Clear: Composition of Likelihoods for Evolve And Resequence Experiments

Arya Iranmehr[1], Ali Akbari[1], Christian Schlötterer[2], and Vineet Bafna[3]

[1]Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA.
[2]Institut für Populationsgenetik, Vetmeduni, Vienna, Austria.
[3]Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA.

Abstract

The advent of next generation sequencing technologies has made whole-genome and whole-population sampling possible, even for eukaryotes with large genomes. With this development, experimental evolution studies can be designed to observe molecular evolution "in-action" via Evolve-and-Resequence (E&R) experiments. Among other applications, E&R studies can be used to locate the genes and variants responsible for genetic adaptation. Existing literature on time-series data analysis often assumes large population size, accurate allele frequency estimates, and wide time spans. These assumptions do not hold in many E&R studies.

In this article, we propose a method–Composition of Likelihoods for Evolve-And-Resequence experiments (Clear)–to identify selection in short-term (as well as long-term), E&R experiments in sexual populations with small size. Clear takes whole-genome sequence of pool of individuals (pool-seq) as input, and properly addresses heterogeneous ascertainment bias resulting from uneven coverage. Clear also provides unbiased estimates of model parameters, including population size, selection strength and overdominance, while being computationally efficient. Extensive simulations show that Clear achieves higher power in detecting and localizing selection over a wide range of parameters, and is robust to variation of coverage. We applied Clear statistic to multiple E&R experiments, including, data from a study of *D. melanogaster* adaptation to alternating temperatures and a study of outcrossing Yeast populations, and identified multiple regions under selection with genomewide significance.

1 Introduction

Natural selection is a key force in evolution, and a mechanism by which populations can adapt to external 'selection' constraints. Examples of adaptation abound in the natural world [27], including for example, classic examples like lactose tolerance in Northern Europeans [12], human adaptation to high altitudes [72, 89], but also drug resistance in pests [19], HIV [30], cancer [35, 90], malarial parasite [5, 56], and others [73]. In these examples, understanding the genetic basis of adaptation can provide actionable information, underscoring the importance of the problem.

Experimental evolution refers to the study of the evolutionary processes of a model organism in a controlled [9, 14, 37, 47, 48, 59, 60] or natural [7, 11, 20, 21, 52, 65, 88] environment. Recent advances in whole genome sequencing have enabled us to sequence populations at a reasonable cost even for large genomes. Perhaps more important for experimental evolution studies, we can now evolve and resequence (E&R) multiple replicates of a population to obtain *longitudinal time-series data*, in order to investigate the dynamics of evolution at molecular level. Although constraints such as small sizes, limited timescales, and oversimplified laboratory environments may limit the interpretation of E&R results, these studies are increasingly being used to test a wide range of hypotheses [44] and have been shown to be more predictive than static data analysis [16, 22, 69]. In particular, longitudinal E&R data is being used to estimate model parameters including population size [43, 62, 78, 83, 84, 86], strength of selection [15, 39, 40, 51, 54, 74, 78], allele age [51] recombination rate [78], mutation rate [8, 78], quantitative trait loci [6] and for tests of neutrality hypotheses [11, 17, 29, 78].

While many E&R study designs are being used [8, 70], we restrict our attention to the adaptive evolution of multi-cellular sexual organisms. For simplicity, we assume fixed population size, and for the most part, positive single locus selection (only one favored mutation). This regime has been considered earlier, typically with *D. melanogaster* as the model organism of choice, to identify adaptive genes in longevity and aging [17, 66] (600 generations), courtship song [81] (100 generations), hypoxia tolerance [91] (200 generations), adaptation to new laboratory environments [32, 59] (59 generations), egg size [42] (40 generations), C virus resistance [53] (20 generations), and dark-fly [41] (49 generations).

The task of identifying genetic adaptation can be addressed at different levels of specificity. At the coarsest level, identification could simply refer to deciding whether some genomic region (or a gene) is under selection or not. In the following, we refer to this task as *detection*. In contrast, the task of *site-identification* corresponds to the process of finding the favored mutation/allele at nucleotide level. Finally, *estimation of model parameters*, such as strength of selection and overdominance at the site, can provide a comprehensive description of the selection process.

In an effort for analyzing E&R selection experiments many authors chose to adopt existing tests that originally used for static data, for scanning dynamic data with two time-points. For instance, Zhu *et al.* [91] used the ratio of the estimated population size of case and control populations to compute test statistic for each genomic region. Burke *et al.* [17] applied Fisher exact test to the last observation of data on case and control populations. –VB note: Something wrong with the name? Orozco-terWengel *et al.* [59] used the Cochran-Mantel-Haenszel (CMH) test [2] to detect SNPs whose read counts change consistently across all replicates of two time-point data. Turner *et al.* [81] proposed the diffStat statistic to test whether the change in allele frequencies of two populations deviate from the distribution of change in allele frequencies of two drifting populations. Bergland *et al.* [11] applied F_{st} to populations throughout time to signify their differentiation from ancestral (two time-point data) as well as geographically different populations. Jha *et al.* [42] computed test statistic of generalized linear-mixed model directly from read counts.

The problem of parameter estimation in time series selection data was first addressed by Bollback *et al.* [15]. They provided a diffusion approximation to the continuous Wright Fisher Markov process and estimated the selection coefficient s numerically for large population sizes. Steinrücken and Song [74] proposed a general diploid selection model which takes into account of dominance of the favored allele and approximates likelihood analytically. Mathieson and McVean [54] adopted HMMs to structured populations and estimated parameters using an Expectation Maximization (EM) procedure on discretized allele frequency. Feder *et al.* [29] modeled increments in allele frequency with a Brownian motion process, proposed the Frequency Increment Test (FIT). More recently, Topa *et al.* [80] proposed a Gaussian Process (GP) for modeling single-locus time-series pool-seq data. Terhorst *et al.* [78] extended GP to compute joint likelihood of multiple loci under null and alternative hypotheses. Recently, schraiber *et al.* [71] proposed a Bayesian framework to estimate parameters using Monte Carlo Markov chain sampling.

While existing methods have been successfully applied to their corresponding application, they make some assumptions which may not hold in E&R studies. First, they assume that the underlying population size is large, so continuous state models can be applied for dynamics of allele frequencies. These methods were originally designed to process wide time spans such as ancient DNA studies.

Finally, they assume that input data is in the form of unbiased allele frequencies.

Here, we consider a model similar to Williamson *et al.* [86] and Bollback *et al.*'s [15] but under a "small-population-size" scenario. Specifically, we use a discrete state (frequency) model. We show that for small population sizes, discrete models can compute likelihood exactly, which improves statistical performance, especially for short time-span experiments. Additionally, we add another level of sampling-noise to the traditional HMM model, allowing for heterogeneous ascertainment bias due to uneven coverage among variants. We show that for a wide range of parameters, our algorithm Clear provides higher power for detecting selection, is robust to ascertainment bias due to coverage heterogeneity, estimates model parameters consistently, and localizes favored allele more accurately compared to the state-of-the-art methods, while being computationally efficient.

2 Materials and Methods

Consider a diploid population with fixed size of N individuals where ν_t denotes allele frequency of the derived allele at generation t. Experimental evolution for R replicates is conducted. Samples of n individuals from each replicate are chosen for pooled sequencing in generations specified by the set $\mathcal{T} = \{\tau_i : 0 \leq \tau_0 < \tau_1, \ldots < \tau_T\}$. To identify the genes and variants that are responding to selection pressure, we use the following procedure:

(i) **Estimating population size.** The procedure starts by estimating the effective population size, $\widehat{N}$, under the assumption that much of the genome is evolving neutrally.

(ii) **Estimating selection parameters.** For each polymorphic site, selection and overdominance parameters s, h are estimated so as to maximize the likelihood of the time series data, given $\widehat{N}$.

(iii) **Computing likelihood statistics.** For each variant, a log-odds ratio of the likelihood of selection model ($s > 0$) to the likelihood of neutral evolution/drift model is computed. Likelihood ratios in a genomic region are combined to compute the Clear statistic for the region.

(iv) **Hypothesis testing.** The Clear (or single locus) statistics are normalized using a z-score computation. A null distribution of the normalized statistic values is computed using a set of whole-genome (single locus, respectively) drift simulations with population size of $\widehat{N}$, and variant starting frequency and coverage matching the experimental data. Given the null distribution, p-values and corresponding False Discovery Rate (FDR) are calculated.

These steps are described in detail below.

2.1 Estimating Population Size

Methods for estimating population sizes from temporal neutral evolution data have been developed [4, 15, 43, 78, 86]. However, our method explicitly addresses the biases that arise in pool-seq data. Specifically, we model the variation in sequence coverage over different locations, and the noise due to sequencing only a subset of the individuals in the population. In addition, many existing methods [15, 29, 78, 80] are designed for large populations, and model frequency as a continuous quantity. However, we show that smooth approximations may be inadequate for small populations, low starting frequencies and sparse sampling (in time) that are typical in experimental evolution (see Results, Fig 3A-C, and Fig 2). To this end, we model the Wright-Fisher Markov process for generating pool-seq data (S1 Fig) via a *discrete* HMM (Fig 1-B). We start by computing a likelihood function for the population size given neutral pool-seq data.

Likelihood for Neutral Model. We model the allele frequency counts $2N\vec{u}_t$ as being sampled from a Binomial distribution. Specifically,

$$\begin{aligned} \nu_0 &\sim \pi, \\ 2N\nu_t|\nu_{t-1} &\sim \text{Binomial}(2N, \nu_{t-1}) \end{aligned}$$

where π is the global distribution of allele frequencies in the base population. Here we simply assume is π is the site frequency spectrum of fixed sized neutral population S2 Fig–VB note: Something is wrong with the reffig macro. Note that π may depend on the demographic history of the founder lines.

To estimate frequency after τ transitions, it is enough to specify the $2N \times 2N$ transition matrix $P^{(\tau)}$, where $P^{(\tau)}[i, j]$ denotes probability of change in allele frequency from $i/2N$ to $j/2N$ in τ generations:

$$\begin{aligned} P^{(1)}[i, j] &= \Pr\left(\nu_{t+1} = \frac{j}{2N} \middle| \nu_t = \frac{i}{2N}\right) = \binom{2N}{j} \nu_t^j (1 - \nu_t)^{2N-j}, && (1) \\ P^{(\tau)} &= P^{(\tau-1)} P^{(1)} && (2) \end{aligned}$$

Finally, in typical E&R experiments, $n < N$ individuals are randomly selected for sequencing. The sampled allele frequencies, $\{y_t\}_{t\in\mathcal{T}}$, are also Binomially distributed

$$2ny_t \sim \text{Binomial}(2n, \nu_t) \tag{3}$$

We introduce the $2N \times 2n$ sampling matrix Y, where $Y[i, j]$ stores the probability that the sample allele frequency is $i/2n$ given that the true allele frequency is $i/2N$–VB note: One of sample or true allele frequencies needs to be $j/2N$.

