Title
Detecting recent selective sweeps while controlling for mutation rate and background selection

Journal
Molecular Ecology, 25(1)

ISSN
0962-1083

Authors
Huber, Christian D
DeGiorgio, Michael
Hellmann, Ines
et al.

Publication Date
2016

DOI
10.1111/mec.13351

MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 142–156 doi: 10.1111/mec.13351

DETECTING SELECTION IN NATURAL POPULATIONS: MAKING SENSE OF GENOME SCANS AND TOWARDS ALTERNATIVE SOLUTIONS

Detecting recent selective sweeps while controlling for mutation rate and background selection

CHRISTIAN D. HUBER,*† ‡ MICHAEL DEGIORGIO,§¶ INES HELLMANN** and RASMUS NIELSEN† †

*Max F. Perutz Laboratory, University of Vienna, Vienna, Austria, †Vienna Graduate School of Population Genetics, University of Veterinary Medicine, Vienna, Austria, ‡Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 621 Charles E. Young Drive South, Los Angeles, CA 90095-1606, USA, §Departments of Biology and Statistics, Pennsylvania State University, University Park, PA, USA, ¶Institute for CyberScience, Pennsylvania State University, University Park, PA, USA, **Department Biologie II, Ludwig-Maximilians-Universität München, Großhaderner Str. 2, 82152, Planegg-Martinsried, Germany, ††Departments of Integrative Biology and Statistics, University of California, Berkeley, CA, USA

Abstract

A composite likelihood ratio test implemented in the program SWEEPFINDER is a commonly used method for scanning a genome for recent selective sweeps. SWEEPFINDER uses information on the spatial pattern (along the chromosome) of the site frequency spectrum around the selected locus. To avoid confounding effects of background selection and variation in the mutation process along the genome, the method is typically applied only to sites that are variable within species. However, the power to detect and localize selective sweeps can be greatly improved if invariable sites are also included in the analysis. In the spirit of a Hudson–Kreitman–Aguadé test, we suggest adding fixed differences relative to an out-group to account for variation in mutation rate, thereby facilitating more robust and powerful analyses. We also develop a method for including background selection, modelled as a local reduction in the effective population size. Using simulations, we show that these advances lead to a gain in power while maintaining robustness to mutation rate variation. Furthermore, the new method also provides more precise localization of the causative mutation than methods using the spatial pattern of segregating sites alone.

Keywords: background selection, Hudson–Kreitman–Aguadé test, population bottlenecks, sweep detection, SWEEPFINDER

Received 31 March 2015; revision received 31 July 2015; accepted 17 August 2015

Introduction

Rapid advances in sequencing technology during the past few years have facilitated studies using genome-wide molecular data for detecting signatures of selective sweeps (Akey *et al.* 2002; Carlson *et al.* 2005; Kelley *et al.* 2006; Voight *et al.* 2006; Wang *et al.* 2006; Kimura *et al.* 2007; Sabeti *et al.* 2007; Tang *et al.* 2007; Williamson *et al.* 2007; Xia *et al.* 2009; Qanbari *et al.* 2012; Chávez-Galarza *et al.* 2013; Long *et al.* 2013; Ramey *et al.* 2013; Huber *et al.* 2014), and a large number of computational methods have been developed for this purpose (e.g. Fu & Li 1993; Kim & Stephan 2002; Sabeti *et al.* 2002, 2007; Kim & Nielsen 2004; Nielsen *et al.* 2005; Voight *et al.* 2006; Jensen *et al.* 2007; Boitard *et al.* 2009; Chen *et al.* 2010; Pavlidis *et al.* 2010, 2013; Li 2011). The various methods differ in the assumptions that they make about the selective sweep. For example, the extended haplotype test and its derivatives are powerful in cases where the beneficial mutation has not yet reached fixation in the population (Sabeti *et al.* 2002, 2007; Voight *et al.* 2006). Methods based on measures of population subdivision rest on the assumption that a selective sweep in geographically structured populations has a locally confined effect on genetic diversity,

Correspondence: Christian D. Huber, Fax: (310) 206 0484; E-mail: chuber53@ucla.edu

which increases population differentiation at the position of the sweep (Akey *et al.* 2002; Chen *et al.* 2010). More recently, statistics have been developed specifically for the detection of soft sweeps, that is a pattern caused by multiple haplotypes sweeping to high frequencies (Ferrer-Admetlla *et al.* 2014; Garud *et al.* 2015).

In this study, we are solely concerned with the model of a classical hard selective sweep in a single population, and we assume that the beneficial mutation has reached fixation not too long ago. The methods usually applied in this scenario aim to detect deviations in the shape of the site frequency spectrum (SFS), which can be quantified with simple summary statistics like Tajima's D or Fay and Wu's H. In addition, more powerful statistics have been developed that explicitly model the effect of a selective sweep on the SFS in a likelihood ratio framework (Kim & Stephan 2002; Nielsen *et al.* 2005). Kim & Stephan (2002) proposed a composite likelihood ratio statistic based on calculating the product of marginal likelihood functions for all sites on a chromosome under models with and without a selective sweep at a particular position, and under the assumption of a panmictic population of constant size. The resulting composite likelihood ratio is then computed for each position of interest to evaluate the evidence for a sweep at those positions. This method, therefore, does not only incorporate information regarding the SFS, but does so in a way that uses the spatial distribution (along the chromosome) at segregating alleles of different frequencies. The null distribution of the test statistic is approximated using simulations. An extension to this test was proposed by Nielsen *et al.* (2005). In this method, the overall genomic SFS is used as the neutral, or background, model instead of using the standard neutral model as the null. The distribution of the SFS under the alternative hypothesis of selection is derived by considering the way a selective sweep would modify the observed background distribution of allele frequencies. This leads to a computationally fast method, facilitating genomewide analyses. Nielsen *et al.* (2005) also argued that the use of the overall genomic SFS to represent the neutral case leads to increased robustness, and showed that the method was robust to a two-epoch growth model and an isolation–migration model with population growth in both populations, with parameters estimated from human single nucleotide polymorphism (SNP) data (Marth *et al.* 2004). Since then, it has become clear that, while this method may be more robust than some previous SFS-based approaches, it can produce a high proportion of false positives if there has been a strong recent bottleneck in population size, but a standard neutral model is used to calculate critical values (Jensen *et al.* 2005; Pavlidis *et al.* 2008).

If invariable sites are included in the analysis, then both the methods of Kim & Stephan (2002) and Nielsen *et al.* (2005) may be sensitive to assumptions regarding selective constraint and mutation rates. A region with strongly reduced levels of variation due to selective constraint or reduced mutation rate may be misinterpreted as a region that has experienced a recent selective sweep (Nielsen *et al.* 2005; Boitard *et al.* 2009; Pavlidis *et al.* 2010). For these reasons, Nielsen *et al.* (2005) proposed using only polymorphic sites, an option that became incorporated as default in both SWEEPFINDER (Nielsen *et al.* 2005) and SweeD (Pavlidis *et al.* 2013).

Background selection can also lead to locally reduced levels of neutral variation (Charlesworth *et al.* 1993, 1995; Hudson & Kaplan 1994, 1995; Nordborg *et al.* 1996; Charlesworth 2012; Cutter & Payseur 2013) and cannot be ignored for the study of neutral polymorphisms in many cases (Williford & Comeron 2010; Cutter & Payseur 2013; Messer & Petrov 2013). Cutter & Payseur (2013) argue that the inevitability and prevalence of deleterious mutations necessitates the incorporation of background selection in the null model when identifying positive selection. There is a well-developed mathematical framework for quantifying the strength of background selection given the genomewide mutation rate, recombination rate, position of functional elements and distribution of fitness effects (Hudson & Kaplan 1995; Nordborg *et al.* 1996; Nicolaisen & Desai 2013). As data sets and methods for estimating the effect of background selection for each position in the genome are becoming available (McVicker *et al.* 2009; Comeron 2014), the objective of developing methods for detecting positive selection that can take background selection into account is becoming tenable. However, it is unknown to what degree those currently available maps of background selection are also affected by recurrent selective sweeps (McVicker *et al.* 2009), which could lead to overcorrection when using those maps.

