

COMALE: Analyzing Adaptive Experimental Evolution with Pooled Sequencing Data

Arya Iranmehr¹, Ali Akbari¹, Christian Schlötterer², and Vineet Bafna³

¹Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA.

²Institut für Populationsgenetik, Vetmeduni, Vienna, Austria.

³Computer Science & Engineering, University of California, San Diego, La Jolla, CA 92093, USA

Abstract

Experimental evolution (EE) studies are powerful tools for observing molecular evolution “in-action” in populations sampled in controlled and natural environments. The advent of sequencing technologies has made whole-genome and whole-population sampling possible even for eukaryotic organisms with large genomes, and allowed us to locate the genes and variants responsible for genetic adaptation. While many computational tests have been developed for detecting regions under selection, they are mainly designed for static (single time) data, and work best when the favored allele is close to fixation.

EE studies provide samples over multiple time points, underscoring the need for tools that can exploit the data. At the same time, EE studies are constrained by the limited time span since onset of selection, depending upon the generation time for the organism. This constraint impedes adaptation and optimization studies, as the population can only be propagated for a small number of generations, relative to the fixation time of the favored allele. Moreover, sequencing depths of pooled experiments vary across replicates and time points, leading to ascertainment bias.

In this article, we directly address these issues while developing tools for identifying selective sweep in short-term and pool-sequenced experimental evolution of sexual organisms and propose Composite Of MArkovian Likelihoods for Experimental evolution (COMALE) statistic. Extensive simulations show that COMALE achieves higher detection power over a wide range of parameters, including both soft and hard sweep scenarios, and is many orders of magnitude faster than other methods, making genomic scans feasible. We apply the COMALE statistic to the controlled experimental evolution of *D. melanogaster* to detect adaptive genes/alleles under alternating cold and hot temperatures, and identify many genes that are adapting to the selection constraints.

1 Introduction

Natural selection is the key force in evolution, and a mechanism by which populations can adapt to external ‘selection’ constraints. Examples of adaptation abound in the natural world, including for example, classic examples like lactose tolerance in Northern Europeans [10], human adaptation to high altitudes [62, 77], but also drug resistance in pests [15], HIV [25], cancer [27, 78], malarial parasite [4, 46], and other antibiotic resistance [63]. In each of these examples, understanding the genetic basis of adaptation can provide actionable information, underscoring the importance of the problem.

Modern experimental evolution refers to the study of the evolutionary processes of a model organism at genomic level in a controlled [8, 11, 29, 38, 39, 49, 50] or natural [6, 9, 16, 17, 42, 54, 76] environment. Recent advances in whole genome sequencing have enabled us to sequence populations

at a reasonable cost even when the genomes are large. Perhaps more important for experimental evolution studies, we can now obtain and sequence *longitudinal time-series data*, sampling populations over a period of time, and investigating the dynamics of evolution at molecular level. Although constraints such as small population sizes, limited timescales, and oversimplified laboratory environments limit the interpretation of EE results, these studies are increasingly being used to test hypotheses regarding mutation rate, inbreeding, environmental variability, sexual selection & conflict, kin selection and cooperation, life history and sex allocation, sexual reproduction and mating systems, behavior and cognition, and host-parasite interactions [35]. In many cases, they provide more accurate inferences than static data analysis [13, 18, 60]. Dynamic data is also being used to estimate model parameters including population size [51, 67, 72–74], strength of selection [12, 31, 32, 41, 44, 64, 67], allele age [41] recombination rate [67], mutation rate [7, 67], quantitative trait loci [5] and for tests of neutrality hypotheses [9, 14, 24, 67].

While different types of EE experiments have been proposed [7, 61] our focus here is the adaptive evolution of multi-cellular sexual organisms. For simplicity, we assume continuous culture, fixed population size, and for the most part, positive single locus selection (only one favored mutation). This regime has been extensively applied, often with *Drosophila* as the model organism of choice. It has been used to identify adaptive genes in longevity and aging [14, 55] (600 generations), courtship song [70] (100 generations), hypoxia tolerance [79] (200 generations), adaptation to new temperatures [49, 68] (59 generations), egg size [34] (40 generations), C virus resistance [43] (20 generations), and dark-fly [33] (49 generations) experiments. Our methods can be generalized to other sexual populations, but are separate from the analysis of asexual populations. In sexual populations, the mutation responding to the selection constraint is linked to neighboring mutations, but the linkage decays rapidly due to recombination. Together, this provides a strengthened signal around the favored mutation. Therefore, methods for identifying genomic regions under selection in sexual populations often focus on analyzing genomic regions, rather than single sites.

The task of identification of a selection event in natural or experimental evolution can be addressed at different levels of specificity. At the coarsest level, identification could simply refer to deciding whether some genomic region 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 itself. Finally, *estimation of model parameters* such as strength of selection and over-dominance at the site can provide a comprehensive description of the selection process.

A wide range of computational methods [71] have been developed to detect regions under positive selection. A majority of the existing methods focus on static data analysis—analyzing a single sample of the population at a specific time, either during the sweep, or subsequent to fixation of the favored allele. For instance, reduction in genetic diversity [23, 57, 66] in allele-frequency data, prevalence of long haplotypes [58, 71] in haplotype (phased) data, population differentiation [14, 30] in multiple-population data and others. An important component of many methods is the analysis of the Site Frequency Spectrum (SFS) for an example, see Fig. S11, to identify departure from neutrality. Classical examples including Tajima’s D [66], Fay and Wu’s H [23], Composite Likelihood Ratio [48], were all shown to be weighted linear combination of the SFS values [1]. However, the SFS of a sample under selection changes with time since onset of selection, the frequency of the favored allele at the onset of selection, (high initial frequency of the favored allele is also referred to as standing variation) and other parameters. Not surprisingly, the power of these SFS based tests depends upon different selection parameters. This insight can be exploited to learn the right weights for each regime and improve detection accuracy [57].

While successful, these methods are prone to both, false negatives [45], as also false-discoveries due to confounding factors such as demography, including bottleneck and population expansions,

and ascertainment bias [3, 45, 47, 52, 53]. Nevertheless, SFS based tests are simple and inexpensive and continue to be used, often in combination with other tests [3, 71]. It remains an open question if the analysis of dynamic time-series data can improve the power of SFS based methods.