We denote the pool-seq data for that variant as $\{x_t = \langle c_t, d_t\rangle\}_{t\in\mathcal{T}}$ where d_t, c_t represent the read depth, and the read count of the derived allele, respectively, at time τ_t. Let $\{\lambda_t\}_{t\in\mathcal{T}}$ be the sequencing coverage at different generations, then, the observed data are sampled according to

$$d_t \sim \text{Poisson}(\lambda_t), \qquad\qquad c_t \sim \text{Binomial}(d_t, y_t) \tag{4}$$

The emission probability for a observed tuple $x_t = \langle d_t, c_t\rangle$ is

$$\mathbf{e}_i(x_t) = \binom{d_t}{c_t} \left(\frac{i}{2n}\right)^{c_t} \left(1 - \frac{i}{2n}\right)^{d_t - c_t}. \tag{5}$$

For $1 \leq t \leq T, 1 \leq j \leq 2N$, let $\alpha_{t,j}$ denote the probability of emitting $x_1, x_2, \ldots, x_t$ and reaching state j at τ_t. Then, α_t can be computed using the forward-procedure [23]:

$$\alpha_t^T = \alpha_{t-1}^T P^{(\delta_t)} \text{diag}(Y\mathbf{e}(x_t)) \tag{6}$$

where $\delta_t = \tau_t - \tau_{t-1}$. The joint likelihood of the observed data from R independent observations is given by

$$\mathcal{L}(N|\{\boldsymbol{x}^{(r)}\}, n) = \prod_{r=1}^{R} \mathcal{L}(N|\boldsymbol{x}^{(r)}, n) = \Pr(\{\boldsymbol{x}^{(r)}\}|N, n) = \prod_{r=1}^{R} \sum_i \alpha_{T,i}^{(r)} \tag{7}$$

where $\boldsymbol{x} = \{x_t\}_{t\in\mathcal{T}}$. The graphical model and the generative process for which data is being generated is depicted in Fig 1-B and S1 Fig, respectively.

Finally, the last step is to compute an estimate $\widehat{N}$ that maximizes the likelihood of all M variants in whole genome. Let $\boldsymbol{x}_i^{(r)}$ denote the time-series of the i-th variant in replicate r. Then,

$$\widehat{N} = \arg\max_N \prod_{i=1}^{M} \prod_{r=1}^{R} \mathcal{L}(N|\boldsymbol{x}_i^{(r)}) \tag{8}$$

2.2 Estimating Selection Parameters

Likelihood for Selection Model. Assume that the site is evolving under selection constraints $s \in \mathbb{R}$, $h \in \mathbb{R}_+$, where s and h denote selection strength and overdominance parameters , respectively. By definition, the relative fitness values of genotypes $0|0$, $0|1$ and $1|1$ are given by $w_{00} = 1$, $w_{01} = 1 + hs$ and $w_{11} = 1 + s$. Recall that ν_t denotes the frequency of the site at time $\tau_t \in \mathcal{T}$. Then, ν_{t+}, the frequency at time $\tau_t + 1$ (one generation ahead), can be estimated using:

$$\begin{aligned}\hat{\nu}_{t+} = \mathbb{E}[\nu_{t+}|s, h, \nu_t] &= \frac{w_{11}\nu_t^2 + w_{01}\nu_t(1-\nu_t)}{w_{11}\nu_t^2 + 2w_{01}\nu_t(1-\nu_t) + w_{00}(1-\nu_t)^2} \\ &= \nu_t + \frac{s(h + (1-2h)\nu_t)\nu_t(1-\nu_t)}{1 + s\nu_t(2h + (1-2h)\nu_t))}.\end{aligned} \tag{9}$$

The machinery for computing likelihood of the selection parameters is identical to that of population size, except for transition matrices. Hence, here we only describe the definition transition matrix $Q_{s,h}$ of the selection model. Let $Q^{(\tau)}_{s,h}[i,j]$ denote the probability of transition from $i/2N$ to $j/2N$ in τ generations, then (See [25], Pg. 24, Eqn. 1.58-1.59):

$$Q^{(1)}_{s,h}[i,j] = \Pr\left(\nu_{t+} = \frac{j}{2N} \middle| \nu_t = \frac{i}{2N}; s, h, N\right) = \binom{2N}{j} \hat{\nu}^{j}_{t+}(1 - \hat{\nu}_{t+})^{2N-j} \tag{10}$$

$$Q^{(\tau)}_{s,h} = Q^{(\tau-1)}_{s,h} Q^{(1)}_{s,h} \tag{11}$$

The maximum likelihood estimates are given by

$$\widehat{s}, \widehat{h} = \arg\max_{s,h} \prod_{r=1}^{R} \mathcal{L}(s, h|\boldsymbol{x}^{(r)}, \widehat{N}) \tag{12}$$

Using grid search, we first estimate N (Eq. 8), and subsequently, we estimate parameters s, h (Eq. 12). By broadcasting and vectorizing the grid search operations across all variants, the genome scan on millions of polymorphisms can be done in significantly smaller time than iterating a numerical optimization routine for each variant(see Results and Fig 4).

2.3 Empirical Likelihood Ratio Statistics

The likelihood ratio statistic for testing directional selection, to be computed for each variant, is given by

$$H = -2\log\left(\frac{\mathcal{L}(\bar{s}, 0.5|\{\boldsymbol{x}^{(r)}\}, \widehat{N})}{\mathcal{L}(0, 0.5|\{\boldsymbol{x}^{(r)}\}, \widehat{N})}\right), \tag{13}$$

where $\bar{s} = \arg\max_{s} \prod_{r=1}^{R} \mathcal{L}(s, 0.5|\boldsymbol{x}^{(r)}, \widehat{N})$. Similarly we can define a test statistic for testing if selection is over-dominant as:

$$D = -2\log\left(\frac{\mathcal{L}(\widehat{s}, \widehat{h}|\boldsymbol{x}^{(r)}, \widehat{N})}{\mathcal{L}(\bar{s}, 0.5|\boldsymbol{x}^{(r)}, \widehat{N})}\right). \tag{14}$$

While extending the single-locus WF model to a multiple linked-loci can improve the power of the model [78], it is computationally and statistically expensive to compute exact likelihood–VB note: Please provide plain text from the git repository to allow me to correct this. haplotype resolved data, which pool-seq does not provide. Instead, similar to Nielse *et al* [58], we calculate Composite Likelihood Ratio score for a genomic region.

$$\mathcal{H} = \frac{1}{|L|} \sum_{\ell \in L} H_\ell. \tag{15}$$

where L is a collection of segregating sites and H_ℓ is the likelihood ratio score based for each variant ℓ in L. The optimal value of the hyper-parameter L depends upon a number of factors, including initial frequency of the favored allele, recombination rates, linkage of the favored allele to neighboring variants, population size, coverage, and time since the onset of selection (duration of the experiment). However, we provide a heuristic choose size of L for an experiment –VB note: WE cannot use this argument. You are saying that L is very dependent on these parameters and then you say that it can have large impact on the final answer. In that case why should the reviewer allow us to choose arbitrary value? We need to say that the answer *does not* change when L is chosen in a reasonable range. This needs to be discussed and rewritten..

In general, as selection acts locally in the genome, size of L have a direct effect on the power. For instance, when L is chosen to be a large region (e.g. chromosome), power will be degraded since distribution of null and alternative $\mathcal{H}$ statistics converge together. Hence, we choose L to be the largest, such that it provides enough discoveries that satisfies experiment's FDR.

2.4 Hypothesis Testing

Single-Locus tests. Under neutrality, Log-likelihood ratios can be approximated by $\mathcal{X}^2$ distribution [85], and p-values can be computed directly. However, Feder *et al.* [29] showed that when the number of independent samples (replicates) is small, $\mathcal{X}^2$ is a crude approximation to the true null distribution and underestimates FDR. Following their suggestion, we compute p-values based on the empirical distribution of statistic on simulations using the estimated population size. (See S1 Fig for details). The empirical distribution of statistic H is used to compute p-values as fraction of null values that exceed the test score. Finally, we use Storey and Tibshirani's method [76] to control for False Discovery Rate in multiple testing.

Composite likelihood tests. As selection is expected to have local effect on the genome, we normalize $\mathcal{H}$ with respect to each chromosome both in simulated and experimental data:

$$\mathcal{H}_i^* = \frac{\mathcal{H}_i - \mu_\mathcal{C}}{\sigma_\mathcal{C}}, \qquad \forall i \in \mathcal{C}, \tag{16}$$

where $\mu_\mathcal{C}$ and $\sigma_\mathcal{C}$ are the mean and standard deviation of $\mathcal{H}$ values in a large (entire chromosome) region. The normalize $\mathcal{H}$ scores are used to compute p-values and FDR using the methodology for single locus analysis. After discovering intervals that exceed the cut-off for the desired FDR, we further select variants within selected intervals that have significant individual scores based on single-locus tests, and identify the genes spanning those variants.

2.5 Simulations

We performed extensive simulations using parameters that have been used for *D. melanogaster* experimental evolution [46]. See also Fig 1-A for illustration. To implement real world pool-seq experimental evolution, we conducted simulations as follows:

I. **Creating initial founder line haplotypes.** Using `msms` [26], we created neutral populations for F founding haplotypes with command `$./msms <F> 1 -t <2`μ`LNe> -r <2rNeL> <L>`, where $F = 200$ is number of founder lines, $N_e = 10^6$ is effective population size,

$r = 2 \times 10^{-8}$ is recombination rate, $\mu = 2 \times 10^{-9}$ is mutation rate and $L = 50K$ is the window size in base pairs which gives $\theta = 2\mu N_e L = 200$ and $\rho = 2N_e r L = 2000$.

II. **Creating initial diploid population.** To simulate experimental evolution of diploid organisms, initial haplotypes were first cloned to create F diploid homozygotes. Next, each diploid individual was cloned N/F times to yield diploid population of size N.

III. **Forward Simulation.** We used forward simulations for evolving populations under selection. We also consider selection regimes which the favored allele is chosen from standing variation (not *de novo* mutations). Given initial diploid population, position of the site under selection, selection strength s, number of replicates $R = 3$, recombination rate $r = 2 \times 10^{-8}$ and sampling times $\mathcal{T} = \{0, 10, 20, 30, 40, 50\}$, `simuPop` [61] was used to perform forward simulation and compute allele frequencies for all of the R replicates. For hard sweep (respectively, soft sweep) simulations we randomly chose a site with initial frequency of $\nu_0 = 0.005$ (respectively, $\nu_0 = 0.1$) to be the favored allele.

IV. **Sequencing Simulation.** Give allele frequency trajectories we sampled depth of each site identically and independently from Poisson(λ), where $\lambda \in \{30, 100, 300\}$ is the coverage for the experiment. Once depth d is drawn for the site with frequency ν, the number of reads c carrying the derived allele are sampled according to Binomial(d, ν). For experiments with finite depth the tuple $\langle c, d\rangle$ is the input data for each site.

3 Results

Modeling Allele Frequency Trajectories in Small Populations. We first tested the goodness of fit of the discrete versus continuous models in modeling allele frequency trajectories, under general E&R parameters. For this purpose, we conducted 100K simulations with two time samples $\mathcal{T} = \{0, \tau\}$ where $\tau \in \{1, 10, 100\}$ is the parameter controlling the density of sampling in time. In addition, we repeated simulations for different values of starting frequency $\nu_0 \in \{0.005, 0.1\}$ (i.e., hard and soft sweep) and selection strength $s \in \{0, 0.1\}$ (i.e., neutral and selection). Then, given initial frequency ν_0, we computed the expected distribution of the frequency of the next sample ν_τ under two models and compared them with empirical distributions calculated from simulated data. Fig 2A-F shows that Brownian motion (continuous model) is inadequate when ν_0 is far from 0.5, or when sampling times are sparse ($\tau > 1$). If the favored allele arises from standing variation in a neutral population, it is unlikely to have frequency close to 0.5, and the starting frequencies are usually much smaller (see S2 Fig). Moreover, in typical *D. melanogaster* experiments for example, sampling is sparse. Often, the experiment is designed so that $10 \le \tau \le 100$ [32, 46, 59, 91].

In contrast to the Brownian motion results, discrete Markov chain can provide predictions when the the allele is under selection. In addition Fig 2A-M also shows that Markov chain predictions (Eq. 11) are highly consistent with empirical data for a wide range of simulation parameters.

Detection Power. We compared the performance of Clear against other methods for detecting selection. For each method we calculated detection power as the percentage of true-positives identified with false-positive rate ≤ 0.05. For each configuration (specified with values for selection coefficient s, starting allele frequency ν_0 and coverage λ), power of each method is evaluated over 2000 distinct simulations, half of which modeled neutral evolution and the rest modeled positive selection.