Here, we explore the potential for improving the composite likelihood ratio test of SWEEPFINDER (Nielsen *et al.* 2005) by either including invariant sites that differ with respect to an out-group (i.e. fixed differences), or all invariant sites, in addition to polymorphic sites. When only including fixed differences, the method incorporates the information typically represented in a Hudson–Kreitman–Aguadé (HKA) test (Hudson *et al.* 1987), but adds the information from the spatial distribution of allele frequencies. We show that this approach is robust to variation in mutation rate across the genome, and also develop an approach for incorporating estimates of the strength of background selection into the SWEEPFINDER framework. Using the reduction in

diversity relative to divergence as a necessary hallmark of a selective sweep in our model also helps to reduce false positives, *for example* in the case of a recent population bottleneck. Finally, we compare results of both the old and the new version of the likelihood ratio test applied to human genetic data.

Materials and methods

Including invariant sites into the SWEEPFINDER *framework*

Starting with n aligned DNA sequences, each of length L, we wish to determine whether a selective sweep has occurred at some defined position along the sequence. Based on results of Durrett & Schweinsberg (2004), Nielsen *et al.* (2005) derived an approximate formula for p_k*, the probability of observing k derived alleles, $k \in \{1,2,\ldots,n-1\}$, in a sample of size n, immediately after a selective sweep, for a site at a particular distance (d) from the selected mutation. For each k, p_k* is a function of d, the background allele frequency distribution $\mathbf{p} = (p_1,p_2,\ldots,p_{n-1})$, and the parameter $\alpha = r\ \ln(2\ N_e)/s$. Here, r is the per-base per-generation recombination rate, s is the selection coefficient, and N_e is the effective population size. The parameter p_k is the expected proportion of sites, not affected by the sweep, in which the derived allele has a frequency of k/n in the sample. The vector $\mathbf{p}$ is commonly estimated as the observed SFS from the whole genome, under the assumption that only a small and therefore negligible proportion of positions are affected by selection. The parameter α quantifies the relative influence of recombination and selection, with small values of α indicating strong sweeps.

The equations in Nielsen *et al.* (2005) allow for the incorporation of invariant sites that may or may not be fixed differences relative to an out-group, using $\mathbf{p} = (p_0, p_1,\ldots,p_n)$ as the definition of $\mathbf{p}$, and with the modification that the upper limit of the sum in equation (5) of Nielsen *et al.* (2005) is n and not n–1. The quantity p_k* is a function of the probability of a lineage escaping a selective sweep, $P_e = 1–\exp(-\alpha d)$, where d is the distance between the polymorphic site and the sweep location. Our new version of SWEEPFINDER allows distances between sites to be defined as genetic distance. This is achieved by allowing d to be defined by a recombination map rather than by physical distance as in the previous version. As in Nielsen *et al.* (2005), we then define the composite likelihood ratio statistic $CLR = 2[\log(CL_{sweep}) - \log(CL_{background})]$, where CL_{sweep} is the composite likelihood maximized over alpha, and $CL_{background}$ is the composite likelihood calculated under the assumption of $\alpha = \infty$. This is a composite likelihood ratio, and not a full likelihood ratio, because sites in the genome are not independent, but correlated due to linkage disequilibrium. One thing to notice, about which there has existed some confusion in the literature, is that this approach is not window based but in theory incorporates information from all SNPs in the genome to inform the CLR calculated for a single point in the genome. However, for computational efficiency SWEEPFINDER uses a cut-off for distances from the focal SNP to include in the calculation. As distances become large, the contribution to the likelihood ratio approaches zero. The value used for the cut-off in SWEEPFINDER is $\alpha d = 12$, corresponding to a probability of a lineage escaping a sweep of 0.999994. Furthermore, SWEEPFINDER calculates probabilities on a grid of recombination distances and uses a smooth interpolation to approximate probabilities for a particular point.

The effect of including invariant sites on the SFS is illustrated in Fig. 1. In a region close to the site of the selective sweep, variability is reduced because almost all the probability mass is concentrated on fixed alleles. Notice also that under the infinite sites assumption, as the mutation rate affects all categories proportionally, a change in the mutation rate will not change the SFS defined on $\{1, 2,\ldots,n\}$ (Fig. 1a, b). This statement does not hold true when invariant sites that do not differ from the out-group (Fig. 1c) are incorporated.

Correcting for background selection

A B-value (B) is the factor by which the effective population size is expected to be reduced due to background selection, that is $N_e^* = N_eB$, where N_e and N_e* are the effective population sizes with and without background selection, respectively (Charlesworth 2012). We will assume that a reasonable estimate of the 'B-value map', the value of B for each site in the genome, is available (see, e.g., McVicker *et al.* 2009; for humans). We note that this assumption limits the use of our method to organisms for which such estimates have been obtained. We also note that we only model the main effect of background selection: the well-known reduction in effective population size. However, background selection can also affect the distribution of allele frequencies (Charlesworth *et al.* 1993, 1995; Hudson & Kaplan 1994; Lohmueller *et al.* 2011a; Zeng & Charlesworth 2011; Nicolaisen & Desai 2013), an effect that is ignored here.

Based on the B-value map, the expected site frequency spectrum can be adjusted simply by multiplying all categories in the spectrum, except for the zero and the n (fixed differences) category, by B, that is, by setting $p_k^{(B)} = Bp_k$ for $1 < k < n$–1, as the expected diversity reduction is proportional to B. The n category can be adjusted as described in the next section. The zero

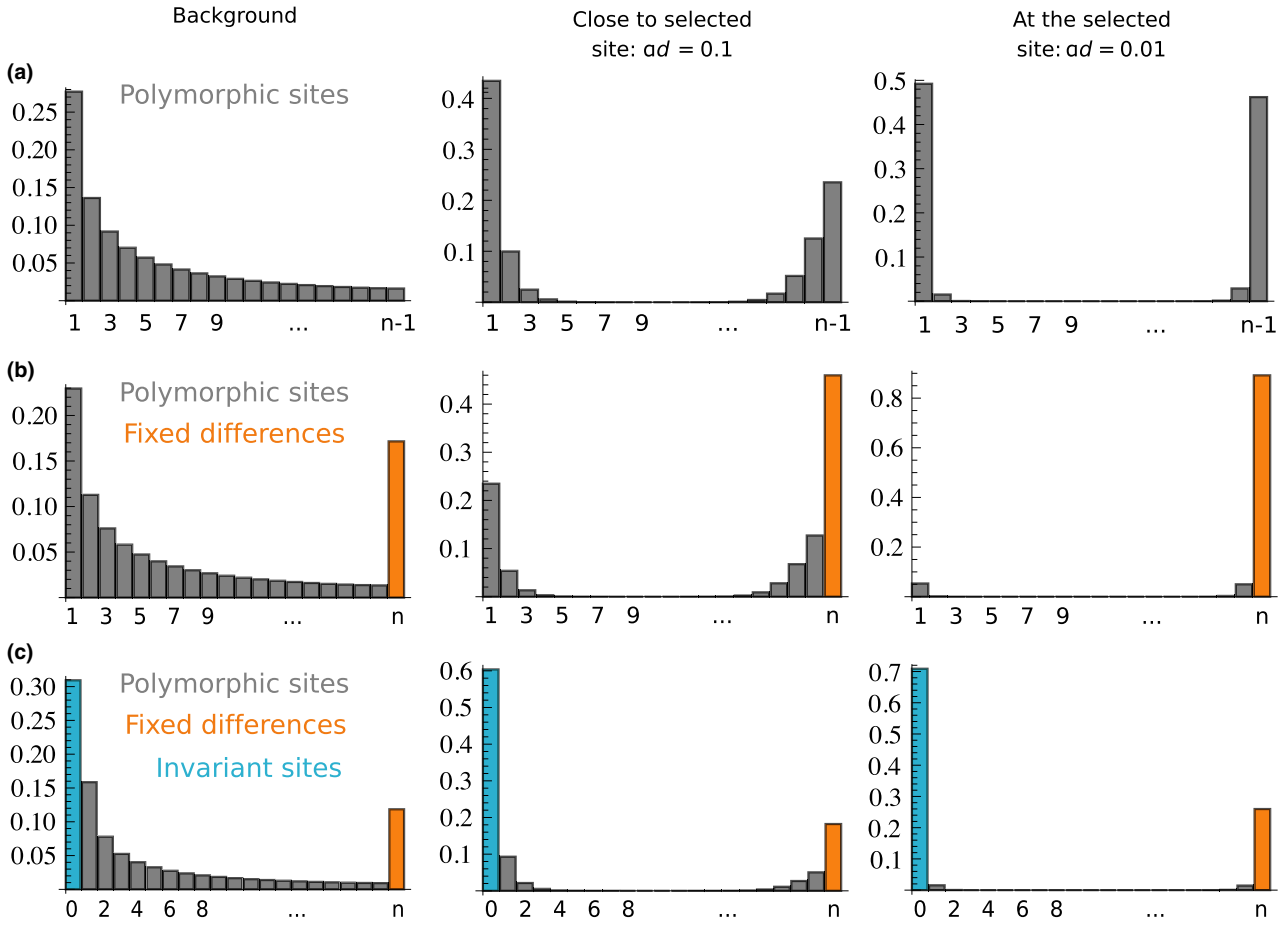

Fig. 1 Effects of a selective sweep on the expected site frequency spectrum (SFS). The horizontal axis of each plot shows the derived allele frequency in a sample of size $n = 20$; the vertical axis shows the proportion of sites with that frequency. (a) The expected SFS of a standard neutral background and of a neutral site linked to a selective sweep, assuming different distances between the neutral site and the sweep locus. (b) The same expectations for a SFS that is extended to include the class of fixed differences (sites that are invariant in the sample, but different to an out-group species). (c) The same expectations for a SFS that is extended for the class of fixed differences and invariant sites that do not differ from the out-group species. All expectations are calculated with the formulas in Nielsen *et al.* (2005).

