A Population Genetic Hidden Markov Model for Detecting Genomic Regions Under Selection

Andrew D. Kern*,[1] and David Haussler[2,3,4]

[1]Department of Biological Sciences, Dartmouth College
[2]Department of Biomolecular Engineering, University of California, Santa Cruz
[3]Center for Biomolecular Science and Engineering, University of California, Santa Cruz
[4]Howard Hughes Medical Institute, University of California, Santa Cruz

***Corresponding author:** E-mail: andrew.d.kern@dartmouth.edu.

Associate editor: Carlos Bustamante

Abstract

Recently, hidden Markov models have been applied to numerous problems in genomics. Here, we introduce an explicit population genetics hidden Markov model (popGenHMM) that uses single nucleotide polymorphism (SNP) frequency data to identify genomic regions that have experienced recent selection. Our popGenHMM assumes that SNP frequencies are emitted independently following diffusion approximation expectations but that neighboring SNP frequencies are partially correlated by selective state. We give results from the training and application of our popGenHMM to a set of early release data from the Drosophila Population Genomics Project (dpgp.org) that consists of approximately 7.8 Mb of resequencing from 32 North American *Drosophila melanogaster* lines. These results demonstrate the potential utility of our model, making predictions based on the site frequency spectrum (SFS) for regions of the genome that represent selected elements.

Key words: population genomics, machine learning, HMM, selection.

Introduction

Genetic variation within and between species is jointly determined by the actions of selection, demography, and random drift. A major goal of population genetics has been to infer with what relative strength these processes act in natural populations using a combination of empirical studies and explicit theoretical models (Lewontin 1974; Kimura 1983; Gillespie 1991). Identifying the genomic targets of natural selection using molecular population genetic data is of particular interest as these regions are expected to represent those DNA elements that have been functionally relevant, at least historically, to the species under examination. Systematic identification of the targets of natural selection within our own human genome has now begun in earnest (e.g., Akey et al. 2002; Clark et al. 2003; Stajich and Hahn 2005; Voight et al. 2006), and accordingly, numerous methods have been proffered that aim to uncover the footprint of natural selection in molecular population genetic data (reviewed in Nielsen et al. 2007).

Natural selection changes patterns of DNA variation in both deterministic and stochastic ways (Gillespie 1991, 2000; Barton 2000). For example, selection against deleterious mutations leads to a relative excess of rare alleles within populations, whereas recurrent directional selection on beneficial mutations can lead to an excess of high-frequency alleles (Wright 1938; Sawyer and Hartl 1992; Gillespie 1993, 1994a, 1994b; Gillespie 1997). Thus, the distribution of allele frequencies (also know as the SFS) at a locus contains information about the selective forces at work at that locus. This intuition lead to classical population genetic tests such as Tajima's *D* (Tajima 1989) and its brethren (Fu and Li 1993; Fay and Wu 2000), which summarize information from the SFS to make statements about neutrality. Modern likelihood-based approaches to estimation and hypothesis testing of natural selection instead use all the information in the SFS (e.g., Bustamante et al. 2001; Williamson et al. 2004, 2005).

Recently, the use of hidden Markov models (HMMs) in computational molecular biology has expanded. Currently, HMMs are being applied broadly from partitioning of genomes with respect to base composition (Churchill 1989) to sequence profiling and multiple alignment of proteins (Haussler et al. 1993; Baldi et al. 1994) and even to phylogenetic inference (Yang 1995; Felsenstein and Churchill 1996; Siepel and Haussler 2004; Siepel et al. 2005). HMMs are a popular class of machine learning algorithms that have been used with great success to derive insights about "hidden" parameters underlying data (Baum and Petrie 1966). In the case of genomic sequence data, these hidden parameters could be the rate of evolution or the probability of a given nucleotide at a site. A natural extension of this class of models is to utilize population genetics theory of allele frequency distributions in an HMM framework to map variation in underlying population genetic parameters over a stretch of DNA sequence or over a genome. These models take various forms but generally can be called population genetic hidden Markov models (popGenHMMs) (e.g., Hobolth et al. 2007; Boitard et al. 2009). A number of models, such as STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) and FRAPPE (Tang et al. 2005), incorporate popGenHMM approaches for inferring genomic regions of differing ancestry (see Bryc et al. 2010 to a recent application). In these example, ancestry is a hidden parameter, and genotype frequencies are emitted conditional on those unobserved states. One problem plaguing such methods is dealing with background linkage disequilibrium that will induce conditional dependencies among sites. More recently, a Markov-Hidden-Markov model (MHMM) approach to this problem has been introduced (Tang et al. 2006) that partly overcomes the difficulty of linkage disequilibrium (LD) by accounting for first-order Markovian dependencies along a chromosome (see also Sankararaman et al. 2008).

Here, we develop and implement a fully probabilistic popGenHMM, which fuses the methodology of HMMs to the probabilistic framework of population genetics. Our aim was to introduce a general class of models for detecting genomic regions under the influence of natural selection using population genetic data (i.e., single nucleotide polymorphism (SNP) frequency information). We develop a m-state model, where states are either selected or neutral, and derived allele frequencies are emitted from each state with probabilities determined by sampling from the stationary distribution of allele frequencies as determined via diffusion approximations (Wright 1969; Sawyer and Hartl 1992). By using the standard tools for HMM parameter estimation and decoding (Viterbi 1967; Baum et al. 1970), one can effectively color a sequence partitioning sites into the neutral and selected states that our popGenHMM specifies.

The Model

Formal Definition

Formally, we define an m-state popGenHMM $\Phi = (\boldsymbol{S}, \boldsymbol{\Psi}, \boldsymbol{A}, \boldsymbol{b})$ to be composed of the set of states $\boldsymbol{S} = \{s_1, \ldots, s_m\}$, their corresponding population genetic models $\boldsymbol{\Psi} = \{\psi_1, \ldots, \psi_m\}$ from which emissions are calculated, a matrix of state transition rates $\boldsymbol{A} = \{a_{i,j}\}$ for $(1 \leqslant i,j \leqslant m)$, and a vector of initial state probabilities, $\boldsymbol{b} = (b_1, \ldots, b_m)$. Specifically, for all k, $a_{i,j}(1 \leqslant i,j \leqslant m)$ is the conditional probability of visiting state j at site k given that state i was visited at site $k - 1$ and $b_i(1 \leqslant i \leqslant m)$ is the probability of the first state visited is state i. These parameters satisfy $\sum_i b_i = 1$ and for all i, $\sum_j a_{i,j} = 1$. We consider the data, $\boldsymbol{D}$, to be the matrix of alignment columns from a population sample of DNA of length L, where the ith alignment column is denoted as $\boldsymbol{D}_i(1 \leqslant i \leqslant L)$. In this manuscript, we will include a mutation parameter in our model, thus allowing the inclusion of monomorphic sites; however, it is worth noting that our model could be based on segregating sites alone.

Of central interest is the probability that an alignment column $\boldsymbol{D}_i$ is emitted by state s_j, which can be expressed as the conditional probability of $\boldsymbol{D}_i$ given the corresponding population genetic model, $\Pr(\boldsymbol{D}_i|\psi_j)$. Each of the population genetic models, ψ_j, are specified by diffusion approximations of population genetic models (Wright 1969; Sawyer and Hartl 1992) and their associated parameters. Here, we are primarily interested in detecting regions under selection, so we continue by defining popGenHMMs whose state spaces differ according to selective regime. In practice, we might be interested in other state spaces, such as regions of high mutation or recombination; popGenHMMs can readily be developed for these cases as well.

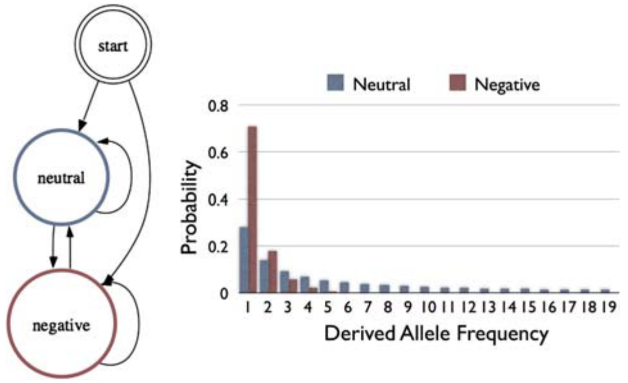

FIG. 1. A two-state popGenHMM. To the left, a graphical representation of the popGenHMM is shown with states depicted by nodes and transitions among states shown with the unlabeled edges. As the model is Markovian, the sum of all transition probabilities exiting a node sum to 1. To the right, a histogram representing the expected SFS from each of the states of the model is given, assuming a sample size of $n = 20$. Two states are shown, a neutral state (blue) that emits allele frequencies based on the neutral SFS shown to the right and a state labeled negative (red) that represents a selected state. In this case, selection is negative ($\alpha = -10$) and the corresponding SFS for emissions is shown to the right. Note that for a two-state popGenHMM, the selected state need not be negative.