We compared the power of Clear with Gaussian process (GP) [78], FIT [29], and CMH [2] statistics. FIT and GP convert read counts to allele frequencies prior to computing the test statistic.

Clear shows the highest power in all cases and the power stays relatively high even for low coverage (Fig 3 and S1 Table). In particular, the difference in performance of Clear with other methods is pronounced when starting frequency is low. The advantage of Clear stems from the fact that favored allele with low starting frequency might be missed by low coverage sequencing. In this case, incorporating the signal from linked sites becomes increasingly important. We note that methods using only two time points, such as CMH, do relatively well for high selection values and high coverage. However, the use of time-series data can increase detection power in low coverage experiments or when starting frequency is low. Moreover, time-series data provide means for estimating selection parameters s, h (see below). Finally, as Clear is robust to change of coverage, our results (Fig 3B,C) suggest that taking many samples with lower coverage is preferable to sparse sampling with higher coverage.

Site-identification. In general, localizing the favored variant, using pool-seq data is a nontrivial task [79]. We used the simple approach of ranking each site in a region detected as being under selection. The variants were ranked according to the likelihood ratio scores (Eqn. 13). For each setting of ν_0 and s, we conducted 1000 simulations and computed the rank of the favored mutation in each simulation. The cumulative distribution of the rank of the favored allele in 1000 simulation for each setting (Fig 5) shows that Clear outperforms other statistics.

An interesting observation is revisiting the contrast between site-identification and detection [49, 79]. When selection coefficient is high, detection is easier (Fig 3A-F), but site-identification is harder due to the high LD between hitchhiking sites and the favored allele (Fig 5A-F). Moreover, site-identification is harder in hard sweep scenarios relative to soft sweeps. For example, when coverage $\lambda = 100$ and selection coefficient $s = 0.1$, the detection power is 75% for hard sweep, but 100% for soft sweep (Fig 3B-E). In contrast, the favored site was ranked as the top in 14% of hard sweep cases, compared to and 95% of soft sweep simulations.

Estimating Parameters. Clear computes the selection parameters $\hat{s}$ and $\hat{h}$ as a byproduct of the hypothesis testing. We computed bias of selection fitness $(s - \hat{s})$ and overdominance $(h - \hat{h})$ for of Clear and GP in each setting. The distribution of the error (bias) for 100× coverage is presented in Fig 6 for different configurations. S4 Fig and S5 Fig provide the distribution of estimation errors for 30×, and infinite coverage, respectively. For hard sweep, Clear provides estimates of s with lower variance of bias (Fig 6A). In soft sweep, GP and Clear both provide unbiased estimates with low variance (Fig 6B). Fig 6C-D shows that Clear provides unbiased estimates of h as well.

Running Time. As Clear does not compute exact likelihood of a region (i.e., does not explicitly model linkage between sites), the complexity of scanning a genome is linear in number of polymorphisms. Calculating score of each variant requires and $\mathcal{O}(TRN^2)$ computation for $\mathcal{H}$. However, most of the operations are can be vectorized for all replicates to make the effective running time for each variant. We conducted 1000 simulations and measured running times for computing site statistics H, FIT, CMH and GP with different number of linked-loci. Our analysis reveals (Fig 4) that Clear is orders of magnitude faster than GP, and comparable to FIT. While slower than CMH on the time per variant, the actual running times are comparable after vectorization and broadcasting over variants (see below).

These times can have a practical consequence. For instance, to run GP in the single locus mode on the entire pool-seq data of the *D. melanogaster* genome from a small sample (≈1.6M variant sites), it would take 1444 CPU-hours (≈ 1 CPU-month). In contrast, after vectorizing and broadcasting operations for all variants operations using `numba` package, Clear took 75 minutes to perform an scan, including precomputation, while the fastest method, CMH, took 17 minutes.

3.1 Analysis of a *D. melanogaster* Adaptation to Cold and Hot Temperatures

We applied Clear to the data from a study of *D. melanogaster* adaptation to alternating temperatures [32, 59], where 3 replicate samples were chosen from a population of *D. melanogaster* for 59 generations under alternating 12-hour cycles of hot (28°C) and cold (18°C) temperatures and sequenced. In this dataset, sequencing coverage is different across replicates and generations (see S2 Fig of [78]) which makes variant depths highly heterogeneous (S3 Fig).

We first filtered out heterochromatic, centromeric and telomeric regions [31], and those variants that have collective coverage of more that 1500 in all 13 populations: three replicates at the base population, two replicates at generation 15, one replicate at generation 23, one replicate at generation 27, three replicates at generation 37 and three replicates at generation 59. After filtering, we ended up with 1,605,714 variants.

Next, we estimated population size $\hat{N} = 250$ using all genomic variants (Fig 7). The likelihood curves of Clear is sharper around the optimum compared to Bollback et. al [15] (see Supplementary Fig. 1 in [59]). Also, chromosomes 3L and 3R appear to have smaller population size Fig 7-D, $\widehat{N} = 200, 150$, respectively.

Using the general estimated population $\hat{N} = 250$, we computed ML estimates of s, and computed the normalized test statistic $\mathcal{H}^*$ on sliding windows of size of 500 SNPs and step size of 100 variants over the genome. We computed null distribution of $\mathcal{H}^*$ by creating 100 chromosome simulations using experimental data parameters and length of 20Mbp. After correcting for multiple testing, only 16 intervals Fig 8 satisfy FDR$\leq$ 0.05. The 16 intervals encompass 5 contiguous regions covering 2,829 polymorphic sites. To focus on the strongest signals, we selected 174 variants with FDR $\leq$ 0.01% within selected regions using single locus hypothesis testing. To compute p-values of H statistics, we calculated single locus Wright-Fisher simulations for $\widehat{N}$ =250, initial frequencies and variant depths of real data (see S1 Fig). We repeated forward simulations 50 times for whole genome, to collect ≈90M pool-seq trajectories with starting frequencies and coverage of real data. Then, p-value of each variant in the real data is calculated as the fraction of null statistics that are greater than or equal of test statistic(see S8 Fig)–VB note: We need to change this a bit. The selected 174 variants fall within 32 genes S3 Table.

Finally, we tested if variants showing signal of overdominance, we computed D statistic on simulated and experimental data, and computed p-values accordingly. After correcting for multiple testing, 96 variants discovered with FDR$\leq$ 0.01 Fig 9.

Here is UCSC track for this data.

3.2 Analysis of Outcrossing Yeast Populations

We also applied Clear to outcrossing Yeast populations [18], with 12 replicates where samples are taken at generations $\mathcal{T} = \{0, 180, 360, 540\}$. While this experiment is being conducted with larger set of replicates, population size, and number of generations, it appears that a number of replicates undergoing severe demographic events S9 Fig. Hence we chose seven replicates $r \in \{3, 7, 8, 9, 10, 11, 12\}$ that exhibit consistent genome-wide site-frequency spectrum over the whole experiment S10 Fig.

We estimated population size to be $\widehat{N} = 2000$ haplotypes, and computed $\widehat{s}$, $\widehat{h}$ and H statistic accordingly. To compute p-values, we created 1M single-locus neutral simulations according to experimental data's initial frequency and coverage. By setting FDR cutoff to 0.05, only 18 and 16 variants show significant signal for directional and overdominant selection, respectively, see Fig 9.

Here is UCSC track for this data.

4 Discussion

We developed a computational tool, Clear, that can detect regions and variants under selection E&R experiments of sexual populations. Using extensive simulations, we show that Clear outperforms existing methods in detecting selection, locating the favored allele, and estimating selection parameters. Also, while being computationally efficient, Clear provide means for estimating populations size and hypothesis testing.

Many factors such as small population size, finite coverage, linkage disequilibrium, finite sampling for sequencing, duration of the experiment and the small number of replicates can limit the power of tools for analyzing E&R. Here, by an discrete modeling, Clear estimates population size, and provides unbiased estimates of s, h. It adjusts for uniform and heterogeneous ascertainment bias of pooled-seq data, and exploits presence of high linkage within a region to compute composite likelihood ratio statistic.

It should be noted that, even though we outlined Clear for small populations for small and fixed N, it can be adjusted for such scenarios. For instance, for a *known* changing population sizes, transition probabilities can be readily computed. For large populations, transitions can be computed for a fixed size of frequencies.

The comparison of hard and soft sweep scenarios showed that initial frequency of the favored allele can have an nontrivial effect on the statistical power for identifying selection. Interestingly, while in stronger selections it is easier to detect regions of selection, it is difficult to locate favored allele in those regions.

There are many directions to improve the analyses presented here. In particular, we plan to focus our attention on other organisms with more complex life cycles, experiments with variable population size and longer sampling-time-spans. As evolve and resequencing experiments continue to grow, deeper insights into adaptation will go hand in hand with improved computational analysis.

Software and Data Availability. The source code and running scripts for Clear are publicly available at `https://github.com/bafnalab/clear`.

D. melanogaster data originally published [32, 59]. The dataset of the *D. melanogaster* study, until generation 37, is obtained from Dryad digital repository (`http://datadryad.org`)under accession DOI: 10.5061/dryad.60k68. Generation 59 of the *D. melanogaster* study is accessed from European Sequence Read Archive (http://www.ebi.ac.uk/ena/) under the project accession number: PRJEB6340. The dataset containing experimental evolution of Yeast populations [18] is downloaded from `http://wfitch.bio.uci.edu/~tdlong/PapersRawData/BurkeYeast.gz` (last accessed 01/24/2017). UCSC browser tracks for *D. melanogaster* and Yeast data analysis are found in Suppl. Data 1 and 2, respectively.

Acknowledgments

AI, AA, and VB were supported by grants from the NIH (1R01GM114362) and NSF (DBI-1458557 and IIS-1318386). CS is supported by the European Research Council grant ArchAdapt.

Conflict of interest

VB is a co-founder, has an equity interest, and receives income from Digital Proteomics, LLC (DP). The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. DP was not involved in the research presented here.

