Journal of Heredity 2012:103(2):287–296
doi:10.1093/jhered/esr128
Advance Access publication January 12, 2012

© The American Genetic Association. 2012. All rights reserved.
For permissions, please email: journals.permissions@oup.com.

Recent Progress in Polymorphism-Based Population Genetic Inference

Jessica L. Crisci, Yu-Ping Poh, Angela Bean, Alfred Simkin, and Jeffrey D. Jensen

From the Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01655 (Crisci, Poh, Bean, Simkin, and Jensen); and the École Polytechnique Fédérale de Lausanne (EPFL), School of Life Sciences, Lausanne, Switzerland 1007 (Jensen).

Address correspondence to Jeffrey D. Jensen at the address above, or e-mail: jeffrey.jensen@epfl.ch.

Abstract

The recent availability of whole-genome sequencing data affords tremendous power for statistical inference. With this, there has been great interest in the development of polymorphism-based approaches for the estimation of population genetic parameters. These approaches seek to estimate, for example, recently fixed or sweeping beneficial mutations, the rate of recurrent positive selection, the distribution of selection coefficients, and the demographic history of the population. Yet despite estimating similar parameters using similar data sets, results between methodologies are far from consistent. We here summarize the current state of the field, compare existing approaches, and attempt to reconcile emerging discrepancies. We also discuss the biases in selection estimators introduced by ignoring the demographic history of the population, discuss the biases in demographic estimators introduced by assuming neutrality, and highlight the important challenge to the field of achieving a true joint estimation procedure to circumvent these confounding effects.

Key words: *Bayesian statistics, demography, likelihood estimation, positive selection*

One of the principle aims of population genetics is to characterize the relative roles of adaptive and nonadaptive processes in shaping patterns of genomic variation. With the advent of next-generation sequencing, data are becoming increasingly sufficient to afford power to large-scale statistical inference. Researchers have become increasingly focused on quantifying the distribution of selection coefficients of newly arising, segregating, and fixed mutations and describing the neutral demographic history of the population under consideration.

The last decade has seen a tremendous increase in computational approaches for the estimation of population genetic parameters. These methodologies are both likelihood based and approximate Bayesian computation (ABC) based and rely on summary statistics, which are based on expected patterns in the site frequency spectrum (SFS), linkage disequilibrium (LD), or divergence. Although intimately related, these estimators have largely been addressed as 3 separate fields: demographic estimation, genomic scans for adaptive fixations, and recurrent hitchhiking (RHH) estimation.

The first challenge in comparing between existing estimators is achieving an understanding of the parameters of interest. This is not always straightforward—with different nomenclature (e.g., the population selection coefficient is designated as either γ and α) and different population scaling between publications (e.g., $2N$ or $4N$). A summary of the parameters being estimated, and the variety of nomenclature used to name those parameters, is given in Table 1. Even after rescaling and renaming to make estimators comparable, a number of obvious inconsistencies between statistics begin to arise. As a motivating example, among RHH estimators—statistics designed to estimate the distribution of selection coefficients and the rate of recurrent selection—results differ wildly, even when considering identical data sets. Sella et al.. (2009) found that among 5 commonly used estimators applied to data from *Drosophila melanogaster*, estimates of the mean selection coefficient differ by 3 orders of magnitude, and estimates of the mean rate of adaptation differ by 2 orders of magnitude. Reasons for these discrepancies are unclear, though could be attributable to the different approach each estimator takes for dealing with the underlying demographic history of the population. Additionally, there appears to be a correlation with the type of data used for estimation—with divergence-based approaches consistently estimating smaller selection coefficients than polymorphism-based approaches (Figure 1).

Thus, our aim here is 2-fold. First, we will review the current state of population genetic inference, discussing the similarities and differences of existing approaches and presenting what is known about their relative performance. This is intended to serve as a useful reference for empiricists attempting to determine which methodology

American Genetic Association
The Journal of Heredity
www.jhered.oxfordjournals.org
Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

Table 1 Summary of common parameter nomenclature

Symbol(s)	Estimated selection parameters
N_e	Effective population size
θ	Population mutation rate = $4N_e\mu$
θ_W	Watterson's estimate of θ
ρ	Population recombination rate = $4N_e r$
D	Tajima's test of the equilibrium neutral model
H	Fay and Wu's test of the equilibrium neutral model
π	Average pairwise difference between nucleotides
s	Selection coefficient
α, γ	Population selection coefficient; ($2N_e s$)
F_{ST}	Measure the amount of heterozygosity within a subpopulation compared with the entire population
K_A/K_S	Ratio of the rate of nonsynonymous substitutions to the rate of synonymous substitutions, also referred to as dN/dS
ω_{MAx}	Finds a window that maximizes sweep-like LD patterns for a given region
ω	Modified ω_{MAx} statistic; using a variable sliding window approach
X	Location of selected site
$2N\lambda$	Rate of beneficial fixation (per base pair, per $2N$ generations)
α	Fraction of advantageous amino acid divergence
$\nu(\Delta N_e)$	Ancestral population size/current population size
$\lambda(\Delta N_e)$	Population size during polymorphic phase/ population size during divergence phase
M	Population migration rate = $4N_e m$
T	Time of historical population split
t	Time of historical population size change
f	Severity of historical population size change
d	Duration of historical population size change

is appropriate for their specific data set and question. Second, we will make a case for the importance of merging demographic and selection inference, highlighting their strong interdependence and emphasizing the need to develop true joint estimators of selection and demography. This is intended as a call to theoreticians, as this need will become increasingly pronounced as whole-genome data sets continue to accumulate.