A two-state popGenHMM, with a state representing sites under selection (directional or purifying), and a state representing neutral sites are shown in figure 1. Assuming an infinite site Wright–Fisher model, with no linkage among sites, and no interference among mutations, the stationary distribution of the frequency p of a newly arisen mutation under selection can be expressed as follows:,

$$\phi(p|\alpha,\theta) = \frac{1 - e^{-\alpha(1-p)}}{1 - e^{-\alpha}} \frac{2\theta}{p(1-p)}, \tag{1}$$

where $\alpha = 2Ns$ is the product of the haploid effective population size, N, and the selection coefficient of the new mutation, s, and $\theta = 4Nu$ is mutation rate (Fisher 1930; Wright 1969). Under neutrality, the analogous limiting distribution for allele frequency, p, is as follows:

$$\phi(p|\alpha = 0, \theta) = \frac{2\theta}{p} \tag{2}$$

(Kimura 1971; Sawyer and Hartl 1992).

To model the observed data, we add a layer of binomial sampling to express the probability of observing derived alleles segregating in i out of n individuals $(1 \leqslant i < n)$. Let D be the random variable denoting the number of derived alleles at a site, where $D \in 0, 1, \ldots, n$. Then,

$$Pr(D = i|n, \alpha, \theta) = \int_0^1 \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \phi(p|\alpha,\theta) dp. \tag{3}$$

For simplicity, we model the "folded" combination of $D = 0$ and $D = n$ as a single class of monomorphic sites, so that the probability of being monomorphic, $\Pr(\text{mono}|n,\alpha,\theta)$, is

$$\begin{aligned}\Pr(\text{mono}|n,\alpha,\theta)\\ &= \Pr(D=0|n,\alpha,\theta)+\Pr(D=n|n,\alpha,\theta) \quad (4)\\ &= 1-\sum_{j=1}^{n-1}\Pr(D=j|n,\alpha,\theta) \quad (5)\end{aligned}$$

assuming the mutation parameter $\theta \ll 1$ as it is when expressed on a per site basis, so that the quantity (5) is positive (Sawyer and Hartl 1992; Kim and Stephan 2002). Unfolding $D = 0$ and $D = n$ into separate classes using outgroup information would surely add power but would require the inclusion of divergence parameters. Thus, the counts of derived alleles (D) act as sufficient statistics for the alignment columns ($\boldsymbol{D}_i$) emitted by our HMM, and we use these probabilities as surrogate emission probabilities as our popGenHMM goes from site to site. So for simplicity, even the alignment data $\boldsymbol{D}$, we define the probability of the ith alignment column, $\boldsymbol{D}_i$, which is segregating j nucleotides out of n, being emitted by the kth state, ψ_k, by

$$\Pr(\boldsymbol{D}_i|\psi_k) = \Pr(D=j|n,\alpha_{\psi_k},\theta), \quad (6)$$

where α_{ψ_k} refers to the selection coefficient associated with the kth state, the only parameter that differs between states in the current model.

To deal with missing data or gaps, we model the correlation between states to be a decreasing function of physical distance between two sites. Let $\boldsymbol{A}$ represent the transition matrix for state transitions in some n-state popGenHMM. If two sites are d nucleotides apart from one another, we then use the transition matrix power $\boldsymbol{A}^d$ to represent state transition probabilities between those two sites.

At this point, it is worth noting that we are assuming that sites emit allele frequencies independently conditional upon their underlying state but that underlying states themselves have an underlying correlation structure induced by HMM. This is a difficult assumption, perhaps first most because we have not dealt with LD present in genomes. This is perhaps even more problematic in light of the canonical selective sweep model, which is expected to create correlations in heterozygosity near sites that have recently swept to fixation. Although our model is currently far from exhaustive in the way it models the many genomic correlations present in real data, we show below that our HMM structure is an effective method with which to discover selection in genomes.

Of utmost interest is the most likely path of states through our popGenHMM. Let a path be defined as a sequence of states $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_L)$ such that $\pi_i \in 1,\ldots,m$, for $1 \leqslant i \leqslant L$. The joint probability of an observed data set $\boldsymbol{D}$ and a state sequence $\boldsymbol{\pi}$ conditional on our popGenHMM

$$\begin{aligned}\Pr(\boldsymbol{\pi},\boldsymbol{D}|\Phi) &= b_{\pi_1}\Pr(\boldsymbol{D}_1|\psi_{\pi_1})\prod_{i=2}^{L}a_{\pi_{i-1},\pi_i}\Pr(\boldsymbol{D}_i|\psi_{\pi_i})\\ &= \prod_{i=1}^{L}\Pr(\boldsymbol{D}_i|\pi_i,\Phi)\Pr(\pi_i|\pi_{i-1},\Phi), \quad (7)\end{aligned}$$

where $\Pr(\pi_1|\pi_0,\Phi) = b_{\pi_1}$, $\Pr(\pi_i|\pi_{i-1},\Phi) = a_{\pi_{i-1},\pi_i}$ for $i \geqslant 2$, and $\Pr(\boldsymbol{D}_i|\pi_i,\Phi) = \Pr(\boldsymbol{D}_i|\psi_{\pi_i})$. The sum over all possible paths through the data provides the likelihood function, $\Pr(\boldsymbol{D}|\Phi) = \sum_\pi \Pr(\pi,\boldsymbol{D}|\Phi)$. The maximum-likelihood path $\hat{\pi}$ is simply

$$\hat{\pi} = \text{argmax}_\pi \Pr(\pi,\boldsymbol{D}|\Phi). \quad (8)$$

These two quantities can be computed used two well-known dynamic programming algorithms, the forward algorithm and the Viterbi algorithm, respectively. Using the backward analog to the forward equation, posterior probabilities of each observation being produced by a given state can be readily calculated using a forward–backward algorithm. Additionally, each of the parameters of the popGenHMM can be estimated using the Baum–Welch algorithm, and modifications thereof, which is a special case of an expectation–maximization (EM) algorithm (Baum and Petrie 1966; Baum et al. 1970; Durbin et al. 1998).

Implementation

To estimate parameters from our model, we use a version of the EM algorithm that is closely related to the Baum–Welch algorithm. During EM, the maximum of the likelihood function is found by iteratively maximizing the so-called estimated complete log-likelihood function. The algorithm proceeds in alternate rounds of (E) expectation and (M) maximization steps, which are guaranteed to converge to a local maximum of the likelihood function (Dempster et al. 1977; Durbin et al. 1998). We define the complete log likelihood to be the log of the joint probability of an observed data set and a specific path through the popGenHMM as

$$\begin{aligned}l(\Phi|\boldsymbol{D},\boldsymbol{\pi}) &= \log(\Pr(\boldsymbol{\pi},\boldsymbol{D}|\Phi))\\ &= \log\prod_{i=1}^{L}\Pr(\boldsymbol{D}_i|\pi_i,\Phi)\Pr(\pi_i|\pi_{i-1},\Phi)\\ &= \sum_{d\in D,\pi\in\boldsymbol{\Psi}}u_{d,\pi}\log(\Pr(d|\pi,\Phi))\\ &\quad+\sum_{\pi,\pi'\in\boldsymbol{\Psi}}v_{\pi,\pi'}\log(\Pr(\pi'|\pi,\Phi)),\end{aligned}$$

where D is the set of distinct combinations of derived allele frequencies at given sample sizes within $\boldsymbol{D}$, the ds are individual instances of D (i.e., a derived allele frequency j, at sample size n), $u_{d,\pi}$ is the count of how many times d is emitted by a state π, and $v_{\pi,\pi'}$ is the count of transitions from state π to π'. These counts, $u_{d,\pi}$ and $v_{\pi,\pi'}$, are sufficient statistics for $l(\Phi|\boldsymbol{D},\boldsymbol{\pi})$ under the independent and identically distributed model. The expected complete log likelihood is the expectation of $l(\Phi|\boldsymbol{D},\boldsymbol{\pi})$ over all possible paths, conditioned upon the data and a particular instance of the parameter values of our model Φ^t. Let $U_{d,\pi}$ and $V_{\pi,\pi'}$ be random variables representing the counts $u_{d,\pi}$ and $v_{\pi,\pi'}$. We can then formalize the expected complete log likelihood as

$$
\begin{aligned}
&E_{\Phi^t}\{l(\Phi|\mathbf{D},\boldsymbol{\pi})\}\\
&\quad= E_{\Phi^t}\left\{\sum_{d\in D,\pi\in\boldsymbol{\Psi}} U_{d,\pi}\log(\Pr(d|\pi,\Phi))\right.\\
&\qquad\left.+\sum_{\pi,\pi'\in\boldsymbol{\Psi}} V_{\pi,\pi'}\log(\Pr(\pi'|\pi,\Phi))\right\} \qquad (9)\\
&\quad= \sum_{d\in D,\pi\in\boldsymbol{\Psi}} E_{\Phi^t}\{U_{d,\pi}\}\log(\Pr(d|\pi,\Phi))\\
&\qquad+\sum_{\pi,\pi'\in\boldsymbol{\Psi}} E_{\Phi^t}\{V_{\pi,\pi'}\}\log(\Pr(\pi'|\pi,\Phi)). \qquad (10)
\end{aligned}
$$