Figures

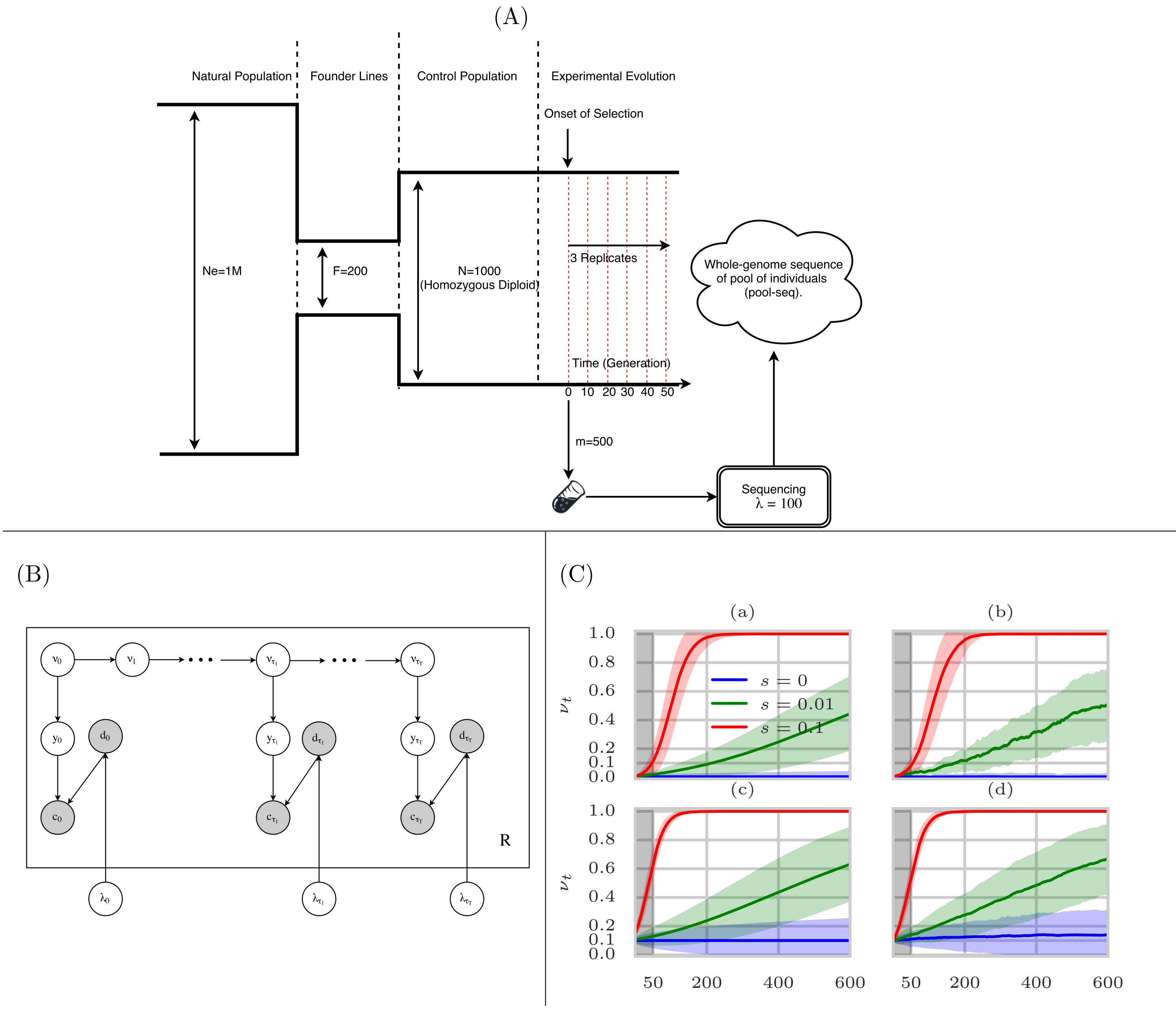


Fig 1: **Evolve and Resequence Selection Experiments on *D. melanogaster* .** (A) Typical configuration in which dynamic data is collected for *D. melanogaster* . (B) Graphical model showing dependence of the random variables in the single-locus model used to compute Clear statistics. Only shaded variables are observed. (C) Mean and the 95% confidence interval of the theoretical (a,b) and empirical (b,d) trajectories of the favored allele for hard (a,b) and soft (c,d) sweep scenarios and $N = 1000$. The first 50 generations are shaded in gray to represent the sampling span of sampling in short-term experiments, implying the difficulty in predicting selection at early stages of selective sweep.

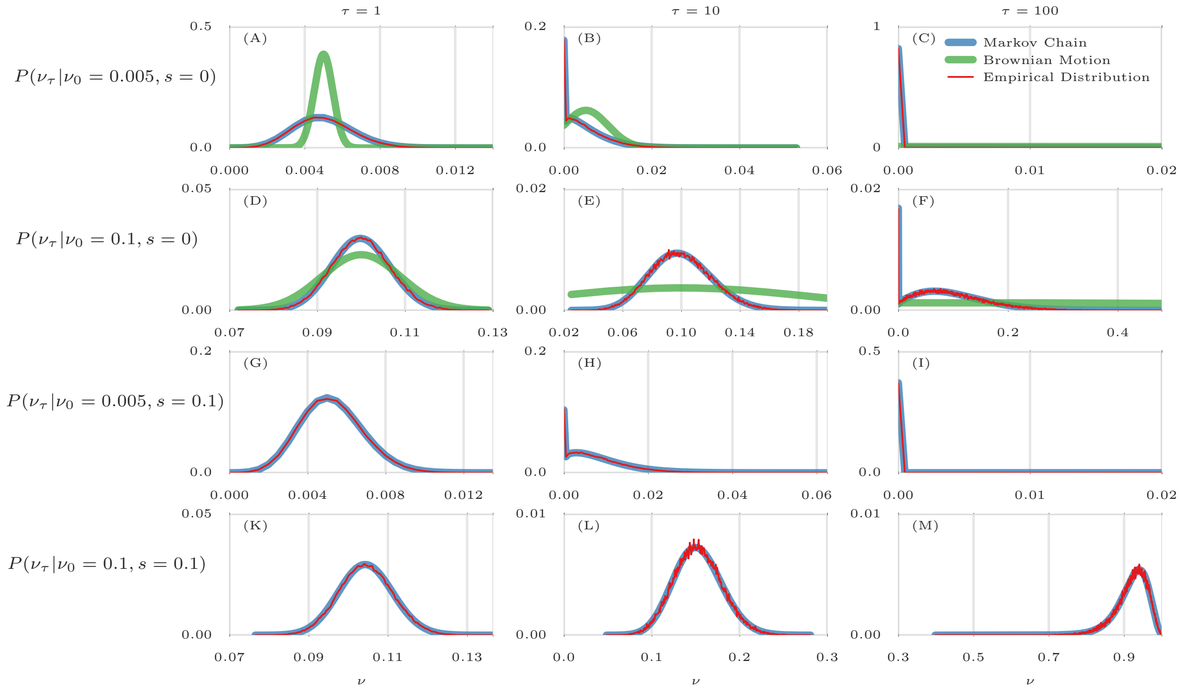


Fig 2: **Comparison of empirical distributions of allele frequencies (red) versus predictions from Brownian Motion (green), and Markov chain (blue).**
Comparison of empirical and theoretical distributions under neutral evolution (panels A-F) and selection (panels G-M) with different starting frequencies $\nu_0 \in \{0.005, 0.1\}$ and sampling times of $\mathcal{T} = \{0, \tau\}$, where $\tau \in \{1, 10, 100\}$. For each panel, the empirical distribution was computed over 100,000 simulations. Brownian motion (Gaussian approximation) provides poor approximations when initial frequency is far from 0.5 (A) or sampling is sparse (B,C,E,F). In addition, Brownian motion can only provide approximations under neutral evolution. In contrast, Markov chain consistently provide a good approximation in all cases.

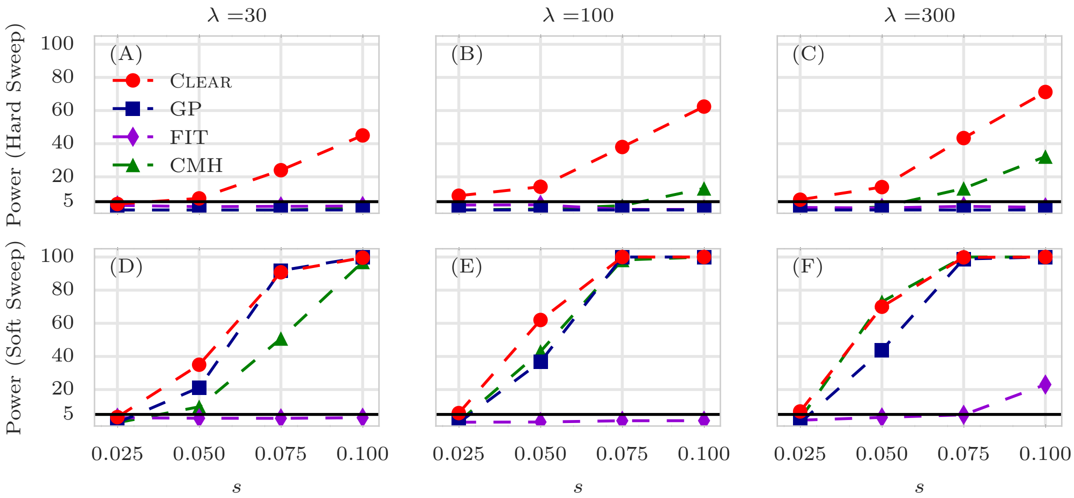


Fig 3: **Power calculations for detection of selection.**
Detection power for Clear($\mathcal{H}$), Frequency Increment Test (FIT), Gaussian Process (GP), and CMH under hard (A-C) and soft sweep (D-F) scenarios. λ, s denote the mean coverage and selection coefficient, respectively. The y-axis measures power – sensitivity with false positive rate FPR $\leq$ 0.05 – for 2,000 simulations of 50Kbp regions. The horizontal line reflects the power of a random classifier. In all simulations, 3 replicates are evolved and sampled at generations $\mathcal{T} = \{0, 10, 20, 30, 40, 50\}$.

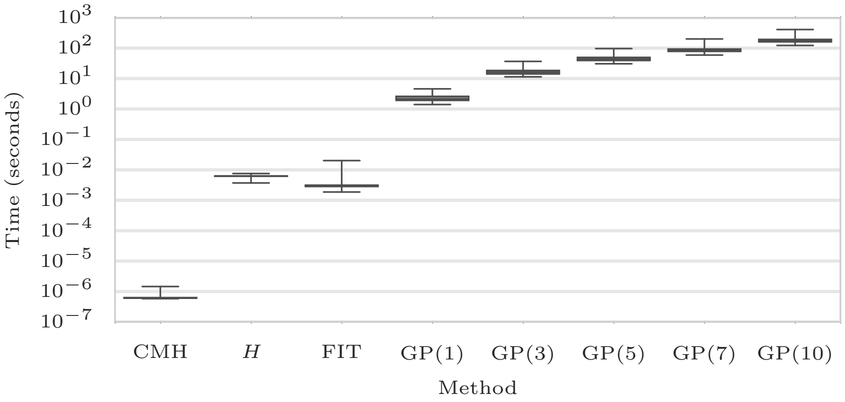


Fig 4: **Running time.**
Box plots of running time per variant (CPU-secs.) of Clear($\mathcal{H}$), CMH, FIT, and GP with single, 3, 5, 7, and 10 loci over 1000 simulations conducted on a workstation with Intel Core i7 processor. The average running time for each method is shown on the x-axis. In all simulations, 3 replicates are evolved and sampled at generations $\mathcal{T} = \{0, 10, 20, 30, 40, 50\}$.

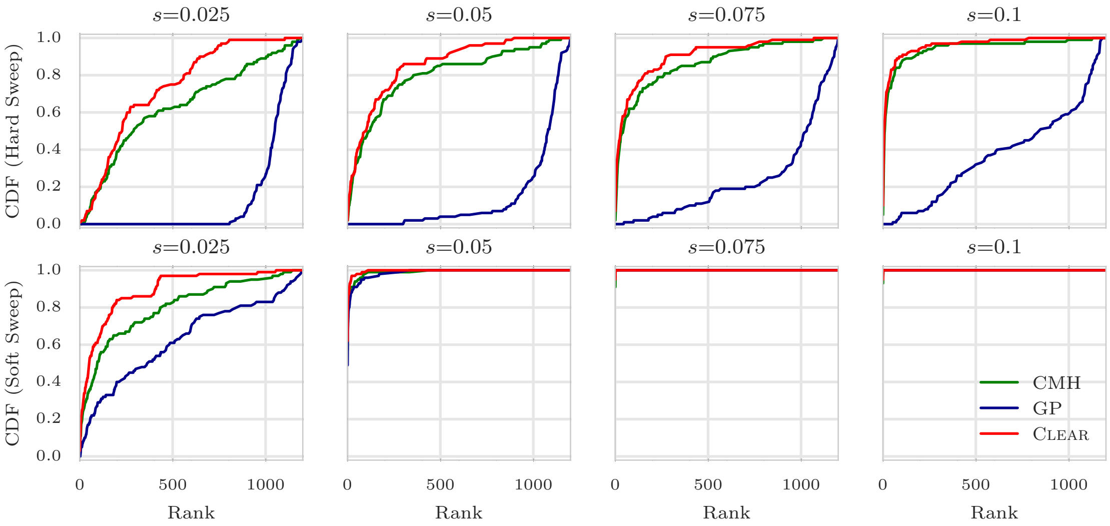


Fig 5: **Ranking performance for 100× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for Clear (H), Gaussian Process (GP), CMH, and Frequency Increment Test (FIT), for different values of selection coefficient s and initial carrier frequency. Note that the individual variant Clear score (H) is used to rank variants. The Area Under Curve (AUC) is computed as an overall quantitative measure to compare the performance of methods for each configuration. In all simulations, 3 replicates are evolved and sampled at generations $\mathcal{T} = \{0, 10, 20, 30, 40, 50\}$.