category can be obtained by standardization, that is $p_0^{(B)} = 1 - \sum_{k=1}^{n} p_k^{(B)}$. If the zero category is not included in the analysis, all included categories will have to be standardized to ensure that the frequencies sum to 1. The calculation of the CLR then proceeds as in Nielsen *et al.* (2005).

Effect of background selection on number of fixed differences

We assume the availability of a sample of n chromosomes and a single chromosome from an out-group species, which split from the in-group species g generations ago. Fixed differences are defined as sites with an allele that is invariant within the in-group sample, but different from the allele at the orthologous position of the out-group chromosome (Fig. 2). The expected number of fixed differences, K, in the sample is then $E[K] = \mu(2T_{\text{anc}} - T_{\text{in}})$, where T_{in} is the time to the most recent common ancestor in the in-group sample, T_{anc} is the divergence time between in-group and out-group, and μ is the per-generation mutation rate. We further assume a standard neutral coalescent model with populations of constant sizes $N_{\text{e,in}}$ and $N_{\text{e,anc}}$ for the in-group population, and ancestral population, respectively (Fig. 2), and that the split time, g, is so large that we can assume $\Pr(T_{\text{in}} > g) \approx 0$. Then $E[T_{\text{in}}] = 4N_{\text{e,in}}(1–1/n)$, where n is the sample size of in-group sequences, and $E[T_{\text{anc}}] = g + 2N_{\text{e,anc}}$. Then, under an infinite sites model

$$E[K] = \mu\left(2(g + 2N_{\text{e,anc}}) - 4N_{\text{e,in}}(1 - 1/n)\right)$$

and the relative number of fixed differences with and without background selection is

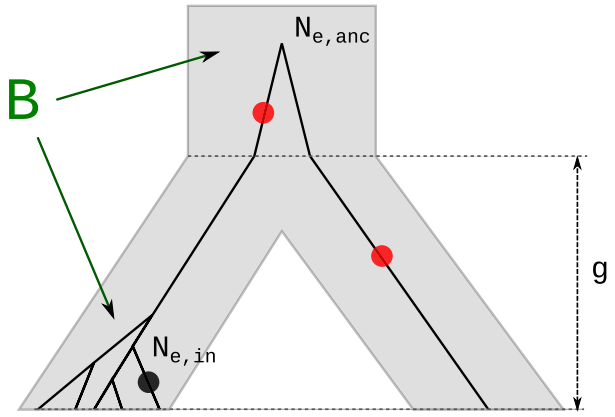

Fig. 2 Definition of fixed differences and polymorphic sites. We assume the infinite sites model, that is every mutation happens on a different site. We define fixed differences (red) as sites that are not polymorphic within the in-group and differ between in-group and out-group. Note that mutations on the lineage to the out-group also count as a fixed difference. Here, the in-group is sampled with 5 chromosomes and the out-group with one chromosome. Background selection influences both the number of fixed differences and the number of polymorphisms.

$$\frac{E[K^{(B)}]}{E[K]} = \frac{\mu\left(2(g+2BN_{e,anc}) - 4BN_{e,in}(1-1/n)\right)}{\mu\left(2(g+2N_{e,anc}) - 4N_{e,in}(1-1/n)\right)} = \frac{g+2BN_{e,anc} - 2BN_{e,in}(1-1/n)}{g+2N_{e,anc} - 2N_{e,in}(1-1/n)}$$

which reduces to

$$\frac{E[K^{(B)}]}{E[K]} = \frac{g+2BN_e/n}{g+2N_e/n}$$

for $N_{e,anc} = N_{e,in} = N_e$. In the limit of large split times ($g \to \infty$), $E[K^{(B)}]/E[K] \approx 1$, and the effect of background selection on fixed differences can generally be ignored if $g \gg N_e/n$. In our new version of SWEEPFINDER, if the *B*-value map is included for sweep detection, estimates of $N_{e,in}$, $N_{e,anc}$ and g have to be provided to the software.

Constant size and bottleneck simulations

Simulations were performed under the model described in Fig. 2 assuming L = 100 kb and n = 30, using *msms* (Ewing & Hermisson 2010).

We set the split time, g, between in-group and out-group to 20 coalescent time units ($2N_e$ generations), resulting in a neutral divergence of 0.1. The scaled mutation rate $\theta = 4N_e\mu$ per site was set to 0.005 and the population scaled recombination rate per site, $4N_er$, to 0.02. Those parameters where chosen to be comparable to the ones in (Nielsen *et al.* 2005). One chromosome was sampled from the out-group species to classify invariant sites into sites that differ or do not differ to the out-group. To analyse the effect of reduced mutation rate in a genomic region compared to the background, we varied the mutation rate between 0.1 and 0.9 times the mutation rate in other regions. Further, we simulated two demographic scenarios, a constant size and a bottleneck population. In simulations with selection, the selected mutation was introduced in the population at specified times (0.01, 0.02, 0.04, 0.06, 0.08, 0.16, 0.24,. . ., 1.2), at a frequency of $1/(2N_e)$ with a population scaled selection coefficient of $2N_es$ = 200. We only kept simulations in which the mutation did not get lost (-SFC option in *msms*).

For the bottleneck simulations, we varied onset (0.004, 0.04 and 0.4), strength (0.05, 0.1 and 0.5) and duration (0.08 and 0.4) and explored all possible combinations of those parameters. To compare different bottleneck scenarios, θ was scaled depending on the bottleneck parameters to keep SNP density constant for all simulations (on average ~1850 SNPs per simulation, see Figs S2 and S3). This was achieved by calculating a scaling factor (f) using the formula of Marth *et al.* (2004) and the approach described in DeGiorgio *et al.* (2014). The recombination rate was scaled to be $4fN_er$ to keep the mutation over recombination rate ratio comparable to the constant size simulations. The split time was also adjusted to g/f.

For the simulations with selective sweeps, we used 200 replicates for each parameter setting and sweep start time and assumed N_e = 10 000. For calculation of the false-positive rate (FPR), we conducted 4000 neutral simulations under each bottleneck condition. For power calculations, we generally assumed that the correct demographic model was known and used to identify critical values for the test, while for investigations of robustness, we used the standard neutral model to estimate critical values. In all cases, the background site frequency spectrum was estimated using 1000 neutral simulations. Note that in our analyses, the significance level is set so that 5% of all simulated 100 kb regions are expected to contain at least one outlier, that is it is an experiment-wise significance level based on our simulated sequence length.

Simulation of background selection

Background selection was simulated with the forward simulation software SFS_CODE (Hernandez 2008). To reduce the computational burden, we simulated relatively small populations of N_e = 250 (Hernandez 2008). We used n = 15 and assumed constant population sizes with neutral and deleterious mutation rate of θ = 0.0025 per bp, $g/(4N_e) = 2$, $4N_er = 0.15$ and L = 100 kb. We fur-

ther assumed a selection coefficient of $2N_es = -50$, reducing the neutral diversity by background selection by 40%. In the middle of the sequence (from 37.5 kb to 62.5 kb), we introduced a 100-fold reduction in recombination rate, which led to a local increase in the effect of background selection and an 80% reduction in SNP density (see Fig. S5). This reduction in recombination rate mimics a selective sweep by locally reducing diversity through the effect of background selection (Fig. S5). While the effect of background selection is more likely to act on a megabase scale (McVicker *et al.* 2009), we simulated strong background selection in a small segment of simulated sequence to keep the data sets small reducing the computational burden of the simulations. However, the difference in scale should not affect the generality of our conclusions.