During the E step of the EM algorithm, we compute the expected counts, $E_{\Phi(t)}\{U_{d,\pi}\}$ and $E_{\Phi(t)}\{V_{\pi_1,\pi_2}\}$, using a forward/backward technique as is standard for HMMs (Durbin et al. 1998), but with emission probabilities calculated from the population genetic diffusion approximations given above. In the M step, a new parameter set for the model, $\Phi^{(t+1)}$, is found by maximizing the expected complete log likelihood. This maximization can be seen to factor into two terms corresponding to the first and second terms of equation (11). The first term is a function of the population genetic parameters associated with our states of interest (e.g., θ, α), whereas the second term is a function of the transition matrix $\boldsymbol{A}$ associated with our HMM. Maximization of the second term is a straightforward problem that can been solved by the canonical Baum–Welch algorithm (Durbin et al. 1998). Maximization of the first term is a bit trickier. The expression to be maximized is a weighted combination of the population genetic models, with weights determined by the expected counts of each state. We rely here on numerical optimization of this term, which we do using either the simplex method or a broyden-fletcher-goldfarb-shanno quasi-Newton algorithm (Press et al. 1992).

After training of the popGenHMM, genomic intervals under selection are then predicted using posterior decoding. These elements are then scored using a log-odds ratio of being in the specified state over a given genomic interval versus being in any other state from that specified over that same genomic interval. Let $\boldsymbol{\Psi}_k$ equal the complete set of states of a popGenHMM excluding the kth state ψ_k. Then the log-odds score of an element being in the state ψ_k from position i to position j is

$$
s_{ij} = \log\frac{\Pr(D_i,\ldots,D_j|\psi_k)}{\Pr(D_i,\ldots,D_j|\boldsymbol{\Psi}_k)}. \qquad (11)
$$

These scores can readily be calculated using the forward algorithm and a reduced state-space version of the popGenHMM. These scores can be readily converted to a likelihood ratio statistic, which is chi-square distributed with degrees of freedom specified by the difference in the number of free parameters between the full and reduced state-space version of the popGenHMM. This would allow for the significance of an individual element to be assessed if the distribution of scores did indeed follow the expected distribution. As shown below, this is not the case, thus we use the composite likelihood ratio testing approach to assess significance as has become popular in the literature (e.g., Kim and Stephan 2002; Zhu and Bustamante 2005). Thus, to assess whether an individual element is indeed significant simulations must be conducted to explore what the null distribution of element scores might be (see below). All code was written in C, making use of many of the methods available in the Gnu Scientific Library (http://www.gnu.org/software/gsl/).

Table 1. Sensitivity of the popGenHMM to Violations of Independence among Sites.

4*Nr*	popGenHMM	PRF
1	0.20	0.33
5	0.16	0.26
10	0.15	0.22
25	0.11	0.13
50	0.09	0.09
75	0.08	0.09
100	0.08	0.08
200	0.06	0.06
500	0.05	0.05
1,000	0.05	0.05

NOTE.—Here, we show the proportion of simulations that show a significant element at the 5% level when data are generated under a SNM with varying levels of recombination. Results labeled "popGenGMM" are from a two-state HMM; those labeled "poisson random field (PRF)" are from Bustamante et al. (2001) LRT shown for comparison.

Application to Simulated Data

To examine the performance of our popGenHMM, we trained it on simulated data, both to assess false-positive rates under neutral and selected models of evolution. Our first concern was that the hidden Markov approach would be sensitive to low levels of recombination, as a priori we make the assumption of independence among sites. It is well known and intuitive that the assumption of independence is nonconservative when applied to data, which is in truth linked. Table 1 shows the percentage of false positives, as defined by an element scored below the $P = 0.05$ critical value assuming that converted scores (i.e., $-2 \times s_{ij}$) are chi-square distributed. To assess the robustness of our popGenHMM to deviations from the assumption of independence among sites, we generated coalescent simulations using Hudson's coalescent simulation software MS (Hudson 2002), assuming a fixed value of $\theta = 30$, but varying levels of recombination $\rho = 4Nr$ under the standard neutral model (SNM). A total of 10^5 simulations were performed for each combination of parameter values and the popGenHMM was both trained and decoded for each simulation replicate. For comparison, we implemented the likelihood ratio test (LRT) of Bustamante et al. (2001) based on the Poisson random field model of Sawyer and Hartl (1992), which should provide a fair comparison as the models are intricately connected. False-positive rates are shown in table 1. Generally, we can see that the model is very sensitive to violations of the site independence assumption;

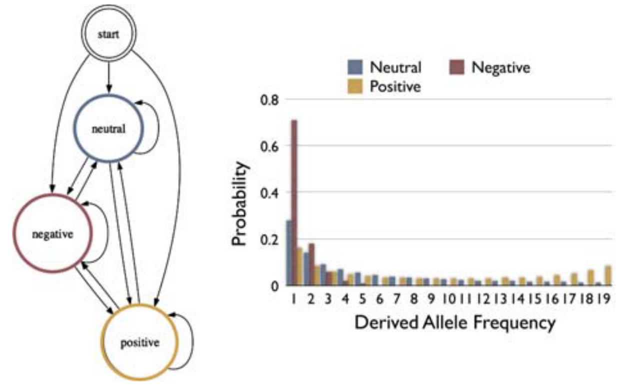

FIG. 2. A three-state popGenHMM. To the left, a graphical representation of the popGenHMM is shown with states depicted by nodes and transitions among states shown with the unlabeled edges. See caption of figure 1 for details. To the right, a histogram representing the expected SFS from each of the states of the model is given, assuming a sample size of $n = 20$. Three states are shown, a neutral state (blue), a selected state labeled negative (red) ($\alpha = -10$), and a second selected state labeled positive (yellow) ($\alpha = 10$).

however, it performs slightly better than the LRT proposed by Bustamante et al. (2001). This sensitivity is to be expected (e.g., Bustamante et al. 2001; Zhu and Bustamante 2005), thus to account for this, we proceed by assessing significance of an element by generated a null distribution of element scores through coalescent simulations. This composite likelihood approach has been used in multiple contexts at this point (Kim and Stephan 2002; Zhu and Bustamante 2005) and should restore the appropriate significance cutoffs if an appropriate null distribution is generated by simulation.

Robustness to Demography

To examine the effect of nonstationary demography on the robustness of our popGenHMM, we performed coalescent simulations that included a bottleneck of varying severity and timing in the history of the sample. The bottleneck model is especially relevant to data from both *Drosophila* (Thornton and Andolfatto 2006) and humans (Marth et al. 2004; Stajich and Hahn 2005) and thus is of great interest to potential application of our model. The bottleneck held constant the length of time at which the population remained at its reduced size ($\tau = 0.015$ units of 4N generations) but varied the severity, $f = [0.05, 0.1]$, and the timing of the bottleneck, $t = [0.005, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5]$. Critical values of the popGenHMM were generated by simulating data under the SNM using θ estimated from the number of segregating sites and the true value of ρ. For comparison, we calculated the composite likelihood ratio test (CLRT) of Kim and Stephan (2002), which has been shown previously to be sensitive to deviations from the assumption of stationarity (Jensen et al. 2005) using the software package CLSW from Yuseob Kim. To derive P values for the CLRT, we also generated null distributions using coalescent simulations of the SNM with θ estimated from the number of segregating sites and the true value of ρ. It is clear that the popGenHMM in its current incarnation is extremely sensitive to population bottlenecks and yields spurious evidence of selection over the entire range of parameters examined here. Figure 3 displays these results graphically.

