9]	0.010 [0.007-0.013]	52.0 [49.2-54.7]
East Asia[b]	0.012 [0.006-0.022]	34.9 [23.1-44.5]	0.011 [0.007-0.017]	54.7 [51.1-57.9]
All[b]	0.013 [0.007-0.023]	35.7 [23.3-45.2]	0.011 [0.007-0.017]	53.4 [49.5-56.9]
Africa[c]	0.014 [0.008-0.023]	38.3 [24.2-48.5]	0.014 [0.007-0.024]	52.0 [46.1-56.6]
Europe[c]	0.018 [0.010-0.032]	26.4 [12.6-38.5]	0.010 [0.007-0.013]	51.5 [48.4-54.1]
East Asia[c]	0.013 [0.006-0.025]	30.0 [17.1-41.2]	0.012 [0.008-0.018]	53.6 [49.8-57.1]
All[c]	0.015 [0.008-0.027]	31.6 [17.9-42.8]	0.012 [0.007-0.018]	52.3 [48.1-55.9]

[a]Point estimates and 95% CIs edges averaged across populations of the same continental origin (95% CIs are indicated in squared brackets). [b]Estimations obtained using the 1000G ORs averaged across chromosomes. [c]Estimations obtained using the 1000G ORs computed merging chromosomes).

Table 3. Comparing with previous estimations of α in the human lineage.

Human lineage	X[a]	[95% CI]	X_g[b]	α[c]	[95% CI]
Africa	68	[46-91]	0.019	0.177	[0.120-0.236]
Africa[d]	58	[37-81]	0.017	0.151	[0.096-0.211]
Uricchio et al 2019				0.135	[0.096-0.170]
Zhen et al 2021[e]				0.060	
Zhen et al 2021[f]				0.160	
Human-chimp					
Zhen et al 2021[e]				0.110	
Zhen et al 2021[f]				0.250	

[a]Point estimates and 95% CIs edges averaged across populations of the same continental origin (95% CIs are indicated in squared brackets). [b]X_g stands for X per generation, i.e., X divided by the 3,500 simulated generations. [c]Predictions using X per generation, 29,925 nonsynonymous and 7,9Ky of human-chimp divergence (95% CIs are indicated in squared brackets). [d]Estimations obtained using the 1000G ORs computed merging chromosomes (Table 1). [e,f]Estimations under the simple and complex models presented in Zhen et al. (2021).

Table 4. Comparing with *Drosophila*.

Human	X	X_g[a]	S	Ns[b]	λ	λS	$2N\lambda$[b]
Africa	68	0.019	0.013	130	6.5E-12[c]	8.4E-14	1.3E-07
Africa[d]	58	0.017	0.014	140	5.5E-12[c]	7.5E-14	1.1E-07
Drosophila[e]							
Jensen et al 2008	-	-	0.011	27,500	7.9E-12	8.7E-14	3.9E-05
Macpherson et al 2007	-	-	0.010	15,000	3.6E-12	3.6E-14	1.1E-05
Jensen et al 2008	-	-	0.002	4,800	4.2E-11	8.4E-14	2.0E-04
Li and Stephan 2006	-	-	0.002	17,200	1.1E-11	2.2E-14	1.9E-04
Andolfatto 2007	-	-	1.2E-05	23	6.9E-10	8.3E-15	2.6E-03

[a] X_g stands for X per generation, i.e., X divided by the 3,500 simulated generations. [b]Mean Ns and $2N\lambda$ computed from the estimations of S and λ using $N = 10{,}000$ as reference for humans effective sizes or, in *Drosophila*, the effective size values drawn from Jensen et al (2008). [c]X per generation divided by the number of base pairs analyzed. [d]Estimations obtained using the 1000G ORs computed merging chromosomes (Tables 1 and 2). [e]Values drawn from Jensen et al (2008).