Inferring the Parameters of Selection

A beneficial mutation that rises in frequency within a population due to positive selection will impact local variation by 1) increasing the frequency of linked neutral alleles—a process known as the hitchhiking effect (Maynard Smith and Haigh 1974; Kaplan et al. 1989), 2) increasing LD on either side of the target (Kim and Nielsen 2004), and 3) reducing local variation—with the extent of this reduction determined by the ratio of the selection coefficient and the local recombination rate (s/r). This distinct genomic pattern is known as a selective sweep and different statistical methods, which will be discussed below, have been developed to detect these characteristic patterns.

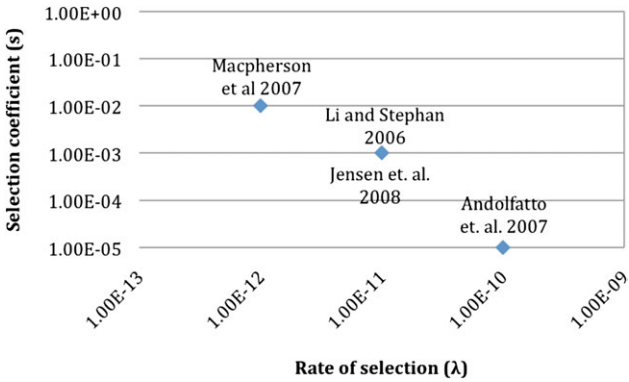

Figure 1. Relationship between the strength and rate of selection in Drosophila. The estimated values for the strength and rate of selection for various Drosophila data sets (see text). As shown, inferred values differ by orders of magnitude for both the rate and strength of RHH. There is perfect overlap between the estimates of Li and Stephan (2006) and Jensen et al., and they appear as one point on the graph.

Frequency Spectrum–Based Estimators

The most common way to detect a pattern that departs from the standard neutral model is to use summary statistics based on the SFS, which represents all of the polymorphisms in a sample and their frequencies. The simplest of these is Watterson's estimate of $\theta = 4N\mu$ (θ_W, Watterson 1975), which represents the number of segregating sites in a population sample. Combining other measures of θ, 2 common tests of the equilibrium neutral model are often used to identify sweep-like patterns: Tajima's D (Tajima 1989), sensitive to the number of rare variants and high frequency variants; and Fay and Wu's H (Fay and Wu 2000), sensitive to high frequency–derived variants. As has been well reviewed in the literature, a selective sweep is expected to reduce variation, resulting in a skew toward rare mutations around the target of selection, and an excess of high-frequency mutations in flanking regions in the presence of recombination (see review of Nielsen 2005).

Frequently, methods for detecting signatures of positive selection in natural populations make use of the fact that the spatial distribution of mutations is expected to be different under neutrality than after a selective sweep. The composite likelihood ratio (CLR) test of Kim and Stephan (2002) considers the probability of observing a derived allele at a given frequency when drawn from a null or sweep distribution. The sweep model is taken from Fay and Wu (2000) and gives the expected frequency spectrum of a derived variant immediately after a hitchhiking event as:

$$\phi(p) = \begin{cases} \frac{\theta}{p} - \frac{\theta}{C} & \text{for } 0 < p < C, \\ \frac{\theta}{C} & \text{for } 1 - C < p < 1, \end{cases} \quad (1)$$

where p is the frequency of the derived variant and C is approximated by $1 - \epsilon^{r/s}$ (ϵ is the frequency of the beneficial

American Genetic Association
The Journal of Heredity
www.jhered.oxfordjournals.org
Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

American Genetic Association

The Journal of Heredity

www.jhered.oxfordjournals.org

mutation before in began increasing deterministically, s is the selection coefficient, and r is the recombination fraction between the selected and neutral locus).

In this case, the null model is the standard equilibrium neutral model and assumes a constant population size and no underlying population structure. As such, Jensen et al. (2005) point out that rejections of this statistic may similarly owe to nonequilibrium demographic histories—showing through demographic simulations that the false-positive rate can be as high as 80% if the population has experienced a recent and severe bottleneck. To this effect, they devise a goodness-of-fit (GOF) test to further distinguish between these 2 models. This test considers the probability of the data given a null model (sweep) versus the probability of an alternative model. Thus, conditional on rejecting equilibrium neutrality with CLR test of Kim and Stephan (2002), the GOF statistic explicitly examines fit to a hitchhiking model, thereby greatly reducing the false-positive rate. An important and inherent assumption of these test statistics is that the selective pressure began acting immediately on a newly arising beneficial mutation and that this mutation reached fixation in the population immediately prior to sampling.

Nielsen et al. (2005) took a different approach to add robustness to CLR test of Kim and Stephan (2002). They modified this likelihood ratio test by drawing the null distribution from the background SFS of the region under consideration, instead of using a standard neutral model. This method proposes that the probability of observing a mutant allele at a given frequency, B, is equal to the probability that all lineages (k) escaped the sweep plus the probability that a certain number of lineages escaped the sweep out of n individuals (summed over all $k < n$, see Equation 6 in Nielsen et al. 2005). Because the neutral expectation is taken from the background SFS, multiple large genomic regions are necessary in order to afford sufficient power. Thus, this method may be used on whole-genome data sets, although CLR test of Kim and Stephan (2002) is intended for subgenomic regions of interest. Alternatively, SweepFinder can draw its null distribution from a user supplied SFS created by simulations, and in this way, it can be used on subgenomic regions. These methods have the most power when there has been a recent fixation within a population.

In a recent study by Williamson et al. (2007), they apply SweepFinder to 1 of every 10 genomic windows of 200 SNPs in 3 human populations (African American, European American, and Chinese). From this analysis, they conclude that as much as 10% of the human genome is affected by selective sweeps through genomic linkage. Because of the reliance on the background SFS, these tests inherently make an assumption about the prevalence of positive selection in the genome—it must be frequent enough that outlier loci indeed represent true positives (i.e., any model will have outliers) and yet must be infrequent enough that the background frequency spectrum may be differentiated from the "true" outliers. As a counterexample to the above results, Hernandez et al. (2011) find that the reduction in variation near human substitutions is not substantial enough to be described by selective sweeps. They suggest that some other form of adaptation must be responsible for the majority of phenotypic differences in humans.