To simulate selective sweeps in conjunction with background selection, a single positively selected mutation was introduced into the population 0.02 coalescence time units ($2N_e$) in the past in the middle of the sequence, with a selection coefficient of $2N_es = 2000$, or 0.1 coalescence time units in the past with a selection coefficient of $2N_es = 200$. Whenever the mutation was lost from the simulation, the output was discarded and the simulation was repeated. For simulations without background selection, we set the deleterious mutation rate to zero. The composite likelihood ratio was calculated using a grid of 40 points for each simulated data set. The neutral simulations described above were used as background site frequency spectrum. For the HKA test, we used nonoverlapping windows of length 5 kb.

Analysis of human data

We used data from nine unrelated European individuals sequenced by Complete Genomics (Drmanac *et al.* 2010). Data and filtering steps were the same as in DeGiorgio *et al.* (2014). We found that, in low complexity regions around the centromeres and elsewhere in the genome, diversity drops to low levels while divergence from chimpanzee stays constant or even increases relative to other regions. Those regions are highly correlated with low values of CRG100, a measure of local alignability, and increased levels of missing data. Therefore, they most likely reflect errors due to poor mappability and not patterns of recent selective sweeps. To filter those regions out, we only retained SNPs and fixed differences with a CRG100 value of 1 and full sample size. We also excluded windows with average CRG100 value of less than 0.9, in 100 kb windows moving by 50 kb. CRG100 values (Derrien *et al.* 2012) were downloaded from the UCSC Genome Browser at http://genome.ucsc.edu/.

We obtained recombination rates between pairs of sites from the sex-averaged pedigree-based human recombination map from deCODE Genetics (Kong *et al.* 2010).

For the sweep scan, we calculated a composite likelihood ratio at grid points with 1 kb spacing. We ran both standard SWEEPFINDER, using only polymorphic sites (CLR1), and our new method using polymorphic sites, fixed differences relative to chimpanzees and the B-values map from McVicker *et al.* (2009) (CLR2B). Each chromosome was run in parallel, taking 1 week for the whole genome. We assume an effective population size of humans and the human–chimpanzee ancestor population of 10 000 and 99 000, respectively, and a split time of 240 000 generations (McVicker *et al.* 2009). To look for overlaps with previous sweep scans, we use the supplementary table from (Akey 2009), compiling SFS-based scans (Carlson *et al.* 2005; Kelley *et al.* 2006; Williamson *et al.* 2007), LD-based scans (Voight *et al.* 2006; Wang *et al.* 2006; Kimura *et al.* 2007; Sabeti *et al.* 2007; Tang *et al.* 2007) and one F_{ST}-based scan (Akey 2009).

Results

Including diversity as a sweep signal increases power and precision

We compare the power and accuracy of the CLR test when including only variable sites (CLR1), variable sites and fixed differences (CLR2), and all sites (CLR3), in the calculation of the composite likelihood ratio. CLR1 is the CLR that is calculated by current sweep detection software (Nielsen *et al.* 2005; Pavlidis *et al.* 2013). We start with a simple scenario of a constant population size with no background selection, and an advantageous mutation in the middle of the sequence, with selection strength of $2N_es = 200$ and varying start times (see Methods).

The power drops quickly with the age of the selected mutation and approaches zero for sweeps that start more than 0.5 coalescence time units ($2N_e$ generations) in the past (Fig. 3a). The root-mean-square error (RMSE) of the estimated location of the sweep also increases for older sweeps (Fig. 3b). At an age of 0.5 coalescent time units, localization using the CLR1 statistic is not better than picking a site at random. In contrast, CLR2 and CLR3 still have power until 0.8 time units in the past. Furthermore, for sweeps that start 0.2 coalescence time units in the past, there is an almost 40% increase in power. We also tested CLR2 and CLR3 on data with less neutral divergence from the out-group (1%, 5%) and do not see a reduction in power (Fig. S1). This suggests that a recent split time between in- and out-group does not negatively affect performance of the tests.

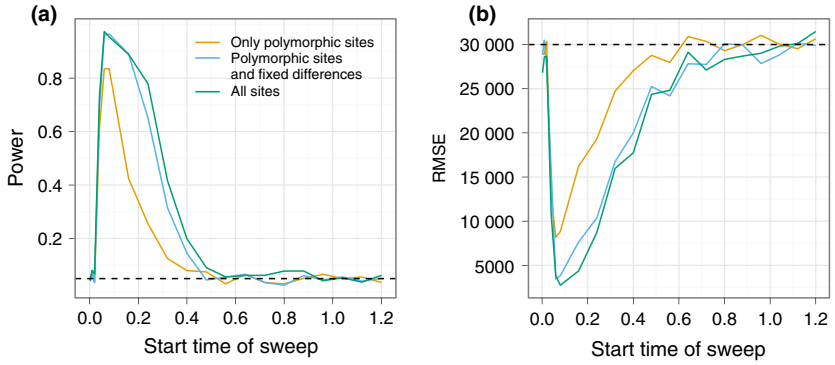

Fig. 3 Power and accuracy comparison of the CLR tests. The power of the selection tests (a) and the root-mean-square error (RMSE) of the estimated location of the sweep (b) is shown as a function of the time since introduction of the beneficial mutation into the population in $2N_e$ generations (*x*-axis). The dashed line in (a) indicates the 5% significance level assumed in the power calculations, and in (b), it indicates the RMSE in case of random (uniform) localization of the sweep position. RMSE is calculated as the standard deviation of estimated minus true position in bp. Each 100 kb simulated region is scored significant if it contains at least one significant outlier CLR at the 5% level.

In summary, both power and accuracy of localization of the selected allele vastly increase when including fixed sites and there is little difference between including all sites (CLR3) and fixed differences (CLR2).

Including only fixed differences maintains robustness against mutation rate variation

We investigated the effect of varying mutation rates on the inference of sweeps. To this end, we use two sets of simulations in 100 kb windows: one set with a population mutation rate of 0.005 and another set of simulations with reduced mutation rates relative to the first set. The likelihood ratio is then calculated using the first set of simulations as the background SFS when calculating the CLR for the second set (see Materials and Methods). The power is estimated by running a third set of simulations, with similarly varying mutation rates as in the second set, but with a beneficial mutation with selection coefficient $2N_es = 200$ arising at 0.08 coalescence units in the past. The selected site is placed in the middle of the simulated region. In both cases, the null distribution of the test statistic is obtained using simulations with a constant high mutation rate of 0.005 and no selective sweeps.

If all sites are used for inference (CLR3), the power is close to 1 irrespective of the mutation rate. However, the FPR increases rapidly with the reduction in the mutation rate, so that at a 60% reduction already half the signals are false positives and at a reduction of 40%, almost all of the signals from the neutral simulations are false positives (Fig. 4). This explains the apparently constant power. The reason for the increase in FPR with decreasing mutation rate is the reduction in the proportion of polymorphic sites relative to all other sites, which replicates what is expected after a selective sweep (Fig. 1c). In contrast to CLR3, the power of both CLR1 and CLR2 reduces with the reduction in mutation rate (Fig. 4). For CLR1, this reduction in power is due to the reduced SNP density. The power for CLR1 is only 80% to begin with and drops to 55% at a reduction in mutation rate by 50%. CLR2 performs much better: the power to detect a sweep is still at 80% with a mutation rate reduction of 50%. The FPR for both CLR1 and CLR2 stays at or below the expected 5% level, as predicted, as decreasing mutation rate does not affect the relative proportion of polymorphic sites to fixed differences. In fact, the tests become extremely conservative when a mutation rate that is too high is used to obtain the null distribution of the composite likelihood ratio. This is because the distribution of the composite likelihood ratio is not invariant with respect to the number of SNPs included in the analysis. Including many more SNPs for generating the null distribution (as a consequence of a higher mutation rate) than used in the analyses of the data will result in a conservative test.