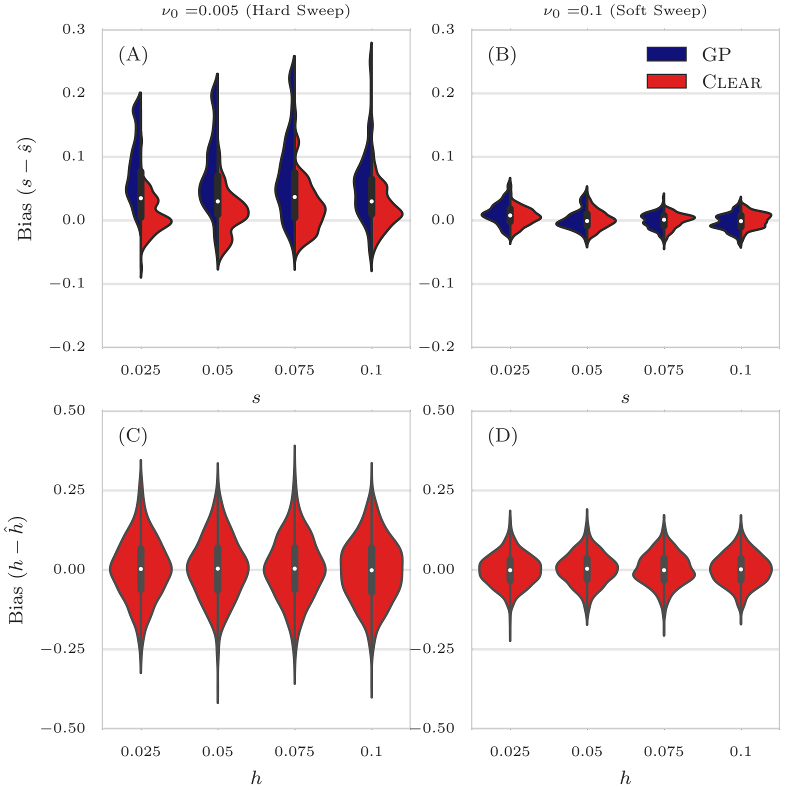


Fig 6: **Distribution of bias for 100× coverage.**

The distribution of bias $(s-\hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and Clear (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = 100$ (Panels A,B). GP and Clear have similar variance in estimates of s for soft sweep, while Clear provides lower variance in hard sweep. Also see S2 Table. Panels C,D show the variance in the estimation of h. In all simulations, 3 replicates are evolved and sampled at generations $\mathcal{T} = \{0, 10, 20, 30, 40, 50\}$.

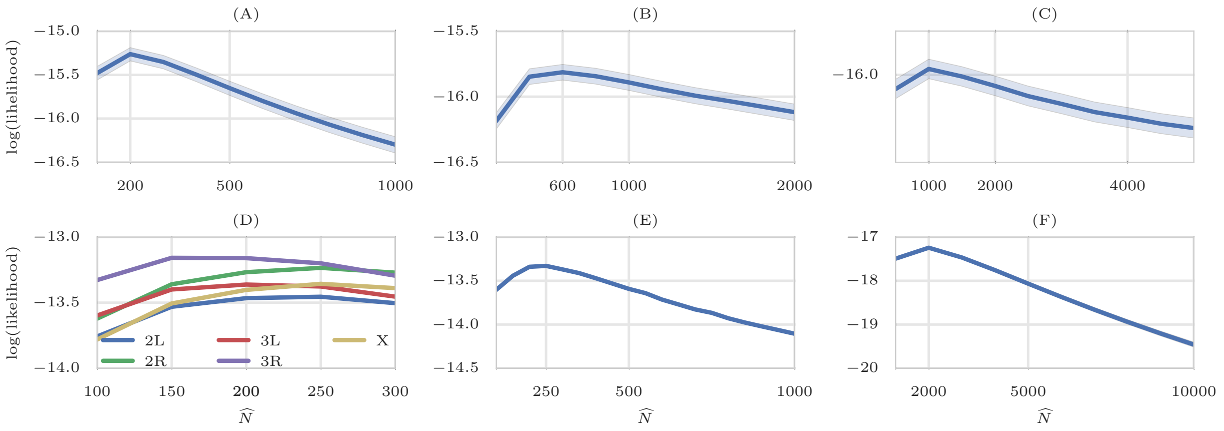


Fig 7: **Estimating N on simulated and real data.** Average and 95% confidence interval if likelihood of simulated data with $N = 200$ (A), $N = 600$(B), and $N = 1000$ individuals, over 100 simulations, shows that estimator is unbiased. Chromosome-wise (D) and genome-wide (E) estimation of population size for data from a study of *D. melanogaster* adaptation to alternating temperatures. Chromosome 3R fits population size of 150, while genome-wide population size is 250. (F) Despite large census population size ($10^6 - 10^7$ [18]), Yeast dataset exhibits much smaller ($\widehat{N} = 2000$) populations size.

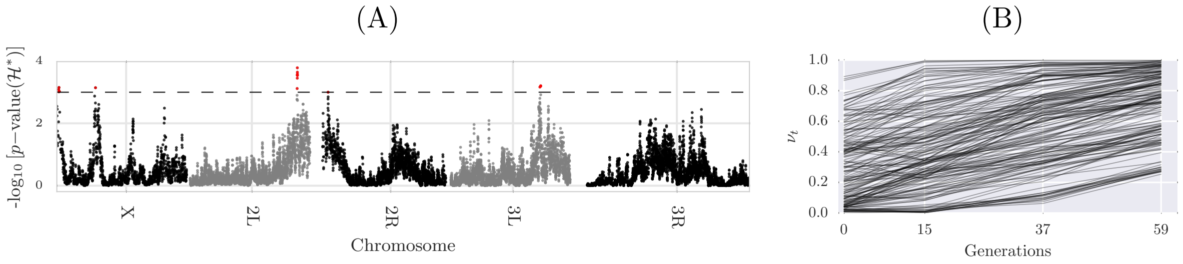


Fig 8: **Scan of composite statistic on data from a study of *D. melanogaster* adaptation to alternating temperatures.** Manhattan plot of scan for $\mathcal{H}^*$ statistic over the genome. The dashed line represents cutoff for genome-wide FDR$\leq$ 0.05, selecting 16 regions (A). Trajectories of the top 174 variants shown in panel (B).

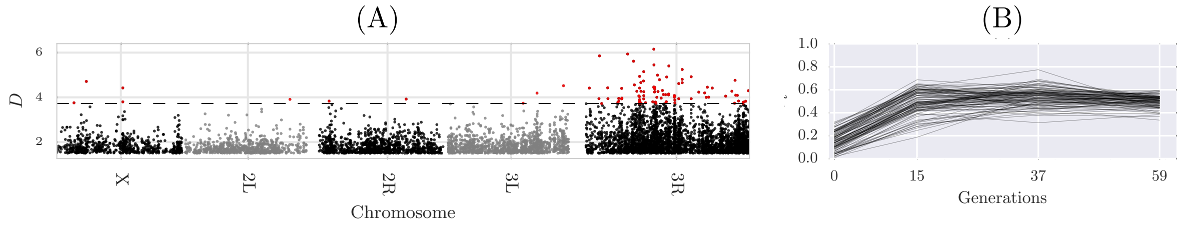


Fig 9: **Single locus analysis of the data from a study of *D. melanogaster* adaptation to alternating temperatures.**

Manhattan plot of scan for testing overdominant selection (A). Significant variants with FDR $\leq$ 0.01 are denoted in red, and their trajectories are depicted in panel (B).

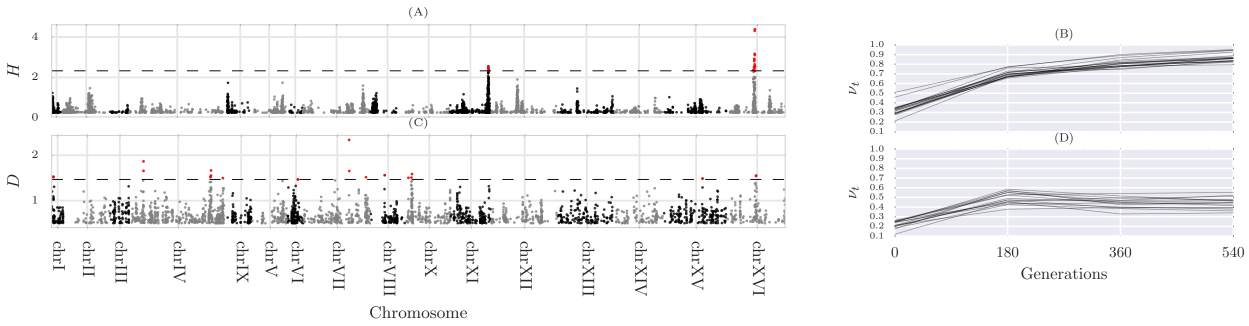


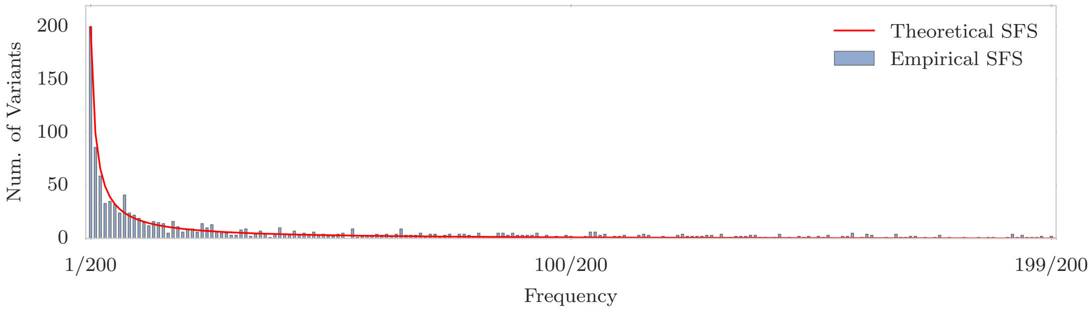
Fig 10: **Single locus analysis of the data from a study of *D. melanogaster* adaptation to alternating temperatures.**

Manhattan plot of scan for testing directional selection (A) and overdominant selection (C). dashed The dashed line represents cutoff for genome-wide FDR≤ 0.05. Trajectories of the selected variants are depicted in panels (B) and (D).