Haplotype-Based Estimators

Another genomic pattern indicative of a selective sweep is LD. Kim and Nielsen (2004) show that LD corresponds with high frequency–derived alleles and that there is increased LD on either side of a beneficial mutation as well as decreased LD across the sweep region. All of these patterns owe to independent recombination events in regions flanking the beneficial mutation. They developed a statistic, ω_{MAx}, which finds a window that maximizes LD for a given region, and demonstrate that this statistic may indeed increase the power to detect selective sweeps. However, like the CLR test of Kim and Stephan (2002), this statistic is meant for subgenomic regions of interest due to computational costs and only has power in a very short time frame following fixation.

The extended haplotype heterozygosity test (EHH) of Sabeti et al. (2002) also utilizes LD to detect sweep signatures, but it is designed for detecting sweeps prefixation. Although this relaxes the assumption of fixation immediately prior to sampling, it has power over a very narrow time scale, owing to the rapid transit time of strongly beneficial mutations. This approach infers the age of high-frequency SNPs and then identifies those that are too young to be consistent with neutrality. There are several methods based on this EHH statistic. In their 2007 publication, Sabeti et al. expand on this method to look for SNPs that have been brought to high frequency in one population but not others (XP-EHH, for cross population EHH). This method operates under the assumption that populations may experience different selective pressures, but it is not robust to neutral demographic histories. Thus, it is useful to combine multiple signals for different tests to achieve a manageable false-positive rate (see below). Chen et al. (2010) develop a similar method to the EHH but use a CLR to compare the allele distributions between populations (XP-CLR), and thus, it is not applicable to a single population.

Voight et al. (2006) observe that the area under the curve of EHH versus distance is much larger under positive selection compared with neutrality. They develop a new test where they plot EHH for the derived and ancestral state of alleles separately and then integrate each of these EHH curves (iHS_D and iHS_A, respectively). They term this new statistic iHS, for integrating haplotype score:

$$\text{unstandardized iHS} = \ln\left(\frac{iHS_A}{iHS_D}\right). \qquad (2)$$

This expression is standardized to have a mean of 0 and a variance of 1. A comparison of the 2 states will produce a negative value when there are long haplotypes carrying the derived state, thus suggesting positive selection. Unlike

Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

American Genetic Association

The Journal of Heredity

www.jhered.oxfordjournals.org

EHH, the iHS test can be used to scan the genome for clusters of SNPs showing evidence of a recent adaptive event.

A common factor of all haplotype-based tests discussed above is that they are suitable for detecting nearly or very recently fixed sweeps because the pattern of LD decays quickly after fixation of the target due to subsequent drift, mutation, and recombination. Human genomic scans using these methods largely identify genes involved in immunity, hair and skin pigmentation, and metabolism in human populations, all of which are argued to be affected by recent adaptations (Sabeti et al. 2002, 2007; Voight et al. 2006; Grossman et al. 2010).

SFS-LD Hybrid Methods

Particular patterns in LD have been shown to increase the robustness in identifying adaptive regions in a nonequilibrium population (e.g., Stephan et al. 2006; McVean 2007). Thus, several new approaches have been developed that combine both frequency- and LD-based statistics in order to improve power to detect selective sweeps. Grossman et al. (2010) develop a Bayesian estimator that combines several statistics in an approach they term CMS or composite multiple signals. This includes iHS and XP-EHH as well as 2 new tests, ΔDAF and ΔiHH. These are sensitive to alleles at high and low frequencies, respectively, and they also incorporate F_{ST} (see Table 1). By utilizing multiple signals, they show increased power to localize causal variants of positive selection.

Pavlidis et al. (2010) modify the ω_{MAx} statistic to be used on whole-genome data sets by implementing a sliding variable-sized window. Additionally, they modify SweepFinder (Nielsen et al. 2005) by including a small fraction of monomorphic sites, arguing that although these sites are left out of the SFS-based test because of their increased computational load, they provide valuable information about the SFS surrounding a putative sweep region. Pavlidis et al. (2010) show that both of these modified statistics perform better than their original counterparts and that using them in tandem has increased power to distinguish single hitchhiking (SHH) events from equilibrium and nonequilibrium neutral models and also reduces false-positive rate.

Also of note, Lin et al. (2011) combine several different summary statistics in a machine-learning approach and find that their power to detect selective sweeps versus bottlenecks is improved when the target of selection is known. A summary of all the methods discussed above can be found in Table 1.

Quantifying RHH

The methods discussed above are intended to detect SHH events, that is, they assume sweeps are strong enough to impact a large genomic region and rare enough to represent outliers in variation. Kaplan et al. (1989) described an RHH model, where the expected number of sweeps (per base pair, per $2N$ generations) is $2N\lambda$, with sweeps occurring at random locations in the genome. The RHH model is most commonly considered for the case of genic selection on new mutations entering the population (e.g., Kaplan et al. 1989; Wiehe and Stephan 1993; Braverman et al. 1995). Under this model, several patterns expected under the single-sweep model no longer apply. For example, the single-sweep model predicts that at some distance from the selected site, coalescent histories are dominated by long internal branches, as some lineages may escape the recent coalescent event via recombination. This results in the widely employed prediction of an excess of high frequency–derived alleles flanking the fixed site (Fay and Wu 2000)—a pattern also utilized in the EHH and iHS class of statistics discussed above. Under RHH models, however, the probability of such a history is small, as sweeps are on average old and high frequency–derived mutations have thus likely drifted to fixation (Przeworski 2002).