Relative to the analysis of static samples, fewer tests-of-selection for dynamic time-series data have been proposed. Often, existing tests for static data are adopted for dynamic data with two time-points. Zhu *et al.* [79] used the ratio of the estimated population size of case and control populations to compute test statistic for each genomic region. Burke *et al.* [14] applied Fisher exact test to the last observation of data on case and control populations. Orozco-Terwengel *et al.* [49] 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.* [70] proposed the diffStat statistic to test whether the change in allele frequencies of two population deviates from the distribution of change in allele frequencies of two drifting populations. Bergland *et al.* [9] 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.* [34] computed test statistic of generalized linear-mixed model directly from read counts. Bollback *et al.* [12] provided diffusion approximation to the continues Wright Fisher Markov process and estimated s numerically and performed standard likelihood ratio test under χ^2 distribution.

It is only recently that direct tests for analyzing time-series data have been developed. Using (continuous-time continuous-state) Brownian motion process, Feder *et al.* [24] proposed the Frequency Increment Test (FIT) for dynamic allele frequency data. More recently, Song *et al.* [67] proposed empirical Gaussian Process (GP) likelihood ratio test for single and multiple loci dynamic allele frequency data. The heavy computational requirements of these methods makes it difficult to apply them in a whole genome setting. Topa *et al.* [69] extended GP to analyze pooled-sequencing read count data. A key contribution of our paper is the development of a direct, and significantly faster method, COMALE, for detecting selection in short-term experimental evolution with pooled sequencing (read count) data (see Fig. S1 for an example). We show for a wide range of parameters that COMALE provides higher power for detecting selection, is robust to ascertainment bias, estimates model parameters consistently, and localizes beneficial allele more accurately compared to the state-of the art methods, while being orders of magnitude faster.

2 Materials and Methods

Notation. Consider a locus with starting derived allele frequency ν_0 . Frequencies are sampled at T distinct generations specified by $\mathcal{T} = \{\tau_i : 1 \leq \tau_1 < \tau_2, \dots < \tau_T\}$, and denoted by $\boldsymbol{\nu} = \{\nu_1, \dots, \nu_T\}$. Moreover, R replicate measurements are made, and we denote the r -th replicate frequency data as $\boldsymbol{\nu}^{(r)}$.

2.1 The COMALE statistic

To identify if the locus is evolving under positive selection, we follow previous approaches to focus on a parametrized likelihood based model that (a) maximizes the likelihood of the time series data w.r.t. selection and overdominance parameters s, h ; and, (b) computes the log-odds ratio of the likelihood of selection model to the likelihood of neutral evolution/drift model.

Likelihood for Neutral Model. To model neutral evolution, it is natural to model the change in frequency ν_t over time via Brownian motion [24] or Gaussian process [67]. Significant deviations from this Null could be indicative of non-neutrality. However, in our experiments, we found that the Brownian motion approximation is inadequate for small population sizes and low starting

frequencies that are typical in experimental evolution (see Results, and Fig. 2). In fact, other continuous models such as Gaussian process for allele frequencies, are also susceptible to this issue.

Instead, by computing likelihood of data using a discrete-time discrete-state-space Wright-Fisher Markov Chain, we turn the problem of small-population size into an advantage. Consider a neutrally evolving diploid population with N individuals. Define a $2N \times 2N$ transition matrix P , where $P^{(\tau)}[i, j]$ denotes probability of change in allele frequency from $\frac{i}{2N}$ to $\frac{j}{2N}$ in τ generations, solely due to genetic drift. P is defined as follows [21]:

$$P^{(1)}[i, j] = \Pr\left(\nu_{t+1} = \frac{j}{2N} \mid \nu_t = \frac{i}{2N}\right) = \binom{2N}{j} \nu_t^j (1 - \nu_t)^{2N-j}, \quad (1)$$

$$P^{(\tau)} = P^{(\tau-1)} P^{(1)} \quad (2)$$

Note that pre-computing and storing $P^{(\tau)}$ is tractable and numerically stable for controlled experimental evolution experiments where $N \leq 2000$. For larger N , we bin the frequencies values.

Likelihood for Selection Model. Assume that the site is evolving under selection constraints $s, h \in \mathbb{R}$, where s and h denote selection strength and dominance, 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 $\tau_t + 1$ can be estimated using:

$$\begin{aligned} \hat{\nu}_{t+} &= \mathbb{E}[\nu_{t+} | s, h, \nu_t] = \frac{w_{11}\nu_t^2 + w_{01}\nu_t(1 - \nu_t)}{w_{11}\nu_t^2 + 2w_{01}\nu_t(1 - \nu_t) + w_{00}(1 - \nu_t)^2} \\ &= \nu_t + \frac{s(h + (1 - 2h)\nu_t)\nu_t(1 - \nu_t)}{1 + s\nu_t(2h + (1 - 2h)\nu_t)}. \end{aligned} \quad (3)$$

For finite populations, let $Q_{s,h}^{(\tau)}[i, j]$ denote the probability of transition from $\frac{i}{2N}$ to $\frac{j}{2N}$ in τ generations. We model Q as follows (See [21], Pg. 24, Eqn. 1.58-1.59):

$$Q_{s,h}^{(1)}[i, j] = \Pr\left(\nu_{t+} = \frac{j}{2N} \mid \nu_t = \frac{i}{2N}; s, h\right) = \binom{2N}{j} \hat{\nu}_{t+}^j (1 - \hat{\nu}_{t+})^{2N-j} \quad (4)$$

$$Q_{s,h}^{(\tau)} = Q_{s,h}^{(\tau-1)} Q_{s,h}^{(1)} \quad (5)$$

For $s = 0$, Eq. 4 and 5 are identical to Eq. 1 and 2, respectively. The likelihood of observing the trajectory $\boldsymbol{\nu}$ is computed using:

$$\mathcal{L}_M(s, h | \boldsymbol{\nu}) = \Pr(\boldsymbol{\nu}; \nu_0, s, h) = \prod_{t=1}^T \Pr(\nu_t | \nu_{t-1}; \nu_0, s, h) = \prod_{t=1}^T Q_{s,h}^{(\delta_t)}[\hat{i}, \hat{j}], \quad (6)$$

where, $(\hat{i}, \hat{j}) = (\lfloor 2N\nu_{t-1} \rfloor, \lfloor 2N\nu_t \rfloor)$, and $\delta_t = \tau_t - \tau_{t-1}$. Combining the likelihood over independent replicate samples $\boldsymbol{\nu}^{(r)}$, we get:

$$\mathcal{L}_M(s, h | \{\boldsymbol{\nu}^{(r)}\}) = \prod_r \mathcal{L}_M(s, h | \boldsymbol{\nu}^{(r)}). \quad (7)$$