Figure 1

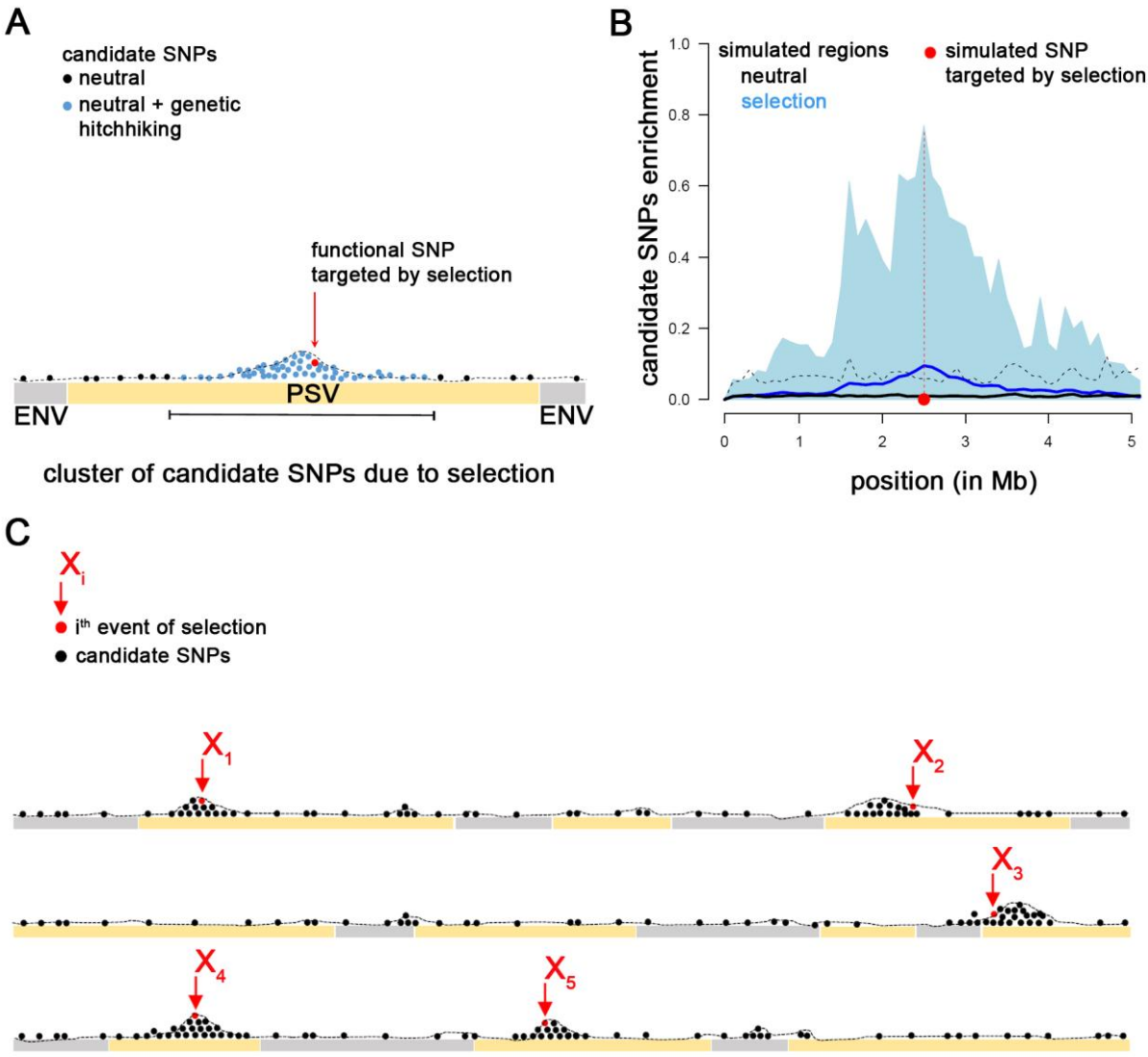

A
candidate SNPs
neutral
neutral + genetic hitchhiking
functional SNP targeted by selection
PSV
ENV
ENV
cluster of candidate SNPs due to selection
B
simulated regions
neutral
selection
simulated SNP targeted by selection
candidate SNPs enrichment
position (in Mb)
C
X_i
i^{th} event of selection
candidate SNPs
X_1
X_2
X_3
X_4
X_5