Wiehe and Stephan (1993) showed that under an RHH model, for a given recombination rate, the expected level of heterozygosity at linked sites relative to neutral expectations is dependent on the compound parameter $(s)(2N\lambda)$, where $2N\lambda$ is the rate of fixation of beneficial mutations and s is the average strength of selection. This result implies that the 2 parameters are confounded (much like the effective population size, N_e, and mutation rate, μ, in $\theta = 4N_e\mu$) as their effect on expected levels of diversity depends on their product. In *D. melanogaster* and *D. simulans*, lower than expected levels of nucleotide diversity are observed in regions of reduced recombination (Begun and Aquadro 1992) and in the coding sequences of rapidly evolving proteins (Andolfatto 2007; Macpherson et al. 2007). These findings are compatible with either strong but infrequent positive selection (i.e., large s and small $2N\lambda$) or weak but common positive selection (i.e., small s and large $2N\lambda$) (Wiehe and Stephan 1993; Kim 2006).

A number of methods have been proposed for quantifying s and $2N\lambda$ separately using divergence and polymorphism data. These approaches typically make strong assumptions regarding the possible distribution of selection coefficients, the number of adaptive substitutions between species, or the timing of selection. For example, Li and Stephan (2006) examined 250 noncoding regions from an East African population of *D. melanogaster*. Using a likelihood approach, they estimate that approximately 160 beneficial mutations have fixed in this population over the last ~60 000 years (corresponding to $2N\lambda = 1.9E{-}04$), with mean selection coefficient $\hat{s}{\sim}0.002$. This inference is achieved by effectively assuming that the timing of all sweeps is known (and the time since the sweep, $\tau = 0$). Under a recurrent sweep model, this assumption may bias the estimation of s and $2N\lambda$. Additionally, as this method relies on first fitting a demographic model to noncoding DNA polymorphisms, it is possible that the effects of purifying selection on the SFS of noncoding DNA (Andolfatto 2005) may strongly affect the estimates.

Using synonymous polymorphism data in *D. melanogaster*, and divergence to *D. simulans*, at 137 X-linked loci, Andolfatto (2007) employed a maximum likelihood

Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

American Genetic Association

The Journal of Heredity

www.jhered.oxfordjournals.org

approach to estimate the joint parameter $2N\lambda s$, followed by a McDonald–Kreitman-based method to separately estimate $2N\lambda$ and s (McDonald and Kreitman 1991). Based on these calculations, Andolfatto (2007) estimated that most beneficial amino acid substitutions are very weakly advantageous on average (with average $\hat{s}\sim 1.2\text{E}-5$ and $2N\hat{\lambda}\sim 2.6\text{E}-03$). Macpherson et al. (2007), using polymorphism data from *D. simulans* (and divergence to *D. melanogaster*), propose a method to infer the rate and strength of selection from the spatial scale of variation in polymorphism and divergence. In contrast to Andolfatto's (2007) estimates, Macpherson et al. (2007) estimate a much stronger average selection coefficient ($\hat{s}\sim 0.01$) and less frequent selection ($2N\hat{\lambda}\sim 1\text{E}-05$). However, they note that their method is more likely to detect strong selection, so the effects of many weakly beneficial mutations may be missed. Additionally, by assuming fixed selection coefficients—as opposed to distributions—their estimator is upwardly biased (Jensen et al. 2008).

Jensen et al. (2008) took an approximate Bayesian approach (and see Thornton 2009) to estimate the rate and strength of recurrent positive selection. They utilized the relationship of these parameters with the means and standard deviations of common polymorphism summary statistics, including the mean average pairwise diversity (π), the number of segregating sites (S), θ_H, and ZnS (a measure of LD, see Kelly 1997). Calculating these summary statistics from the observed data and from simulated data with parameters drawn from uniform priors, they implement the regression approach of Beaumont et al. (2002), which fits a local-linear regression of simulated parameter values to simulated summary statistics and substitutes the observed statistics into a regression equation. For inferences on selection parameters, they assume exponential distributions of $2N\lambda$ and s, such that each draw from the prior represents the mean of the distribution. As shown in Figure 1, these 4 approaches achieve very different estimates for the strength and rate of recurrent selection.

As a separate but related approach aimed at identifying the fraction of positively selected amino acid mutations, Eyre-Walker and Keightley (2009) use information from both the SFS and divergence. This approach estimates both this proportion as well as a simple demographic model (by assuming that the population begins at equilibrium and experiences a step change in size t generation ago). The fraction of advantageous amino acid divergence (α) is estimated as:

$$\alpha = \frac{d_N - d_S \int_0^\infty 2Nu(N,s) f(s|a,b)\mathrm{d}s}{d_N}, \qquad (3)$$

where $f(s|a, b)$—the distribution of effects of deleterious mutations—is a gamma distribution with scale parameter a and shape parameter b. N is the effective population size, u is the mutation rate per site, and thus, $2Nu(N, s)$ gives the rate of fixation from recurrent mutation. They use synonymous sites to define a neutral class (i.e., $s = 0$), and d_N and d_S are the numbers of selected (i.e., nonsynonymous) and neutral (i.e., synonymous) substitutions per site, respectively. The difference between the observed and expected, as determined from the neutral class, rate of selected substitution corresponds to the estimate of the proportion of adaptive substitutions.