Let \hat{s}, \hat{h} denote the parameters that maximize the likelihood. The simplest form of the test statistic for each variant is given by

$$M = \log\left(\frac{\mathcal{L}_M(\hat{s}, \hat{h} | \{\boldsymbol{\nu}^{(r)}\})}{\mathcal{L}_M(0, 0 | \{\boldsymbol{\nu}^{(r)}\})}\right). \quad (8)$$

Accounting for ascertainment bias using an HMM. In the discussion so far, we assumed that the exact allele frequencies are supplied. However, in most cases, allele frequencies are estimated from genotype data with finite samples or pooled-sequencing data with finite depth of coverage (See Fig. S2). Moreover, the depth at a site varies for different replicates, and different time samples (Fig. S3). To account for this heterogeneity, we extend the Markov chain in Eq. 7 to a Hidden Markov Model for pooled-seq data.

Consider a variant position being sampled at time point $\tau_t \in \mathcal{T}$. We denote the pooled-seq data for that variant as $x_t = \langle c_t, d_t \rangle$ where d_t, c_t represent the read depth, and the count of the derived allele, respectively, at time τ_t . The dynamic data is represented by the sequence $\mathbf{x} = x_1, x_2, \dots, x_T$. Define an HMM with $2N + 1$ states. State i ($0 \leq i \leq 2N$) corresponds to allele frequency $\frac{i}{2N}$. The HMM is stationary in that transition and emission distributions do not change over time. Therefore, at any time step, the probability that state i emits $x = \langle d, c \rangle$ is given by

$$\mathbf{e}_i(x) = \binom{d}{c} \left(\frac{i}{2N} \right)^c \left(1 - \frac{i}{2N} \right)^{d-c}.$$

For $1 \leq t \leq T$, let $\alpha_{t,i}$ denote the probability of emitting the x_1, x_2, \dots, x_t and ending in state i at τ_t . Then, $\alpha_{t,i}$ can be computed using the forward-procedure [19]:

$$\alpha_{t,i} = \left(\sum_{1 \leq j \leq 2N} \alpha_{t-1,j} Q_{s,h}^{(\delta_t)}[j, i] \right) \mathbf{e}_i(x_t) . \quad (9)$$

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

$$\mathcal{L}_H(s, h | \{\mathbf{x}^{(r)}\}) = \prod_{r=1}^R \mathcal{L}_H(s, h | \mathbf{x}^{(r)}) = \prod_{r=1}^R \sum_i \alpha_{T,i}^{(r)} . \quad (10)$$

Similar to Eq. 8, let s^*, h^* denote the parameters that maximize likelihood. The modified likelihood statistic for pool-seq data is given by

$$H = \log \left(\frac{\mathcal{L}_H(s^*, h^* | \{\mathbf{x}^{(r)}\})}{\mathcal{L}_H(0, 0 | \{\mathbf{x}^{(r)}\})} \right) . \quad (11)$$

Composite Likelihood. Consider a genomic region defined by a collection of segregating sites L , with little or no recombination between sites. To test if some site in the region is under selection, we would like to compute composite likelihood functions under selection and drift regimes. The dynamic frequency of a site $\ell \in L$ is governed by selection, drift, and also linkage with other sites. Computing exact likelihood requires estimating the linkage throughout time which is computationally expensive.

Instead, we simplify the computation by multiplying individual SNP likelihoods, to get a Composite Likelihood Ratio score (CLR) [48, 75]. CLR is a popular and effective method for detecting regions under selection [71].

Let $\Lambda_M(\ell)$ (or, $\Lambda_H(\ell)$) denote the likelihood ratio score for each site ℓ in L . Under positive selection, a site ℓ increases in frequency, and sites on the same lineage correspondingly increase in frequency too, at the expense of sites on other branches, which decrease or drift in frequency. In computing composite likelihood for a region, we can choose to only include sites whose likelihood ratio score is above a certain threshold. For percentile cut-off π , let $L_\pi \subseteq L$ denote the set of sites

whose likelihood ratio scores had percentile π or better. For all π , the modified CLR statistic for Markov chain and HMM is computed using:

$$\mathcal{M}_\pi = \frac{1}{|L_\pi|} \sum_{\ell \in L_\pi} M_\ell , \quad (12)$$

$$\mathcal{H}_\pi = \frac{1}{|L_\pi|} \sum_{\ell \in L_\pi} H_\ell . \quad (13)$$

Note that $\mathcal{M}_{100}, \mathcal{H}_{100}$ (respectively, $\mathcal{M}_0, \mathcal{H}_0$) correspond to the CLR after choosing the best (respectively, all) site(s) in the region.

Estimating parameters. Depending on data (read count or allele frequency) the optimal value of the parameters can be found by

$$\hat{s}, \hat{h} = \arg \max_{s,h} \sum_r^R \log \left(\mathcal{L}_\mathcal{M}(s, h | \boldsymbol{\nu}^{(r)}) \right), \quad \text{or,} \quad (14)$$

$$\hat{s}, \hat{h} = \arg \max_{s,h} \sum_r^R \log \left(\mathcal{L}_\mathcal{H}(s, h | \boldsymbol{x}^{(r)}) \right). \quad (15)$$

where likelihoods are defined in Eq. 7 and Eq. 10, respectively. Objective functions Eqs. 14, 15 are optimized using a simple grid search and the operations are vectorized for all variants. Using this simple optimization, solution is obtained in a reasonable time (see Fig. 5).