Robustness to population bottlenecks

We simulated several bottleneck scenarios, varying onset, duration and strength of the bottleneck (Fig. 5a) and calculated the FPRs for the three sweep statistics (CLR1-3). The background SFS is calculated from neutral simulations under the respective bottleneck model. Critical values for a 5% significance level were obtained from simulations with a constant size population. For each bottleneck scenario, we adjusted mutation rate, recombination rate and split time to

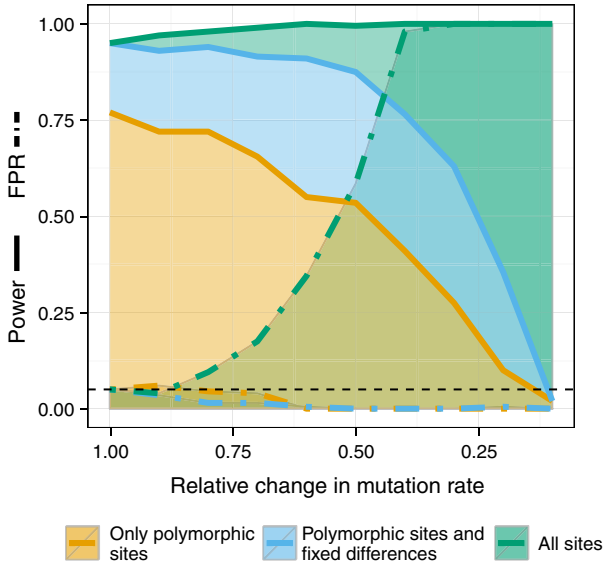

Fig. 4 False-positive rate (FPR) and power with reduced mutation rate. FPR and power, at a nominal significance level of 5%, as a function of the reduction in mutation rate of the sequence under investigation, relative to the mutation rate of the sequence that is used to calculate the background site frequency spectrum. Both power and FPR are calculated by assuming a nominal significance level that is derived from simulations with no reduction in mutation rate (relative reduction = 1). Each 100 kb simulated region is scored as significant if it contains at least one significant outlier CLR at the 5% level.

the out-group, so that the expected number of SNPs as well as divergence from the out-group is comparable for all bottleneck models and for the constant size model (see Methods). This is equivalent to adjusting mutation rate and recombination rate in the simulations used to obtain critical values to match the observed data.

In a scenario with recent (onset = 0.004 or 0.04) and strong (strength = 0.05) or intermediate (strength = 0.1) bottlenecks, this generates a large proportion of false positives (>87%) if the population size is assumed to be constant (Fig. 5b). The proportion of false positives is smaller if the bottleneck is old, as most lineages coalesce before or during the bottleneck. This is true for all three CLR statistics. However, by including invariant sites or fixed differences in the CLR framework, we increase robustness to bottlenecks whenever the chance of surviving the bottleneck is relatively small, for example when the bottleneck is strong (5%) or intermediate (10%) and has a long duration (0.2). We also conducted simulations under the bottleneck scenarios of Fig. 5, but also varied mutation rate relative to the background mutation rate. We observe the same qualitative relationship as in Fig. 4. In particular, CLR1 and CLR2 consistently show decreasing levels of FPR with decreasing mutation rate, whereas CLR3 consistently shows increasing levels of FPR with decreasing mutation rate (Fig. S4).

As a specific example, European humans are assumed to have experienced a bottleneck during colonization of Europe. Estimated bottleneck parameters (Lohmueller *et al.* 2011b) indicate a relatively recent, short, but strong bottleneck (onset = 0.055, duration = 0.02, strength = 0.05). Simulating data under this scenario results in a proportion of false positives of 0.21 for CRL1 and CLR2 and 0.24 for CLR3, suggesting that constant population size is not a suitable demographic model for calculating significance thresholds for any of the three CLR tests.

False positives due to background selection are prevented by including a B-value map

A strong reduction in diversity relative to divergence in regions of the genome can be caused not only by selective sweeps, but also by the effects of deleterious mutations on linked neutral variation, that is background selection. We adapted SWEEPFINDER to enable the inclusion of genomewide estimates of this effect, the *B*-value map, to account for this type of variation. To evaluate the method, we simulated a genomic region with increased background selection, that is a local reduction in diversity due to background selection (Figs S5, 6 and Methods). The background SFS used to calculate the CLR statistics was based on otherwise identical neutral simulations. To evaluate power in the presence of background selection, we simulated data with both background selection and a recently completed selective sweep located in the middle of the sequence (Fig. 6). The nominal FPR, which is used to determine the nominal significance level, was estimated from neutral simulations without background selection.

The HKA test and the uncorrected CLR2 and CLR3 cannot distinguish background selection from selective sweeps, as is evident from the nearly 100% false positives under our strong background selection scenario (Fig. 7a). If only polymorphic sites (CLR1) are used, the test does not suffer from an elevated level of false positives, indicating that CLR2 and CLR3 mainly pick up on the diversity reduction. However, if the diversity reduction due to background selection is factored in using a *B*-value map, the statistics return to the desired behaviour in that the FPR corresponds to the nominal significance level, while maintaining increased power as compared to CLR1. The same results are found for simulations with background selection and a recent population bottleneck (Fig. S7), assuming bottleneck parameters that were estimated for European humans (Lohmueller *et al.* 2011b).

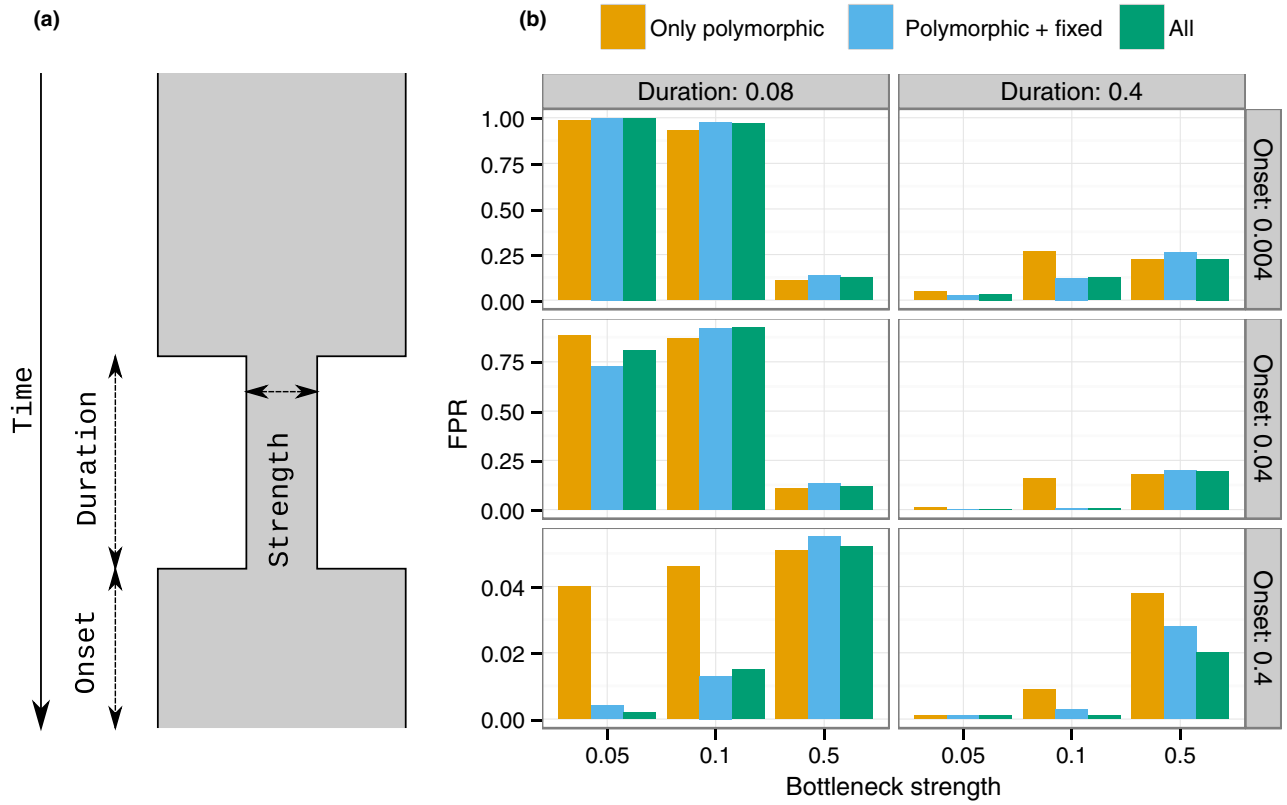

Fig. 5 Robustness to population bottlenecks. (a) Illustration of the bottleneck model used for the simulations, with varying onset time, duration and bottleneck strength leading to population size changes over time. 'Strength' is defined as $N_{e(b)}/N_e$ the effective population size during the bottleneck ($N_{e(b)}$) divided by the effective population size before or after the bottleneck (N_e), 'duration' is measured in number of generations divided by $2N_e$, and 'onset' is number of generations as the bottleneck started divided by $2N_e$. (b) Proportion of false positives (probability of observing at least one wrongly inferred sweep) for bottleneck models if the null model for calculating statistical significance is based on a wrong constant size model with the same average number of SNPs and the same mutation to recombination ratio (see Methods for details). Each 100 kb simulated region is scored significant if it contains at least one significant outlier CLR at the 5% level.