Quantifying Levels of Constraint

To estimate the extent of purifying selection, Loewe et al. (2006) developed a method to characterize the fitness effects of deleterious nonsynonymous mutations, using polymorphism data from 2 species with different effective population sizes. Briefly, the underlying premise is that variants subject to sufficiently strong purifying selection will not increase significantly as effective population size increases, whereas neutral diversity is expected to increase proportionally with population size. Thus, the extent to which nonsynonymous diversity differs between species with different levels of synonymous site diversity should provide information regarding the strength of purifying selection. Thus, for species i, they define $\pi_{S_i} = 4N_{e_i}u$, $\pi_{A_i} = 4c_N N_{e_i} u + (1 - c_N)H_{P_i}$, $K_{S_i} = u$, and $K_{A_i} = c_N u + (1 - c_N)K_{P_i} + c_a u$. Here, H_{P_i} is the mean equilibrium diversity at sites subject to purifying selection, K_{P_i} is the mean substitution rate at these sites, c_N is the fraction of neutral nonsynonymous mutations, u is the mutation rate per site, and c_a measures the substitution as a fraction of all mutations. Assuming a model of strong purifying selection ($N_e s > 1$), the equilibrium diversity contributed by sites subject to purifying selection is well approximated by the deterministic expression $2u/s$ (McVean and Charlesworth 2000). Thus, one can simplify as $\pi_{A_i} = c_n\theta_i + 2(1 - c_n)\frac{u}{s_h}$, where $\theta_i = 4N_{e_i}u$, and s_h is the harmonic mean of selection coefficients (assumed to be the same in both species), and K_{P_i} becomes negligibly small. Thus, $K_{A_i} = c_n u + c_a u$ and $c_n = \frac{\pi_{A_2} - \pi_{A_1}}{\pi_{S_2} - \pi_{S_1}}$. Substituting, selection may be estimated as:

$$2N_{e_1}s_h = \frac{\pi_{S_1}(\pi_{A_1} + \pi_{S_2} - \pi_{A_2} - \pi_{S_1})}{\{\pi_{A_1}(\pi_{S_2} - \pi_{S_1}) - \pi_{S_1}(\pi_{A_2} - \pi_{A_1})\}} \quad \text{and}$$
$$c_a = \frac{K_A}{K_S} - c_N. \qquad (4)$$

In order to account for the confounding effects of population history on the inference of purifying selection, Williamson et al. (2005) proposed a likelihood model-based approach in which data from a putatively neutral class (e.g., synonymous sites) are estimated and fixed in order to perform the estimation of selection on the putatively selected class (nonsynonymous sites). As such, this approach also provides a demographic estimate—a stepwise size change at some time in the past, which may be compared with the above-described approaches. Briefly, given that the expected number of polymorphic sites with i derived alleles segregating in a sample of n is $E[x_i] = \theta_1 F_1(i, n; \tau, \upsilon)$, the probability that a particular SNP is at frequency i out of n is:

Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

American Genetic Association

The Journal of Heredity

www.jhered.oxfordjournals.org

Table 2 Summary of commonly used methodology to infer selection

Method	Estimated selection parameters	Required data set	Model	Inference procedure	Demographic consideration?
Kim and Stephan (2002) (CLR test)	α, X, θ	Polymorphism	SHH	Composite likelihood	No
Jensen et al. (2005) (GOF)	N/A	Polymorphism	SHH	Composite likelihood	Yes
Nielsen et al. (2005) (SweepFinder)	X	Polymorphism	SHH	Composite likelihood	Yes
Kim and Nielsen (2004)	ω_{MA}, X	Polymorphism	SHH	Maximum likelihood	No
Pavlidis et al. (2010)	ω, X	Polymorphism	SHH	Composite likelihood	Yes
Sabeti et al. (2002) (EHH)	X	Polymorphism	SHH	Likelihood	No
Voight et al. (2006) (iHS)	X	Polymorphism	SHH	Likelihood	No
Grossman et al. (2010) (XP-EHH)	X	Polymorphism, 2 populations	SHH	Likelihood	No
Chen et al. (2010) (XP-CLR)	X	Polymorphism, 2 populations	SHH	Composite likelihood	No
Li and Stephan (2006)	s, $2N\lambda$	Polymorphism	RHH	Likelihood	Yes
Andolfatto (2007)	s, $2N\lambda$	Polymorphism, 2 species	RHH	Maximum likelihood	No
Macpherson et al. (2007)	s, $2N\lambda$	Polymorphism, 2 species	RHH	Maximum likelihood	No
Jensen et al. (2008)	s, $2N\lambda$	Polymorphism	RHH	Approximate Bayesian	No
Williamson et al. (2005)	γ, $\nu(\Delta N_e)$	Polymorphism, 2 populations	Purifying selection	Maximum likelihood	Yes
Eyre-Walker and Keightley (2009)	α, $\lambda(\Delta N_e)$	Polymorphism, 2 species	Purifying selection	Diffusion approximation	Yes
Loewe et al. (2006)	θ, π, K_A, K_S, $N_e s$	Polymorphism, 2 species	Purifying selection	Likelihood	No

$$P_1(i, n; \tau, \upsilon) = \frac{F_1(i, n; \tau, \upsilon)}{\sum_{i=1}^{n-1} F_1(j, n; \tau, \upsilon)}, \quad (5)$$

where υ = ancestral population size/current population size and τ = the time of the size change. With selection, we have the function:

$$F_2(i, n; \gamma, \tau, \upsilon) = \int_0^1 \binom{n}{i} q^i (1 - q)^{n-i} f_2(q; \gamma, \tau, \upsilon) \mathrm{d}q, \quad (6)$$

where there is the additional parameter $\gamma = 2Ns$, and the expected number of polymorphic sites segregating at a frequency i in a sample of size n becomes $E[x_i] = \theta_2 F_2(i, n, \gamma, \tau, \upsilon)$. The probability that a particular polymorphic site is at frequency i out of n is:

$$P_2(i, n; \gamma, \tau, \upsilon) = \frac{F_2(i, n; \gamma, \tau, \upsilon)}{\sum_{i=1}^{n-1} F_2(j, n; \gamma, \tau, \upsilon)}. \quad (7)$$

Thus, to estimate the demographic parameters τ and υ, the likelihood function is maximized using class 1 data (synonymous sites). Then, for class 2 data (nonsynonymous sites), these parameters (τ and υ) are fixed in order to maximize the expression and estimate the selection parameter, γ. Inherently, this approach does not account for the effects of linkage on synonymous sites.

