Distinguishing between Selective Sweeps from Standing Variation and from a *De Novo* Mutation

Benjamin M. Peter[1]*, Emilia Huerta-Sanchez[1,2], Rasmus Nielsen[1,2]

1 Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America, **2** Department of Statistics, University of California Berkeley, Berkeley, California, United States of America

Abstract

An outstanding question in human genetics has been the degree to which adaptation occurs from standing genetic variation or from *de novo* mutations. Here, we combine several common statistics used to detect selection in an Approximate Bayesian Computation (ABC) framework, with the goal of discriminating between models of selection and providing estimates of the age of selected alleles and the selection coefficients acting on them. We use simulations to assess the power and accuracy of our method and apply it to seven of the strongest sweeps currently known in humans. We identify two genes, ASPM and PSCA, that are most likely affected by selection on standing variation; and we find three genes, ADH1B, LCT, and EDAR, in which the adaptive alleles seem to have swept from a new mutation. We also confirm evidence of selection for one further gene, TRPV6. In one gene, G6PD, neither neutral models nor models of selective sweeps fit the data, presumably because this locus has been subject to balancing selection.

Introduction

Most organisms harbor large amounts of, mostly neutral or nearly neutral, standing genetic variation [1–4]. As environments change, alleles that previously segregated neutrally, or were only weakly affected by selection, may become targets of strong selection. Examples of a change in environment that could induce such a change include invasion of a new habitat or niche through dispersal, climate changes, and introduction of novel disease agents. This type of selection, in which selection acts on already segregating alleles, is called selection from standing variation (SSV).

We contrast this model with the more commonly assumed model of selection on a *de novo* mutation (SDN). In the SDN model the selection pressure already exists when a new mutation is introduced into the population. In addition, there are several more complicated scenarios of selection. The case where an allele under selection has multiple independent origins has received particular attention [5–7], and is often also referred to as selection from standing variation. In this paper, we focus on the case where all copies of an allele are identical by descent, and do not consider multi-origin alleles.

Of great interest is the question of which mode of selection has been more frequent in the evolution of a species [5,8]. In particular, if we observe a selected variant, which mode of selection is more likely to have occurred? Theoretical results by Hermisson & Pennings [5] find that SDN should be common if selection is strong and mutation rates are low, in all other cases we expect SSV to be more prevalent.

Statistics affected by selection

Detection of selected regions has been a major goal in population genetics in recent years [9–13]. Rather than working with the full data, all of these studies simplified their data by using various statistics designed to detect the signal of selection (see e.g. [12,14]). These statistics may be classified in different categories, based on the information they exploit. First, functional differences between different codon positions, and the substitution rates of synonymous and non-synonymous sites were used by [15,16]. Another approach relies on finding related populations, where selection acts on only one of them. This leads to locus-specific high population differentiation, which may be detected by statistics such as F_{ST} [17] or XP-EHH [18]. A third category of statistics is based on the length of haplotypes associated with a given allele. Haplotypes associated with the selected allele will on average be younger than haplotypes carrying the derived allele, and there will therefore be fewer recombination events that break up the haplotypes. Statistics such as EHH [9] and iHS [19] were developed to detect this pattern. Finally, the site frequency spectrum (SFS) can also be used to detect departures from neutrality and hence selection. SFS based statistics usually compare various estimators of the population mutation rate θ. The first and perhaps most well-known statistic in this category is Tajima's D [20], but the statistic can be generalized [21,22], and other statistics such as Fay and Wu's H [23] belong to the same family.

Distinguishing SSV and SDN

In this study, we are interested in distinguishing the SDN and SSV models of evolution for a single putatively adaptive mutation. Barrett & Schluter [24] identify three possible ways of identifying SSV: i) the selected allele may occur in an ancestral population, ii) an allele is shown to be older than the environment it is adaptive in

Author Summary

Considerable effort has been devoted to detecting genes that are under natural selection, and hundreds of such genes have been identified in previous studies. Here, we present a method for extending these studies by inferring parameters, such as selection coefficients and the time when a selected variant arose. Of particular interest is the question whether the selective pressure was already present when the selected variant was first introduced into a population. In this case, the variant would be selected right after it originated in the population, a process we call selection from a *de novo* mutation. We contrast this with selection from standing variation, where the selected variant predates the selective pressure. We present a method to distinguish these two scenarios, test its accuracy, and apply it to seven human genes. We find three genes, ADH1B, EDAR, and LCT, that were presumably selected from a *de novo* mutation and two other genes, ASPM and PSCA, which we infer to be under selection from standing variation.

and iii) the signature of selection at linked loci, the selected sweep, is different between SSV and SDN. Our approach is based on differences in the genetic signature of selection, but when possible, we will compare to inferences based on i) and ii).

To understand the difference between the SSV model and the SDN model, it is important to realize that all the information regarding selection, and mode of selection, is captured by the allele frequency trajectory through time. In other words, the full allele frequency path through time would be a sufficient statistic for the selection coefficient, if it was known. As selection acts only to change the allele frequency in the selected site, and does not act directly on adjacent sites, the effects of the selection on linkage disequilibrium, haplotype patterns, allele frequencies in linked sites, etc., are only through the effects caused by the change in allele frequency of the selected allele (hitch-hiking effects). This observation is the foundation for standard population genetic theory on selective sweeps (e.g., [27–28]) and forms the basis for several simulation methods, in which the path of the selected mutation is first simulated and then neutral simulations are performed conditional on the allele frequency path [28]. Such simulation methods would be invalid if the allele frequency path did not contain all information regarding the selection coefficient acting on the selected mutation. Similarly, if the path of an allele is the same under the SSV and the SDN model, no additional genomic data could help us distinguish between the two models.

Armed with this insight, we can further explore the differences between the two models. Figure 1a, 1b depict the trajectory, the number of copies of the selected allele through time for an SSV and SDN model. Looking backward in time, the adaptive alleles are selected at first in both models, and during this stage the two models do not differ at all. In the SSV model, however, the mutation stops being advantageous at some point in the past. Backwards from this time point, the mutation in the SSV model acts as a neutral allele, whereas the mutation in the SDN model is under selection.

As selection is the same in the phase when both alleles are selected, the difference between the models is during the phase in which selection is acting on the mutation in the SDN model but not in the SSV model. How big is this difference? It depends on two parameters: the selective advantage of the mutation under the SDN model, and the frequency of the mutation at the time when selection first start acting in the SSV model. A good measure of the

difference might be the allele age distribution at this point, which is plotted in Figure 1c and 1d for a mutation at a frequency of 1% and 5%, respectively. Unfortunately, it turns out that the difference it is rather small: While the allele age of a mutation at a low frequency does depend on the selection coefficient, the difference is very small if selection is weak. Clearly, it will be much easier to distinguish between the two models if selection is strong and if the frequency of the mutation is initially high in the SSV model.

However, we cannot observe the trajectory directly, but only the diversity at linked site. It has been shown that the genetic signature of sweeps from standing variation differs in three important aspects from the signature of sweeps from new mutations [29]: at the same selection coefficient, the signal of selection from standing variation is 1) weaker and 2) affecting a narrower region. As a third difference, we expect an increased variance in both allele age and trajectory. Under the SSV model, the selected allele may be present on several haplotypes when selection starts, and these haplotypes will be affected equally strongly by selection. Thus, there will be more variation compared to SDN, and the, loss-of-diversity signal of selection will be weaker. The fact that the signal of selection affects a narrower region is due to the fact that the selected allele is older in the SSV model, and hence recombination had more time to break it up (Figure 1a, 1b). The increase in variance is evident from the large variance in the neutral phase of the allele trajectory in Figure 1b, and the wider distribution of the allele age of neutral alleles in Figure 1c and 1d. In Figure 1e and 1f we give the expected distribution of Fay and Wu's H [23] and EHH [9], two statistics used to detect selection, and where we show that the signal is indeed expected to be weaker and affecting a narrower region under the SSV model.

The objective of this paper is to develop and explore a statistical method for distinguishing between SSV and SDN models, and for providing associated estimates of relevant parameters. However, the method we develop is not intended as a new method for performing scans for selection in genome-wide data or for quantification of genome-wide levels of selection. For computational reasons, other methods might be more suitable for such genome-wide analyses. We focus on illustrating the method on a few loci previously hypothesized to be under selection in humans, but the method could as well be applied to other human loci or data from other species.

Approximate Bayesian Computation

To exploit the characteristics of selective sweeps discussed in the previous section, we combine different statistics and calculate them for different genomic regions. Using combinations of statistics to improve inference is not a new concept, and has been applied previously (e.g. [30]). Here, we choose an Approximate Bayesian Computation (ABC) framework for combining statistics [31,32]. ABC has the advantage that it extends naturally to allow both model choice and parameter estimates under a given model.