Precomputing Transition Matrices. COMALE requires pre-computation of matrices P and $Q_{s,h}$ for the entire range of s, h values. Pre-computation of 1313 transition matrices for $s \in \{-0.5, -0.49, \dots, 0.5\}$ and $h \in \{-1, -0.75, \dots, 2\}$ took less than 20 minutes (≈ 1 second per matrix) on a desktop computer with a Core i7 CPU and 16GB of RAM.

2.2 Extending Site Frequency Spectrum based tests for time series data

The site frequency spectrum (SFS) is a mainstay of tests of neutrality and selection, and can be computed using allele frequencies (does not need haplotypes). Following Fu, 1995 [26], any linear combination of the site frequencies is an estimate of θ . However, under non-neutral conditions, different linear combinations behave differently. Therefore, many popular tests of neutrality either compute differences of two estimates of θ , or perform cross-population tests comparing the θ estimates in two different populations [1, 57, 59].

We asked if SFS-based tests could be adapted for time-series data. A simple approach is to use cross-population SFS tests on the populations at time 0 (before onset of selection), and at time sample τ_t , for each t . However, these tests are not independent. Evans *et al.* [20] developed diffusion equations for evolution of SFS in time series, but they are difficult to solve. Instead, we derive a formula for computing D_t , the dynamic of Tajima's D at generation t (see Fig. S11), in terms of initial statistic D_0 , initial carrier frequency ν_0 and s , as

$$D_t = D_0 - \log(1 - \nu_t) \frac{W_0}{\log(2N)} - \nu_t^2 \Pi_0 , \quad (16)$$

where W_0 and Π_0 are Watterson and Tajima estimates of θ in the initial generation (Appendix 6.2). Similarly, we show (Appendix 6.3), that the dynamics of the H statistic are directly related to

average of Haplotype Allele Frequency (HAF) score [56], and can be written as a function of ν_t as follows:

$$nH_t = \theta\nu_t \left(\frac{\nu_t + 1}{2} - \frac{1}{(1 - \nu_t)n + 1} \right) + \theta(1 - \nu_t) \left(\frac{n + 1}{2n} - \frac{1}{(1 - \nu_t)n + 1} \right) \quad (17)$$

In both cases, ν_t itself can be written as a function of s, t (Eq. S2). This allows us to compute likelihood functions $\mathcal{L}_S(s, h; \{D_t\})$ or $\mathcal{L}_S(s, h; \{H_t\})$. Then, a likelihood ratio, similar to Eqns. 8, 11 provides a statistic for detecting selection in each window.

However, as ν_0 and D_0 are often unknown in the sampling from natural population experiments, we cannot directly use Eqns. 16, 17. Instead, we heuristically aggregate statistics throughout time to compute time-series score. See details in Appendix 6.4.

2.3 Simulations

To implement real world scenario for experimental evolution we simulated populations as follows (see also, Fig. 1 B for illustration):

- I. **Creating initial founder line haplotypes.** Using msms program [22], we created neutral populations for F founding haplotypes with *default* parameters `$./msms <F> 1 -t <2μLNe> -r <2rNeL> <L>`, where $F = 200$ is number of founder lines, $N_e = 10^6$ is effective population size, $r = 2 * 10^{-8}$ 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 note that all experiments denoted as ‘soft-sweep’ in the results are soft-sweep when selection acts upon standing variation. 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 $\mathcal{T} = \{10, 20, 30, 40, 50\}$, simuPop was used to perform forward simulation and compute allele frequencies for all of the R replicates. Also, to avoid spurious simulation samples, simulation results were constrained to those samples where the favored allele escaped stochastic loss due genetic drift, and established in all replicates. For hard sweep (and soft sweep) simulations we randomly chosen a site with initial frequency of $\nu_0 = 0.005$ (and $\nu_0 = 0.1$ for soft sweep) to be the favored allele.
- IV. **Sequencing Simulation.** Give allele frequency trajectories we sampled depth of each site identically and independently from $\text{Poisson}(\lambda)$, where $\lambda \in \{30, 100, \infty\}$ is the coverage for the experiment. Once depth d are drawn for the site with frequency ν , the number of derived allele read counts c are sampled according to $\text{Binomial}(d, \nu)$. For the experiments with finite depth the tuple (c, d) is the input data for each site, and for infinite depth experiments simply allele frequency is given and Markov Chain is evaluated for the HMM method, i.e. $\mathcal{H} = \mathcal{M}$.

We also conducted simulations to test methods in the sampling-from-natural-populations setting. Unlike the experimental evolution simulation, here mutations after onset of selection are taken into account. msms simulator is used to forward simulate a population with $N_e = 10^4$, $\nu_0 = 10^{-4}$ and record SFS of a 50Kbp region as well as the frequency. The reset of parameters are chosen the same as experimental evolution (see Fig. 1A).

3 Results

Modeling neutral trajectories in finite populations. We tested the closeness of fit for the Markov Likelihood as a model for neutral trajectories, compared to Brownian motion. We performed 150K simulations for different values of ν_0 ($\nu_0 \in \{0.005, 0.1\}$) and time τ generations $\tau \in \{1, 10, 100\}$. Fig. 2 shows that Brownian motion is inadequate when ν_0 is far from 0.5, and when sampling is done after many generations $\tau > 1$. (sampling times are sparse). In most experimental evolution scenarios, a site is unlikely to have frequency close to 0.5, and the starting frequencies are usually much smaller. Moreover, sampling times are sparse. In typical *Drosophila* experiments for example, < 10 time points are samples in a span of 100 generations from the onset of selection [49, 79].

In contrast, Fig. 2A-F also shows that Markov Likelihood predictions (Eq. 5) are highly consistent with empirical data for a wide range of simulation parameters. We also tested the model under selection by conducting 100K simulations with selection strength $s = 0.1$ on a site with initial frequency $\nu_0 = 0.005$ and sampling after τ ($\tau \in \{1, 10, 100\}$) generations. The empirical and theoretical distributions tracked closely (Fig. 2G-I).