Finally, the Eyre-Walker and Keightley (2009) approach described above also allows for estimation of parameters of deleterious mutations while additionally accounting for demography and the presence of beneficial mutations in a stepwise fashion. A summary of the described methods may be found in Table 2.

Inferring the Parameters of Population History

A major complication when attempting to infer the recent action of selection is the demographic history of the population under consideration. Many different demographic scenarios are capable of mimicking hitchhiking patterns (e.g., Tajima 1989; Przeworski 2002; Jensen et al. 2005). Thus, accurate estimation of the underlying demographic model is essential to correctly identifying genomic regions affected by positive (and purifying) selection.

Demographic scenarios involving population subdivision have been a major focus in population genetics and molecular ecology. One of the most common models—"isolation–migration"—considers a population giving rise to 2 populations in the continued presence of gene flow. This model has at least 6 major parameters: the population sizes of the ancestral and 2 extant populations, migration rates between the 2 populations, and the separation time since the ancestral population split (Figure 2). Nielsen and Wakeley (2001) first developed a likelihood/Bayesian framework for estimating the demographic parameters based on a single nonrecombining locus drawn from the 2 populations. Under this model, the ancestral state was inferred by tracing

Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

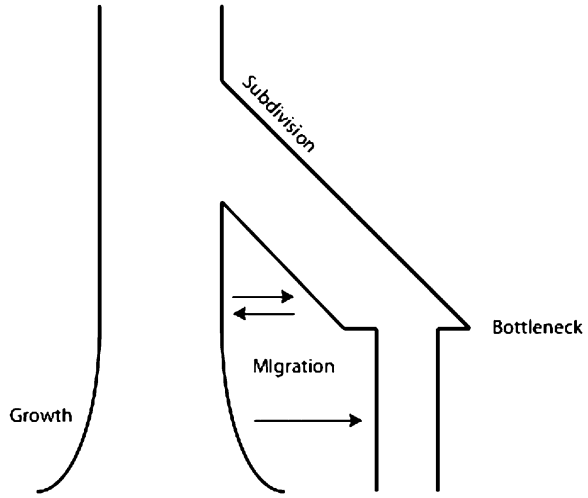

Figure 2. Schematic of the variety of inferred demographic parameters. An ancestral population can experience multiple changes that violate the equilibrium neutral population model. Populations may subdivide. They can experience exponential growth or bottlenecks. Migration can occur between 2 populations both symmetrically and asymmetrically. All of these deviations from a constant-sized randomly mating population can have dramatic impacts on inferences of selection when polymorphism is sampled from present-day populations.

genealogies backward in time under the following coalescent process: Assume that at a given point in time, there are n_1 ancestral gene copies in population 1 and n_2 ancestral copies in population 2. Coalescence events occur in population 1 and population 2 at rate $n_1(n_1 - 1)/2$ and $n_2/(n_2 - 1)/2r$, respectively, where $n_2/n_1 \rightarrow r$. At the same time, migration events are occurring at rates n_1M_1 and n_2M_2, and mutations arise independently on each lineage according to a Poisson process with rate $q/2$. Coalescence events occur until the most recent common ancestor. According to this process, probabilities can be assigned to different genealogies. Nielsen and Wakeley (2001) subsequently implemented a Markov chain Monte Carlo approach (Metropolis et al. 1953; Hastings 1970) to jointly approximate the demographic parameters of the integrated likelihood function.

The application of this model is extended by considering multiple independent unlinked loci simultaneously (Hey and Nielsen 2004), and intralocus recombination has been subsequently introduced into the model (Becquet and Przeworski 2007). Given the clear phylogenetic relationship among populations, this model has also been extended to fit demographic scenarios with more than 2 populations (Hey 2010). However, because this method relies on coalescence simulation, the method quickly becomes computationally intractable as population size increases due to its dependence on population scaled parameters.

Based on a population survey of variation, summary statistics like F_{ST} and Tajima's D may be used to infer changes in historical population size, but these measures only provide a rough picture of the demographic model. Thornton and Andolfatto (2006) implemented an ABC method based on several summary statistics to infer bottleneck parameters in non-African populations of *D. melanogaster*. Although these summary statistics may produce good-fitting models, within this computationally efficient ABC framework, it may indeed be possible to utilize the full frequency spectrum in future work. And although previous work estimated these parameters by maximizing the likelihood function (Wooding and Rogers 2002; Polanski and Kimmel 2003), this approach assumes that no recombination has occurred and becomes computationally infeasible when the data are incompletely linked between variable sites—providing another advantage to an approximate Bayesian framework.

An improved method of maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked sites was developed by Adams and Hudson (2004), and the full-likelihood approach is applicable if the sites are considered independently under the assumption of free recombination. Supposing the full allele frequency spectrum $x = (x_0, x_1, x_2 \ldots x_{n-1})$, where m_0 is the number of sites monomorphic in the sample, and m_i is the number of polymorphic sites in which the derived allele occurred i times in the sample size of n, for L unlinked sites, m is multinomially distributed,

$$\text{Prob}(x) = \binom{L}{x_0, x_1, \ldots, x_{n-1}} \prod_{i=0}^{n-1} P_i^{x_i}, \qquad (8)$$

where P_i is the probability that a site is polymorphic with i derived alleles. Because P_i is a function of the demographic model, we can obtain estimates of the demographic parameters by maximizing the right-hand side of the function. This method involves exploring a large parameter space by using coalescent-based Monte Carlo approaches. Due to computational limitations, these models remain very simple. As an alternative to the full-likelihood approach, the regression-based method of conditional density estimation is introduced in Beaumont et al. (2002), and the likelihood-free approach is implemented to infer population structure and local adaptation in a Bayesian hierarchical model (Bazin et al. 2010).