ABC was developed to estimate parameters of complex models in manageable computer time, and has been widely used in population genetics, most frequently to infer parameters for complex models of demographic history [32–37]. Several implementations of the ABC algorithm have recently been published [38–40], and in the past few years, various variations of the algorithm have been developed [41–43]. ABC is a rejection sampling algorithm used to calculate the posterior distribution of a parameter under a given model, used frequently when the likelihood cannot be calculated analytically. In ABC inference, a large number of data sets are simulated using parameters randomly drawn from a prior distribution. If a simulation does not match the observed data, it is rejected,
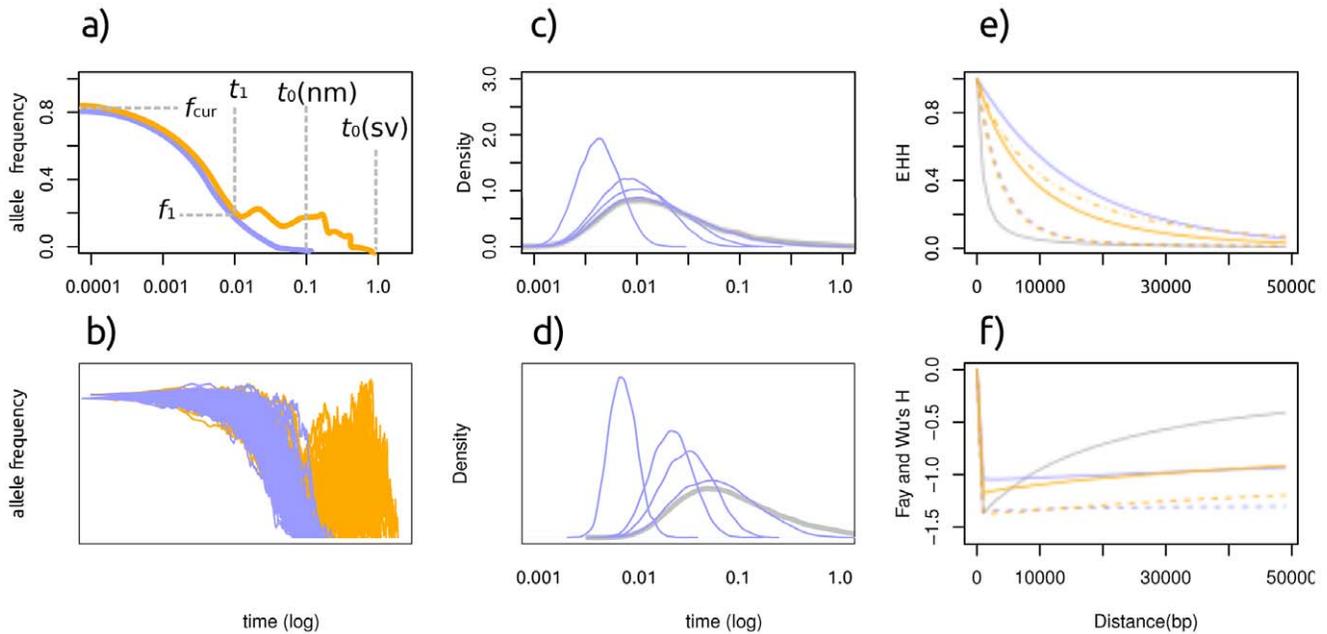
Figure 1. Characteristics of a selective sweep from standing variation. orange: sweep from standing variation blue: sweep from a new mutation, blue: neutral model a: A cartoon of the allele frequency trajectory with relevant parameters: f_1: allele frequency at the time selection started, f_{cur}: allele frequency at the time mutation is observed. t_1: time at which selection started. t_0: time when mutation arose., b: 100 stochastic realizations of the allele frequency trajectory. Panels c,d: Age distribution of an allele at 1% frequency and 5% frequency in a population (log scale). Blue line denotes neutrality, green lines represent selection with $\alpha = 20,100,200$ and 1000 (right to left). Panels e,f: Distribution of the EHH (e) and H (f) statistic under neutrality (blue), a de novo mutation (green) and standing variation (red). Full and dashed lines represent selective pressures of $\alpha = 1,000$ and 200, respectively. The dash-dot line represents $\alpha = 4000$. Note that the slopes of the curves are different for the two scenarios, and the low H value around 0 under neutrality is due to the conditioning on a high frequency derived allele. Times are given in coalescent units and are plotted on a logarithmic scale.
doi:10.1371/journal.pgen.1003011.g001

otherwise it is retained. However, if the data is complex, the probability of a match is prohibitively low, and two important approximation steps are used: First, the data is transformed into a set of summary statistics. If these statistics are sufficient (i.e. retain all the information present in the data), this step is exact. However, in many cases, including this study, no sufficient statistics are known, and this step results in a first approximation step. In many cases, however, this transformation will still result in very low acceptance probabilities. Therefore, the condition of an exact match is relaxed. Specifically, the summary statistics based on the simulations (S) are compared to the summary statistics of observed data (S^*). Using some distance measure δ, simulations are retained if $|\delta(S,S^*)| < \varepsilon$ for an arbitrarily small distance ε. Frequently, some post-sampling adjustment is used in an attempt to correct for the error introduced in the second approximation step, and posterior distributions are estimated from the parameters of the retained simulations.

In this study, we propose to use ABC to distinguish between a selective sweep from a new mutation and a selective sweep from standing variation. We use simulations to determine which parts of the parameter space the method has power to make this distinction, and aim to estimate parameters under both models. We then apply our method to seven genes that were previously reported to be under selection.

Results

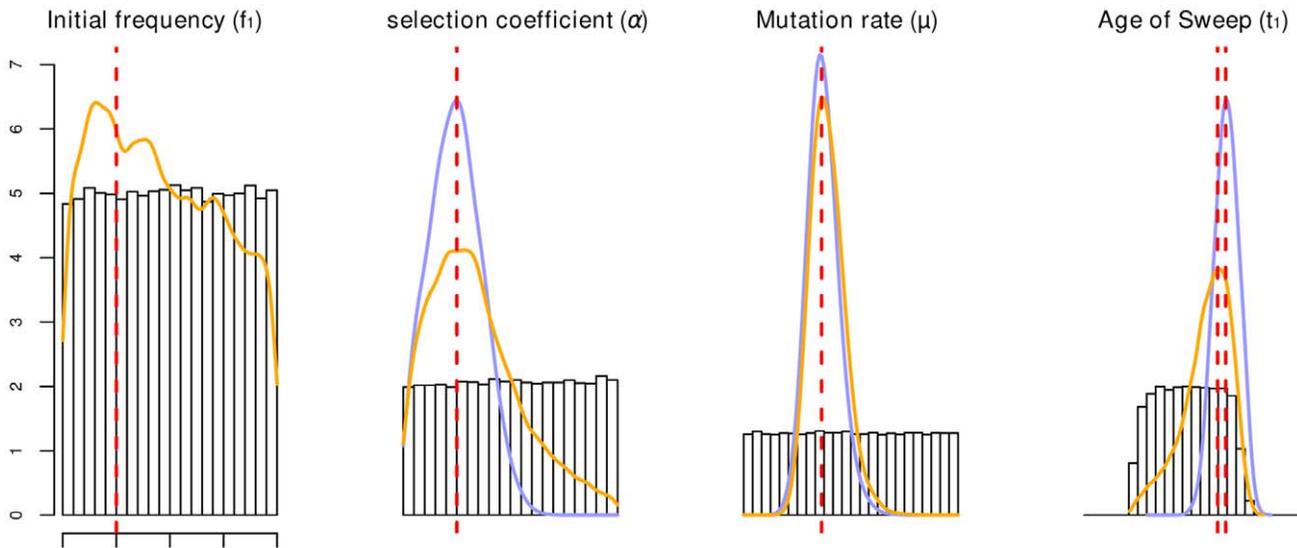### Accuracy of parameter estimates

We first wanted to assess how accurately we can estimate the selection coefficient and the age of the selected mutation from the SSV and SDN models. For this purpose, we performed ABC inference on simulated data sets with known parameter values. Results for a case of

moderately strong selection ($\alpha = 400$) are given in Figure 2, with α being the population scaled selection coefficient $\alpha = 4Ns$. As can be seen from the figure, the mode is an accurate estimator of the true value for both models. However, in the SSV case the posterior distribution is much broader than under the SDN model, and the 95% confidence interval extends to the edges of the prior, indicating low accuracy in the estimate. For the initial frequency parameter, f_1, the posterior differs only marginally from the prior, and therefore this parameter cannot be reliably estimated.

Accuracy of model choice

We aim to identify parameter regions where we can distinguish between the SSV and the SDN model. As a control, we also consider a model of neutral evolution (NT), where an allele increases to high frequency solely due to genetic drift. In particular, we are interested in three parameters that are expected to have a strong influence on model choice accuracy: the selection parameter α, the frequency of the mutation when it became selective advantageous, f_1, and the current frequency of the selected allele f_{cur}. In Figure 3 and Figure S1, we explore the accuracy of our model choice procedure in three series as a function of α, f_1, and f_{cur}.

We find (Figure 3a) that in cases where $\alpha < 100$; the method cannot reliably distinguish between selection and a neutral model. This is not surprising, as for such values of α, standard neutrality tests have little or no power to detect selection [44,45]. For selection coefficients of $\alpha = 100$ and $\alpha = 200$ the neutral model has a very low posterior probability and would be rejected, but we still do not have sufficient power to distinguish the signals from SSV from SDN. Only under strong selection ($\alpha = 1,000$) do we have

Figure 2. Parameter estimation accuracy under SSV and SDN model. Prior distributions are given as histograms; the orange and blue lines depict the average posterior distribution from 100 replicates of the parameters under the SSV and SDN model, respectively. The vertical dashed red line gives the parameters used for the simulation: $\alpha = 400$, $\mu = 2.5e\text{-}8$, $f_1 = 0.05$, $\log(t_1) = -1.51$ (SSV)/-1.36(SDN). Estimates for the SSV model are less accurate for all parameters except μ, and 95% confidence intervals of estimates under the SSV model span the entire prior range for f_1, α and t_1. The age of the sweep is given in coalescence units.
doi:10.1371/journal.pgen.1003011.g002

reasonable power to distinguish between SSV and SDN. Thus, we find that there is a parameter range of α between 100 and 500, in which selection can be reliably detected, but the two models of selection are statistically indistinguishable.

In the second series (Figure 3b), we vary the initial allele frequency (f_1). We find that simulations under the SSV model, with $f_1 = 1\%$, are identified as SDN models, but that the accuracy in model choice increases with f_1. For larger values of f_1, we can detect selection when selection is strong ($\alpha = 1,000$). For high initial allele frequency ($f_1 = 20\%$) we correctly infer the true mode of selection even when α is 200. This suggests that the ability to distinguish the two models increases with f_1. Furthermore, we also find a negative relationship between the estimated value of f_1 for a data set and the posterior probability of the SDN model (Figures S2, S3 and S4): As we would expect, the larger the estimate of f_1, the lower is the posterior probability of the SDN model, and we find a strong negative correlation ($R^2 = 0.51$) between these two quantities based on 1,000 simulations.

In the third series (Figure 3c), we investigate the effect of the current allele frequency f_{cur} on the model comparison. For simulations under the SSV model, we find that the accuracy strongly decreases with f_{cur}. For $f_{cur} = 0.2$, we classify slightly less than half of the data sets correctly. This is in contrast to simulations under the SDN model, where the power to correctly classify simulated data sets gradually increases with f_{cur}. Thus, in studies aimed at detecting selection on standing variation, the false positive rate should depend only slightly on f_{cur}, but the false negative rate is expected to increases drastically when f_{cur} is low.

Figure 4 illustrates how the selection parameter (α) and the initial allele frequency (f_1) affect the accuracy of model choice between the SSV, SDN and NT models for three values of f_{cur} ($f_{cur} = 0.95$, $f_{cur} = 0.8$ and $f_{curr} = 0.5$). As in Figure 3, the number of correctly assigned data sets increases with α, f_1 and f_{cur}. Under the SSV model, the gradient with which the power declines is strongest when f_{cur} is large (95%, Figure 4a), and becomes less pronounced for smaller f_{cur} (see Figure 4c and 4e). For $f_{cur} = 95\%$ (Figure 4a), there is a region with $f_1 > 0.05$ and $\alpha > 1,000$ where

there clearly is very high power to infer the correct model. On the other hand, for $\alpha < 200$ or $f_1 < 0.03$, we make incorrect inferences more than half of the time, indicating that in these regions of the parameter space, the signal of the sweep is too weak to discriminate between the SSV and SDN models. While that global pattern is the same for $f_{cur} = 0.8$ and $f_{cur} = 0.5$ (Figure 4c, 4e), the distinction between regions where we can and cannot assign simulated data sets correctly is less pronounced.