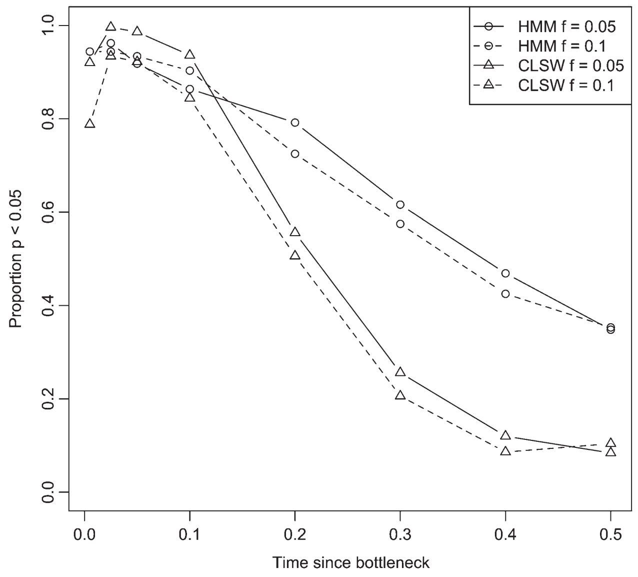

FIG. 3. Proportion of bottleneck simulations rejecting neutrality. Shown is a comparison of the false-positive rate of the popGenHMM and the CLRT of Kim and Stephan (2002). Two severities of the strength of the bottleneck are shown $f = 0.05$ and $f = 0.1$ for each model.

The sensitivity of the popGenHMM to bottlenecks stems from the two problems implicit in our approach. The first is that we are not dealing with the effects of demography on the SFS. This could overcome by modeling the effects of demography directly into the SFS and in so doing to jointly estimate demographic and selection parameters from the data (e.g., Williamson et al. 2005; Boyko et al. 2008). We are currently implementing such a model and it will be the topic of a future study. The second problem is that bottlenecks create a tremendous variance among regions that exceeds that expected under the SNM with recombination. To deal with this issue, one needs to perform simulations under realistic demographic parameters to arrive at an accurate null distribution. This is a reasonable approach as many authors have set out to estimate demographic parameters from multiple populations of both *Drosophila* and humans (e.g., Marth et al. 2004; Thornton and Andolfatto 2006). This is the approach that we will take in our application to data below.

Power to Detect Selection

Although the popGenHMM in its current form is not robust to demographic deviations, it remains to be seen how well it can detect selection operating on a genomic regions. To examine this, we benchmarked the performance of the model by performing simulations under three different selection regimes: 1) a single selective sweep that has occurred in the recent evolutionary history of a region (e.g., Kim and

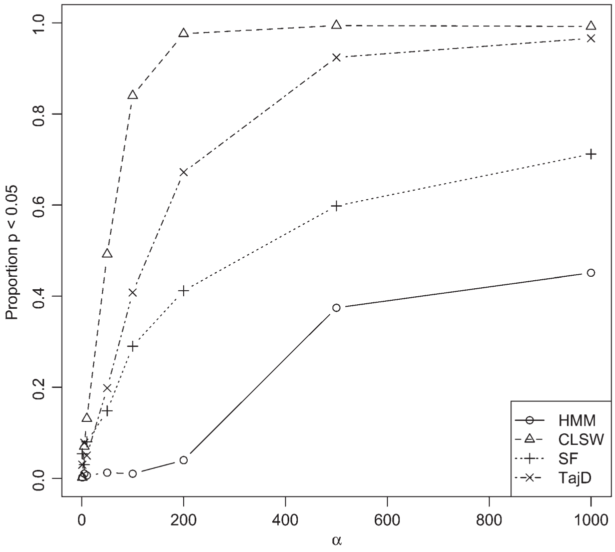

FIG. 4. Power to detect a single selective sweep. Shown is a comparison of the power of our popGenHMM as a function of the strength of selection, $\alpha = 2Ns$, in comparison with CLSW (Kim and Stephan 2002), SweepFinder (Nielsen et al. 2005), and a sliding window implementation of Tajima's D (Tajima 1989). Each point consists of 1,000 coalescent simulations with a single selective sweep (stochastic trajectory) that has finished its sojourn through the population just before the current generation ($\tau = 0$). We simulated 20 kb from samples of size $n = 50$ with $\theta/\text{bp} = 0.01$ and $\rho/\text{bp} = 0.025$.

Stephan 2002), 2) recurrent positive directional selection (the normal shift model—Gillespie 1997), and 3) recurrent negative selection (the exponential shift model—Gillespie 1994a).

Selective Sweeps

To model single selective sweeps, we implemented a coalescent model with the inclusion of a sweep in the history of a sample. Our implementation is similar in spirit to that of Kim and Stephan (2002), allowing for a single site under selection at an arbitrary location in the middle of the sample along with intragenic recombination, but during the sweep phase of the simulation, we chose to use the rejection algorithm of Braverman et al. (1995). In addition, rather than use deterministic trajectories of the selected mutation as in Kim and Stephan (2002) and Braverman et al. (1995), we have implemented stochastic trajectories according to the procedure of Przeworski et al. (2005). The accuracy of these simulations was checked first using deterministic trajectories against software available from Y. Kim (Kim and Stephan 2002) and under stochastic trajectories against software provided to us by K. Thornton (personal communication).

For each parameter set, we generated 1,000 samples of size $n = 50$ of 20 kb with $\theta/\text{bp} = 0.01$ and $\rho/\text{bp} = 0.025$. The selected site was simulated in the exact middle of the locus and the sweep was simulated such that it ended right before the sample was taken (i.e., $\tau = 0$). As before, critical values of the popGenHMM were generated by simulating data under the SNM using θ estimated from the number of segregating sites and the true value of ρ. For comparison, we also analyzed our simulated sweep data using CLSW (Kim and Stephan 2002) and SweepFinder (Nielsen et al. 2005), which is similar to CLSW but does not assume stationarity of the population, as well as with a sliding window version of Tajima's D (Tajima 1989). In each case, test statistics were compared with a null distribution we generated as above to achieve a proper critical value for testing. Both SweepFinder and our sliding window implementation of Tajima's D require a choice of window size of the number of SNPs—ours was chosen in such a way as to maintain windows that were approximately equal to 10% of the data, thus allowing for at least ten independent observations. This is motivated by empirical realities in which one does not have a way to choose window size to maximize power a priori.

Figure 4 shows the power of the popGenHMM to detect such a sweep and compares its power to other methods. The popGenHMM lags behind all other methods examined here in its power to detect a single historical sweep. This is perhaps not surprising—we have not explicitly modeled the effect of sweeps on the spatial patterns of linked neutral polymorphism as do the CLSW and SweepFinder methods. Instead, we have used a model that only considers the actual sites under selection. For the parameter space examined, CLSW has the greatest power followed by the sliding window version of Tajima's D. We had expected the SweepFinder method to outperform Tajima's D, given that SweepFinder is aimed specifically to detect the spatial pattern around selective sweeps. However, we did not attempt to maximize the power of the test as in Nielsen et al. (2005). It seems that our popGenHMM in its current form will only have power to detect the strongest single sweeps in a genome.

Normal Shift Model

To examine the power of our method to detect recurrent positive directional selection, we simulated samples drawn from the normal shift model (Gillespie 1997). The normal shift model assumes that the fitness of new mutations is the sum of its parent allele's fitness and a normally distributed random variable with mean 0 and variance σ^2. The strength of selection in the normal shift model is measured as $\alpha = 2N\sigma$. The normal shift model is similar to the recurrent sweep model of Braverman et al. (1995) except it allows for selected sites to interfere with one another (i.e., so-called clonal interference). The model cannot be simulated readily in the coalescent framework, thus we performed forward population genetic simulations to draw samples from populations undergoing this type of selection. The details of our forward simulations can be found in Kern et al. (2004) and Kern (2009), but briefly, we model a population consisting of alleles that undergo selection, drift, mutation, and recombination in that order. Populations are allowed to burn-in for $40N$ generations before sampling begins and then samples are taken each $10N$ generations. Such spacing between sampling provides independence among our samples, and this has been confirmed by examining the auto-correlation in summary statistics drawn from the evolving population.

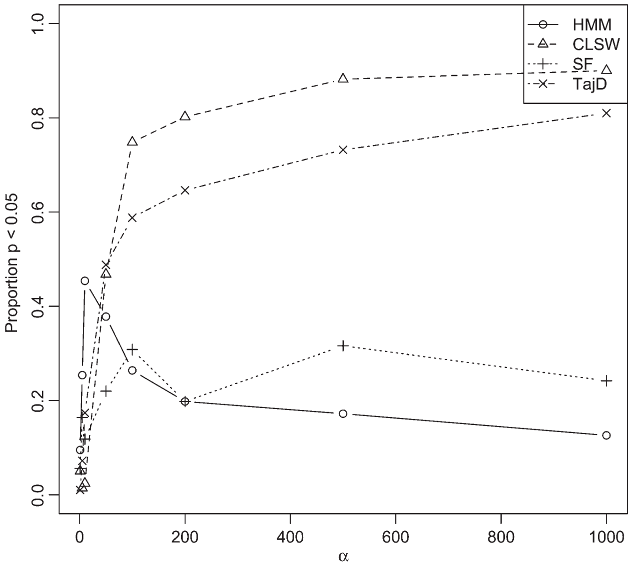

FIG. 5. Power to detect a locus undergoing recurrent directional selection. Shown is a comparison of the power of our popGenHMM as a function of the strength of selection, $\alpha = 2N\sigma$, in comparison with CLSW (Kim and Stephan 2002), SweepFinder (Nielsen et al. 2005), and a sliding window implementation of Tajima's D (Tajima 1989) to detect selection on a locus evolving according to the normal shift model. Each point consists of 1,000 samples drawn from forward population genetic simulations in which we simulated 20 kb from samples of size $n = 50$ with $\theta/\text{bp} = 0.01$ and $\rho/\text{bp} = 0.025$. Selected sites occur in the middle fifth of the simulated locus (bases 8,000–12,000) and θ is constant across the entire region.