To efficiently simulate the SFS for model-based comparison, a diffusion approach was adopted to approximate the allele distribution in the population. The distribution of allele frequency f at an arbitrary time t is approximated by the general solution to the forward Kolmogorov equation

$$\frac{\mathrm{d}}{\mathrm{d}t} f(q,t) = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}q^2}\{V(q)f(q,t)\} - \frac{\mathrm{d}}{\mathrm{d}q}\{M(q)f(q,t)\}, \quad (9)$$

where $M(q)$ and $V(q)$ are the mean and variance of the change of allele frequency over 1 unit of time, respectively. Kimura (1955) found a transient solution for a Wright–Fisher population having undergone 2 epochs of population size over recent evolutionary history. Since the diffusion

American Genetic Association
The Journal of Heredity
www.jhered.oxfordjournals.org
Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

approximation is computationally efficient, this allows for the usage of more flexible demographic models. Williamson et al. (2005) utilize this property, considering increasingly realistic demographic scenarios. Gutenkunst et al. (2009) further develop and apply the diffusion-based approach to approximate the joint multipopulation frequency spectrum (implemented in *dadi*, Gutenkunst et al. 2009). Given the infinite-sites model and Wright–Fisher populations in each generation, the dynamics of distribution of allele frequency, f, for P populations could be modeled by a linear diffusion equation:

$$\frac{\partial}{\partial \tau} f = \frac{1}{2} \sum_{i=1,2,\ldots,P} \frac{\partial^2}{\partial^2 x_i} \frac{x_i(1-x_i)}{v_i} f - \sum_{i=1,2,\ldots,P} \frac{\partial}{\partial x_i} [\gamma_i x_i (1-x_i) + \sum_{j=1,2,\ldots,P} M_{i \leftarrow j} (x_j - x_i)] f, \quad (10)$$

where $t = t/(2N_{\text{ref}})$, t is the time in generations and N_{ref} is a reference (ancestral) effective population size. $M_{i \leftarrow j}$ is the scaled migration rate from population j to population i per generation. $\gamma_i = 2N_{\text{ref}} s_i$, where s_i is the relative selective fitness in population i. In order to estimate the demographic parameters, Q, we would like to estimate from the observed multipopulation joint frequency spectrum, $S[d_1, d_2, \ldots d_P]$. Assuming no linkage between polymorphisms, each entry in the joint frequency spectrum is an independent Poisson variable, with mean $M[d_1, d_2, \ldots d_P]$. The likelihood function can be written as follows:

$$L(\Theta|S) = \prod_{i=1,\ldots,P} \prod_{d_i=0,\ldots,n_i} \frac{e^{-M[d_1,d_2,\ldots,d_P]} M[d_1, d_2, \ldots, d_P]^{S[d_1,d_2,\ldots,d_P]}}{S[d_1, d_2, \ldots, d_P]!}. \quad (11)$$

In general, *dadi* calculates the expected allele frequency spectrum M under a specific demographic model by a diffusion approach. Then the demographic parameters Q can be estimated by maximizing the likelihood function. The demographic models maintain great flexibility and can be used to model complicated demographic scenarios among multiple populations. A summary of discussed methodology may be found in Table 3.

Conclusion

Comparing between existing estimators, a number of notable discrepancies arise. As discussed in the introduction, estimates of the strength and rate of RHH differ by orders of magnitude when applied to similar data sets. There are, however, 2 notable correlations between the different estimates: 1) small estimates of s seem to correspond to analyzing small region sizes (i.e., Andolfatto (2007) estimates $s = 10^{-5}$, when analyzing coding regions) and large estimates of s when analyzing large regions (i.e., Macpherson et al. (2007) estimates $s = 10^{-2}$, when analyzing 100-kb sliding windows) and 2) divergence-based approaches tend to estimate smaller s than polymorphism-based approaches. However, given that divergence-based methods may be counting the effect of the fixation of many weakly selected mutations over longer evolutionary time, whereas polymorphism-based methods may be most impacted by the recent fixation of strongly beneficial mutations (i.e., impacting large genomic regions), it is indeed possible that these estimates are not incompatible—but rather simply estimating different tails of the true underlying distribution of selection coefficients.

Additionally, outlier-based genome scans utilizing either SFS or LD approaches identify different genomic regions as targets of positive selection with inconsistent overlap (see Enard et al. 2010). For example, Williamson et al. (2005) and Voight et al. (2006) find a similar number of genomic regions experiencing selection (444 and 460, respectively) but only 41 of the regions overlap. Conversely, Carlson et al. (2005) use a Tajima's D-based approach and find 986 positively selection genomic regions, of which 217 are shared by Williamson et al. and 71 are shared with Voight et al. (2006).

And yet one commonality between approaches is the inability to adequately account for the demographic history of the population in question (see Table 1). Nonequilibrium models are well known to mimic patterns of positive selection in polymorphism data (e.g., see reviews of Nielsen 2005; Thornton et al. 2007), and there is accumulating evidence that they may also similarly impact divergence-based approaches (e.g., Andolfatto 2008) as well as cause them to be conservative (see Parsch et al. 2009). Although attempts at joint estimation have been made (e.g.,

Table 3 Summary of commonly used methodology to infer demography

Method	Inference procedure	Estimated demographic parameters	Nonneutral consideration?
Nielsen and Wakeley (2001)	MCMC	θ, M, T, N	No
Hey and Nielsen (2004)	MCMC	θ, M, T	No
Becquet and Przeworski (2007)	MCMC	θ, M, T	No
Hey (2010)	MCMC	θ, M, T	No
Thornton and Andolfatto (2006)	Approximate Bayesian	θ, ρ, t, f, d	No
Adams and Hudson (2004)	Maximum likelihood	T, t, f, d	No
Bazin et al. (2010)	Bayesian hierarchical	Flexible	Yes
Williamson et al. (2005)	Maximum likelihood	M, T, t, f, d	Yes
Gutenkunst et al. (2009)	Maximum likelihood	M, T, t, f, d	No

American Genetic Association
The Journal of Heredity
www.jhered.oxfordjournals.org
Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

Williamson et al. 2005; Li and Stephan 2006; Eyre-Walker and Keightley 2009), they are accomplished in a stepwise manner. Thus, the demographic model is likely to be overfit to the data, accounting for much of the selection signature in the genome. Similarly, in the absence of demographic estimation, selection models are likely to be biased toward higher rates and strengths of adaptation in an attempt to fit the diversity-reducing and frequency spectrum–skewing effects produced by the underlying population history.