Quite surprisingly, however, we find that for $f_{cur} = 0.8$, the number of correctly assigned data sets increases when selection is low. The same trend holds for $f_{cur} = 0.5$ (Figure 4e), however here the influence of selection is even weaker, and inference becomes quite ambiguous, with posterior probabilities ranging from 60% to 80% in the entire parameter space.

In contrast, the pattern is much simpler for simulations under the SDN model (Figure 4b, 4d and 4f), where the probability to correctly identify the model increases with decreasing f_{cur}. When f_{cur} is set to 0.95, we need a selection coefficient of $\alpha = 1,500$ to make confident inferences. For f_{cur} of 0.8 and 0.5, this value decreases to 900 and 300, respectively.

In summary, a high current allele frequency increases the power to distinguish between SSV from SDN (Figure 3c, Figure 4). The frequency with which the SDN model is correctly inferred increases slightly with decreasing f_{cur}, presumably because the selected phase makes up a larger proportion of the trajectory.

Applications

We illustrate our model choice procedure by analyzing seven genes that have previously been identified as candidates for being under selection. These genes are ADH1B, ASPM, EDAR, G6PD, LCT, PSCA and TRPV6. The genes were selected using the following set of criteria: i) there is evidence for selection from a previous study, ii) a putative causal mutation has been identified and iii) the putative causal site has reached a high frequency in at least one population, but has not yet reached fixation. In addition, we also analyzed four regions that were noncoding and presumably neutral. We retrieved polymorphism data from the 1000 Genomes Project low coverage data [46] using *tabix*

a) Selection strength
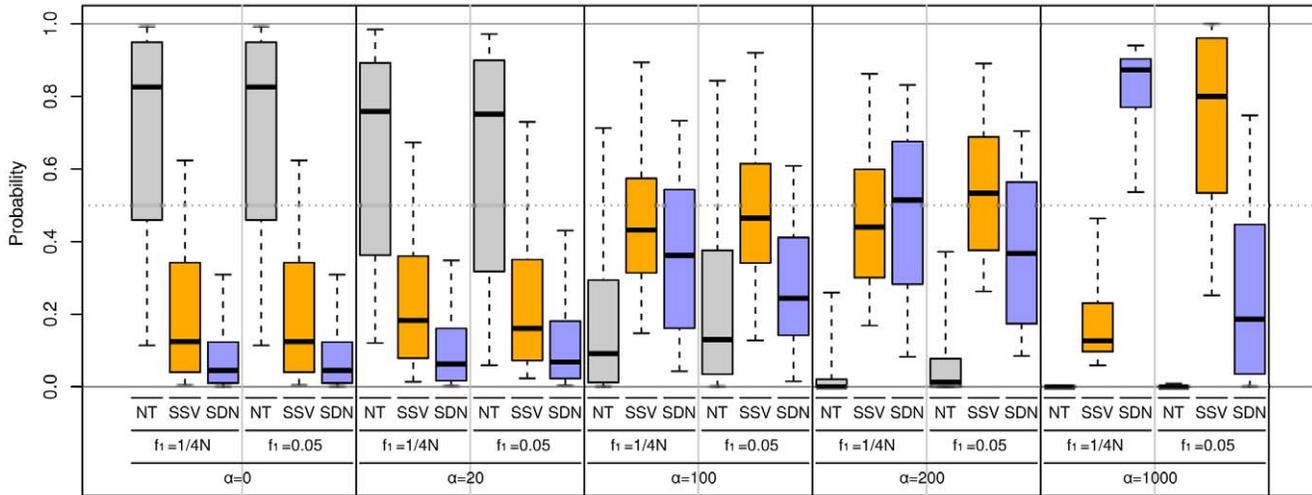
b) Initial frequency
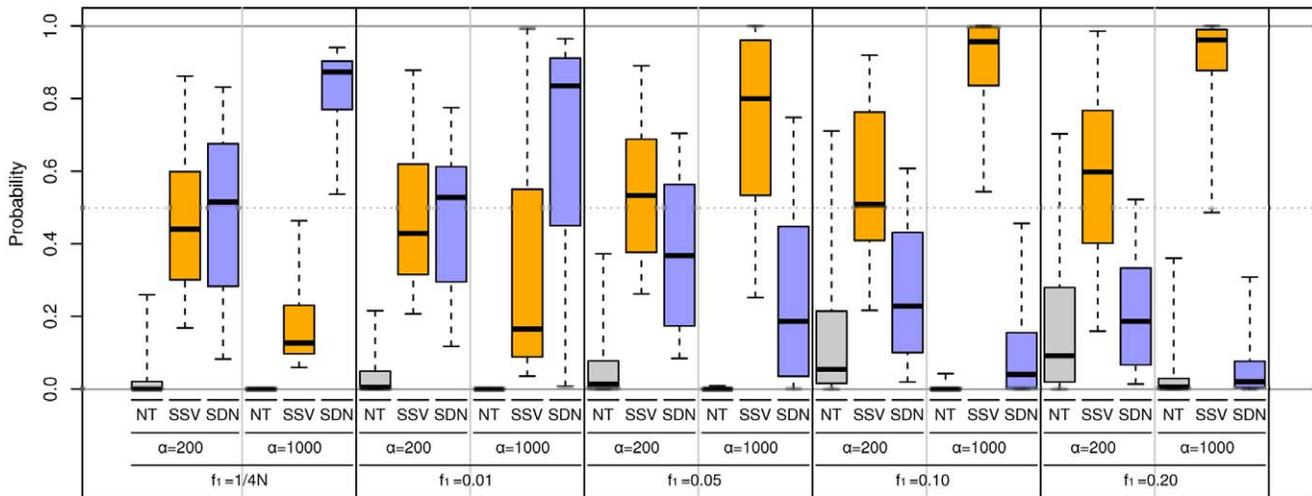
c) Current frequency

Figure 3. Simulation results for ABC model choice procedure. We simulated data using the fixed parameter values given in the lower part of the figure. The boxplots show the lower and upper quartiles, the median and the limits of a 95% interval of the posterior probabi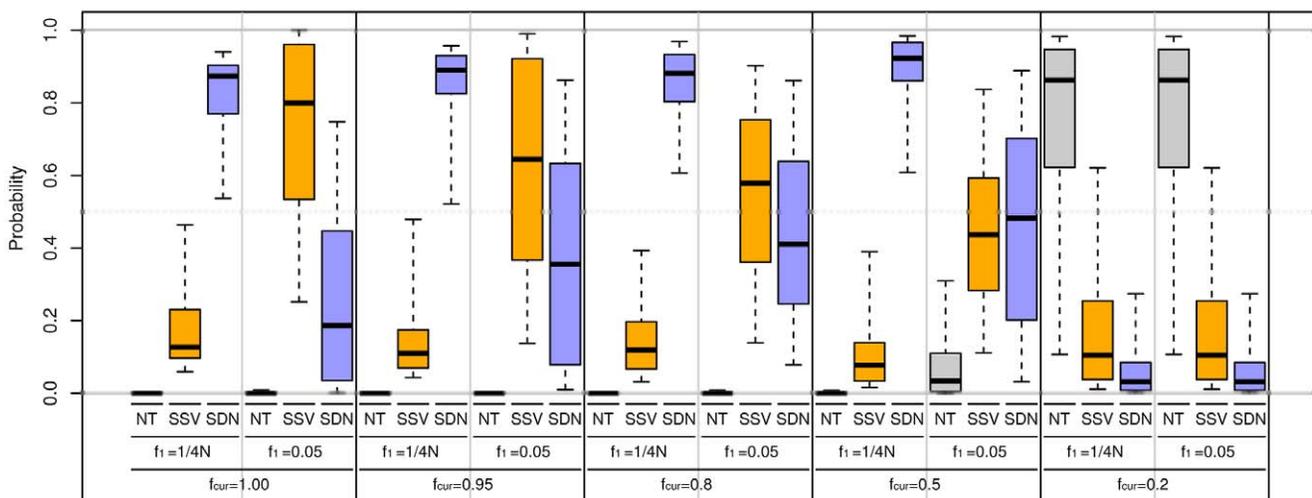lity for the NT (blue), SSV (red) andSDN(green) models, respectively. Panel a: We compare the effect of the increasing selection coefficient α. Panel b: The effect of increasing initial frequency f_1. Panel c: The effect of the current frequency f_{cur}, In panels a,b f_{cur} was set to 0.95, and in panel c, α = 1,000.
doi:10.1371/journal.pgen.1003011.g003

[47]. Ancestral genotypes were inferred by comparison to the homologous chimpanzee allele. If a signal of selection was present in more than one population, we used data for the population where the selected site was most frequent, to facilitate inference. Model choice and parameter estimation were performed using the procedures described in the methods section. In contrast to the inference on simulated data sets, here we explicitly model varying recombination rates and the complex demographic history of the human population.

Results for the sample genes are given in Figure 5 and Figure S4, as well as Table 1. For six of the seven genes analyzed, the neutral scenario was strongly rejected with a posterior probability of less than 1%, and we can confirm the prior evidence that these genes are under selection. Three of those genes, ADH1B, EDAR and LCT, were found to be under selection from a new mutation and one gene, TRPV6 could not be assigned with any significant probability to either model. Two genes, ASPM and PSCA, were found to be under selection from standing variation. Finally, none of the three models provided a good fit to observed data in the G6PD gene, suggesting that neither of the models is appropriate for this gene. In the following paragraphs, we will discuss each gene in some detail, and give estimates for selection coefficient and time when appropriate. All estimates are given with a point estimate for the mode, and the lower and upper bound of a 95% Highest Posterior Density interval in brackets. Estimates in years were made assuming a generation time of 25 years.

Discussion

Applications

ADH1B. The ADH1B gene encodes one of three subunits of the Alcohol dehydrogenase (ADH1) protein, a major enzyme in the alcohol degradation pathway that catalyzes the oxidization of alcohols into aldehydes. ADH1B is part of a 60 kb gene cluster on chromosome 4, encoding for all three ADH1 subunits. Selection on the major ADH gene complex has received major attention as it is suggested to be one of the major genetic causes of alcoholism risk [48], and a possible cause of the "alcohol flush" phenotype prevalent in many Asian populations, where individuals turn red due to increased acetaldehyde levels in the blood after alcohol consumption [49]. As a result, the genes are well studied and several non-synonymous polymorphisms are known to have various effects on enzyme activity [50,51]. One particular allele, Arg47His, has been proposed to be under selection based on several lines of evidence: First, the derived Histidine allele results in an increased enzymatic activity. Second, age estimates of the derived allele based on its frequency correlate with the onset of rice domestication [48,49] and the availability of fermented beverages [52].