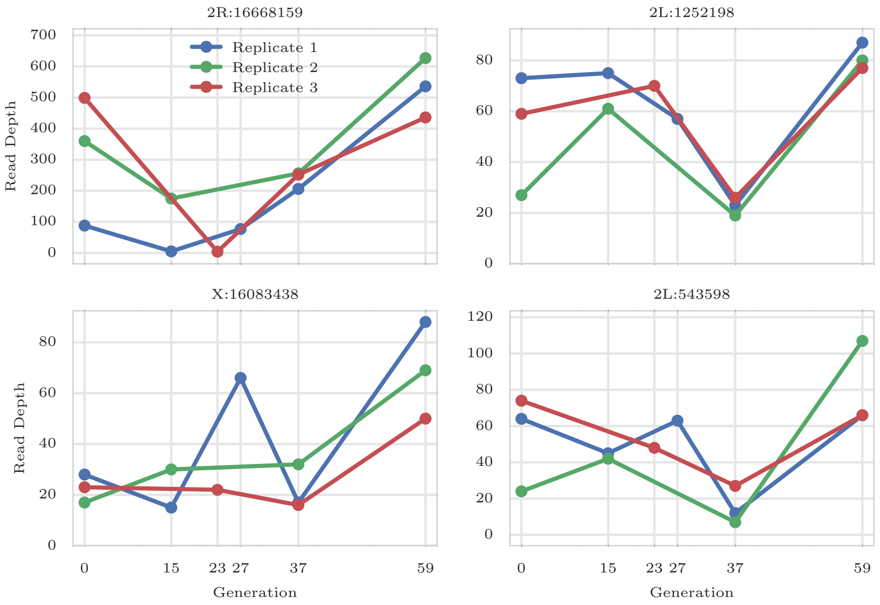
Generative Process 1: The Generative Process for Dynamic Pool-seq Data.

```
Input: $N, n, R, \{\lambda_{\tau_0}, \dots, \lambda_{\tau_T}\}, \mathcal{T} = \{\tau_0, \dots \tau_T\}$
Output: Time-series pool-seq data for $R$ replicates of a single locus $\{\mathbf{c}^{(r)}\}$ and $\{\mathbf{d}^{(r)}\}$.
for $r \leftarrow 1$ to $R$ do
    for $t \leftarrow \tau_0$ to $\tau_T$ do
        $2N\nu_t \sim \text{Binomial}(2N, \nu_{t-1})$;
        if $t \in \mathcal{T}$ then
            $d_t^{(r)} \sim \text{Poisson}(\lambda_{\tau_i})$ ;
            $2ny_t \sim \text{Binomial}(2n, \nu_t)$;
            $c_t^{(r)} \sim \text{Binomial}(d_t^{(r)}, y_t)$;
        end
    end
end
```

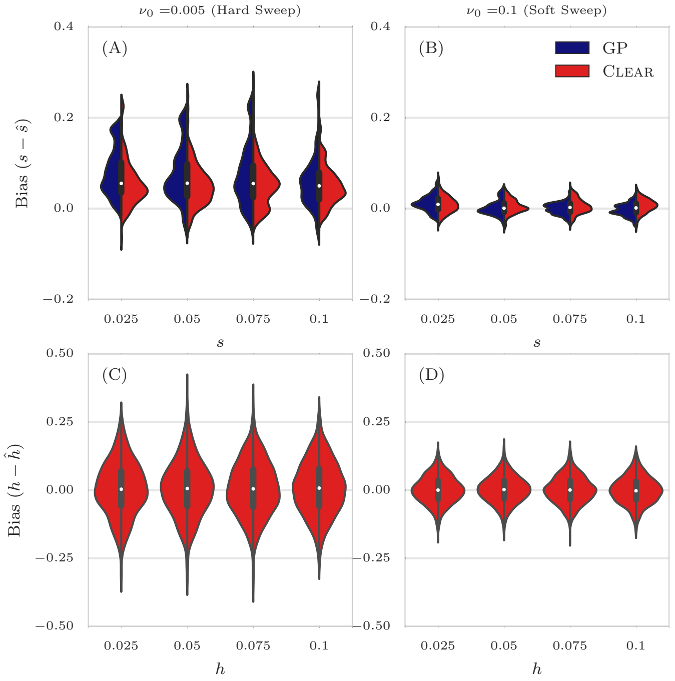
S1 Fig: **The Generative Process for Dynamic Pool-seq Data.**



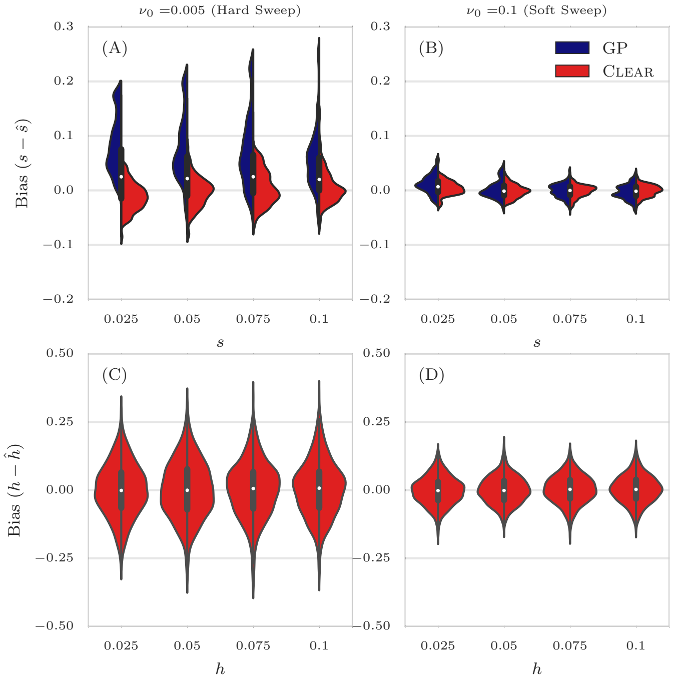
S2 Fig: **Site Frequency Spectrum.**
Theoretical and Empirical SFS in a 50Kbp region for a neutral population of 200 individuals when $N_e = 10^6$ and $\mu = 10^{-9}$. The x-axis corresponds to site frequency, and the y-axis to the number of variants with a specific frequency. In a neural population, majority of the variations stand in low frequency.



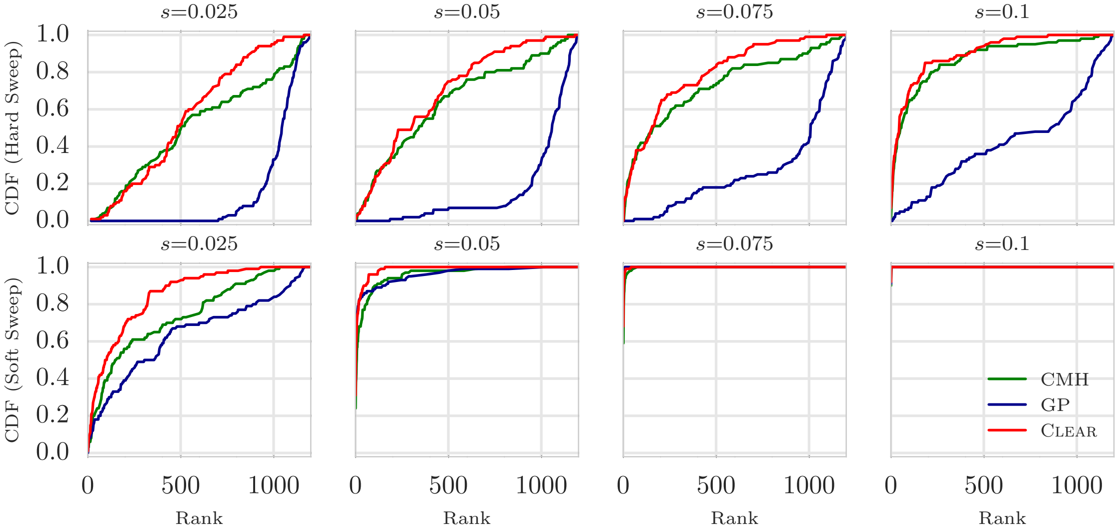
S3 Fig: **Coverage heterogeneity in time series data.**
Each panel shows the read depth for 3 replicates of the data from a study of *D. melanogaster* adaptation to alternating temperatures data (see section 3.1). Heterogeneity in depth of coverage is seen between replicates, and also at different time points, in all 4 variants. None of these sites pass the the hard filtering with minimum depth of 30.



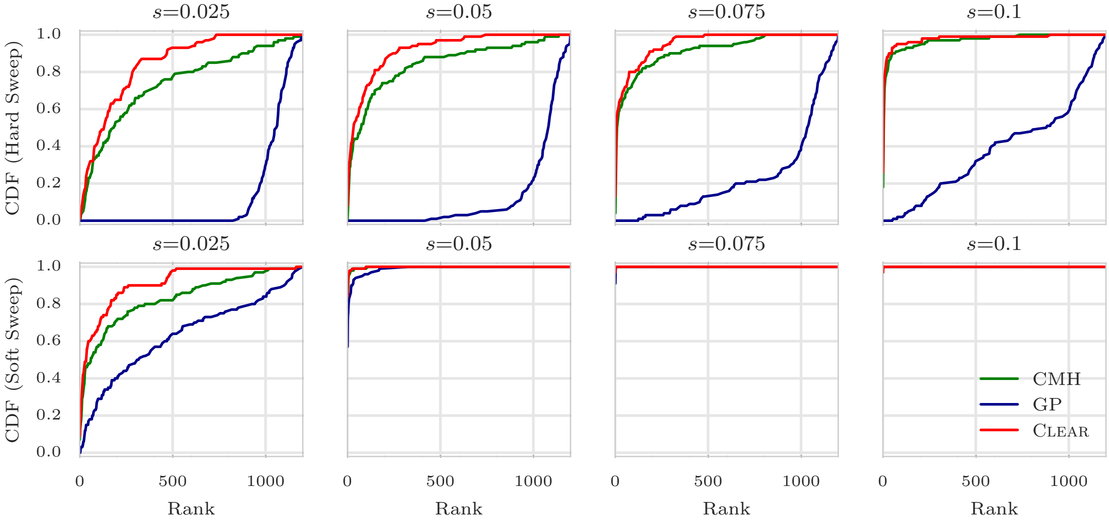
S4 Fig: **Distribution of bias for 30× coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and Clear (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = 30$ (Panels A,B). GP and Clear have similar variance in estimates of s for soft sweep, while Clear provides lower variance in hard sweep. Also see S2 Table. Panels C,D show the variance in the estimation of h.



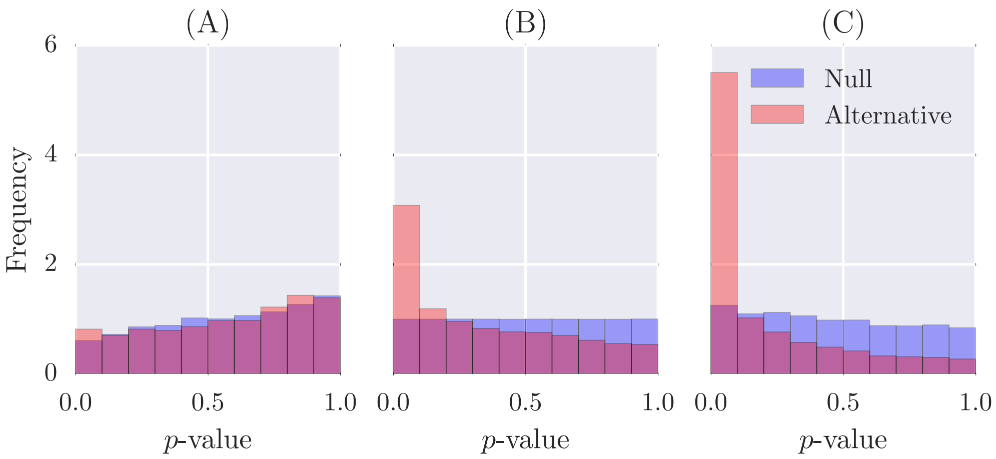
S5 Fig: **Distribution of bias for infinite coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and Clear (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = \infty$ (Panels A,B). GP and Clear have similar variance in estimates of s for soft sweep, while Clear provides lower variance in hard sweep. Also see S2 Table. Panels C,D show the variance in the estimation of h.