Figure 2

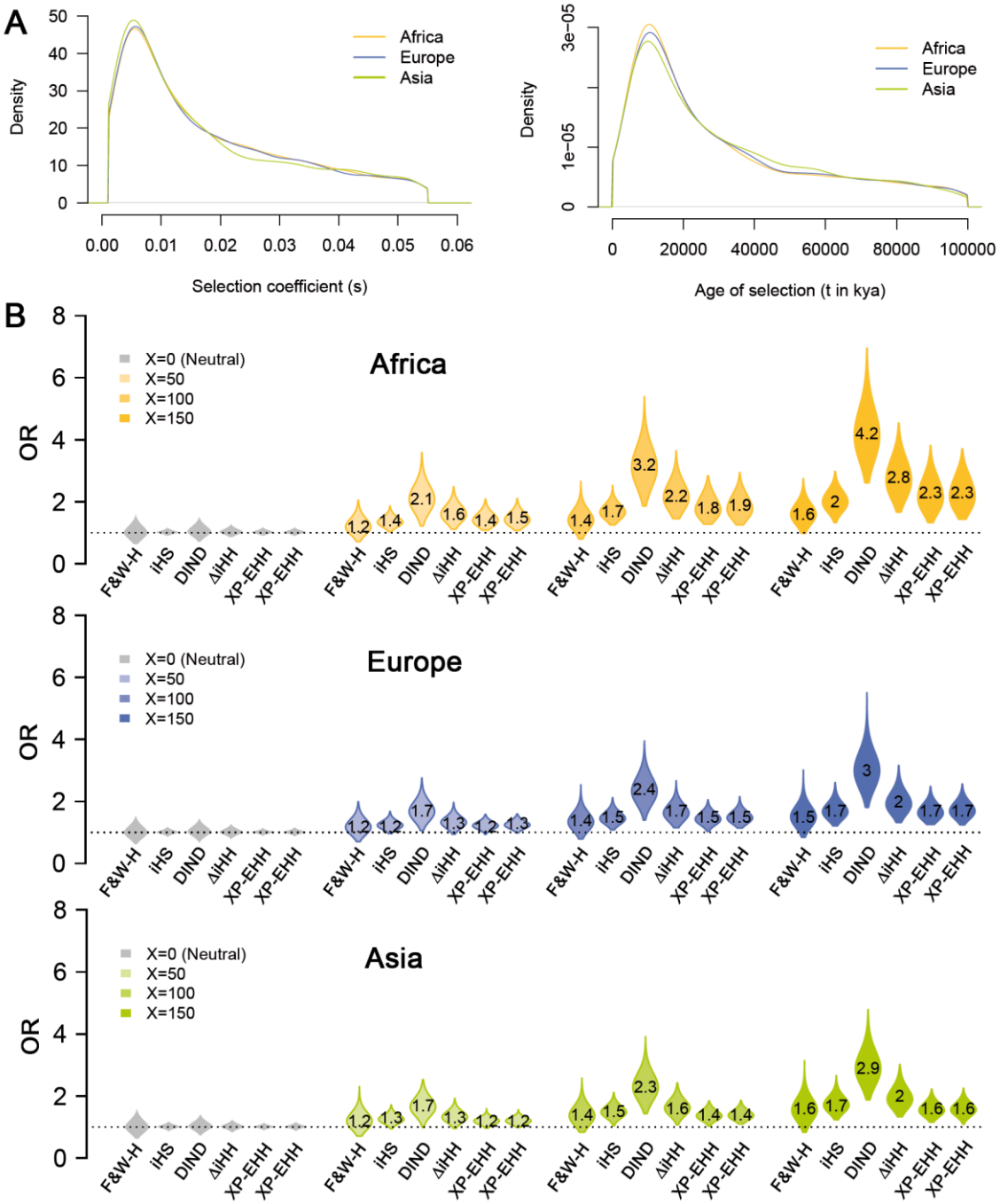

A
Density
Africa
Europe
Asia
Selection coefficient (s)
Age of selection (t in kya)
B
OR
X=0 (Neutral)
X=50
X=100
X=150
Africa
Europe
Asia
F&W-H
iHS
DIND
ΔiHH
XP-EHH
XP-EHH