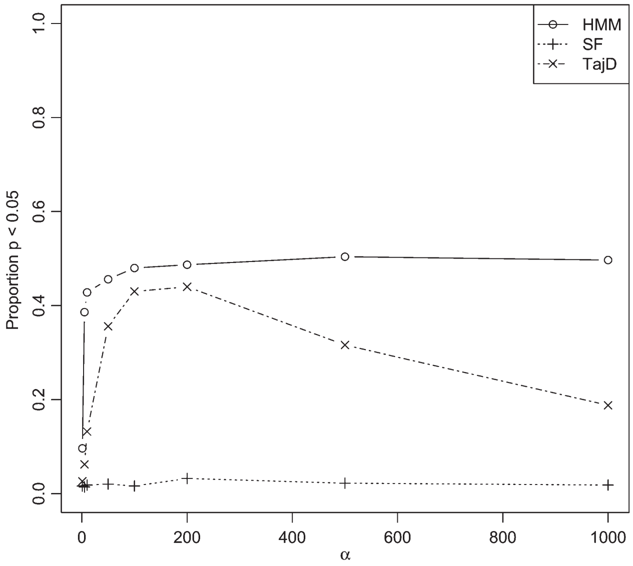

FIG. 6. Power to detect a locus undergoing recurrent negative selection. Shown is a comparison of the power of our popGenHMM as a function of the strength of selection, $\alpha = 2N\sigma$, SweepFinder (Nielsen et al. 2005) and a sliding window implementation of Tajima's D (Tajima 1989) to detect selection on a locus evolving according to the exponential shift model. Each point consists of 1,000 samples drawn from forward population genetic simulations in which we simulated 20 kb from samples of size $n = 50$ with $\theta/\text{bp} = 0.01$ and $\rho/\text{bp} = 0.025$. Selected sites occur in the middle fifth of the simulated locus (bases 8,000–12,000) and θ is constant across the entire region.

The accuracy of our implementation has been validated against expectations of the substitution process in a no-recombination setting (Gillespie J, personal communication), and the accuracy of our recombination routine has been error checked against results obtained from coalescent simulation.

Figure 5 presents power of the popGenHMM in comparison with other methods in samples drawn from the normal shift model under various strengths of selection. For each data point in that figure, we simulated 1,000 samples of sample size $n = 50$ consisting of 20 kb with $\theta/\text{bp} = 0.01$ and $\rho/\text{bp} = 0.025$. In the middle of the locus, from bases 8,000–12,000, there is an embedded locus under selection. We then carried out the various statistical procedures as above to attempt to detect selection acting at that embedded locus.

As can be seen, the popGenHMM performs best under relatively weak normal shift selection and then power drops off from there. This performance is the result of very strong recurrent selection affecting the locus as a whole, thus although the HMM did find evidence of switching to the selected state, it was increasingly outside of the actual selected locus. This phenomenon seems to be hurting the SweepFinder method as well, which had worse performance than the popGenHMM under weaker selection, but began to do slightly better as α increased. Conversely CLSW and Tajimas D performed well over the parameter space examined, although under moderate selection ($\alpha < 20$), the popGenHMM performed best.

Exponential Shift Model

One feature of our popGenHMM that is unique is our ability to explicitly scan population genetic data for areas undergoing "negative" or purifying selection. To examine the performance of our popGenHMM to detect such regions of the genome, we simulated samples drawn from the exponential shift model (Gillespie 1994a). As in the normal shift model, under the exponential shift model, new mutations receive a selection coefficient that is the sum of the parental allele's fitness and a random variable—in this case, fitness of the parental allele is decreased by an exponential deviate with mean σ. The strength of selection under the exponential shift model is measured as in the normal shift model, $\alpha = 2N\sigma$. Also as in the normal shift model we use forward simulations to draw samples from this model we parameters and simulation conditions identical to those described above.

Figure 6 presents power of the popGenHMM to detect selection in a region undergoing exponential shift selection. For comparison, we present power from the sliding window version of Tajima's D and from SweepFinder. Under the entire parameter space examined, our popGenHMM has greater power than either of the other methods examined. Although Tajima's D does have good power to detect a region of the genome undergoing moderately strong negative selection, its power decreases as the strength of selection increases past $\alpha > 200$. SweepFinder has nearly no ability to detect a locus undergoing negative selection, which is

Table 2. popGenHMM Model Comparison.

Chrom	States	LogLik	Params.	Obs	BIC
2L	2	−402,659.796	6	2,884,699	805,408.842
	3	−401,712.001	13	2,884,699	803,617.375
X	2	−253,841.601	6	2,567,728	507,771.752
	3	−252,722.130	13	2,567,728	505,636.121

NOTE.—Chrom is the chromosome for the given region of data. States is the number of states in the popGenHMM trained. LogLik is the final trained log likelihood of the popGenHMM. Params. is the number of free parameters in the model. Obs is the number of observations used for training. BIC is the Bayesian information criterion. Models with lower values of BIC are preferred. For both the chromosomes X and 2L, the three-state popGenHMM fits the data significantly better.

of course not surprising—the SweepFinder method is based explicitly on the expected patterns of spatial variation surrounding a sweep. Our popGenHMM thus seems to be a promising method for scanning the genome for negative selection using population genetic data exclusively.

Application to Empirical Data

To test our implementation of the popGenHMM, we applied it to a large-scale resequencing data set from *Drosophila*. The data set used is an early release (release 0.5) data set from the *Drosophila* Population Genomics Project (http://www.dpgp.org/) consisting of 7 Mb sequenced from 32 inbred lines of *Drosophila melanogaster* from North Carolina. These data were collected from chromosomes 2L and X and resequencing was performed on a high-density oligonucleotide microarray-based resequencing platform (see http://www.dpgp.org/ for details and data used).

Two- and three-state popGenHMMs (figs. 1 and 2) were trained via EM on 1 Mb (maximum) chunks of the data. This was done for efficiency, as training on complete chromosomal (approximately 3.5 Mb) regions was computationally untenable. Once maximum-likelihood estimates of parameters were found for each 1-Mb region, weighted averages of the parameter estimates were used to find approximate maximum-likelihood estimates for parameters at the chromosomal level. We keep chromosomal parameter estimates distinct as we expect a priori that the X chromosome of *Drosophila* should have different parameters of evolution than autosomal regions of the genome due to different effective population sizes and properties of selection. Parameter estimates from the trained two-state and three-state popGenHMMs are shown in table 3.

One immediate result is that our estimates of the strength of selection, independent of the size of our popGenHMM, suggest that the strength of selection on the X chromosome is considerably stronger than that on chromosome 2L. This is not surprising given the hemizygous nature of *Drosophila* X chromosomes and its subsequent impact of selected mutations (e.g., Betancourt et al. 2004). To choose between two-state and three-state popGenHMMs, we employ a Bayesian information criterion (BIC) (Schwarz 1978), which accounts for both the likelihood of the model and the number of free parameters trained. Three-state models fit the data on both chromosomes significantly better using the BIC as can be seen in table 2, thus we only consider this model for the remainder of the present report.

Table 3. popGenHMM Parameter Estimates "States" Refers to the Number of States in the popGenHMM Trained.

States	Chrom	θ	α
2	X	0.01087	−104.994
2	2L	0.0074	−22.908
3	X	0.0041	−20.026, 2.779
3	2L	0.0043	−9.95, 0.927

NOTE.—Chrom is the chromsome from which the data come. θ is the mutation parameter shared between states in the model. α refers to the selection coefficient or set of selection coefficients for the selection parameters of the model.