In our analysis, we analyzed the CHB population where the allele is found at a frequency of 0.71 in the 1000 genomes data. For this data set, we could clearly reject the neutral model, with a posterior probability of 1e-8. The SDN and SSV models have posterior probabilities of 78.3% and 21.7%, respectively, indicating slightly stronger evidence in favor of the SDN model. Under this model, we estimate a selection coefficient of $s = 0.036$ (0.009–0.19), and an age of the mutation of 11,100 (1,900–42,900) years It is remarkable that this age corresponds very well with the arrival of rice agriculture and the availability of fermented beverages in China around 10,000 year ago [49]. Our finding of evidence for a *de novo* sweep is conflicting with the fact that the derived 47His allele also occurs at a high frequency in Western Asian populations, but only at low frequencies in Central Asian and Indian populations [48], a pattern of genetic variation that has previously been suggested to be a result of selection on standing variation [48].
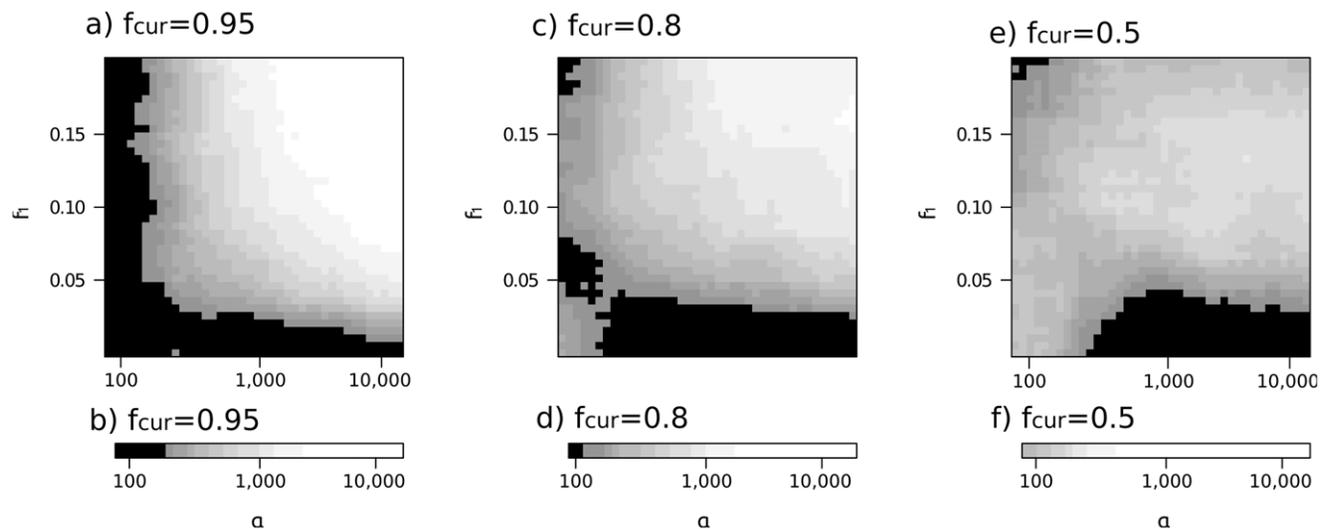
Figure 4. Parameter regions where distinction between models is possible. On x and y axes are the prior ranges for selection coefficient and initial frequency of a selective sweep, respectively. Panels a, c and e give simulations under the SSV model, panels b, d and f for the SDN model. The different panels represent different current frequencies: In Panels a, b f_{cur} is 0.95, in c, d f_{cur} = 0.8 and in panels e and f f_{cur} = 0.5. Color gives the proportion of simulated data sets that were assigned to the correct model, when compared to the two alternative models. Black areas correspond to regions where this proportion is less than 50%, white areas to parameter regions where 95% or more of the data sets are correctly assigned. Each shade of grey corresponds to a 5% increase in the number of correctly assigned data sets.
doi:10.1371/journal.pgen.1003011.g004
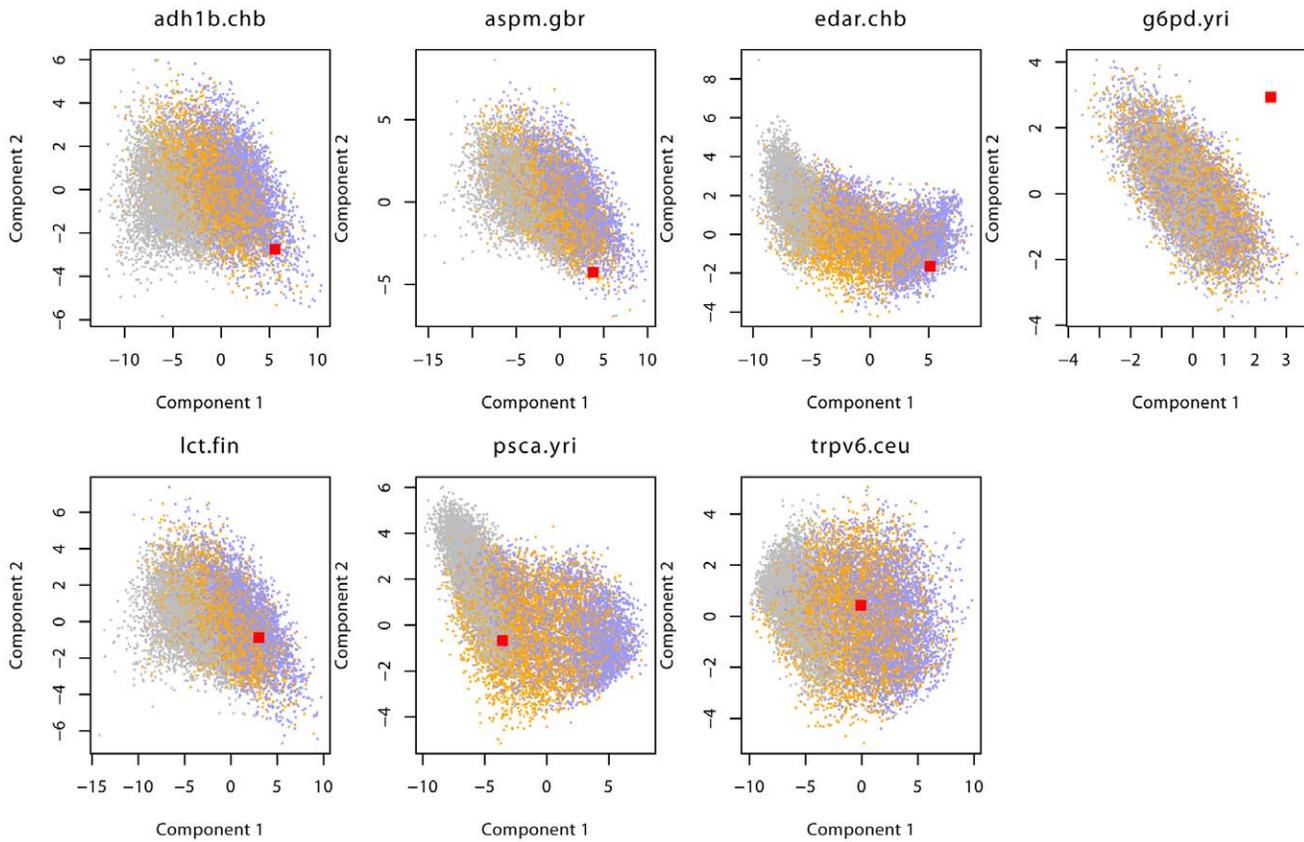
Figure 5. Distribution of summary statistics of 7 genes. This figure shows the observed (red) and prior predictive distribution of the first two PLS-DA components. Neutral simulations are shown in grey, SSV in orange and SDN in blue. For G6PD we show components 2 and 3 to highlight the finding that none of the three models analyzed is able to model the data for this gene.
doi:10.1371/journal.pgen.1003011.g005

ASPM. The ASPM (abnormal spindle-like microcephaly associated) gene has been identified as a major determinant of brain size [53]. Much attention has been focused on the difference between humans and chimpanzee in that gene, and several studies [54,55] have quantified these differences and found an unusual high amount of fixed substitutions between these two species, indicating positive selection on the branch between humans and chimps. In addition, recent ongoing selection was proposed based on the finding that a single haplotype was unusually frequent in several populations [56]. However, the interpretation of their results stirred considerable debate [57–60], with researchers pointing out that the haplotype distribution found by [56] is not that unusual [60] and that neutral demographic scenarios are able to produce haplotype distributions similar to the one observed in ASPM [58].

Table 1. Genes analyzed in this study.

Gene	chr	function	pop	Model	S	t_1 (years)	t_0 (years)	References
						Estimates		
ADH1B	4	Alcohol metabolism	CHB	SDN(0.78)	0.036 (0.009–0.192)	-	11,100 (1,900–42,900)	[50]
ASPM	1	microcephalism	GBR	SSV(0.87)	0.029(0.003–0.17)	17,400 (800–56,400)	79 (17–288) ky	[56]
EDAR	2	NF-κB Activation	CHB	SDN(0.88)	0.14 (0.07–0.31)	-	11,400 (4,300–43,700)	[61]
G6PD	X	malaria resistance	YRI	-	-	-	-	[66]
LCT	2	lactase persistence	FIN	SDN(0.99)	0.025 (0.004–0.20)	-	11,200 (1500–64,900)	[73,74]
PSCA	8	Involved in bladder & pancreas cancer	YRI	SSV(0.86)	0.035 (0.004–0.015)	8,000 (1,000–54,900)	191 (50–698) ky	[86]
TRPV6	7	Calcium absorption	CEU	SSV (0.55)	0.032 (0.005–0.25)	7.600 (900–43,300)	211 (29–697)ky	[90]
				SDN (0.45)	0.023 (0.007–0.08)	-	23,400 (6,400–70,400)	

Chr: chromosome, pop: population we analyzed using the population code from the 1000 genomes project; For each gene, we give the favored model(s) and in brackets the posterior probability for that model.
doi:10.1371/journal.pgen.1003011.t001

We used the non-synonymous SNP A44871G (rs41310927) for our study, which was identified in [56] as a putative causal variant in our analysis. We found evidence for selection on standing variation (Pr(SSV) = 0.87), with little support for the neutral and SDN model with posterior probabilities of 0.13 and 2e-7, respectively.

We estimate a rather weak selection coefficient of 0.029 (0.003–0.170), and estimate that selection started to act 17,412 (771–56,443) years ago, and an age of the mutation of 97 (17–289) ky. This is considerably older than the estimate of 5,800 years for the most recent common ancestor of the selected allele by Mekel-Bobrov et al. [56], a difference that might be due to the fact that we assume a different demographic history.