Figure 3

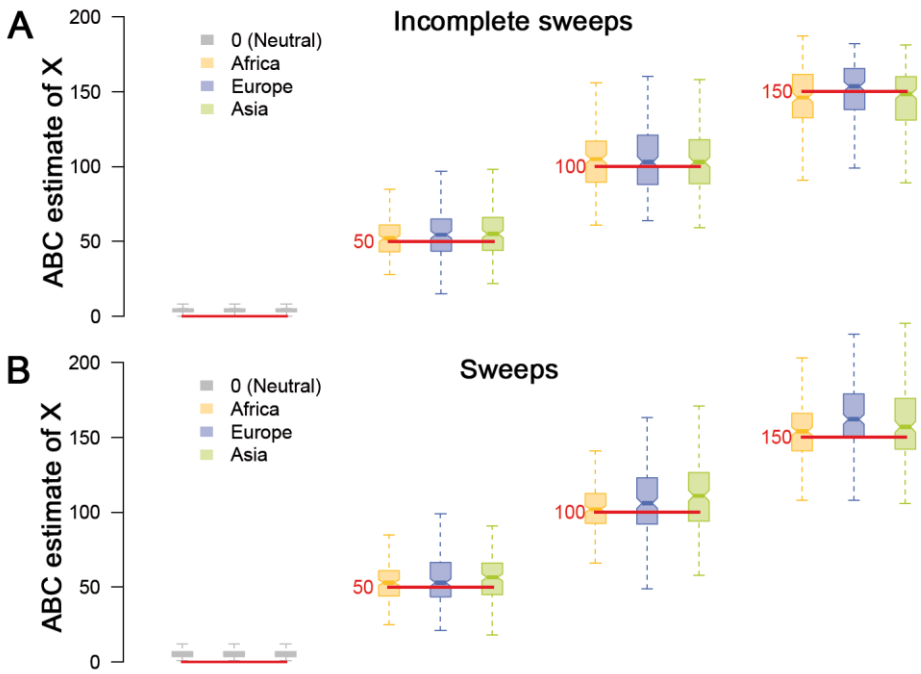

A
Incomplete sweeps
ABC estimate of X
200
150
100
50
0
0 (Neutral)
Africa
Europe
Asia
50
100
150
B
Sweeps
ABC estimate of X
200
150
100
50
0
0 (Neutral)
Africa
Europe
Asia
50
100
150