Thus, the challenge to the field is clear—it is essential to develop an estimator capable of jointly inferring the action of both nonneutral and nonequilibrium models simultaneously. This will require at least 2 components: 1) the ability to identify patterns that distinguish selective from demographic effects. As discussed above, the most promising avenue in this regard seems to be patterns in LD that appear to be largely robust to demographic perturbation (e.g., Stephan et al. 2006; Jensen et al. 2007; Pavlidis et al. 2010), though false positives may arise in the presence of gene conversion (Jones and Wakeley 2008). Additionally, the combination of polymorphism- and divergence-based inference may be effectively used to estimate different tails of the true underlying distribution; and 2) a computational framework capable of handling whole genomes worth of data, a large number of summary statistics, and the accurate inference of multiple parameters of interest. Indeed, great progress has been made toward estimating increasingly complex (but neutral) demographic models within Kimura's diffusion framework (Williamson et al. 2005; Gutenkunst et al. 2009).

A good deal of recent work (e.g., Wegmann et al. 2009; Bazin et al. 2010) seems to suggest ABC-based approaches to be the most likely way forward for combining demographic and selective inference. This framework appears capable of handling the large number of summary statistics necessary for joint estimation of the parameters of interest (i.e., parameters of RHH, parameters of population size change, and parameters of subdivision with migration) while simultaneously being computationally efficient enough to handle the type of multiple whole-genome data sets that are currently being generated. Despite the difficulty of this problem, it is perhaps the central issue facing the field today—as understanding the relative roles of adaptive and nonadaptive processes in the evolution of natural populations is truly at the core of population genetics.

References

Adams AM, Hudson RR. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics. 168(3):1699–1712.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature. 437(7062):1149–1152.

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

Andolfatto P. 2008. Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. Genetics. 180(3):1767–1771.

Bazin E, Dawson KJ, Beaumont MA. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics. 185(2):587–602.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics. 162(4):2025–2035.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res. 17(10):1505–1519.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 356(9):519–520.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 140(2):783–796.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. 15(11):1553–1565.

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

Enard D, Depaulis F, Roest Crollius H. 2010. Human and non-human primate genomes share hotspots of positive selection. PLoS Genet. 6(2):e1000840.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26(9):2097–2108.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics. 155(3):1405–1413.

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

Gutenkunst R, Hernandez RD, Williamson S, Bustamante CD. 2009. Inferring the joint desmographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5(10):e1000695.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 57(1):97–109.

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

Hey J. 2010. Isolation with migration models for more than two populations. Mol Biol Evol. 27(4):905–920.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics. 167(2):747–760.

Jensen JD, Kim YH, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 170(3):1401–1410.

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

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(4):2371.

Jones DA, Wakeley J. 2008. The influence of gene conversion on linkage disequilibrium around a selective sweep. Genetics. 180(2):1251–1259.

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

Kelly JK. 1997. A test of neutrality based on interlocus associations. Genetics. 146(3):1197–1206.

American Genetic Association
The Journal of Heredity
www.jhered.oxfordjournals.org
Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012

Kim Y. 2006. Allele frequency distribution under recurrent selective sweeps. Genetics. 172(3):1967–1978.

Kim YH, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics. 167(3):1513.

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

Kimura M. 1955. Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci U S A. 41(3):144–150.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2(10):e166.

Lin K, Li H, Schlötterer C, Futschik A. 2011. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics. 187(1):229–244.

Loewe L, Charlesworth B, Bartolomé C, Nöel V. 2006. Estimating selection on nonsynonymous mutations. Genetics. 172(2):1079–1092.

Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics. 177(4):2083–2099.

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

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

McVean G. 2007. The structure of linkage disequilibrium around a selective sweep. Genetics. 175(3):1395–1406.

McVean GA, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics. 155(2):929–944.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. J Chem Phys. 21(6):1087–1092.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197–218.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics. 158(2):885–896.

Nielsen R, Williamson S, Kim YH, Hubisz MJ, Clark AG, Bustamante CD. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15(11):1566–1575.

Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. Mol Biol Evol. 26(3):691–698.

Pavlidis P, Jensen JD, Stephan W. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics. 185(3):907–922.

Polanski A, Kimmel M. 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 165(1):427–436.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics. 160(3):1179.

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

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

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the Drosophila genome? PLoS Genet. 5(6):e1000495.

Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 172(4):2647–2663.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123(3):585–595.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics. 172(3):1607–1619.

Thornton KR. 2009. Automating approximate Bayesian computation by local linear regression. BMC Genet. 10:35.

Thornton KR, Jensen JD, Becquet C, Andolfatto P. 2007. Progress and prospects in mapping recent selection in the genome. Heredity. 98(6):340–348.

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

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7(2):256–276.

Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics. 182(4):1207–1218.

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

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 102(22):7882–7887.

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

Wooding S, Rogers A. 2002. The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics. 161(4):1641–1650.

Received July 21, 2011; Revised October 7, 2011;
Accepted October 18, 2011

Corresponding Editor: Mohamed Noor

American Genetic Association
The Journal of Heredity
www.jhered.oxfordjournals.org
Downloaded from http://jhered.oxfordjournals.org/ at UniversitÃ© & EPFL Lausanne on May 31, 2012