Analysis of a human genetic variation data set

We screened the data from nine unrelated European individuals sequenced by Complete Genomics (Drmanac *et al.* 2010) for selective sweeps to prove the utility of our improvements to SWEEPFINDER. We compare the composite likelihood ratio across the whole genome, calculated using only polymorphic sites (CLR1), with our new approach by including fixed differences with respect to chimpanzees into the calculation (CLR2). To account for varying diversity across the genome due to background selection, we also incorporate the *B*-value map from McVicker *et al.* (2009) into the calculation of CLR2, henceforth referred to as CLR2B.

Due to the complex human demography and the added complication of background selection, we do not calculate critical values, but report the 0.2% most extreme regions in Table 1. This approach has previously been used in other selection scans (e.g. Voight *et al.* 2006) under the argument that it is an outlier approach, although we notice that no formal testing has been carried out here or in Voight *et al.* (2006) to determine the degree to which the most extreme values indeed are outlying with respect to some parametric distribution. We note however that, based on neutral simulations under a simple bottleneck model with parameters taken from Lohmueller *et al.* (2011b), we would expect 8 sweep signals genomewide above the CLR2B threshold of 270, suggesting 33 true positives amongst our 41 candidates in Table 1.

The strongest sweep signal is on chromosome 4, 33.6 Mbp, a region without any annotated genes. The closest gene, *ARAP2*, is 2.15 Mbp downstream from the CLR2 peak. This sweep region has a *B*-value close to one and a strong reduction in diversity relative to divergence. The peak in CLR1 shows that this region is characterized by a sweep-like site frequency spectrum. This region was also listed as a candidate region in LD-based (Voight *et al.* 2006; Wang *et al.* 2006; Kimura *et al.* 2007; Sabeti *et al.* 2007) and SFS-based sweep scans (Carlson *et al.* 2005; Kelley *et al.* 2006; Williamson *et al.* 2007).

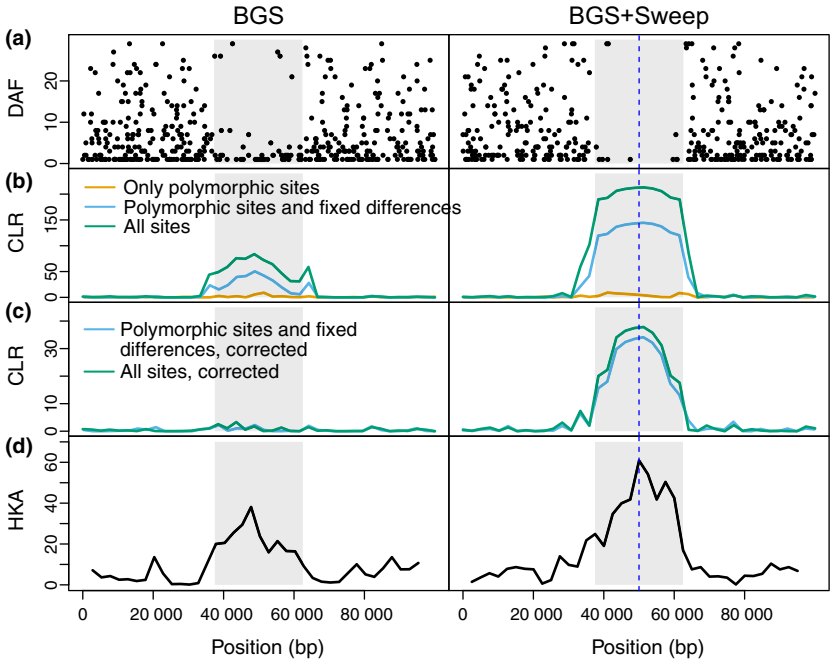

Fig. 6 Two examples of simulation results from forward simulations. Plotted are (a) the derived allele frequency (DAF) of each SNP across the 100 kb sequence, (b) CLR without correction for background selection, (c) CLR corrected for background selection, (d) Hudson–Kreitman–Aguadé test statistic (signed chi-square statistic in nonoverlapping windows). There is a uniform deleterious mutation rate across the 100 kb sequence. A 100-fold reduction in recombination rate in the middle part of the sequence (grey box) generates a larger background selection effect in that part compared to the surrounding sequence (see also Fig. S5). Left: only background selection (BGS). Right: background selection together with a recently fixed selective sweep in the middle of the sequence (BGS+Sweep).

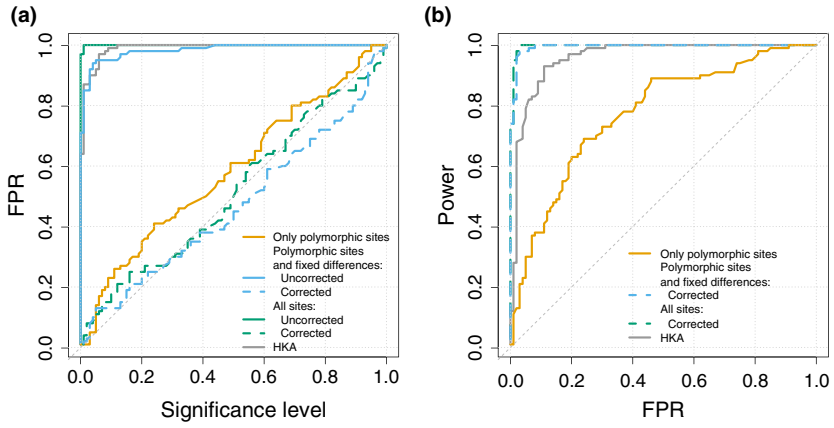

Fig. 7 False-positive rate (FPR) and power under background selection. (a) The observed proportion of false positives in case of simulations with background selection plotted against the nominal FPR (significance level). The nominal FPR is estimated from neutral simulations without background selection. (b) The power to detect a recently fixed selective sweep with $2N_e s$ = 2000 as a function of the proportion of false positives (see Fig. S6b for results with $2N_e s$=200).

The gene with the strongest CLR2B signal is *KIAA1217*, which was suggested to affect lumbar disc herniation susceptibility (Karasugi *et al.* 2009). The gene is also an outlier for haplotype-based sweep statistics for detecting incomplete soft or hard sweeps, in an African population (Ferrer-Admetlla *et al.* 2014). This may suggest that the variant is fixed, or at very high frequency in Europe, but still polymorphic in Africa. Another gene in one of the outlier regions, *HERC2*, is known to modulate iris colour and blonde hair (Wilde *et al.* 2014). This candidate has previously been identified in a screen for population-specific sweeps using XP-CLR (Chen *et al.* 2010). Analyses of ancient DNA suggest that strong selection has been operating on

Table 1 A list of sweep regions, using an outlier approach. Only regions with CLR2B values larger than the genomewide 99.8% quantile are shown. Consecutive outlier CLR2B values are merged to a single sweep region. The overlap with previous scans is tabulated using compiled data from Akey (2009)