EDAR. The EDAR gene region has been suggested to be under selection in East Asians based on multiple genome scans [10,11,19] and has been studied in more detail by Bryk et al. [61]. EDAR encodes a cell-surface receptor that activates a transcription factor [61,62], and, among other phenotypes, has been associated with the development of distinct hair and teeth morphologies [62,63]. A non-synonymous SNP (rs3827760, V370A) has been associated with these phenotypes, and has been confirmed in an in vitro study to enhance the activity of the EDAR gene [61]. The rs3827760 SNP lies in a DEATH-domain that is highly conserved within mammals [18], and is found at a very high frequency in East Asian and American individuals, but is absent from all European and African populations [61].

In the 1000 genomes data, EDAR shows the strongest signal of selection for EHH, Tajima's D and Fay & Wu's H among all genes we analyzed. This is reflected in our model choice analysis, where we find a 88.5% probability that the V370A polymorphism originated from a new mutation. The probability for the SSV model was 13.3%, and the neutral model did not receive any measurable support. We estimated a very high selection coefficient of $s = 0.15$ (0.04, 0.31), and an origin of the mutant allele 3,000 (1,400, 6,900) years ago. This estimate is most likely too young, as the allele is also present in Native American population and so is strongly expected to have been present before the colonization of America. A possible explanation for this is that selection does not act codominantly on EDAR. Comparing our codominant model with a model where the dominance parameter h was allowed to vary between 0 and 1 resulted in a strong favor for the more complex model (Bayes Factor = 36). Under this model we estimate a selection coefficient of $s = 0.14$ (0.07–0.31), but a much older age of the allele of 11,400 (4,300–43,700) years. This is at the lower end of estimates for the time of colonization of the Americas [33,64], indicating that the derived allele might have moved into the American populations at a low frequency. This hypothesis is consistent with the very high divergence of the EDAR region between the Mexican and Chinese populations, where we find an F_{ST} of 0.36 (excluding the conserved DEATH-domain), which is much higher than the genome-average F_{ST} of 0.069 between these two populations [65]. This may indicate that the 370A allele has risen in frequency largely independently between these two populations.

This is in contrast to the analysis of Bryk et al. [61], who estimated that the derived 370A allele has been fixed 10,740 years ago. However, both in the 1,000 genomes data and the data of [61], the site is still segregating within the CHB population. While we cannot exclude the scenario of fixation and recent reintroduction of the ancestral allele, the high divergence between Native Americans and East Asians seems to favor a more recent sweep.

G6PD. The G6PD gene is located on the X chromosome, and is one of the best studied cases of selection in humans [9,66–68]. The G6PD gene encodes the Glucose-6-phosphate dehydrogenase

protein, the first enzyme in the pentose phosphate pathway. The G6PD gene has long been associated with reduced-efficiency erythrocytes [69,70], and several hundred variants causing various levels of reduction in catalytic activity have been discovered [71], leading to a significantly reduced fitness in affected individuals. As a benefit, however, G6PD deficiency provides resistance to malaria [72] and therefore even strongly deleterious alleles rise to considerable frequencies in populations where malaria infections are epidemic. Due to these antagonistic selective pressures, G6PD in populations affected by malaria is one of the best examples of balancing selection described in the human genome.

We use the A/A- polymorphism (rs1050828), identified by [66] as the putative site under selection. When applying our method, however, none of the models provided a good fit to the data, indicating that the models we used are too simplistic for the complicated history of G6PD (see Figure 5). The combination of summary statistics with a low EHH, very low IHS and high, non-significant values for Tajima's D and Fay and Wu's H cannot be captured by either of our models. This is not surprising given that the selection on the G6PD locus cannot be described as a selective sweep, but is the effect of balancing selection. It is encouraging the method in this case indirectly, through a poor model fit, helps determine that the simple selective sweep models considered here are not appropriate for this locus.

LCT. In most mammals, the ability to digest lactose, a common disaccharide in milk, decreases when they stop being milk-fed. In contrast, in many humans the main enzyme used to digest lactose into monosaccharides, continues to be expressed even in adults, a phenotype known as lactase-persistence [73–76]. Several presumably independent alleles have been identified that confer the same phenotype [76] in different populations. The first and possibly best-characterized allele is the C/T-13910 polymorphism (rs4988235) that is particularly prevalent in Northern European populations and has been shown in Finnish populations to be 100% associated with the lactase phenotype [77]. Further evidence that the T-13910 allele is causal for the persistence phenotype is given by in vitro analyses [78,79] that found increased enhancer activity.

We analyzed the FIN population from the June 2011 data release of the 1,000 Genomes Project, using the C/T-13910 polymorphism as the selected site. We found a 98.7% posterior probability for the SDN model and only a 1.2% posterior probability for the SSV model, indicating that this particular LCT allele most likely was under selection shortly after it arose. We estimated a rather low selection coefficient of 0.025 (0.003–0.19), and an origin of the mutation 11,200 (1,500–64,900) years ago. Our estimate is much older than the estimates by Bersaglieri et al. [75], who estimated a selection coefficient between 0.09 and 0.19, and an age of the mutation of 1,625–3,188 years using a deterministic approximation based on the observed frequency of the allele. The fact that they used a deterministic approximation may explain the fact that we have wider confidence intervals. Our estimate is more consistent with the estimate of Tishkoff et al., [76] who used the width of the sweep region to date the selected allele to an age of 7,998 years and obtained an estimated selection coefficient of 0.069. Our estimates are also in good concordance with the estimate of Itan et al. [80]. In their study, they modeled the spread of lactase persistence through Europe using a spatially explicit ABC model, which takes advantage of the arrival of dairy farming in various locations. They estimated a selection coefficient of 0.095 in dairy farmers and a slightly older age for the selected allele (7,441 years). While all studies suggest a more recent origin of the selected allele, we note that the confidence intervals on both the selection coefficient and age of the sweep overlap between all four studies.

A complimentary approach to dating the age of an allele, and estimating selection coefficients from modern DNA data, is the usage of ancient DNA [81–83]. Indeed, the derived allele of the LCT the C/T-13910 polymorphism as was found in a single copy in a 5,000 year old sample from Sweden [82], and at a higher frequency of 27% in the Basque country in a sample of approximately the same age [83]. In contrast, the derived allele was absent from an Eastern European sample roughly 7,000 years old [81]. These findings are in good agreement with our estimates and other estimates on genetic data. Based on this ancient DNA evidence, it has been speculated that the LCT allele may have swept from standing variation [83], mainly due to the fact that the derived allele is found at a rather high frequency only two millennia after the introduction of agriculture in that population. However, if the allele was mostly neutral before the arrival we would expect it to be rather old, and in particular we might also expect to see the derived T allele in African populations, which is not the case. Calculating the expected age of an allele at a frequency of 27% [84,85] results in expected ages of 480 ky and 6,500 (2,500–36,000) years for neutrality and selection, respectively, using our estimated selection coefficient and an effective population size of 10,000. While these estimates based solely on allele frequencies should be interpreted with great caution, they nevertheless show that our estimate of a *de novo* selected mutation is consistent both with the observed allele frequencies around 5,000 years ago and an assumed origin of dairy farming 11,000–12,000 years ago.

PSCA. The prostate stem cell antigen gene (PSCA) on chromosome 8 has been proposed to be under selection by [86] based on an analysis of population differentiation in a global array of human populations. A non-synonymous SNP in PSCA (rs2294008) is known to be involved in various forms of cancer [87,88], and we therefore used it as the causal site in our analysis. Interestingly, the derived allele is present in all human populations although the frequency varies considerably between different populations [86]. The highest derived allele frequencies of more than 75% are reported in West African and East Asian populations, whereas some sub-Saharan African and most Native American populations have allele frequencies below 20%. This worldwide distribution of the allele was interpreted as evidence of selection from standing variation [86].

Our analysis confirmed this hypothesis based on analyses of data only from the Yoruban population, with the SSV model receiving a posterior probability of 86.0%, compared to a posterior probability of 23.9% for the SDN model, and 1.2% for a neutral model Under the SSV model, we estimate a selection coefficient of 0.035 (0.004–0.15), with selection having started 8,000 (1,000–54,900) years ago, and the allele being 191 (50–698) thousand years old. The fact that the mutation is distributed globally supports our inference of a sweep based on standing variation.

TRPV6. TRPV6 is in the heart of a 115 kb region on chromosome 7 that has been reported to be under selection [10,89] and has been closely investigated by Akey *et al.* [90]. TRPV6 codes for a protein subunit that encodes cation pores, particularly for calcium ions [90,91]. TRPV6 was found to be in a region of accelerated evolution on the human lineage, as indicated by an elevated ratio of non-synonymous to synonymous fixed differences [90]. In particular, three non-synonymous mutations segregating in humans were found, with a striking diversity pattern; the derived allele was at an intermediate frequency in all African population, but at frequencies of 90% and more in the rest of the world. In addition, both Tajima's D and Fay and Wu's H statistics were significantly negative for non-Africans and European-Americans. For this reason, we restricted our analysis to the

CEU population, and used the first of the non-synonymous SNP (rs4987682) as the focal site for our analysis, as it was the only one that was in the N-terminal region of the TRPV6 protein, the suggested target of selection [90].

While the neutral model could be rejected with a posterior probability close to zero (8e-7), the separation of SSV and SDN model remained inconclusive, with posterior probabilities of 0.55 and 0.45 respectively. The estimate of the selection coefficient was very similar for both models $s_{SSV} = 0.032$ (0.005–0.25), $s_{SDN} = 0.023$ (0.007–0.08), but the confidence interval is much smaller under the SDN model, as expected. Furthermore, the estimated age of the allele differed between models: Under the SSV model, the mutation is inferred to be 211 (29–697) ky old, but became selected only 7,600 (900–43,300) years ago. Under the SDN model, the mutation arose and became selected 23,400 (6,400–70,400) years ago. These findings are in good concordance with the patterns of diversity found previously [10,90], and in particular the evidence that the signature of selection is shared between all non-African populations and thus selection started likely less than 100,000 years ago. Also, the estimate under the SSV model that selection started less than 10,000 years ago is concordant with the role of TRPV6 in absorbing calcium [90].

Neutral regions. In addition to these genes, we also analyzed four putatively neutral regions that were 5 Mb away from our candidate genes. This distance should be big enough that the neutral region are not impacted by the selective sweep, but are likely influenced by the same mutational processes as the selected regions. For all these regions, the neutral model had the highest posterior probability, with posterior probabilities of 0.758, 0.932, 0.994 and 0.999 for the four regions. This indicates we are indeed able to discern selected from neutral regions.