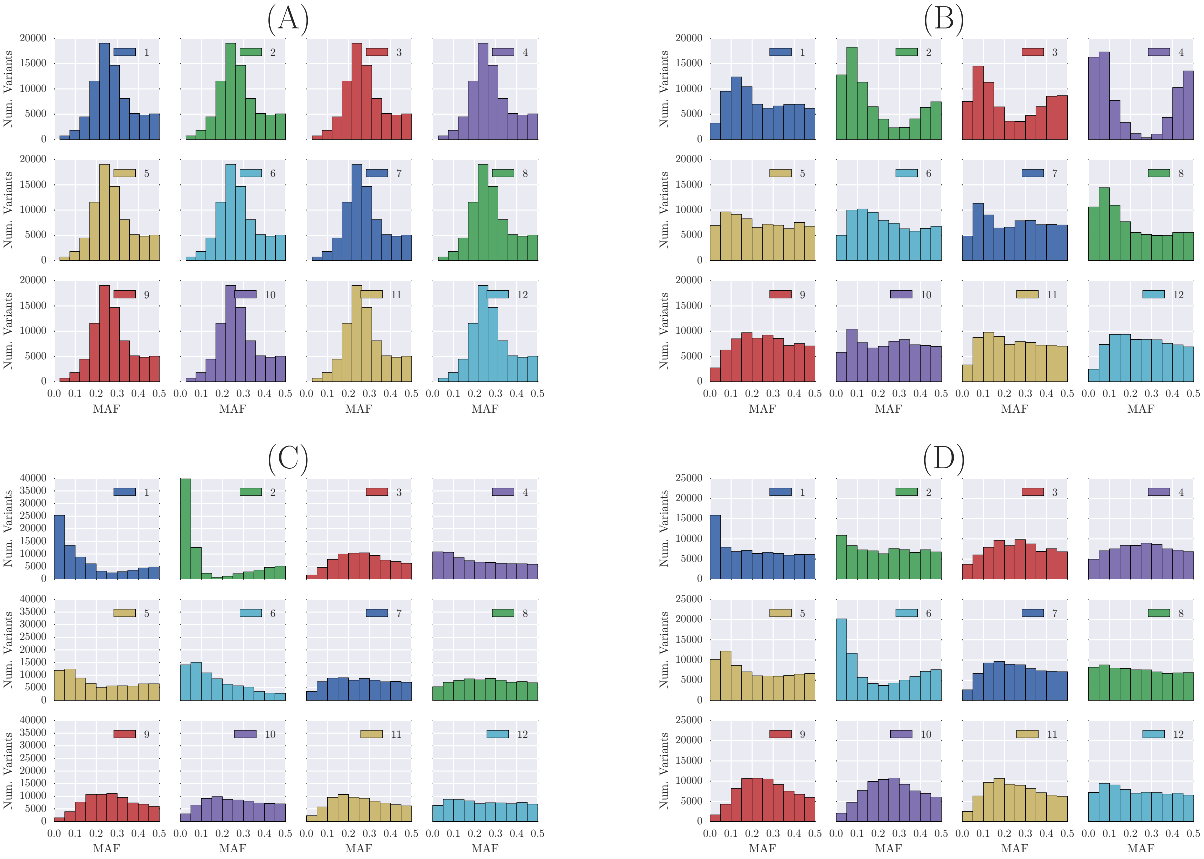
S6 Fig: **Ranking performance for 30× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (H score), Gaussian Process (GP), and Cochran Mantel Haenszel (CMH), for different values of selection coefficient s and initial carrier frequency.



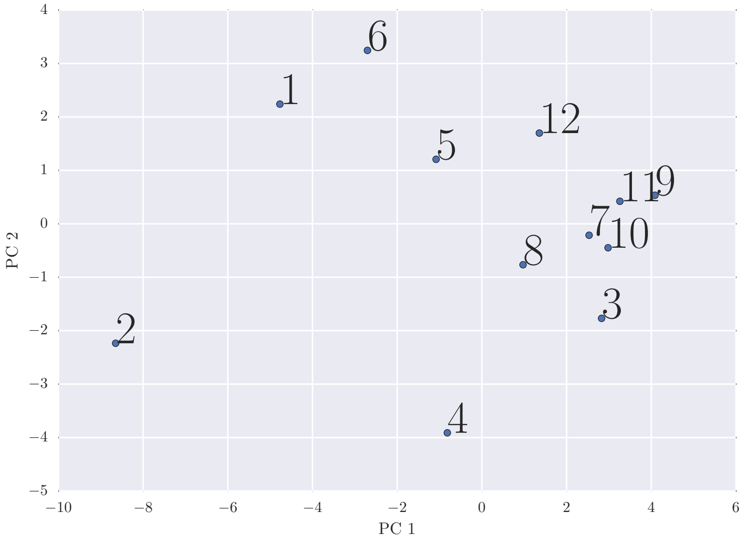
S7 Fig: **Ranking performance for 300× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (H score), Gaussian Process (GP), and Cochran Mantel Haenszel (CMH), for different values of selection coefficient s and initial carrier frequency.



S8 Fig: **Distribution of p-values.** Distribution of p-values of Clear in null simulations and experimental data when $N = 250$. Panel (A),(C) shows the effect of under estimations ($\widehat{N} = 100$) and over-estimation ($\widehat{N} = 500$) of population size in computing p-values, and panel (B) shows the distribution of $p-$values when unbiased estimate is used to create simulations. .



S9 Fig: **Site frequency spectrum of the Yeast dataset.** Whole-genome site frequency spectrum of the Yeast dataset at generations 0 (A), 180 (B), 360 (C) and 540 (D). Some replicates, e.g. replicate 2, undergoing severe demographic events.



S10 Fig: **Population similarity.** Principle component analysis of the 12 replicates throughout the experiment, showing that some populations exhibiting distinct frequency spectra.

S1 Table: **Average of power for detecting selection.**

Hard Sweep			Soft Sweep		
λ	Method	Avg Power	λ	Method	Avg Power
300	Clear	34	300	Clear	69
300	CMH	12	300	CMH	69
300	FIT	2	300	GP	61
300	GP	0	300	FIT	8
100	Clear	31	100	Clear	67
100	CMH	4	100	CMH	60
100	FIT	2	100	GP	59
100	GP	0	100	FIT	1
30	Clear	20	30	Clear	57
30	FIT	2	30	GP	53
30	CMH	0	30	CMH	39
30	GP	0	30	FIT	3

Average power is computed for 8000 simulations with $s \in \{0.025, 0.05, 0.075, 0.1\}$. Frequency Increment Test (FIT), Gaussian Process (GP), Clear ($\mathcal{H}$ statistic) and Cochran Mantel Haenszel (CMH) are compared for different initial carrier frequency ν_0. For all sequencing coverages, Clear outperform other methods. When coverage is not high ($\lambda \in \{30, 100\}$) and initial frequency is low (hard sweep), Clear significantly perform better than others.

S2 Table: **Mean and standard deviation of the distribution of bias ($s - \hat{s}$) of 8000 simulations with coverage $\lambda = 100\times$ and $s \in \{0.025, 0.05, 0.075, 0.1\}$.**

Method	ν_0	Mean	STD
GP	0.005	0.073	0.061
Clear	0.005	0.016	0.035
GP	0.1	0.002	0.016
Clear	0.1	0.002	0.013

S3 Table: **Overlapping genes with the 174 candidate variants.**

index	FBgn	CHROM	start	end	name
1	FBgn0052832	2L	16878326	16879290	CG32832
2	FBgn0032618	2L	16879517	16886319	CG31743
3	FBgn0085342	2L	16879517	16886319	CG34313
4	FBgn0040985	2L	16887109	16887966	CG6115
5	FBgn0261671	2L	16888490	16917052	tweek
6	FBgn0026150	2L	16908229	16910418	ApepP
7	FBgn0262355	2L	16944723	16945374	CR43053
8	FBgn0053179	2L	16973091	16993984	beat-IIIb
9	FBgn0040674	2R	2725579	2726560	CG9445
10	FBgn0265935	2R	2749506	2760223	coro
11	FBgn0033110	2R	2760501	2763324	CG9447
12	FBgn0033113	2R	2768500	2770912	Spn42Dc
13	FBgn0028988	2R	2770785	2772378	Spn42Dd
14	FBgn0033115	2R	2773057	2775767	Spn42De
15	FBgn0050158	2R	2779265	2810118	CG30158
16	FBgn0036421	3L	14362025	14362807	CG13481
17	FBgn0262580	3L	14375013	14376399	CG43120
18	FBgn0036422	3L	14393869	14395825	CG3868
19	FBgn0087007	3L	14405928	14529376	bbg
20	FBgn0036426	3L	14510925	14511575	CG9592
21	FBgn0036427	3L	14512860	14514790	CG4613
22	FBgn0023531	X	1567143	1586801	CG32809
23	FBgn0023130	X	1587648	1589922	a6
24	FBgn0025378	X	1602839	1604215	CG3795
25	FBgn0025391	X	1629978	1648098	Scgdelta
26	FBgn0261548	X	1667752	1747700	CG42666
27	FBgn0026086	X	1667758	1682098	Adar
28	FBgn0026090	X	1751922	1753004	CG14812
29	FBgn0023522	X	1821716	1824550	CG11596
30	FBgn0029941	X	7175122	7299830	CG1677
31	FBgn0029944	X	7218247	7222839	Dok
32	FBgn0029946	X	7240292	7241539	CG15034

References

[1] Guillaume Achaz. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258, 2009.

[2] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[3] Joshua M Akey. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome research*, 19(5):711–722, 2009.

[4] Eric C Anderson, Ellen G Williamson, and Elizabeth A Thompson. Monte Carlo evaluation of the likelihood for Ne from temporally spaced samples. *Genetics*, 156(4):2109–2118, 2000.

[5] Frédéric Ariey, Benoit Witkowski, Chanaki Amaratunga, Johann Beghain, Anne-Claire Langlois, Nimol Khim, Saorin Kim, Valentine Duru, Christiane Bouchier, Laurence Ma, and Others. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature*, 505(7481):50–55, 2014.

[6] James G Baldwin-Brown, Anthony D Long, and Kevin R Thornton. The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. *Molecular biology and evolution*, page msu048, 2014.

[7] Rowan D H Barrett, Sean M Rogers, and Dolph Schluter. Natural selection on a major armor gene in threespine stickleback. *Science*, 322(5899):255–257, 2008.

[8] Jeffrey E Barrick and Richard E Lenski. Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827–839, 2013.

[9] Jeffrey E Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E Lenski, and Jihyun F Kim. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243–1247, 2009.

[10] Mark A Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

[11] Alan O Bergland, Emily L Behrman, Katherine R O'Brien, Paul S Schmidt, and Dmitri A Petrov. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS Genet*, 10(11):e1004775, 2014.

[12] Todd Bersaglieri, Pardis C Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F Schaffner, Jared A Drake, Matthew Rhodes, David E Reich, and Joel N Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[13] Pierre Berthier, Mark A Beaumont, Jean-Marie Cornuet, and Gordon Luikart. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics*, 160(2):741–751, 2002.

[14] Jonathan P Bollback and John P Huelsenbeck. Clonal interference is alleviated by high mutation rates in large populations. *Molecular biology and evolution*, 24(6):1397–1406, 2007.

[15] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of 2Nes from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.

[16] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, and Others. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5):e1000083, 2008.

[17] Molly K Burke, Joseph P Dunham, Parvin Shahrestani, Kevin R Thornton, Michael R Rose, and Anthony D Long. Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, 467(7315):587–590, 2010.

[18] Molly K Burke, Gianni Liti, and Anthony D Long. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. *Molecular biology and evolution*, page msu256, 2014.

[19] P Daborn, S Boundy, J Yen, B Pittendrigh, and Others. DDT resistance in Drosophila correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. *Molecular Genetics and Genomics*, 266(4):556–563, 2001.