Detection Power. To evaluate performance of each method, define power as the fraction of true positives identified with false-positive rate ≤ 0.05 (Fig. S4). Before comparing against other methods, we first evaluated the use of HMM and Markov chain composite likelihoods (\mathcal{H} and \mathcal{M}) in COMALE with different percentile-cutoffs π (Eq. 13) under different sequence coverage settings. We chose the sequence depth of each marker by identically and independently sampling from a Poisson distribution with parameter $\lambda \in \{30, 100, \infty\}$. We computed power for \mathcal{H}_π and \mathcal{M}_π , with $\pi \in \{0, 0.99, 100\}$. See Fig. 3. The performance of HMM is robust with coverage, while the Markov chain's power decays for low coverage values. Also, the composite likelihood for all variants ($\pi = 0$) shows the highest power (Fig. 3, and Table S1). Therefore, we used $\mathcal{H} = \mathcal{H}_0$ as the default statistic for all subsequent calculations.

We compared the power of \mathcal{H} , \mathcal{M} Gaussian process (GP) [67], FIT [24] statistics. All methods other than COMALE convert read counts to allele frequencies and compute their test statistic. For each experiment, (specified with values for selection coefficient s , starting allele frequency ν_0 , coverage λ , sampling time schedule \mathcal{T} , and number of replicates R), we conducted 1000 simulations. Half of these modeled neutral evolution and the rest were under selection. COMALE shows the highest power across these range of parameters (Fig. 4). However, for perfect coverage ($\lambda = \infty$), and high selection coefficient $s = 0.1$, GP has comparable power.

Running Time. As COMALE does not compute full likelihoods, or explicitly model linkage between sites, the complexity of computing likelihoods is $\mathcal{O}(TR)$, and can be efficiently vectorized for multiple replicates and loci. Therefore, it is expected to be faster than other approaches like Gaussian Process (GP) [67]. We conducted 1000 simulations and measured running time for COMALE and GP. COMALE is $\sim 10^3 \times$ faster than single locus GP (Fig. 5), while maintaining high power. These times have a practical consequence. To run GP in the single locus mode on the entire *Drosophila* genome of a small sample (2M variant sites), would take 1444 CPU-hours (≈ 1 CPU-month). In contrast, COMALE ran in 22 CPU-hours in addition to 30 mins. for precomputation of transition matrices. In addition, core computations of COMALE are matrix products, which can be efficiently vectorized. We vectorized operations using `numba` package, which reduced run time of over 1.5M variant sites to less an hour.

SFS for Detection in Natural Samples. We did not show the SFS based statistics in Fig. 4 as they did not perform better than random. In many experimental evolution settings, we sample

a restricted set of F founder lines, where $F \ll N_e$ (Fig. 1B). This creates a severe bottleneck, confounding SFS. The Supp. Fig. S5 demonstrates the effect of experimental evolution on different SFS statistics under neutral evolution for 1000 simulations. A second problem with using SFS for experimental evolution is that the sampling starts right after the onset of experimentally induced selection, and the favored allele may not reach high enough frequency to modify the site frequency spectrum (Suppl. Fig. S6).

However, in experiments involving naturally occurring populations, even if the span of the time-series is small, the onset of selection might occur many generations prior to sampling. To test performance of SFS-based statistics in natural evolution, we conducted 200 (100 neutral and 100 sweep) forward simulations for different values of s, λ using $N_e = 10K$ and accumulating new mutations. The start of sampling was done at a randomly picked time subsequent to the onset of selection in two distinct scenarios. Let $t_{\nu=x}(s, N_e)$ denote the expected time (in generations) required to reach carrier frequency x in a hard sweep and $U[a, b]$ denote discrete uniform distribution in the interval $[a, b]$. First we considered the case when start of sampling is chosen throughout the whole sweep. i.e., $\tau_1 \sim U[1, t_{\nu=1}(s, N_e)]$ (Fig. 6A). Next, we considered sampling start time chosen nearer to fixation of the favored allele, i.e., $\tau_1 \sim U[t_{\nu=0.9}(s, N_e), t_{\nu=1}(s, N_e)]$ (Fig. 6B). In both scenarios, sampling was done over 5 time points within 50 generations of τ_1 . We compared \mathcal{H} , GP, FIT with both static and dynamic SFS based statistics of SFSelect and Tajima's D. Fig. 6A shows that SFS based statistics are outperformed by single locus and CLR methods. However, when sampling is performed close to fixation, i.e., when the favored allele has frequency of 0.9 or higher, SFS based statistics perform significantly better than GP, FIT and \mathcal{H} (Fig. 6b). Moreover, dynamic SFS statistics outperform static SFS statistics, demonstrating that in these scenarios SFS based statistics should be used to detect selection.

Locating the Adaptive Mutation. The secondary task in identifying selection is to locate the position of the adaptive allele. We simply consider the site with highest score in the window as the locus of the favored allele. For each setting of ν_0 and s , we conducted 500 simulations and computed the rank of the favored mutation in each simulation. We plotted the cumulative distribution of the rank in Fig. 7. In all configuration COMALE ranks favored allele higher than GP. In particular when selection is strong ($s = 0.1$), COMALE confidently picks the beneficial allele, i.e., when $s = 0.1$, the beneficial allele is ranked first in 99% of the soft sweep simulations and ranked first in 95% of the simulations.

Strength of Selection. As the COMALE likelihood calculation is model based, we can also compute the model parameters \hat{s} that maximized likelihood (10). We computed bias, $s - \hat{s}$ for each experiment of COMALE, and compared it against GP. The distribution of the error (bias) for 100X coverage is presented in Fig. 8 for different configurations. Also Fig. S16, S17, ?? depict distribution of estimation error for 30X, 300X and infinity coverage, respectively. In general, both GP and COMALE have biased results for weak selection, where genetic drift dominates. However, for stronger sweeps, e.g. $s = 0.1$, COMALE provides estimates with smaller bias and variance. Standard deviation of the bias of \mathcal{H} is 0.028 and 0.011 in hard and soft sweep scenarios, while it is 0.036 and 0.013 for GP, respectively.