Based on our parameter estimates, positively and negatively selected elements were then identified and scored based on posterior decoding of our trained popGenHMM. A total of 13,636 negatively selected elements and 3,956 positively selected elements (hereafter negative and positive elements) were identified after screening out regions of questionable data quality and repeat masking. These elements also represent those which surpass the critical value established by simulation under a neutral model with a single bottleneck in the past using the demographic model estimated by Thornton and Andolfatto (2006), thus these elements can be thought of as statistically significant. Both length distributions are geometric with means of 362 and 12 bps for negative and positive elements, respectively. There is a strong positive correlation between length of a predicted element and logarithm of the odds (LOD) score in both negative elements (Spearman's rank correlation: $\rho = 0.87, P < 2.2^{-16}$) and positive elements (Spearman's rank correlation: $\rho = 0.34, P < 2.2^{-16}$). This is to be expected, however, as LOD scores grow roughly linearly with length even when no signal is present. Clearly, there are large differences in the distributions of element lengths predicted between negative and positive elements. This is particularly clear when we examine the distribution of length normalized scores between selective classes (fig. 7), which shows the clear shift toward smaller elements in the positively selected predictions. The very short average length of positively elements is worrisome at first blush—we should expect relatively large regions to hitchhike along with positively selected mutations. However, our popGenHMM is not designed to pick up hitchhiking regions, but instead, it is meant to find the positively selected mutations themselves. This could perhaps explain the difference is element length distributions, but there is the formal possibility that positive elements are often simply the result of spurious switching between states for our HMM or be the result of incorrect polarization of derived alleles. To examine this possibility, we calculated Tajima's D in our predicted element regions to see if our predictions indeed represented enriched proportions of the empirical distributions as we should expect. For each element set, we generated randomized element sets that conditioned on the size of elements and the coverage in each of the regions to draw up a null distribution of Tajima's D.

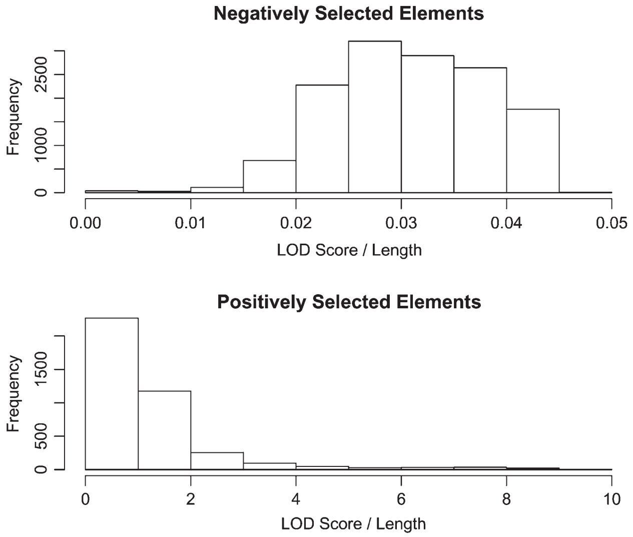

FIG. 7. Distribution of length normalized scores for elements. Shown are histograms of LOD scores/length (s) for each of the elements predicted. Negatively selected elements are shown in the top panel and positively selected elements in the bottom panel. The very different distributions is a function of the short lengths of positive elements predicted.

Mean observed Tajima's D in negative elements is $D = -0.75$ versus a mean of $D = -0.32$ for our randomized elements, and our observed mean value is more extreme than any of our 1,000 randomized mean values (i.e., $P < 0.001$). For the positive elements, mean observed Tajima's D is $D = 0.87$ versus a mean of $D = -0.11$ for randomized elements. Again, this is greater than each of the 1,000 randomized mean values ($P < 0.001$). These correlations with standard population genetic summary statistics imply that our popGenHMM model is recovering elements that we intend it to recover; however, the short lengths of positive elements is still troublesome, as on average, positive elements carry only 2.5 segregating sites per element. Thus, inference of their inclusion into the positively selected state is going to be based on only a few high frequency–derived alleles. Such elements would be susceptible to false prediction if ancestral inference caused a mispolarization of derived and ancestral states, thus we should take care not to over interpret our positive element set.

Given our set of elements, it is interesting to examine their coverage of the genome and the annotation sets that they overlap. Negative elements cover 72.1% of the sampled *Drosophila* genome and positive selected elements cover only 0.6% of the sample genome (recall our data set consists of ~7 Mb of the *D. melanogaster* genome or ~3.9% of the whole genome). Our negative element coverage is much higher than recent estimates from Siepel et al. (2005) based on cross-species comparisons rather than population data, which found 44.5% of the *Drosophila* genome to be evolutionarily conserved using a phylo-HMM framework (Siepel et al. 2005). Comparable estimates of positive selection are not currently available; however, a back of the envelope calculation is possible. Recently, it has been proposed that ~50% of noncoding divergence between *D. melanogaster* and *D. simulans* is adaptive (Andolfatto 2005). Assuming that approximately 90% of the *Drosophila* genome is noncoding (Adams et al. 2000), and furthermore assuming that average noncoding divergence along the *D. melanogaster* lineage because its most recent common ancestor with its sister species *D. simulans* is ~2% (Begun et al. 2007), then we expect ($0.9 \times 0.02 \times 0.5 =$) 0.9% of the genome to be under positive selection. In comparison with our estimate of 0.6%, this is considerably greater, but within an order of magnitude. Of course, the calculation is estimating the proportion of causal variants in the *Drosophila* genome (i.e., sites under selection), whereas our estimate should be composed of a mixture of causal and hitchhiking positions in the genome (i.e., both neutral and selected sites). That is to say, the actually number of causal variants in our set should be much lower, thus our estimated proportion might not be as close to divergence-based estimate as one would think at first blush. Although this is so, we note that the probability of fixation of a beneficial mutation is of course much larger than a neutral or deleterious mutation. This means that we should not necessarily expect the proportion of adaptive substitutions to give us an accurate estimate of the portion of the genome under positive selection at any given time.

It is of considerable interest to examine the coverage of known annotations by our element predictions. The 13,636 negatively selected elements cover 72.9% of the bases in coding regions, 75.8% of bases in 5'-untranslated regions (UTRs), and 74.2% of bases in 3'-UTRs. These coverages show highly significant enrichment for annotation coverage when compared with the expected coverage if the predicted elements were distributed randomly over the genomic regions considered conditioning on element sizes and the subset of genomic sequence examined (table 4). At the level of individual exons, this translates to roughly 90.0% of known protein-coding exons, 87.2% of 5'-UTR exons, and 88.1% of 3'-UTR exons are overlapped by a predicted negative elements. Inasmuch, coverage of known functional elements of the genome is very good for our predicted negatively selected elements. In addition to these protein-coding regions of the genome, predicted negative elements are also enriched in their coverage of transcribed regions of the genome (72.9% of transcribed bases covered) as well as regions known to be regulatory elements (86.4% of ORegAnno annotated base pairs; Montgomery et al. 2006).

Our set of predicted negative elements is also seen to be statistically depleted from regions of the *Drosophila* genome that are impoverished in known functional elements. In particular, negative elements are depleted from intergenic ($P = 0.002$) and from untranscribed regions ($P < 0.0001$). This is consistent with negative elements being predicted in functional regions of the genome.

The 3,956 positively selected elements have less consistent patterns with respect to their annotation enrichment. Positive elements cover 0.6% of the genome, corresponding to 0.55% of the bases in coding regions, 0.42% of bases in 5'-UTRs, and 0.53% of bases in 3'-UTRs. Positively selected elements are only significantly enriched for coding

Table 4. Negative Element Annotation Enrichment.

Annotation	Coverage (%)	*P*
CDS	72.9	<0.0001
5′-UTR	75.8	0.003996
3′-UTR	74.2	<0.0001
Messenger RNA	72.6	<0.0001
Transcribed	72.9	0.001998
ORegAnno	86.4	<0.0001

NOTE.—Shown are the percentages of basepairs covered by predicted negatively selected elements along with *P* values for coverage enrichment. *P* values are determined using a permutation procedure, whereby predicted elements are randomly assigned locations in the subset of the genome studied and coverage of a given annotation is recorded. A total of 10^4 permutations are performed to determine a null distribution of coverage. CDS, coding sequence.

Table 5. GO Enrichment for Genes Overlapped by Negative Elements.