Chromosome	Position (Mbp)	Max CLR2B	Percentile rank CLR2B	Genes with outlier CLR2B	Gene closest to CLR2B peak	Distance between CLR2B peak and closest gene (Mbp)	Max CLR1	Percentile rank CLR1	Overlap LD-scans	Overlap SFS-scans	Overlap F_{ST}-scans	Overlap Any
4	33.6	1026	0.0000		*ARAP2*	2.154	681	0.0001	Yes		Yes	Yes
10	23.9	622	0.0001	*OTUD1, KIAA1217*	*OTUD1*	0.105	642	0.0002				
19	20.3	582	0.0005	*LOC284441, ZNF826*	*ZNF826*	0.026	17	0.0884				
7	119.4	566	0.0005	*KCND2*	*KCND2*	0.269	382	0.0021	Yes		Yes	Yes
2	195.0	503	0.0005		*SLC39A10*	1.204	511	0.0009				
7	72.6	496	0.0005	*FZD9, BAZ1B, BCL7B*	*BCL7B*	0	24	0.0749				
4	60.7	485	0.0006		*LPHN3*	1.348	599	0.0005	Yes			Yes
1	1.2	451	0.0006	*SCNN1D, PUSL1, GLTPD1, TAS1R3, UBE2J2, LOC100128842, ACAP3, CPSF3L, DVL1*	*SCNN1D*	0	367	0.0024				
3	98.7	443	0.0007	*EPHA6*	*EPHA6*	0	173	0.0126	Yes		Yes	Yes
15	27.2	435	0.0007	*GOLGA8G, GOLGA8F, WHAMML2, APBA2, HERC2, FAM189A1*	*FAM189A1*	0	144	0.0166	Yes	Yes	Yes	Yes
8	16.2	429	0.0007		*MSR1*	0.141	241	0.0070	Yes	Yes		Yes
16	32.4	429	0.0007		*SLC6A10P*	0.347	1	0.3293				
8	50.6	421	0.0007	*SNTG1*	*SNTG1*	0.345	361	0.0025	Yes	Yes	Yes	Yes
13	64.6	407	0.0007		*PCDH9*	1.155	341	0.0029				
12	78.9	406	0.0007	*PPP1R12A*	*PPP1R12A*	0.028	90	0.0288				
5	17.7	402	0.0008		*BASP1*	0.359	8	0.1178				
8	43.5	400	0.0008	*CHRNB3, HOOK3, FNTA, SGK196, HGSNAT, POTEA, CHRNA6, THAP1, RNF170*	*POTEA*	0.138	338	0.0029				
3	155.7	390	0.0008		*GPR149*	0.115	387	0.0021				
16	46.2	359	0.0009	*PHKB*	*PHKB*	0	443	0.0014	Yes	Yes		Yes
2	13.2	357	0.0009		*TRIB2*	0.381	290	0.0044				
15	43.2	350	0.0010	*DUOX2*	*DUOX2*	0	472	0.0013	Yes	Yes		Yes
19	23.1276	339	0.0010	*ZNF492, ZNF99*	*ZNF91*	0.205	356	0.0026				
5	21.9167	333	0.0010	*CDH12*	*CDH12*	0	241	0.0070	Yes			Yes
12	87.623	332	0.0010		*KITLG*	0.125	313	0.0036	Yes	Yes	Yes	Yes
16	34.4685	325	0.0011	*LOC283914, LOC146481*	*LOC283914*	0	271	0.0052				
8	52.5704	320	0.0011	*PXDNL*	*PXDNL*	0	302	0.0040	Yes		Yes	Yes

Table 1 *Continued*

Chromosome	Position (Mbp)	Max CLR2B	Percentile rank CLR2B	Genes with outlier CLR2B	Gene closest to CLR2B peak	Distance between CLR2B peak and closest gene (Mbp)	Max CLR1	Percentile rank CLR1	Overlap LD-scans	Overlap SFS-scans	Overlap F_{ST}-scans	Overlap Any
5	23.4958	312	0.0012		*PRDM9*	0.048	402	0.0019				
2	21.6707	310	0.0012		*APOB*	0.550	148	0.0160	Yes			Yes
19	47.7082	306	0.0012	*CXCL17, CEACAM1*	*CEACAM1*	0	261	0.0058	Yes	Yes	Yes	Yes
10	110.931	297	0.0013		*XPNPEP1*	0.683	169	0.0131				
2	88.9951	295	0.0013		*FLJ40330*	0.108	2	0.1873				
9	3.55652	287	0.0015	*RFX3*	*RFX3*	0.041	142	0.0168				
8	54.8804	285	0.0015	*ATP6V1H*	*ATP6V1H*	0	201	0.0098				
16	22.2754	285	0.0015	*CDR2*	*CDR2*	0	6	0.1289	Yes			Yes
4	71.9799	284	0.0015		*MOBKL1A*	0.007	401	0.0019				
8	36.2623	280	0.0016		*UNC5D*	0.491	13	0.1012	Yes	Yes	Yes	Yes
13	67.3157	277	0.0016		*PCDH9*	0.613	52	0.0464	Yes			Yes
5	145.12	272	0.0018	*PRELID2*	*PRELID2*	0	118	0.0213				
19	11.9993	272	0.0018	*ZNF433*	*ZNF433*	0	300	0.0041	Yes		Yes	Yes
9	98.6497	271	0.0018	*ZNF782*	*ZNF782*	0	263	0.0056	Yes		Yes	Yes
13	88.619	271	0.0018		*SLITRK5*	1.489	219	0.0084				

HERC2 in western Eurasia during the past 5000 years (Wilde *et al.* 2014).

About half of our outlier regions in Table 1 overlap with at least one candidate region of previous sweep scans in humans (Akey 2009), and most of them are also outlier regarding CLR1. However, there are some notable exceptions: one example is the sweep region on chromosome 7, at 72.6 Mbp, with the genes *BCL7B*, *FZD9* and *BAZ1B*. This region has a small CLR2B percentile rank of 0.0005, but a much larger CLR1 percentile rank (0.071), and is not listed in Akey (2009).

In conclusion, we show that CLR2B shows enrichment for previously detected candidates, but also identifies novel sweep signals. These previously undetected sweeps are likely to be enriched for sweeps that started between 0.2 and 0.8 N_e generations ago and thus escaped detection with LD-, F_{ST}- or SFS-based methods.

Discussion

We evaluated the performance of a composite likelihood ratio test for detecting selective sweeps (Nielsen *et al.* 2005) when including fixed differences in the likelihood ratio in addition to SFS information, using extensive simulations. We show that there can be a marked increase in power as well as a reduction in FPR for a number of different scenarios in several different models of mutation rate variation, population bottlenecks and background selection. We also show that estimates of the strength of background selection can be included into the framework, to prevent false positives in regions with strong, long-term background selection. By applying the method to human genetic data, we detect novel regions that are not identified as candidate regions with the standard SWEEPFINDER approach.

Using invariant sites increases power and robustness

Given that both diversity and divergence change proportionally with mutation rate, we integrate variation in mutation rates by including a measure of divergence to an out-group species. More specifically, we include sites that are not polymorphic within the species under investigation, but differs from an out-group sequence, that is inferred fixed differences. If the SWEEPFINDER CLR is calculated including all sites (CLR3), variation in mutation rates can create false positives (Fig. 4). However, if only fixed differences are added to the SFS (CLR2), the power, but not the FPR, increases. This strongly suggests using CLR2 instead of CLR3 when out-group information is available.

Furthermore, including invariant sites can increase robustness to certain bottleneck scenarios if the bottleneck is of intermediate to high strength, but not too

recent (Boitard *et al.* 2009; Pavlidis *et al.* 2010). However, like many other methods for detecting selective sweeps (Barton 1998; Jensen *et al.* 2005; Voight *et al.* 2006; Boitard *et al.* 2009; Pavlidis *et al.* 2010; Crisci *et al.* 2013), the CLR test can suffer from a disturbingly high FPR in the presence of recent bottlenecks in population size. The use of an empirically derived demographic background SFS does not eliminate the sensitivity to demographic assumptions, because the CLR does not model the correlation in coalescence times along the sequence correctly irrespective of the demographic model. A bottleneck will force many lineages to coalesce in a short amount of time. If the duration of the bottleneck is such that at least some lineages escape the bottleneck in most regions, the few regions in which all lineages coalesce during the bottleneck may very much resemble regions that have been affected by a selective sweep. Realistic demographic models should be used if assigning *P*-values to individual sweeps.

Background selection as a null model for sweep detection

What is often neglected in previous discussions of diversity-based sweep detection methods is variation in diversity across the genome that is not caused by variation in mutation rate (or conservation level), but by variation in background selection, that is by the effect of deleterious mutations on linked neutral variation (Charlesworth *et al.* 1993; Hudson & Kaplan 1995; Charlesworth 2012; Cutter & Payseur 2013). A locally increased level of background selection will lead to a reduction in diversity similar to that expected after a selective sweep.

As data sets and methods for estimating the effect of background selection for each position in the genome are becoming available (McVicker *et al.* 2009), the objective of developing methods for detecting positive selection that can take background selection into account is becoming tenable. We present the first such method by including a map of predicted *B*-values in the calculation of the CLR. McVicker *et al.* (2009) provide such a *B*-value map for humans by defining functional elements based on mammalian sequence conservation, and fitting parameters to phylogenetic data. Therefore, reductions in neutral diversity in regions of the human data do not influence the local estimation of *B*. Our approach considers a local reduction in diversity as evidence for a selective sweep only if it is not also predicted by a local drop in *B*-values, that is background selection is our evolutionary null model (Cutter & Payseur 2013). We simulated background selection levels typical for humans (McVicker *et al.* 2009), and by accounting for background selection, we could effectively prevent false positives without loosing power. If one does not account for background selection, the proportion of false positives is large and similar to that of a HKA test (Fig. 7a).