3.1 *D. melanogaster* adaptation to alternating temperatures experimental evolution

We applied COMALE to a controlled experimental evolution experiment [49], where 3 replicate samples were chosen from a population of *Drosophila melanogaster* for 59 generations under alternating 12-hour cycles of hot (28°C) and cold (18°C) temperatures and sequenced. The sampling

read depths are highly heterogeneous (Suppl. Figs. S7, S3). Filtering low coverage sites from data can dramatically reduce data available for analysis. For example, by setting minimum read depth at a site to be 30, allowing the site to be retained only if the depth for all time points, and all replicates exceeded 30, the number of sites analyzed would drop from 1,544,374 to 10,387. Instead, COMALE computes \mathcal{H} statistic for all sites where the sites with low coverage will take lower scores by COMALE.

We computed the COMALE \mathcal{H} statistic for sliding window of 50Kbp with steps of 10Kbp over the whole genome. We observed a spurious negative correlation between \mathcal{H} and the number of SNPs in each window (Suppl. Fig. S8). To make this precise, we computed the Pearson correlation between the number of variants w in a window, and \mathcal{H} as $\rho(w, \mathcal{H}) = -0.03$ overall. However, when restricted to the candidate regions with a high \mathcal{H} value, we found $\rho(w, \mathcal{H}) = -0.27$. This nine-fold increase in correlation indicates that \mathcal{H} statistic takes more extreme values when SNPs are depleted. Note that the expected number of variants in a 50Kbp window equals 1175 by Watterson's estimate. Therefore, we filtered out genomic regions containing fewer than 500 SNPs. In repeating the test, the average PCC for all windows was computed to be $\rho(w, \mathcal{H}) = -0.08$, and for windows with a high value of \mathcal{H} , it remained close at $\rho(w, \mathcal{H}) = -0.09$. Fig. 9 shows a Manhattan plot of scores. Our results are consistent with previous studies [49] in showing an over-representation of significant variants on Chromosome 3R.

The 1%-ile cut-off reveals 25 distinct intervals (Table S3) spanning 243 of the 13,965 genes that encoded variants. We found 36 GO terms associated with "Biological Process" to be over-represented in the 243 genes using a Fisher exact test for significance. Table 1 describes all terms that were significant, and contained at least 3 genes. The effect of high temperature on metabolism is profound, and it is not surprising that the enriched GO terms include many generic stress and stress response genes. They also include 5 of 7 genes with aminoacylase activity, and 4 of 17 genes involved in peroxidase activity. Stress induced formation of reactive oxygen species (ROS) can have deleterious effects. In many plants, including strawberry, high temperature adaptation led to an increase in expression of ROS scavenging genes, including peroxidase, and a reduction in total protein content due to protein denaturation, and inhibited synthesis [28], and these effects are also observed in potato. The genes include Pxd (CG3477), the peroxidase component of the chorion [37]. Chorion peroxidases are conjectured to play a role in ROS defense during egg formation in mosquitoes [40].

One of the difficulties in genome-wide association studies is that the number of polymorphisms are different between genes, i.e., longer genes contain more SNPs and larger number of polymorphisms increase the chance of showing association for larger genes. Although COMALE statistic for genes does not favor longer genes, we here perform SNP-based GO enrichment to compare enrichment of two different approaches. We tested associations of top 1%-ile of the SNPs within the 21 candidate intervals (Table S3) using Gowinda [36] and found 325 GO terms (see supplemental table S4) with FDR0.001. Also, to show that both analysis lead to the similar results, we performed Fisher exact test on the 13 previously enriched (Table 1) and 325 newly enriched GO terms. Given total GO terms and intersection of 9, Fisher exact test provides p -value of 10^{-11} .

Software. The source code and running scripts for COMALE are available at <https://github.com/bafnalab/comale>.

4 Discussion

We developed a computational tool COMALE, that can detect regions under selection experimental evolution experiments of sexual populations. Using extensive simulations, we show that COMALE

outperforms existing methods in detecting selection, locating the favored allele, and estimating selection parameters. Importantly, we make design choices that make COMALE very fast in practice, allowing us to scan all variants in genome-wide studies.

Many factors play a role in adaptation during experimental evolution studies. The statistics used by COMALE perform well because they account for many of these aspects, including complete and exact modeling of time-series data, not restricted to two-time point time-series, correction of ascertainment bias for finite-depth pooled-seq or finite sample variant data, and linkage within a region using composite likelihood statistics. COMALE uses s, h as model parameters in its likelihood calculation, and provides optimized estimates of these parameters.

In general, while EE studies are powerful, they also pose some challenges that are not adequately considered by other tools. One serious constraint is the *sampling-time-span*, the gap between the first and last sampled generations, which depends upon the generation time of the organism. It can be very small relative to the time of fixation of the favored allele. In *Drosophila* for example, 30-50 generations are typical, although there are some notable exceptions [79]. Therefore, unless the selection coefficient is very strong the time series data will only capture a ‘partial sweep’. This limitation is more critical in the controlled experimental evolution, where the sampling often starts at the onset of selection, and favored allele grows in frequency even more slowly if the initial frequency of the favored allele is close to zero (hard sweep). Through careful modeling, COMALE performs better than competing tools in these constrained scenarios.

In controlled experimental evolution experiments, populations are evolved and inbred. As this scenario involves picking a small number of founders, the effective population size significantly drops from the large number of wild type (e.g., for *Drosophila*, $(N_e \approx 10^6)$) to a small number of founder lines $F (\approx 10^2)$ for Experimental Evolution, and the evolution includes a severe population bottleneck. This bottleneck confounds SFS-based statistics and makes it difficult to fit a model or test a hypothesis (Fig. S5). Hence we exclude SFS-based statistics for experimental evolution of controlled populations when the initial sampling time is close to the onset of selection. However, in time-series sampling of Natural populations, the time of onset of selection may not be known. Even in this setting COMALE outperforms other tools including SFS based statistics, except in the case where beneficial allele is close to fixation. For this specific scenario, we developed time-dependent SFS based tests, and show excellent performance compared to all other tests. Both COMALE and SFS based tests are implemented in the COMALE software. In the default mode, COMALE computes scores for all tests, to identify regions under selection.