GO term	*P*	Description
GO:0007456	3.63927×10^{-13}	Eye development (sensu Endopterygota)
GO:0007398	5.37182×10^{-13}	Ectoderm development
GO:0001745	1.06236×10^{-12}	Compound eye morphogenesis
GO:0046329	2.7023×10^{-12}	Negative regulation of JNK cascade
GO:0007391	2.7023×10^{-12}	Dorsal closure
GO:0005509	4.9945×10^{-11}	Calcium ion binding
GO:0003700	6.59308×10^{-11}	Transcription factor activity
GO:0007472	3.44757×10^{-10}	Wing disc morphogenesis
GO:0030528	9.43686×10^{-09}	Transcription regulator activity
GO:0045449	2.36004×10^{-08}	Regulation of transcription
GO:0042067	3.85541×10^{-08}	Establishment of ommatidial polarity
GO:0048100	4.41014×10^{-08}	Wing disc anterior/posterior pattern formation
GO:0009993	1.00948×10^{-07}	Oogenesis (sensu Insecta)
GO:0007476	2.49744×10^{-07}	Wing morphogenesis
GO:0046667	2.50434×10^{-07}	Retinal cell programmed cell death
GO:0007268	3.02455×10^{-07}	Synaptic transmission
GO:0048066	4.24759×10^{-07}	Pigmentation during development
GO:0016202	6.90402×10^{-06}	Regulation of striated muscle development
GO:0008632	6.90402×10^{-06}	Apoptotic program
GO:0007498	2.04508×10^{-05}	Mesoderm development

NOTE.—GO terms that are enrichments for negatively selected elements. Elements are assigned to individual genes if they occur within 50 kb of a given gene. *P* values are calculated using a binomial expectation of observing the observed number of elements surrounding a gene. The top 20 term enrichments are shown after Bonferonni correction.

sequence (CDS) bases ($P = 0.004$). At the level of individual exons, this translates to roughly 15.2% of known protein-coding exons, 6.6% of 5′-UTR exons, and 13.3% of 3′-UTR exons are overlapped by a predicted positive elements. Positive elements do not seem to correspond well to any particular genomic annotation, even though there is a slight enrichment in protein-coding regions. Of the annotations we looked at, positive elements had only one significant depletion. Positively selected elements are significantly depleted from predicted cross-species conserved phastCons elements (Siepel et al. 2005). This is what one would expect a priori; however, bear in mind that the current predictions are not based on divergence data between species, but instead on population genetic variation. Perhaps this speaks to an unrealistic modeling assumption in our popGenHMM. Namely, it is not clear that we should expect clustering of positively selected sites in the genome from a biological perspective. Although conserved (i.e., negatively selected) sites would be expected to cluster in functionally constrained stretches of sequence, it could be that runs of positively selected sites are rare and potentially only seen in a handful of regions under consistent diversifying selection.

Given the significant enrichment of genic regions among our negatively selected set of elements, we used the Gene Ontology (GO) database (Ashburner et al. 2000) to examine the molecular functions and biological processes associated with genes overlapping predicted negative elements in our data set. Although the number of genes in our data set is small, there are significant enrichments for about 20 GO terms (table 5) after correction for multiple tests. Interestingly, many of the GO terms enriched in the negatively selected set correspond to GO terms enriched for conserved regions of the genome (Siepel et al. 2005) often corresponding to processes involved in transcriptional regulation and development. The GO terms enriched for genes overlapping negative elements include "morphogenesis," "development," and "regulation of transcription." This is particularly promising given the sparsity of our data set, as a priori these are the sorts of genes that one expects to be under consistent negative selection.

Two examples of predicted elements as visualized in the UCSC genome browser are shown in figures 8 and 9. Figure 8 shows a representative predicted negative element. In this case, the element has an strong associated score (LOD = 33.1). Shown below the element track are tracks measuring divergence between *D. simulans* and *D. melanogaster* (labeled Div) and sequence polymorphism within *D. melanogaster* as measured by nucleotide diversity (Tajima 1983; labeled Pi). For both of these tracks darker colors represent greater amounts of divergence and polymorphism respectively. The predicted negative element can be seen to overlap regions with reduced polymorphism within species and reduced polymorphism and divergence between species. Whereas the 5′ portion of the predicted element overlaps the gene CG4691, surprisingly the 3′ portion of the prediction covers noncoding sequence. The conservation track for a 12-way multiz alignment (http://genome.ucsc.edu) shows that there are perhaps as many as three conserved noncoding elements in this region overlapped by the predicted negative element.

Figure 9 shows an example browser shot of a positively selected element prediction. In this case the score of the element is LOD = 29.6. Both polymorphism and divergence are above average in this region, as expected under a model of diversifying selection. This prediction overlaps exonic sequence of the gene CG6108, for which an amino acid alignment among 12 insect genomes is shown below the conservation track. This exon seems to be evolving quite rapidly among the *Drosophila* species included in the alignment. Thus, although we have predicted this element to be under positive selection within *D. melanogaster*, it also may be under positive selection in a broader evolutionary context.

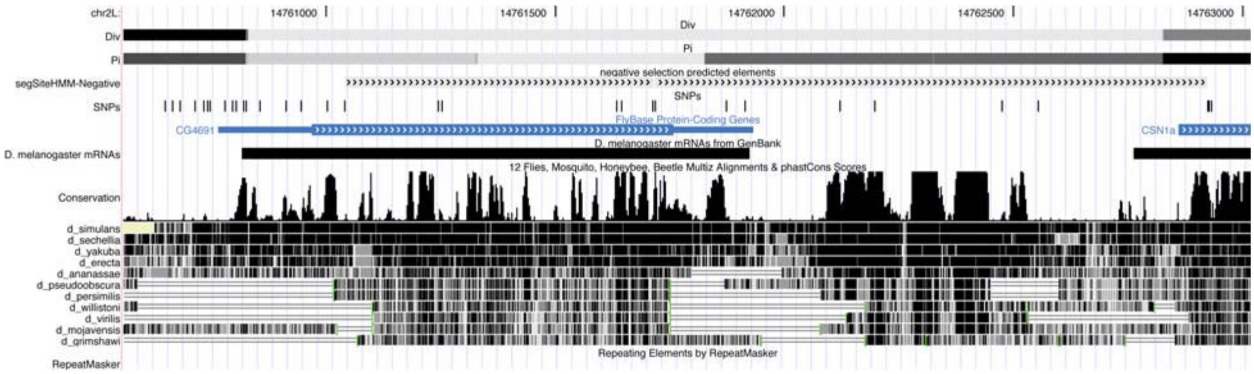

FIG. 8. Browser shot of a negatively selected element prediction. This negative element prediction is shown as the top browser track for this region of the genome. This element has a strong prediction corresponding to a LOD score of 33.1. Shown below the prediction are two tracks corresponding to divergence between *Drosophila simulans* and *D. melanogaster* (labeled Div) and nucleotide diversity (π; Tajima 1983) within *D. melanogaster* (labeled Pi). For both these tracks, darker colors represent greater relative levels of divergence and polymorphism. See text for details.

Discussion

Patterns of genomic variation are influenced by the stochastic forces of genetic drift and demography as well as by the deterministic force of natural selection. As whole-genome surveys of polymorphism data now become available (e.g., Begun et al. 2007), the challenge remains as to how best to identify the targets of natural selection within genomes. This is a goal with broader implications than just a better understanding of the evolutionary forces at work in shaping genetic variation, as the targets of natural selection are expected to represent a set of important functional elements. Indeed with the recent introduction of cheap “next generation” sequencing technologies, an explosion of genome-level resequencing is expected in the coming years within humans and other important model organisms.

In anticipation of these whole-genome polymorphism data sets, we introduce a general class of explicit popGenHMMs, which make predictions about where along genomes underlying population genetic parameters vary. In this manuscript, we use popGenHMMs to estimate selective regimes underlying regions of the genome. Our current implementation has good power to detect negative selection acting in genomes but less power that more specific tests at identifying recent selective sweeps. In addition, we show that our current implementation is highly sensitive to false positives induced by bottleneck models of demography. In our application to *Drosophila* data, our popGenHMM shows some promise, with an enrichment of CDS, UTR, and regulatory regions in the predicted set of negatively selected elements. Given the simplicity of the three-state model, the limitations of the data set, and the fact that existing biological annotations are not expected to completely correlate with negatively selected areas, it is perhaps not surprising that these enrichments are rather weak. However, the results confirm that population genetic data alone harbors some information for making predictions of functional elements solely on the basis of allele frequency variation across the genome.