Figure 4

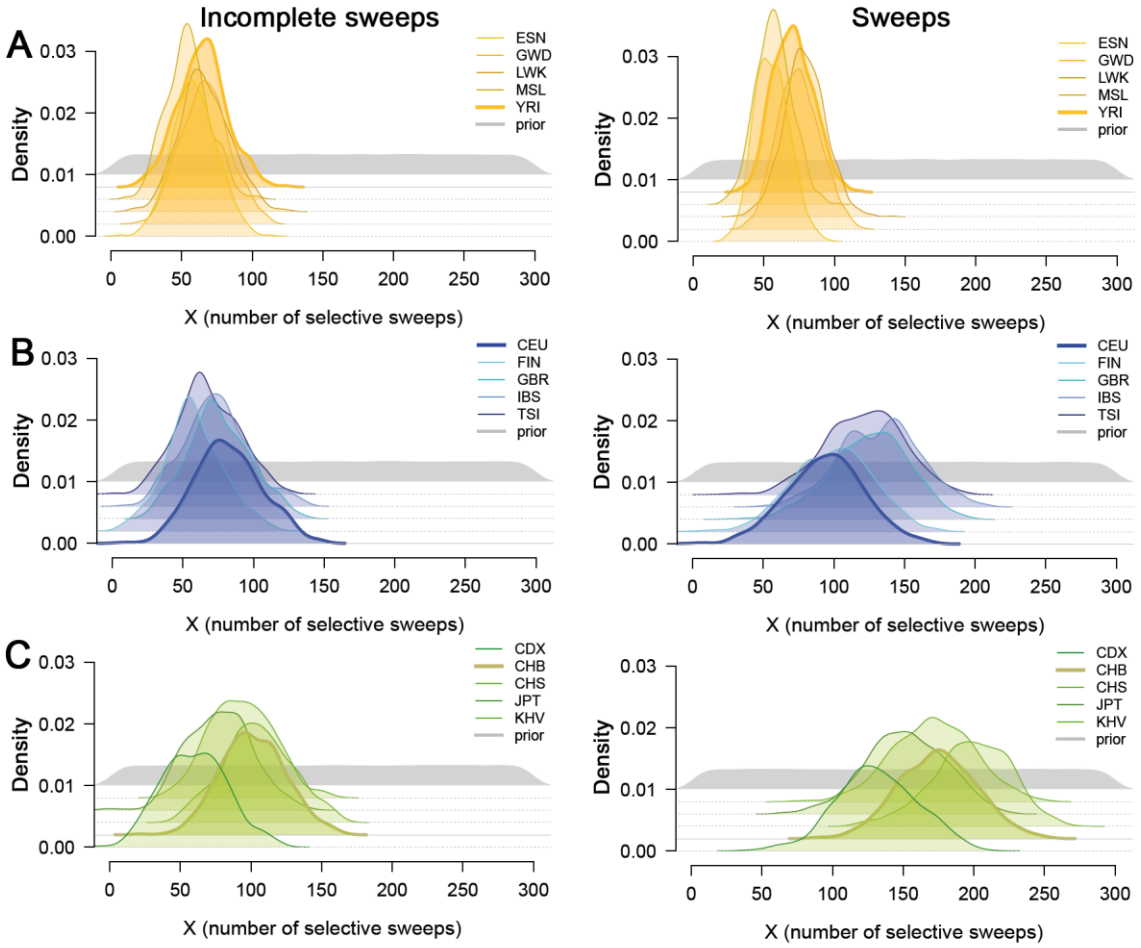
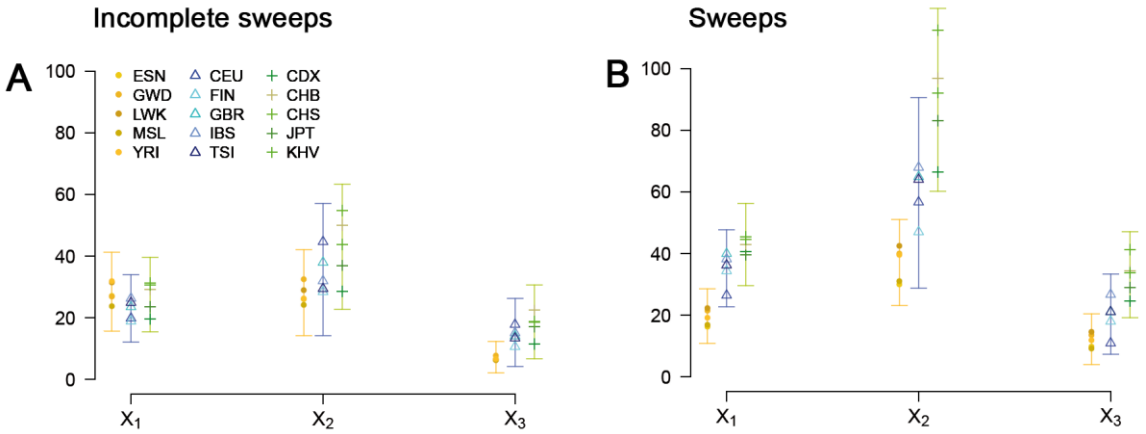

Incomplete sweeps
Sweeps
A
B
C
Density
X (number of selective sweeps)
ESN
GWD
LWK
MSL
YRI
prior
CEU
FIN
GBR
IBS
TSI
CDX
CHB
CHS
JPT
KHV

Figure 5


Incomplete sweeps
A
100
80
60
40
20
0
ESN
GWD
LWK
MSL
YRI
CEU
FIN
GBR
IBS
TSI
CDX
CHB
CHS
JPT
KHV
X1
X2
X3
Sweeps
B
100
80
60
40
20
0
X1
X2
X3