Conclusions on data applications

The distribution of summary statistics in Figure S5 illustrates the impact of choice of summary statistic for model inference [92]. Very high values of EHH are clearly indicative of the SDN model, at both a 10 kb and 20 kb distance. Both the SSV and SDN models are associated with low IHS values, whereas the neutral regions have IHS values closer to zero. Tajima's D and Fay and Wu's H are both very informative for model comparison, with SDN genes having very low D values, SSV genes having D values close to zero and neutral regions having positive D values. The main exception is the LCT gene, however, which we inferred to be selected from a *de novo* mutation, but which has a high D. The signal for SDN apparently comes more from the high EHH and low IHS values.

In general, our results are highly concordant with previous studies of these genes. Our estimates tend to gene, G6PD, we could not make any inferences, because we could not reproduce the observed pattern of diversity using simulations of positive directional selection. G6PD shows an extremely narrow region of reduced diversity, surrounded by a region of high diversity. This may be due to balancing selection between malaria resistance and reduced efficiency oxygen transport introducing a signal that cannot be reproduced by our simple model of directional selection. Alternatively, the X-linked mode of inheritance of this locus is not concordant with the assumptions of our model. This also highlights one of the dangers of ABC: It is crucial that the models investigated are able to reproduce the data observed; otherwise false inferences may be drawn. This danger inherent to any ABC approach is also highlighted by the fact that misidentification of the selected site will bias model choice results towards SSV (Figure S6). This can be explained by the fact that even if the neutral site is closely linked to the selected site, it is likely to "escape" the sweep

by recombining away from the selected haplotype, thus giving the signal of selection from standing variation. Similarly, analyzing data simulated under a population bottleneck under a constant size model will bias the results towards stronger selection and SDN (Figure S7), presumably due to the younger age of mutations being taken as evidence of strong selection.

Model choice accuracy

We have shown that it is much more difficult to estimate the model parameters α, and t_1 from the SSV model than from the SDN model. This is unsurprising, as the SSV model has been shown to have a higher variance in allele age, which results in a higher expected variance for most summary statistics [8]. We further show that there is not enough information to estimate the initial frequency of the sweep f_1. This is unsurprising, as the exact position of the switching point has likely only a minor effect on the data, especially as the effect of selection on the trajectory is weak when the allele frequency of the beneficial allele is low [27].

We further notice that the accuracy of our model choice procedure decreases when the signal of selection is weak. Consistent with previous findings, selection is very hard to detect if α is below about 100 [5,11,45]. This is also the point where our method gains power to distinguish between SVN from SDN. The initial frequency required to detect standing variation is moderate at around 3% for weak selection and 2% for stronger selection. However, selection has to be rather strong, at around $\alpha = 1,000$ and initial frequencies have to be above 5% to allow accurate inference. Presumably, this is because below this threshold, the stochasticity of the trajectory is very large even under selection, and the difference between the two scenarios is small (see also Figure 1c). These findings are not particularly surprising, as selection scans based on summary statistics have been shown in general to have low power under these conditions [45].

These findings certainly limit the scope of our approach. Could we do better with a different strategy? As discussed in the introduction, the ABC approach simplifies data in two ways. First, instead of using the full data, we use an array of summary statistics. Second, we substitute an exact match between observation and simulations with an approximate match, depending on "close" simulations. Regarding the use of summary statistics, we note that summary statistics have been widely used to detect selection from genetic data [9–11,14,19], and currently provide the only way to detect selection from DNA sequence data. No full likelihood based method is available to detect selection from DNA sequence data that could be adapted to distinguish between the two sweep models entertained here.

The second simplification step is based on the number of simulations performed and the tolerance interval and is imposed by computational constraints. We examine the effect of different numbers of simulations and tolerance cutoffs on our results by calculating relative error rates of the posterior mean and the false negative rate of the model choice. We show in Table S1 that increasing the number of simulation by a large amount or changing the rejection parameter does not significantly improve our results, indicating that we do not lose a lot of information at this stage. This shows that the ABC approach reliably estimates the posterior based on the summary statistics, and as such use all the information available in these statistics. Statistics such as EHH, iHS, Tajima's D, etc, do not contain information that will allow us to provide more reliable estimates. It the light of this, it may appear disappointing that our method does not provide more accurate parameter estimates and more power to distinguish between models. However, it is important to realize that as previously argued, all information regarding selection is in the

frequency path of the selected allele [26,27]. For relatively small selection coefficients and/or small initial frequencies of the selected allele, the paths are very similar for the SSV and SDN models. Even if a full likelihood method could be developed, it is unlikely that it had much more power to distinguish between models.

A further simplification in our method is the restriction to a single population. Population differentiation measures, such as F_{ST}, are one of the most successful ways to detect sweeps from standing variation [24,86], and the inclusion of more realistic models of demography may improve our accuracy. Such models, however, require an additional estimation of multi-population demographic history, which greatly increases the complexity of the model.

While we applied our method only to human candidate loci, it should be possible to easily translate it to other species. In particular, as our simulation results suggest that we have more power to distinguish SDN and SSV if selection is strong, species with large population sizes, such as e.g. *Drosophila* or many microorganisms may be very promising targets for a similar study. Another possible target might be species with very strong artificial selection, such as domesticated animals or plants, where we may gain valuable insights on the domestication history of these species. Of course, our approach could also be combined with ancient DNA (e.g. [83]), which could provide much narrower confidence intervals on time estimates and also help improve estimates of selection coefficients.

The two selection models we consider here, the SSV and the SDN models, are nested models, Setting $f_1 = 1/2N$ in the SSV model recovers the SDN model. To facilitate Bayesian model choice we assign positive probability to $f_1 = 1/2N$, and base our inferences on a choice between $f_1 = 1/2N$ and $f_1 \sim U(0,0.2)$ (See Methods). ABC based model choice has recently been criticized and been shown to be biased in some cases where the statistics used are not sufficient [93,94]. While some of the specific issues raised by [91] are not applicable in our setting because we consider nested models, we do not base our inference on sufficient statistics and the statistical properties of our model choice procedure are, therefore, largely unknown. To address this issue, and in general to validate our approach, we use a method introduced in [95]. We show in Figure S8 that our estimated probabilities only show bias for very small values of the Bayes factor, where there appears to be a bias towards inference of the SDN model for simulations generated under the SSV model with very low values of f_1.

Methods

Models

In order to keep our problem simple, we condition on two important parameters: We assume that the exact site under selection is known from extraneous information, and we furthermore assume that the allele frequency f_{cur} of that site at the time of sampling, $t_{cur} = 0$ is known. The interpretation of the parameters is depicted graphically in Figure 1a.

Unless noted otherwise, we assume a panmictic diploid population of size N with an additive selection model where the ancestral homozygous, heterozygous and derived homozygous genotypes have fitness 1, $1+ s/2$ and $1 + s$, respectively. However, the methodology applied here can easily be adapted to more complex scenarios, e.g. models involving multiple populations, more sophisticated demographic models, and other models of selection. For most simulated data sets, we will report the population scaled mutation rate $\alpha = 4Ns$, as the shape of the allele

frequency trajectory depends only on that compound parameter [25]. However, for most of the genes we analyze previous estimates were made directly on s rather than the compound parameter. To facilitate comparisons, we report s for the genes we analyzed.

Sweep from a *de novo* mutation model. The sweep from a *de novo* mutation (SDN) models a single selective sweep and has two parameters: the mutation rate μ, and the selection coefficient s. For all simulations, we follow [80] and record the time t_0 when the mutation arose, as t_0 depends stochastically on s. The prior distributions we use for this model were $\mu \sim \mathrm{U}(0.5e\text{-}8, 6e\text{-}8)$ and $\log_{10}(s) \sim \mathrm{U}(-3, -0.5)$, where U is a uniform distribution.

Sweep from standing variation model. The sweep from standing variation (SSV) model is identical to the *de novo* mutation model, with the exception that we define a frequency f_1 at which the mutation becomes selected. Unless noted otherwise, the priors for μ and s are the same as in the SDN model, and the prior for f_1 is $f_1 \sim \mathrm{U}(0, 0.2)$. In addition to t_0, which is defined analogous to the SDN model, we are also interested in t_1, the time when the mutation becomes selectively advantageous (i.e. the time when the mutation reaches frequency f_1).

Neutral model. We also consider a neutral model (NT), without any selection. The only free parameter in this model is the mutation rate μ, with the same prior distribution as described under SDN model. As under the selection model, however, we still condition on one site having reached a final allele frequency of f_{cur}, so this model does not correspond to the classical neutral coalescent.

Approximate Bayesian Computation

We use a standard ABC approach [31,32], using a post-sampling adjustment in the form of a GLM [96]. We used the *ABCToolbox* package [40], for specifying priors, rejection sampling and post-sampling adjustment. Unless specified otherwise, we perform 10^5 simulations per model, and retained the 100 (0.1%) simulations with associated Euclidean distance between observed and simulated summary statistics closest to zero. To assess how the number of simulations and acceptance rates influence our results, we analyze 10,000 random data sets with up to 10^7 simulations and varying acceptance rates. We show that these parameters have very little impact on the relative error for both the model choice and parameter estimates in Table S1.

Details of statistics used. We use a diverse array of summary statistics, with the goal of maximizing the information captured, while not including any statistics that just add noise. The statistics we used may be broadly classified into statistics based on haplotype patterns, and statistics based on the site frequency spectrum. The haplotype based statistics we used were iHS [19] and EHH [9]. We recorded EHH in a 10 kb, 20 kb and 50 kb window, centered on the selected site. For the SFS based statistics, we used Tajima's D [20], Fay & Wu's H [23], the average number of pairwise differences π, and the number of segregating sites S as statistics. All these statistics are calculated for three regions: A central region of 20 kb around the selected site, an intermediate region consisting of all sites 20–50 kb away from the selected site, and a faraway region consisting of all sites further than 50 kb away from the selected site. Following [97], we linearize all statistics using a Box-Cox-transformation [98]. To choose a set of informative summary statistics, we used a Partial Least Squares Discriminant Analysis (PLS-DA) [99,100]. PLS-DA is a variant of Partial Least Squares regression, that, similarly to principal component analysis, extracts orthogonal components from a high-dimensional data set (in this case the summary statistics). In contrast to PCA, in PLS-DA these components are chosen such

that the covariance between summary statistics and models is maximized [see e.g. 101]. We did our computations using the 'plsda' function of the mixOmics package for R, and kept the five first PLS-DA axes [100].

ABC has two crucial parameters independent of the model it is applied to: The number of simulations n_S and the acceptance rate ε. To assess the effect of these parameters on inference, we calculate the accuracy of our model choice estimates for various values of n_S and ε (Table S1).