Although we focus on selection in this use of popGenHMMs, the class of models is general, so in practice,

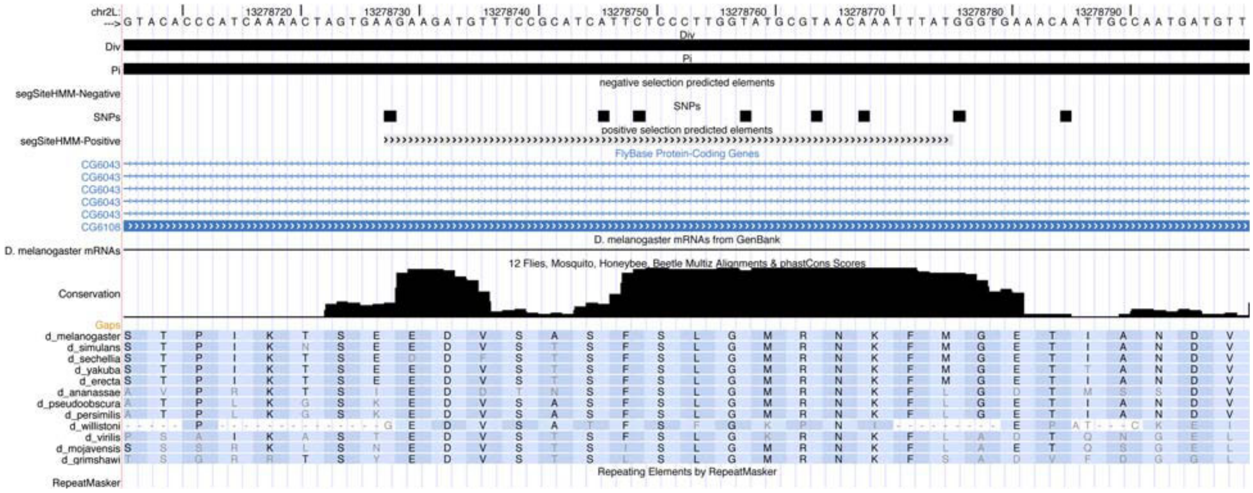
chr2L:
13278720
13278730
13278740
13278750
13278760
13278770
13278780
13278790
---> G T A C A C C C A T C A A A A C T A G T G A A G A A G A T G T T T C C G C A T C A T T C T C C C T T G G T A T G C G T A A C A A A T T T A T G G G T G A A A C A A T T G C C A A T G A T G T T
Div
Pi
segSiteHMM-Negative
negative selection predicted elements
SNPs
segSiteHMM-Positive
positive selection predicted elements
FlyBase Protein-Coding Genes
CG6043
CG6043
CG6043
CG6043
CG6043
CG6108
D. melanogaster mRNAs
D. melanogaster mRNAs from GenBank
12 Flies, Mosquito, Honeybee, Beetle Multiz Alignments & phastCons Scores
Conservation
Gaps
d_melanogaster
d_simulans
d_sechellia
d_yakuba
d_erecta
d_ananassae
d_pseudoobscura
d_persimilis
d_willistoni
d_virilis
d_mojavensis
d_grimshawi
Repeating Elements by RepeatMasker
RepeatMasker

FIG. 9. Browser shot of a positively selected element prediction. This positive element prediction is shown as the fourth browser track from the top for this region of the genome. This element has a prediction score of LOD = 29.6. See caption of figure 5 for details.

any population genetic parameter (e.g., mutation rate, recombination rate) could be the focus. It would also be possible and desirable to extend these models to incorporate divergence data between species directly according to known diffusion approximation expectations (Sawyer and Hartl 1992). We expect this extension to increase predictive power of popGenHMMs significantly and in so doing, bridge the gap between phylo-HMMs (cf. Siepel et al. 2005) and popGenHMMs. In addition, it should be possible to extend popGenHMMs to consider the two-locus allele frequency expectations directly. This extension will move popGenHMMs away from the free recombination allele frequency emission distributions used here and toward conditional emissions that account for linkage relationships among sites. In addition, it is straightforward to extend the model here to include considerations of ascertainment biases underlying variation data (Nielsen et al. 2004) or demographic considerations (Williamson et al. 2005; Boyko et al. 2008) and we expect the later to make vast improvements in our prediction in nonstationary populations.

Toward the goal of identifying genomic regions under natural selection, popGenHMMs used in conjunction with whole-genome polymorphism data sets will produce base-by-base probabilities of selected states. Inasmuch, one should be able to quantitatively describe the portion of sites under selection in a given genome. This is of considerable interest as such an estimate has been a long-term goal of the field of population genetics. Application of phylo-HMMs to multiple alignment data has proven invaluable for the identification of conserved elements of genomes (Siepel et al. 2005), and so too do we expect the application of popGenHMMs to provide a powerful tool for predictive genomics.

Acknowledgments

We thank Adam Siepel for many helpful discussions during the early part of this project. We also thank Matt Hahn, Corbin Jones, Sol Katzman, and Daryl Thomas for their input/suggestions during this work. Carlos Bustamante and four anonymous reviewers improved the quality of this manuscript substantially with their comments. We also thank Christian Schlotterer and Simon Boitard for sharing their unpublished work. A.D.K. was supported by the National Institute of Health/National Human Genome Research Institute and is currently supported by Dartmouth College.

References

Adams MD, Celniker SE, Holt RA, et al. (193 co-authors). 2000. The genome sequence of Drosophila melanogaster. *Science* 287(5461):2185–2195.

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

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

Ashburner M, Ball CA, Blake JA, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25(1):25–29.

Baldi P, Chauvin Y, Hunkapiller T, McClure MA. 1994. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci U S A.* 91(3):1059–1063.

Barton NH. 2000. Genetic hitchhiking. *Philos Trans R Soc Lond B.* 355:1553–1562.

Baum LE, Petrie T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Ann Math Stat.* 37:1554–1563.

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat.* 41:164–171.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.* 5(11):e310.

Betancourt AJ, Kim K, Orr HA. 2004. A pseudohitchhiking model of x vs. autosomal diversity. *Genetics* 168(4):2261–2269.

Boitard S, Schlötterer C, Futschik A. 2009. Detecting selective sweeps: a new approach based on hidden Markov models. *Genetics* 181(4):1567–1578.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4(5):e1000083.

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.

Bryc K, Auton A, Nelson MR, et al. (11 co-authors). 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A.* 107(2):786–791.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159(4):1779–1788.

Churchill GA. 1989. Stochastic models for heterogeneous DNA sequences. *Bull Math Biol.* 51(1):79–94.

Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–1963.

Dempster A, Laird N, Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc B.* 39:1–38.

Durbin R, Eddy S, Krogh A, Mithison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164(4):1567–1587.

Fay JC, Wu CI. 2000. Hitchhiking under positive darwinian selection. *Genetics* 155:1405–1413.

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol.* 13(1): 93–104.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Clarendon Press.

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

Gillespie JH. 1991. The causes of molecular evolution. New York: Oxford University Press.

Gillespie JH. 1993. Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. *Genetics* 134:971–981.

Gillespie JH. 1994a. Substitution processes in molecular evolution. II. Exchangeable models from population genetics. *Evolution* 48:1101–1113.

Gillespie JH. 1994b. Substitution processes in molecular evolution. III. Deleterious alleles. *Genetics* 138:943–952.

Gillespie JH. 1997. Junk ain't what junk does: neutral alleles in a selected context. *Gene* 205:291–299.

Gillespie JH. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155:909–919.

Haussler D, Krogh A, Mian IS, Sjolander K. 1993. Protein modeling using hidden Markov models: analysis of globins. In: Mudge TN, Milutinovic V, Hunter L, editors. Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences. Los Alamitos (CA): IEEE.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* 3(2):e7.

Hudson RR. 2002. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

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

Kern AD. 2009. Correcting the site frequency spectrum for divergence-based ascertainment. *PLoS One* 4(4):e5152.

Kern AD, Jones CD, Begun DJ. 2004. Molecular population genetics of male accessory gland proteins in the Drosophila simulans complex. *Genetics* 167(2):725–735.

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

Kimura M. 1971. Theoretical foundation of population genetics at the molecular level. Theor. *Popul Biol.* 2:174–208.

Kimura M. 1983. The neutral theory of molecular evolution. New York: Cambridge University Press.

Lewontin RC. 1974. The genetic basis of evolutionary change. New York: Columbia University Press.

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(1):351–372.

Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM. 2006. Oreganno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* 22(5):637–640.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet.* 8(11):857–868.

Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168(4):2373–2382.

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

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical recipes in C. Cambridge: Cambridge University Press.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945–959.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312–2323.

Sankararaman S, Kimmel G, Halperin E, Jordan MI. 2008. On the inference of ancestries in admixed populations. *Genome Res.* 18(4):668–675.

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132:1161–1176.

Schwarz G. 1978. Estimation the dimension of a model. Ann *Stat.* 6(2):461–464.

Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15(8):1034–1050.

Siepel A, Haussler D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol.* 11(2–3):413–428.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. *Mol Biol Evol.* 22(1):63–73.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

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

Tang H, Coram M, Wang P, Zhu X, Risch N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet.* 79(1):1–12.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol.* 28(4):289–301.

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.

Viterbi AJ. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory*. IT-13:260–269.

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

Williamson S, Fledel-Alon A, Bustamante CD. 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* 168(1):463–475.

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.

Wright S. 1938. The distribution of gene frequencies under irreversible mutation. *Proc Natl Acad Sci U S A.* 24(7):253–259.

Wright S. 1969. Evolution and the genetics of populations. Vol. 2: the theory of gene frequencies. Chicago (IL): Columbia University Press.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139(2):993–1005.

Zhu L, Bustamante CD. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* 170(3):1411–1421.