In our simulations, we found that the power of detection can be severely affected by sampling schedule as well as initial frequency of the favored allele. Many experimental evolution methods start sampling at the onset of selection, and continue up to 50 or so generations. For small values of the selection coefficient, this may not be sufficient, and other experiments might be possible that span for the entire duration of the sweep. However, even if it were possible to sample over a larger time-span, many methods, especially the ones that compute full likelihoods, would simply not scale to allow computation of evolutionary trajectories over a large time-span. In contrast COMALE precomputes the transition matrices, and scales linearly with number of samples, irrespective of the time-span in which they were acquired.

Sequence coverage is a practical consideration that is often ignored by other tools. In COMALE we use HMMs to explicitly model variation in coverage. Consequently, its performance is robust with decrease in coverage relative to other tools.

The performance of tools in hard versus soft (standing variation) scenarios is a bit counter-intuitive. On the one hand, linkage is strongest in hard-sweep scenarios and a composite likelihood ratio would achieve better power. On the other hand, the growth of the favored allele (and other hitchhiking alleles) is sigmoidal, and we are closer to the exponential growth while sampling if ν_0 is

higher. We model these contrasting issues by introducing the parameter π , the percentile cut-off for including an allele in the composite statistic. While $\pi = 0$ (including all variants) is reasonably robust in all scenarios, it works best in a hard-sweep case. However in soft-sweep scenarios, where $\nu_0 \gg 0$ (close to one), the beneficial allele is at low LD with its vicinity, and larger π works better by filtering out sites that are not linked with the favored allele. The automated choice for the best value of π will be a subject of future research. Here we show results fixing $\pi = 0$. Contextually modifying these parameters helps COMALE achieve better performance in locating the favored allele, and future work will be devoted to identifying the favored allele with higher power.

Here, we applied COMALE to a single example of experimental evolution–analysis of temperature sensitive adaptation in *Drosophila*. However, as EE experiments with sampling of genomic sequence increase in popularity, computational tools for identifying the genetic basis of adaption will play a key role in observing evolution in action.

Acknowledgments

The authors were supported by grants from the NIH (1R01GM114362) and NSF (DBI-1458557 and IIS-1318386).

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.

5 Figures

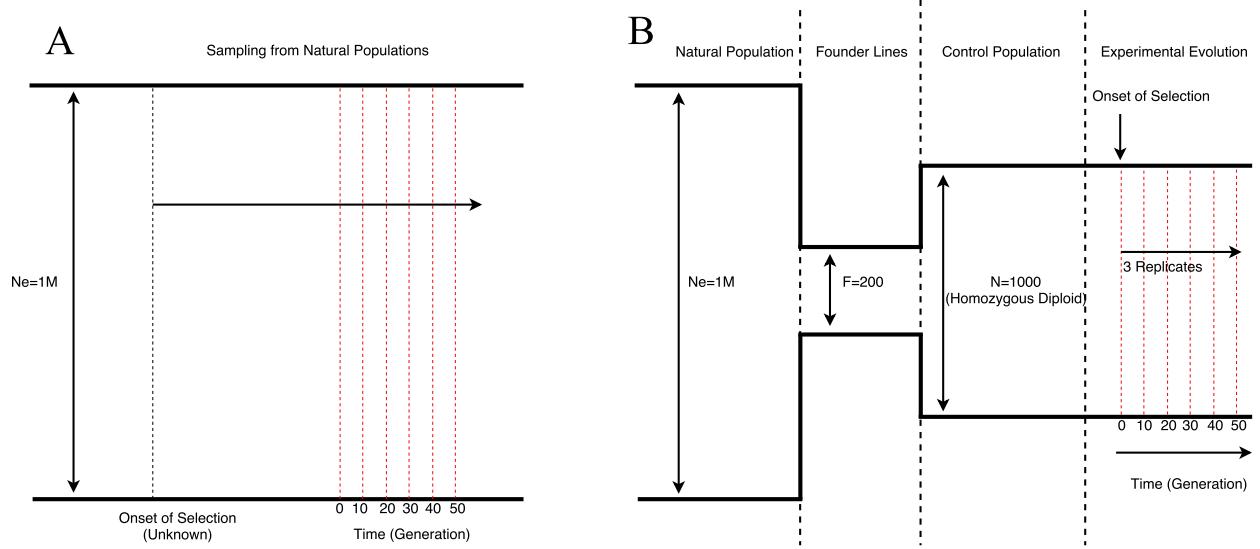


Fig. 1: Two settings for collecting genomic time series data. Different settings in which dynamic data is collected are depicted with typical parameters for *D. melanogaster*. In both settings, 6 samples (vertical red dashed lines) are taken every 10 generation. When sampling from naturally evolving populations (A), the time of onset of selection is unknown, and population size is larger. For (controlled) experimental evolution, founder lines are first sampled from a natural population to create a homogeneous population. Multiple replicates of this population are evolved and sampled over time.