Simulations

All our data sets used for both the ABC inference and the assessment of our procedure are simulated using a modified version of the coalescent simulator *mbs* [102]. *Mbs* allows simulation of genetic data sets with a single selected site using the structured coalescent [26]. *mbs* first simulates the allele frequency trajectory of the site under selection, and then generates a data set conditional on that trajectory. We simulate allele frequency trajectories using Euler's method on the unscaled backwards diffusion equation with selection (eq 7.1 in [103]). This equation makes it very easy to incorporate population size changes by just changing the variance term. To simulate sweeps from standing variation, we set the selection coefficient (s) to zero the first time the trajectory reaches f_1. To analyze simulated data sets, we generally simulate a 100 kb region with a recombination rate of 1.5 cM/Mb. For the human genes, we simulate the gene and a 50 kb flanking region on both sides, resulting in regions that are usually between 100 kb and 150 kb wide. Recombination rates and hotspots are modeled by using the HapMap recombination map [104] in the application to selected genes. For all simulated data sets we assume a constant-sized population. For the analysis of human genes, we use the population history estimated by [105]. Specific regions and details of the used regions are given in Table S2. To ensure that our method does not suffer from a high false positive rate, we also analyze regions 5 Mb downstream from the candidate genes, as they are presumably neutral. For three of genes (ASPM, G6PD, and PSCA), no data was available for these downstream regions, so we analyzed the remaining loci. Candidate loci for selection were chosen using the following criteria: i) they were required to have a derived allele frequency between 0.7 and 0.9 and ii) to be as closely to 5 Mb away from the actual candidate locus in the upstream gene as possible.

We estimate parameters from our models using the standard ABC procedure described above. The parameters we estimate are the mutation rate μ, the age of the sweep t_1 the selection coefficient s and, only under the SSV model, the initial frequency f_1 for the SSV model. In particular we want to determine if our posteriors are unbiased, and if we were able to get reasonable confidence in our estimates. To do this, we simulate data sets with fixed parameters and plotted the average posterior distribution for all parameters in Figure 2.

Model choice

For model choice, our main goal is to calculate the relative probabilities of the models given the data, i.e. Pr(SSV | data), Pr(SDN | data) and Pr(NT | data), which we calculate using the marginal densities as proposed by [96]. To identify parameter regions where there is power to distinguish between the models, we simulate 1,000 data sets each under 30 different scenarios in three series, corresponding to three parameters of interest: The strength of selection α, the frequency when the mutation became selective advantageous $f1$ and f_{cur}, the frequency at which the mutation is observed.

To test the algorithm for approximating Bayes factors, we also use a simulation approach. The estimator of the posterior probability from k simulations, $\hat{p}_k(m|x)$, should have the property $\lim_{k \to \infty} \hat{p}_k(m|x) = p(m|x)$ where m is a model indicator functions for a specific model. Also, for a particular draw from the posterior, $m^{(0)}$, we expect $p(m^{(0)} = m|x) = p(m|x)$, if the simulation algorithm works properly. In other words, $\hat{p}_k(m|x)$, should asymptotically equal $p(m^{(0)} = m|x)$, i.e. if $\hat{p}_k(m|x) = c$, we expect a proportion c of simulations to have been obtained from model m. Equivalently, for an estimated log Bayes factor, $\log_{10} = c$, we expect a proportion $10^c/(1+10^c)$ of draws to be from model m. This prediction is tested in Figure S8, based on 10,000 random data sets from both the SDN and SSV model.

Supporting Information

Figure S1 ROC plots. This figure gives ROC plots for the same data as in Figure 3. As we have three models, the first two columns compare both selection models with a neutral model, and the last two columns compare the two selection model, with the model better characterized as the null model plotted on the x-axis. The lines give the percentage of simulation assigned to the model on the y-axis (sensitivity), given a proportion of models assigned to the x-axis (specificity). Parameters used for the simulations are given above the plot and in the legend box.
(EPS)

Figure S2 Joint posteriors of f_1 and t_1 of simulations under the SDN when analyzed under the SSV model. Inferred joint posterior distribution of nine replicate simulation with parameters of $\alpha = 400$, $\mu = 2.5e-8$ are shown. Red and blank areas correspond to areas with zero probability, yellow areas indicate high probability densities. Notice that for most simulation the inferred initial frequency is below 2%.
(EPS)

Figure S3 Joint posteriors of f_1 and t_1 of simulations under the SSV when analyzed under the SSV model. Inferred joint posterior distribution of nine replicate simulation with parameters $f_1 = 0.1$, of $\alpha = 400$, $\mu = 2.5e-8$ are shown. Red and blank areas correspond to areas with zero probability, yellow areas indicate high probability densities. Notice that for most simulation the inferred initial frequency is above 5%, but the inferred probability of f_1 is often very inaccurate.
(EPS)

Figure S4 Joint posteriors of f_1 and t_1 for analyzed genes. Inferred joint posterior distribution of all seven genes analyzed in this paper. Red and blank areas correspond to areas with zero probability; yellow areas indicate high probability densities.
(EPS)

Figure S5 Observed summary statistic distributions. We show the observed untransformed summary statistics for all genes and genomic regions we analyzed in this study (see Table S2). Colors indicate the most likely mode of evolution: neutral evolution (grey), SDN (blue), SSV (orange) and undetermined (black). TD = Tajima's D, FWH = Fay and Wu's H. The suffix "global" indicates that the statistic was calculated for the entire gene, the suffix "close" indicates the statistic calculated on a 20kb window around the selected site.
(EPS)

Figure S6 Effect of misidentification of selected site: We show the posterior probabilities for SSV (orange), SDN (blue) and NT (grey) for simulations done from a *de novo* mutation(left panel) and standing variation (right panel), if we misidentify the selected allele. Simulations were done with selection strength $\alpha = 1,000$, sample size $n = 100$, mutation rate $\mu = 2.5e-8$ and recombination rate $\rho = 3e-8$. For the SSV simulation, f_1 was set to 0.1 X-axes give the distance between the "true" selected allele from the site for which the summary statistics were calculated. If the distance is larger than 50 kb, we find a bias towards inferring SSV.
(EPS)

Figure S7 Bias in model choice due to a population bottleneck. We show the inferred posterior probabilities for SSV (orange), SDN (blue) and NT (grey) under a constant size model for simulations done under a bottleneck model. The bottleneck started 400 generations ago and lasted for 2,000 generations, which might be similar to the human out-of-Africa bottleneck Simulations were done with selection strength $\alpha = 1,000$, sample size $n = 100$, mutation rate $\mu = 2.5e-8$ and recombination rate $\rho = 3e-8$. For the SSV simulation, f_1 was set to 0.1 X-axes give the strength of the bottleneck as a proportion of the current effective population size. Unaccounted demographic history results in a bias towards estimates of stronger selection.
(EPS)

Figure S8 Model choice bias. B denotes the Bayes factor in favor of the SSV model, $B = \Pr(SSV)/\Pr(SDN)$. We simulated 10,000 data sets under both the SSV and SDN model, and performed our model choice procedure on each data set, and divided the distribution into discrete bins. The figure gives the observed (bars) and expected (red line) proportion of simulations from the SSV in each bin. As can be seen, there is a slight excess of simulations from the SSV on the lower end of the graph. The leftmost bin contains only 28 simulations, two of which were simulated under the SSV model. Both of these simulations had a f_1 below 0.005, corresponding to a parameter region where the SSV and SDN models are very similar. The first and second row of numbers below the figure denote the number of simulations simulated under the SSV and SDN model, respectively.
(EPS)

Table S1 Relative Error for different numbers of simulations and acceptance rates. In this table, we give the relative error of the mean and the false negative rate of the model choice for 1000 data sets randomly simulated under the SSV model with varying number of simulation nSim and proportion of accepted simulations δ. FN = False negative rate in model choice. Chr: chromosome, pop: population we analyzed. All positions given are on the hg19 build of the human genome.
(DOCX)

Table S2 Details of genes and neutral regions analyzed in this study.
(DOCX)

Acknowledgments

Author Contributions

Conceived and designed the experiments: BMP EH-S RN. Performed the experiments: BMP. Analyzed the data: BMP EH-S. Wrote the paper: BMP EH-S RN.

References

1. Kimura M (1985) The Neutral Theory of Molecular Evolution. Cambridge University Press. 388 p.

2. Ohta T (1992) The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics 23: 263–286.

3. Hurst LD (2009) Genetics and the understanding of selection. Nat Rev Genet 10: 83–93. doi:10.1038/nrg2506.

4. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic Selective Sweeps Were Rare in Recent Human Evolution. Science 331: 920–924. doi:10.1126/science.1198878.

5. Hermisson J, Pennings PS (2005) Soft Sweeps. Genetics 169: 2335–2352. doi:10.1534/genetics.104.036947.

6. Pennings PS, Hermisson J (2006) Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. Molecular Biology and Evolution 23: 1076–1084. doi:10.1093/molbev/msj117.

7. Pennings PS, Hermisson J (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2: e186. doi:10.1371/journal.pgen.0020186.

8. Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America 101: 10667–10672. doi:10.1073/pnas.0401720101.

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

10. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, et al. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2: e286. doi:10.1371/journal.pbio.0020286.

11. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. PNAS 102: 7882–7887.

12. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive Natural Selection in the Human Lineage. Science 312: 1614–1620. doi:10.1126/science.1124309.

13. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153–1157. doi:10.1038/nature04240.

14. Nielsen R (2005) Molecular signatures of natural selection. Annual Review of Genetics 39: 197–218.

15. Hudson RR, Kreitman M, Aguadé M (1987) A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116: 153–159.

16. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654. doi:10.1038/351652a0.

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

18. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918. doi:10.1038/nature06250.

19. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72. doi:10.1371/journal.pbio.0040072

20. Tajima F (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123: 585–595.

21. Fu YX (1997) Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics 147: 915–925.

22. Achaz G (2009) Frequency Spectrum Neutrality Tests: One for All and All for One. Genetics 183: 249–258. doi:10.1534/genetics.109.104042.

23. Fay JC, Wu C-I (2000) Hitchhiking Under Positive Darwinian Selection. Genetics 155: 1405–1413.

24. Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. Trends in Ecology & Evolution 23: 38–44. doi:16/j.tree.2007.09.008.

25. Ewens WJ (2004) Mathematical Population Genetics: Theoretical introduction. Springer. 448 p.

26. Hudson RR, Kaplan NL (1988) The Coalescent Process in Models With Selection and Recombination. Genetics 120: 831–840.

27. Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking Effect" Revisited. Genetics 123: 887–899.

28. Spencer CCA, Coop G (2004) SelSim: A Program to Simulate Population Genetic Data with Natural Selection and Recombination. Bioinformatics 20: 3673–3675. doi:10.1093/bioinformatics/bth417.

29. Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59: 2312–2323.

30. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science 327: 883–886. doi:10.1126/science.1183863.

31. Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

32. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

33. Fagundes NJR, Ray N, Beaumont MA, Neuenschwander S, Salzano FM, et al. (2007) Statistical evaluation of alternative models of human evolution. PNAS 104: 17614–17619.