[20] Rachel Daniels, Hsiao-Han Chang, Papa Diogoye Séne, Danny C Park, Daniel E Neafsey, Stephen F Schaffner, Elizabeth J Hamilton, Amanda K Lukens, Daria Van Tyne, Souleymane Mboup, and Others. Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One*, 8(4):e60780, 2013.

[21] Vincent J Denef and Jillian F Banfield. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science*, 336(6080):462–466, 2012.

[22] Michael M Desai and Joshua B Plotkin. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, 180(4):2175–2191, 2008.

[23] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[24] Eyal Elyashiv, Shmuel Sattath, Tina T Hu, Alon Strutsovsky, Graham McVicker, Peter Andolfatto, Graham Coop, and Guy Sella. A Genomic Map of the Effects of Linked Selection in Drosophila. *PLoS Genet*, 12(8):1–24, 2016.

[25] Warren J Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media, 2012.

[26] Gregory Ewing and Joachim Hermisson. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

[27] Shaohua Fan, Matthew E B Hansen, Yancy Lo, and Sarah A Tishkoff. Going global by adapting local: A review of recent human adaptation. *Science*, 354(6308):54–59, 2016.

[28] Justin C Fay and Chung-I Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

[29] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. Identifying signatures of selection in genetic time series. *Genetics*, 196(2):509–522, 2014.

[30] Alison F Feder, Soo-Yon Rhee, Susan P Holmes, Robert W Shafer, Dmitri A Petrov, and Pleuni S Pennings. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *eLife*, 5, jan 2016.

[31] Anna-Sophie Fiston-Lavier, Nadia D Singh, Mikhail Lipatov, and Dmitri A Petrov. Drosophila melanogaster recombination rate calculator. *Gene*, 463(1):18–20, 2010.

[32] Susanne U Franssen, Viola Nolte, Ray Tobler, and Christian Schlötterer. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental Drosophila melanogaster populations. *Molecular biology and evolution*, 32(2):495–509, 2015.

[33] Nandita R Garud, Philipp W Messer, Erkan O Buzbas, and Dmitri A Petrov. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet*, 11(2):e1005004, 2015.

[34] John H Gillespie. *Population genetics: a concise guide*. JHU Press, 2010.

[35] Michael M Gottesman. Mechanisms of cancer drug resistance. *Annual review of medicine*, 53(1):615–627, 2002.

[36] Torsten Günther and Graham Coop. Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1):205–220, 2013.

[37] Matthew Hegreness, Noam Shoresh, Daniel Hartl, and Roy Kishony. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311(5767):1615–1617, 2006.

[38] Kent E Holsinger and Bruce S Weir. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10(9):639–650, 2009.

[39] Christopher J R Illingworth and Ville Mustonen. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics*, 189(3):989–1000, 2011.

[40] Christopher J R Illingworth, Leopold Parts, Stephan Schiffels, Gianni Liti, and Ville Mustonen. Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular biology and evolution*, 29(4):1187–1197, 2012.

[41] Minako Izutsu, Atsushi Toyoda, Asao Fujiyama, Kiyokazu Agata, and Naoyuki Fuse. Dynamics of Dark-Fly Genome Under Environmental Selections. *G3: Genes— Genomes— Genetics*, pages g3—-115, 2015.

[42] Aashish R Jha, Cecelia M Miles, Nodia R Lippert, Christopher D Brown, Kevin P White, and Martin Kreitman. Whole-genome resequencing of experimental populations reveals polygenic basis of egg-size variation in Drosophila melanogaster. *Molecular biology and evolution*, 32(10):2616–2632, 2015.

[43] Ágnes Jónás, Thomas Taus, Carolin Kosiol, Christian Schlötterer, and Andreas Futschik. Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution. *Genetics*, aug 2016.

[44] Tadeusz J Kawecki, Richard E Lenski, Dieter Ebert, Brian Hollis, Isabelle Olivieri, and Michael C Whitlock. Experimental evolution. *Trends in ecology & evolution*, 27(10):547–560, 2012.

[45] Robert Kofler and Christian Schlötterer. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, 28(15):2084–2085, 2012.

[46] Robert Kofler and Christian Schlötterer. A guide for the design of evolve and resequencing studies. *Molecular biology and evolution*, page mst221, 2013.

[47] Gregory I Lang, David Botstein, and Michael M Desai. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188(3):647–661, 2011.

[48] Gregory I Lang, Daniel P Rice, Mark J Hickman, Erica Sodergren, George M Weinstock, David Botstein, and Michael M Desai. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574, 2013.

[49] Quan Long, Fernando A Rabanal, Dazhe Meng, Christian D Huber, Ashley Farlow, Alexander Platzer, Qingrun Zhang, Bjarni J Vilhjálmsson, Arthur Korte, Viktoria Nizhynska, and Others. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature genetics*, 45(8):884–890, 2013.

[50] Anna-Sapfo Malaspinas. Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. *Molecular ecology*, 25(1):24–41, 2016.

[51] Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, 2012.

[52] Frank Maldarelli, Mary Kearney, Sarah Palmer, Robert Stephens, JoAnn Mican, Michael A Polis, Richard T Davey, Joseph Kovacs, Wei Shao, Diane Rock-Kress, and Others. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of virology*, 87(18):10313–10323, 2013.

[53] Nelson E Martins, Vítor G Faria, Viola Nolte, Christian Schlötterer, Luis Teixeira, Élio Sucena, and Sara Magalhães. Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences*, 111(16):5938–5943, 2014.

[54] Iain Mathieson and Gil McVean. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, 2013.

[55] Philipp W Messer and Dmitri A Petrov. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution*, 28(11):659–669, 2013.

[56] Shalini Nair, Denae Nash, Daniel Sudimack, Anchalee Jaidee, Marion Barends, Anne-Catrin Uhlemann, Sanjeev Krishna, François Nosten, and Tim J C Anderson. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Molecular Biology and Evolution*, 24(2):562–573, 2007.

[57] Rasmus Nielsen and James Signorovitch. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical population biology*, 63(3):245–255, 2003.

[58] Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic scans for selective sweeps using SNP data. *Genome research*, 15(11):1566–1575, 2005.

[59] Pablo Orozco-ter Wengel, Martin Kapun, Viola Nolte, Robert Kofler, Thomas Flatt, and Christian Schlötterer. Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular ecology*, 21(20):4931–4941, 2012.

[60] Tugce Oz, Aysegul Guvenek, Sadik Yildiz, Enes Karaboga, Yusuf Talha Tamer, Nirva Mumcuyan, Vedat Burak Ozan, Gizem Hazal Senturk, Murat Cokol, Pamela Yeh, and Others. Strength of selection pressure is an important parameter contributing to the complexity of antibiotic resistance evolution. *Molecular biology and evolution*, page msu191, 2014.

[61] Bo Peng and Marek Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, 2005.

[62] Edward Pollak. A new method for estimating the effective population size from allele frequency changes. *Genetics*, 104(3):531–548, 1983.

[63] Susan E Ptak and Molly Przeworski. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, 18(11):559–563, 2002.

[64] Sebastian E Ramos-Onsins and Julio Rozas. Statistical properties of new neutrality tests against population growth. *Molecular biology and evolution*, 19(12):2092–2100, 2002.

[65] Brian J Reid, Rumen Kostadinov, and Carlo C Maley. New strategies in Barrett's esophagus: integrating clonal evolutionary theory with clinical management. *Clinical Cancer Research*, 17(11):3512–3519, 2011.

[66] Silvia C Remolina, Peter L Chang, Jeff Leips, Sergey V Nuzhdin, and Kimberly A Hughes. Genomic basis of aging and life-history evolution in Drosophila melanogaster. *Evolution*, 66(11):3390–3403, 2012.

[67] Roy Ronen, Nitin Udpa, Eran Halperin, and Vineet Bafna. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–193, 2013.

[68] P C Sabeti, S F Schaffner, B Fry, J Lohmueller, P Varilly, O Shamovsky, A Palma, T S Mikkelsen, D Altshuler, and E S Lander. Positive natural selection in the human lineage. *science*, 312(5780):1614–1620, 2006.

[69] Stanley A Sawyer and Daniel L Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, 1992.

[70] Christian Schlötterer, R Kofler, E Versace, R Tobler, and S U Franssen. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity*, 114(5):431–440, 2015.

[71] Joshua G Schraiber, Steven N Evans, and Montgomery Slatkin. Bayesian inference of natural selection from allele frequency time series. *Genetics*, 203(1):493–511, 2016.

[72] Tatum S Simonson, Yingzhong Yang, Chad D Huff, Haixia Yun, Ga Qin, David J Witherspoon, Zhenzhong Bai, Felipe R Lorenzo, Jinchuan Xing, Lynn B Jorde, and Others. Genetic evidence for high-altitude adaptation in Tibet. *Science*, 329(5987):72–75, 2010.

[73] Brad Spellberg, Robert Guidos, David Gilbert, John Bradley, Helen W Boucher, W Michael Scheld, John G Bartlett, John Edwards, Infectious Diseases Society of America, and Others. The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. *Clinical Infectious Diseases*, 46(2):155–164, 2008.

[74] Matthias Steinrücken, Anand Bhaskar, and Yun S Song. A novel spectral method for inferring general diploid selection from time series genetic data. *The annals of applied statistics*, 8(4):2203, 2014.

[75] Wolfgang Stephan, Yun S Song, and Charles H Langley. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, 172(4):2647–2663, 2006.

[76] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[77] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

[78] Jonathan Terhorst, Christian Schlötterer, and Yun S Song. Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. *PLoS Genet*, 11(4):e1005069, 2015.

[79] Ray Tobler, Susanne U Franssen, Robert Kofler, Pablo Orozco-terWengel, Viola Nolte, Joachim Hermisson, and Christian Schlötterer. Massive habitat-specific genomic response in D. melanogaster populations during experimental evolution in hot and cold environments. *Molecular biology and evolution*, 31(2):364–375, 2014.

[80] Hande Topa, Ágnes Jónás, Robert Kofler, Carolin Kosiol, and Antti Honkela. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, page btv014, 2015.

[81] Thomas L Turner, Andrew D Stewart, Andrew T Fields, William R Rice, and Aaron M Tarone. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS Genet*, 7(3):e1001336, 2011.

[82] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting natural selection in genomic data. *Annual review of genetics*, 47:97–120, 2013.

[83] Jinliang Wang. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical research*, 78(03):243–257, 2001.

[84] Robin S Waples. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391, 1989.

[85] David Williams and David Williams. *Weighing the odds: a course in probability and statistics*, volume 548. Springer, 2001.

[86] Ellen G Williamson and Montgomery Slatkin. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, 152(2):755–761, 1999.

[87] Scott H Williamson, Melissa J Hubisz, Andrew G Clark, Bret A Payseur, Carlos D Bustamante, and Rasmus Nielsen. Localizing recent adaptive evolution in the human genome. *PLoS Genet*, 3(6):e90, 2007.

[88] Mark A Winters, Robert M Lloyd Jr, Robert W Shafer, Michael J Kozal, Michael D Miller, and Mark Holodniy. Development of elvitegravir resistance and linkage of integrase inhibitor mutations with protease and reverse transcriptase resistance mutations. *PloS one*, 7(7):e40514, 2012.

[89] Xin Yi, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E Pool, Xun Xu, Hui Jiang, Nicolas Vinckenbosch, Thorfinn Sand Korneliussen, and Others. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–78, 2010.

[90] Hiba Zahreddine and K L Borden. Mechanisms and insights into drug resistance in cancer. *Front Pharmacol*, 4(28.10):3389, 2013.

[91] Dan Zhou, Nitin Udpa, Merril Gersten, DeeAnn W Visk, Ali Bashir, Jin Xue, Kelly A Frazer, James W Posakony, Shankar Subramaniam, Vineet Bafna, and Gabriel G. Haddad. Experimental selection of hypoxia-tolerant Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 108(6):2349–2354, 2011.