Application to human data

Finally, by applying our method to human genetic variation data, we show that the new method detects novel regions that were not identified as candidates using the standard SWEEPFINDER approach. Based on our simulations, we would expect those regions to be enriched for old selective sweeps that started between 0.2 and 0.8 N_e generations ago, a time range where the power of other SFS-based, F_{ST}- and LD-based methods is low (Sabeti *et al.* 2006). Interestingly, the strongest signal we find, which has been missed by most previous scans, is near *KIAA1217*, a gene affecting lumbar disc herniation susceptibility. We speculate that the selection in this region may possibly be related to changes in human muscular–skeletal function subsequent to the evolution of erect bipedal walk. Increased risk of lumbar disc herniation is a likely consequence of bipedal walk. We may still be evolving to optimize muscular–skeleton functions after this recent, radical change in skeletal structure and function.

Acknowledgements

We gratefully acknowledge Melissa Hubisz for her help with the original SWEEPFINDER software. We also thank the three anonymous reviewers whose comments/suggestions helped improve and clarify this manuscript. This work was supported by the Austrian Science Fund (Vienna Graduate School of Population Genetics, FWF W1225) to C.D.H.

References

Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, **19**, 711–722.

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

Barton NH (1998) The effect of hitch-hiking on neutral genealogies. *Genetical Research*, **72**, 123–133.

Boitard S, Schlotterer C, Futschik A (2009) Detecting selective sweeps: a new approach based on hidden Markov models. *Genetics*, **181**, 1567.

Carlson CS, Thomas DJ, Eberle MA *et al.* (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, **15**, 1553–1565.

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. *Genetics*, **190**, 5–22.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth D, Charlesworth B, Morgan MT (1995) The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**, 1619–1632.

Chávez-Galarza J, Henriques D, Johnston JS *et al.* (2013) Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology*, **22**, 5890–5907.

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

Comeron JM (2014) Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genetics*, **10**, e1004434.

Crisci JL, Poh Y-P, Mahajan S, Jensen JD (2013) The impact of equilibrium assumptions on tests of selection. *Evolutionary and Population Genetics*, **4**, 235.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews. Genetics*, **14**, 262–274.

DeGiorgio M, Lohmueller KE, Nielsen R (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics*, **10**, e1004561.

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R (2015) SWEEPFINDER2: Increased sensitivity, robustness, and flexibility. arXiv:1505.07142 [q-bio].

Derrien T, Estellé J, Marco Sola S *et al.* (2012) Fast computation and applications of genome mappability. *PLoS ONE*, **7**, e30377.

Drmanac R, Sparks AB, Callow MJ *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA Nanoarrays. *Science*, **327**, 78–81.

Durrett R, Schweinsberg J (2004) Approximating selective sweeps. *Theoretical Population Biology*, **66**, 129–138.

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics (Oxford, England)*, **26**, 2064–2065.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, **31**, 1275–1291.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

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

Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.

Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **31**, 3026–3039.

Hudson RR, Kaplan NL (1994) Gene trees with background selection. In: *Non-Neutral Evolution: theories and molecular data* (ed. Golding B), pp. 140–153. Chapman and Hall, New York.

Hudson RR, Kaplan NL (1995) Deleterious background selection with recombination. *Genetics*, **141**, 1605–1617.

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

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics*, **170**, 1401–1410.

Jensen JD, Thornton KR, Bustamante CD, Aquadro CF (2007) On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*, **176**, 2371–2379.

Karasugi T, Semba K, Hirose Y *et al.* (2009) Association of the tag SNPs in the human SKT gene (KIAA1217) with lumbar disc herniation. *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*, **24**, 1537–1543.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, **16**, 980–989.

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. *Genetics*, **167**, 1513–1524.

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

Kimura R, Fujimoto A, Tokunaga K, Ohashi J (2007) A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE*, **2**, e286.

Kong A, Thorleifsson G, Gudbjartsson DF *et al.* (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, **467**, 1099–1103.

Li H (2011) A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Molecular Biology and Evolution*, **28**, 365–375.

Lohmueller KE, Albrechtsen A, Li Y *et al.* (2011a) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genetics*, **7**, e1002326.

Lohmueller KE, Bustamante CD, Clark AG (2011b) Detecting directional selection in the presence of recent admixture in African-Americans. *Genetics*, **187**, 823–835.

Long Q, Rabanal FA, Meng D *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics*, **45**, 884–890.

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, **166**, 351–372.

McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, **5**, e1000471.

Messer PW, Petrov DA (2013) Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences, USA*, **110**, 8615–8620.

Nicolaisen LE, Desai MM (2013) Distortions in genealogies due to purifying selection and recombination. *Genetics*, **195**, 221–230.

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Nordborg M, Charlesworth B, Charlesworth D (1996) The effect of recombination on background selection. *Genetics Research*, **67**, 159–174.

Pavlidis P, Hutter S, Stephan W (2008) A population genomic approach to map recent positive selection in model species. *Molecular Ecology*, **17**, 3585–3598.

Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, **185**, 907–922.
Pavlidis P, Živković D, Stamatakis A, Alachiotis N (2013) SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, **30**, 2224–2234.
Qanbari S, Strom TM, Haberer G *et al.* (2012) A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in Laying Chickens. *PLoS ONE*, **7**, e49525.
Ramey HR, Decker JE, McKay SD *et al.* (2013) Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics*, **14**, 382.
Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
Sabeti PC, Schaffner SF, Fry B *et al.* (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.
Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, **5**, e171.
Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biology*, **4**, e72.
Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proceedings of the National Academy of Sciences, USA*, **103**, 135–140.
Wilde S, Timpson A, Kirsanow K *et al.* (2014) Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*, **111**, 4832–4837.
Williamson SH, Hubisz MJ, Clark AG *et al.* (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, **3**, e90.
Williford A, Comeron JM (2010) Local effects of limited recombination: historical perspective and consequences for population estimates of adaptive evolution. *The Journal of Heredity*, **101**(Suppl 1), S127–S134.
Xia Q, Guo Y, Zhang Z *et al.* (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science (New York, N.Y.)*, **326**, 433–436.
Zeng K, Charlesworth B (2011) The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics*, **189**, 251–266.

C.D.H., I.H. and R.N. conceived the study question and the experimental design. C.D.H. implemented the simulations. C.D.H. and M.D. adapted the software SWEEPFINDER to account for background selection. C.D.H. and M.D. analysed the simulations and the human genetic variation data. C.D.H., M.D., I.H. and R.N. wrote the study.

Data accessibility

We did not generate any new data set for this study. The SWEEPFINDER2 (DeGiorgio *et al.* 2015) software that was used for all CLR calculations is freely available at www.personal.psu.edu/mxd60/sf2.html. The human polymorphism data, the divergence to chimp data, the *B*-value map and the recombination rate data that we used for running SWEEPFINDER2 are available from the Dryad Digital Repository: doi:10.5061/dryad.23d0f.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1. Command line for msms (Ewing & Hermisson 2010) simulations.

Fig. S1 Power of the CLR tests for data with different levels of divergence from the outgroup.

Fig. S2 Boxplot of the distribution of the number of segregating sites for the 18 different bottleneck scenarios, calculated for the simulated 100 kb sequence and 200 replications each.

Fig. S3 Distribution of Tajima's D for the 18 bottleneck scenarios, calculated for the simulated 100 kb sequence and 200 replications each.

Fig. S4 FPR under both population bottleneck and reduced mutation rate.

Fig. S5 Reduction in diversity due to the effect of background selection (B---value map) calculated from forward simulations with SFS_CODE under a constant size model (see Materials and Methods), and under a bottleneck model with parameters for European humans from Lohmueller *et al.* (2011a,b).

Fig. S6 a) The observed proportion of false positives in case of simulations with background selection plotted against the nominal false positive rate (significance level). The nominal false positive rate is estimated from neutral simulations without background selection. b) The power of detecting a recently fixed selective sweep with Nes = 200 as a function of the proportion of false positives.

Fig. S7 FPR and power under both background selection and a population bottleneck.

Fig. S8 Examples from the human genome scan, running both the standard version of CLR using only polymorphic sites (CLR1), and our new version including fixed differences and the B-value map (CLR2).