34. Peter BM, Wegmann D, Excoffier L (2010) Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Molecular Ecology: no–no. doi:10.1111/j.1365-294X.2010.04783.x.

35. Wegmann D, Excoffier L (2010) Bayesian Inference of the Demographic History of Chimpanzees. Molecular Biology and Evolution 27: 1425–1435. doi:10.1093/molbev/msq028.

36. Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol (Amst) 25: 410–418. doi:10.1016/j.tree.2010.04.001.

37. Miller N, Estoup A, Toepfer S, Bourguet D, Lapchin L, et al. (2005) Multiple Transatlantic Introductions of the Western Corn Rootworm. Science 310: 992. doi:10.1126/science.1115871.

38. Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, et al. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24: 2713–2719. doi:10.1093/bioinformatics/btn514.

39. Jobin MJ, Mountain JL (2008) REJECTOR: software for population history inference from genetic data via a rejection algorithm. Bioinformatics 24: 2936–2937.

40. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010) ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11: 116. doi:10.1186/1471-2105-11-116.

41. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences 100: 15324–15328. doi:10.1073/pnas.0306899100.

42. Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences 104: 1760–1765. doi:10.1073/pnas.0607208104.

43. Blum MGB, François O (2009) Non-linear regression models for Approximate Bayesian Computation. Stat Comput 20: 63–73. doi:10.1007/s11222-009-9116-0.

44. Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413.

45. Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Research 16: 702–712. doi:10.1101/gr.5105206.

46. The 1000 Genomes Project Consortium (2010)_A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi:10.1038/nature09534.

47. Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27: 718–719. doi:10.1093/bioinformatics/btq671.

48. Li H, Mukherjee N, Soundararajan U, Tarnok Z, Barta C, et al. (2007) Geographically separate increases in the frequency of the derived AD-H1B*47His allele in eastern and western Asia. Am J Hum Genet 81: 842–846. doi:10.1086/521201.

49. Peng Y, Shi H, Qi X, Xiao C, Zhong H, et al. (2010) The ADH1B Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. BMC Evolutionary Biology 10: 15. doi:10.1186/1471-2148-10-15.

50. Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, et al. (2002) A Global Perspective on Genetic Variation at the ADH Genes Reveals Unusual Patterns of Linkage Disequilibrium and Diversity. The American Journal of Human Genetics 71: 84–99. doi:10.1086/341290.

51. Eng MY, Luczak SE, Wall TL (2007) ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review. Alcohol Res Health 30: 22–27.

52. McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, et al. (2004) Fermented beverages of pre- and proto-historic China. Proceedings of the National Academy of Sciences of the United States of America 101: 17593–17598. doi:10.1073/pnas.0407921102.

53. Bond J, Roberts E, Mochida GH, Hampshire DJ, Scott S, et al. (2002) ASPM is a major determinant of cerebral cortical size. Nat Genet 32: 316–320. doi:10.1038/ng995.

54. Zhang J (2003) Evolution of the Human ASPM Gene, a Major Determinant of Brain Size. Genetics 165: 2063–2070.

55. Kouprina N, Pavlicek A, Mochida GH, Solomon G, Gersch W, et al. (2004) Accelerated Evolution of the ASPM Gene Controlling Brain Size Begins Prior to Human Brain Expansion. PLoS Biol 2: e126. doi:10.1371/journal.pbio.0020126.

56. Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, et al. (2005) Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens. Science 309: 1720–1722. doi:10.1126/science.1116815.

57. Mekel-Bobrov N, Lahn BT (2007) Response to Comments by Timpson et al. and Yu et al. Science 317: 1036. doi:10.1126/science.1143658.

58. Currat M, Excoffier L, Maddison W, Otto SP, Ray N, et al. (2006) Comment on "Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens" and "Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans." Science 313: 172. doi:10.1126/science.1122712.

59. Timpson N, Heron J, Smith GD, Enard W (2007) Comment on Papers by Evans et al. and Mekel-Bobrov et al. on Evidence for Positive Selection of MCPH1 and ASPM. Science 317: 1036. doi:10.1126/science.1141705.

60. Yu F, Hill RS, Schaffner SF, Sabeti PC, Wang ET, et al. (2007) Comment on "Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens." Science 316: 370. doi:10.1126/science.1137568.

61. Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, et al. (2008) Positive Selection in East Asians for an EDAR Allele that Enhances NF-κB Activation. PLoS ONE 3: e2209 doi:10.1371/journal.pone.0002209.

62. Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, et al. (2008) A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. Hum Genet 124: 179–185. doi:10.1007/s00439-008-0537-1.

63. Kimura R, Yamaguchi T, Takeda M, Kondo O, Toma T, et al. (2009) A Common Variation in EDAR Is a Genetic Determinant of Shovel-Shaped Incisors. Am J Hum Genet 85: 528–535. doi:10.1016/j.ajhg.2009.09.006.

64. Waters MR, Stafford TW (2007) Redefining the Age of Clovis: Implications for the Peopling of the Americas. Science 315: 1122–1126. doi:10.1126/science.1137166.

65. The International HapMap 3 Consortium. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52.

66. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, et al. (2001) Haplotype Diversity and Linkage Disequilibrium at Human G6PD: Recent Origin of Alleles That Confer Malarial Resistance. Science 293: 455–462. doi:10.1126/science.1061573.

67. Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, et al. (2002) Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet 71: 1112–1128. doi:10.1086/344345.

68. Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at G6pd and the signature of malarial selection in humans. Genetics 162: 1849–1861.

69. Carson PE, Flanagan CL, Ickes CE, Alving AS (1956) Enzymatic Deficiency in Primaquine-Sensitive Erythrocytes. Science 124: 484–485. doi:10.1126/science.124.3220.484-a.

70. Beutler E (1994) G6PD deficiency. Blood 84: 3613–3636.

71. Nkhoma ET, Poole C, Vannappagari V, Hall SA, Beutler E (2009) The global prevalence of glucose-6-phosphate dehydrogenase deficiency: A systematic review and meta-analysis. Blood Cells, Molecules, and Diseases 42: 267–278. doi:10.1016/j.bcmd.2008.12.005.

72. Ruwende C, Khoo SC, Snow RW, Yates SNR, Kwiatkowski D, et al. (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376: 246–249. doi:10.1038/376246a0.

73. Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, et al. (2001) Lactase haplotype diversity in the Old World. Am J Hum Genet 68: 160–172. doi:10.1086/316924.

74. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, et al. (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30: 233–237. doi:10.1038/ng826.

75. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. (2004) Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am J Hum Genet 74: 1111–1120.

76. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31–40. doi:10.1038/ng1946.

77. Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, et al. (2003) Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. Gut 52: 647–652.

78. Olds LC, Sibley E (2003) Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum Mol Genet 12: 2333–2340. doi:10.1093/hmg/ddg244.

79. Troelsen JT, Olsen J, Møller J, Sjöström H (2003) An upstream polymorphism associated with lactase persistence has increased enhancer activity. Gastroenterology 125: 1686–1694.

80. Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The Origins of Lactase Persistence in Europe. PLoS Comput Biol 5: e1000491. doi:10.1371/journal.pcbi.1000491.

81. Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG (2007) Absence of the Lactase-Persistence-Associated Allele in Early Neolithic Europeans. PNAS 104: 3736–3741. doi:10.1073/pnas.0607187104.

82. Malmström H, Linderholm A, Lidén K, Storå J, Molnar P, et al. (2010) High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. BMC Evolutionary Biology 10: 89. doi:10.1186/1471-2148-10-89.

83. Plantinga TS, Alonso S, Izagirre N, Hervella M, Fregel R, et al. (2012) Low prevalence of lactase persistence in Neolithic South-West Europe. European Journal of Human Genetics. Available:http://www.nature.com/ejhg/journal/vaop/ncurrent/abs/ejhg2011254a.html. Accessed 6 April 2012.

84. Kimura M, Ohta T (1973) The Age of a neutral mutant persisting in a finite population. Genetics 75: 199–212.

85. Maruyama T (1974) The Age of an Allele in a Finite Population. Genetics Research 23: 137–143. doi:10.1017/S0016672300014750.

86. Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, et al. (2011) Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. The American Journal of Human Genetics 89: 368–381. doi:10.1016/j.ajhg.2011.07.025.

87. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer (2008). Nat Genet 40: 730–740. doi:10.1038/ng.152.

88. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, et al. (2009) Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet 41: 991–995. doi:10.1038/ng.421.

89. Stajich JE, Hahn MW (2005) Disentangling the Effects of Demography and Selection in Human History. Molecular Biology and Evolution 22: 63–73. doi:10.1093/molbev/msh252.

90. Akey JM, Swanson WJ, Madeoy J, Eberle M, Shriver MD (2006) TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. Human Molecular Genetics 15: 2106–2113. doi:10.1093/hmg/ddl134.

91. Birnbaumer L, Yidirim E, Abramowitz J (2003) A comparison of the genes coding for canonical TRP channels and their M, V and P relatives. Cell Calcium 33: 419–432. doi:10.1016/S0143-4160(03)00068-X.

92. Marin J-M, Pillai N, Robert CP, Rousseau J (2011) Relevant statistics for Bayesian model choice. arXiv:11104700. Available:http://arxiv.org/abs/1110.4700. Accessed 24 July 2012.

93. Didelot X, Everitt RG, Johansen AM, Lawson DJ (2010) Likelihood-free estimation of model evidence.

94. Robert CP, Cornuet J-M, Marin J-M, Pillai N (2011) Lack of confidence in ABC model choice. 11024432. Available:http://arxiv.org/abs/1102.4432. Accessed 25 August 2011.

95. Cook SR, Gelman A, Rubin DB (2006) Validation of software for Bayesian models using posterior quantiles. Journal of Computational and Graphical Statistics 15: 675–692.

96. Leuenberger C, Wegmann D (2010) Bayesian Computation and Model Selection Without Likelihoods. Genetics 184: 243–252. doi:10.1534/genetics.109.109058.

97. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. Genetics 182: 1207–1218. doi:10.1534/genetics.109.102509.

98. Box G, Cox D (1964) An analysis of transformations. JR Stat Soc, Ser B 26: 211–243.

99. Tenenhaus M (1998) La régression PLS: théorie et pratique. Editions TECHNIP. 274 p.

100. Lê Cao K-A, González I, Déjean S (2009) integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics 25: 2855–2856. doi:10.1093/bioinformatics/btp515.

101. Boulesteix A-L, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinformatics 8: 32–44. doi:10.1093/bib/bbl016.

102. Teshima KM, 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.

103. Kimura M (1964) Diffusion models in population genetics. Journal of Applied Probability 1: 177–232.

104. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science 310: 321–324. doi:10.1126/science.1117196.

105. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature. Available:http://www.nature.com/nature/journal/vaop/ncurrent/full/nature10231.html. Accessed 5 April 2012.