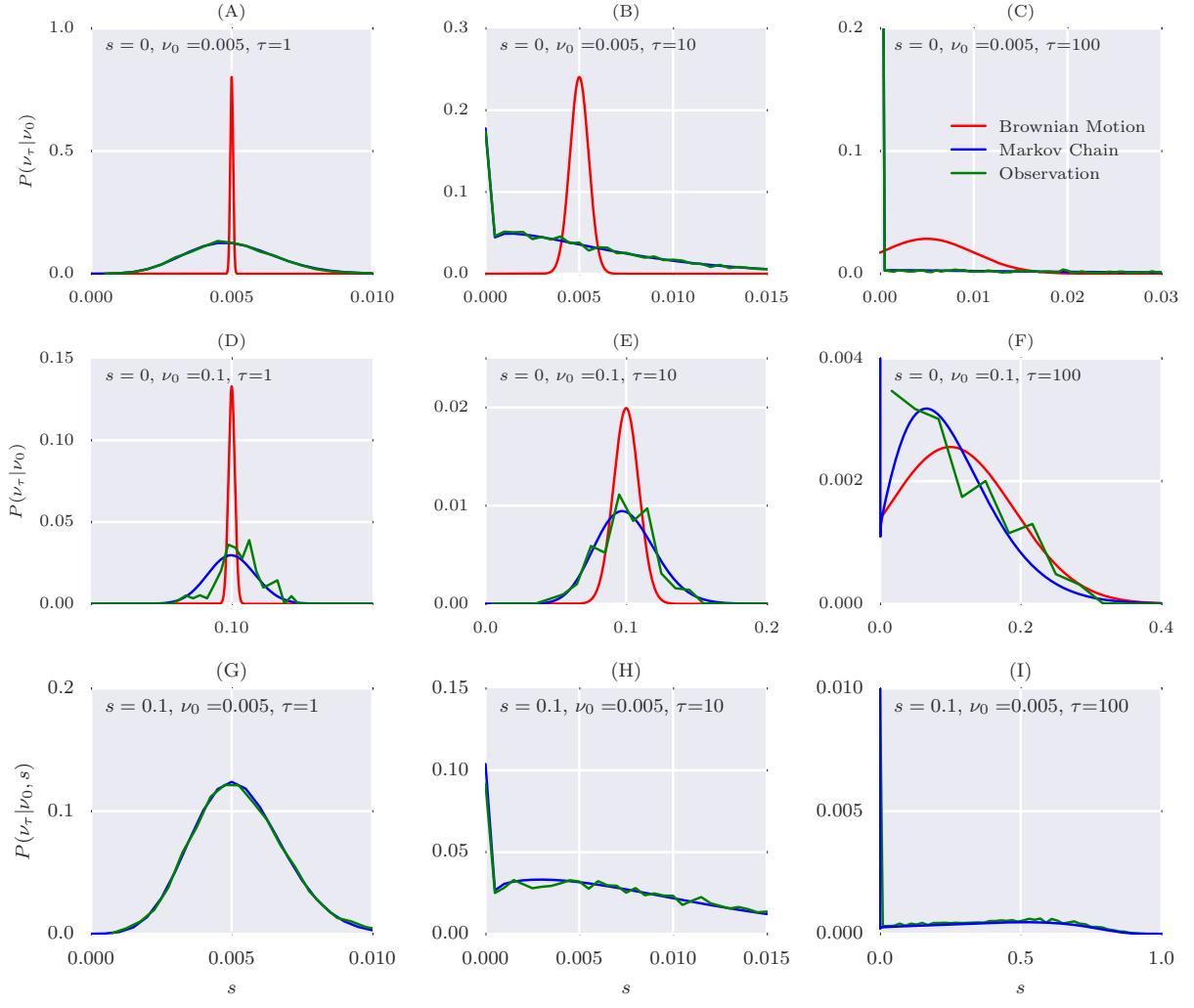


Fig. 2: Comparison of empirical distributions of allele frequencies (green) versus predictions from Brownian Motion (red), and Markov Chain (blue). Panels A-F: Experiments were conducted under neutral evolution with different starting frequencies $\nu_0 \in \{0.005, 0.1\}$ and sampling times $\tau \in \{1, 10, 100\}$ generations. The empirical distribution was computed by sampling 143,900 sites with $\nu_0 = 0.005$ and 47,500 variants with $\nu_0 = 0.1$. Panels G,H,I: Comparisons of Empirical and Markov chain based allele frequency distributions under a selection regime with $s = 0.1$. Initial frequency was chosen as $\nu_0 = 0.005$ and sampling performed after τ generations for $\tau \in \{1, 10, 100\}$.

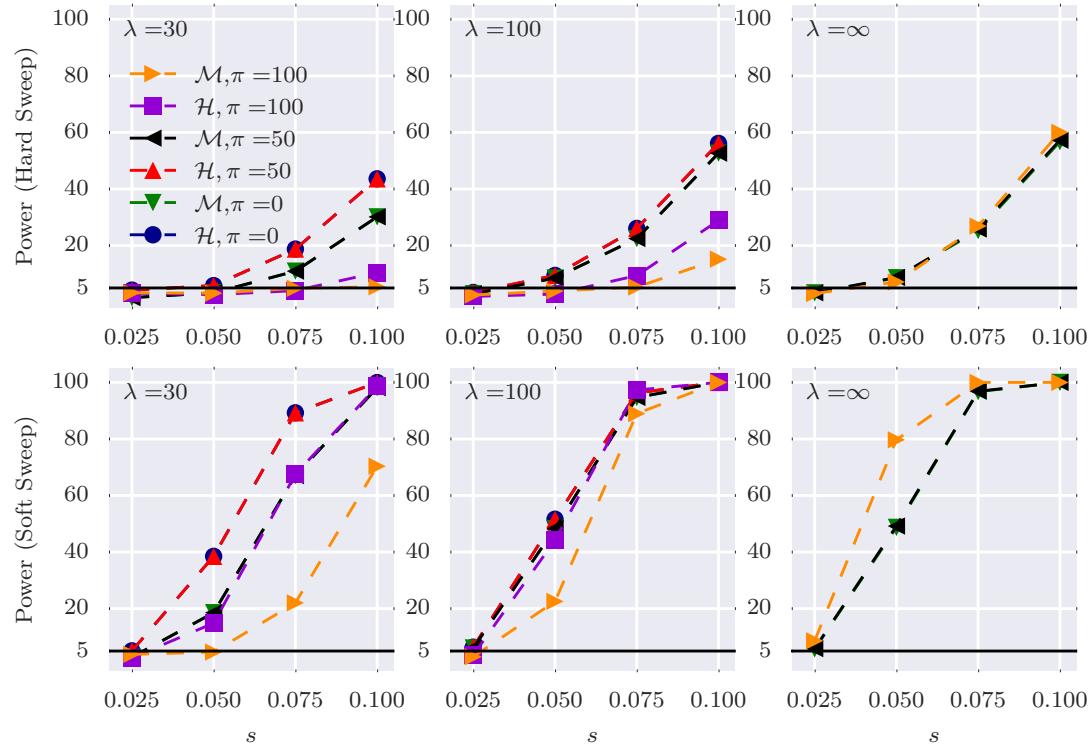


Fig. 3: Detection power for Markov chain and HMM with different composite statistics. Detection power for Markov chain (\mathcal{M}) and HMM (\mathcal{H}) under hard (top) and soft sweep (bottom) scenarios, for different settings of mean coverage λ and selection strength s . The y -axis measures power – sensitivity with false positive rate $FPR \leq 0.05$ – for 1,000 simulations of 50Kbp regions.

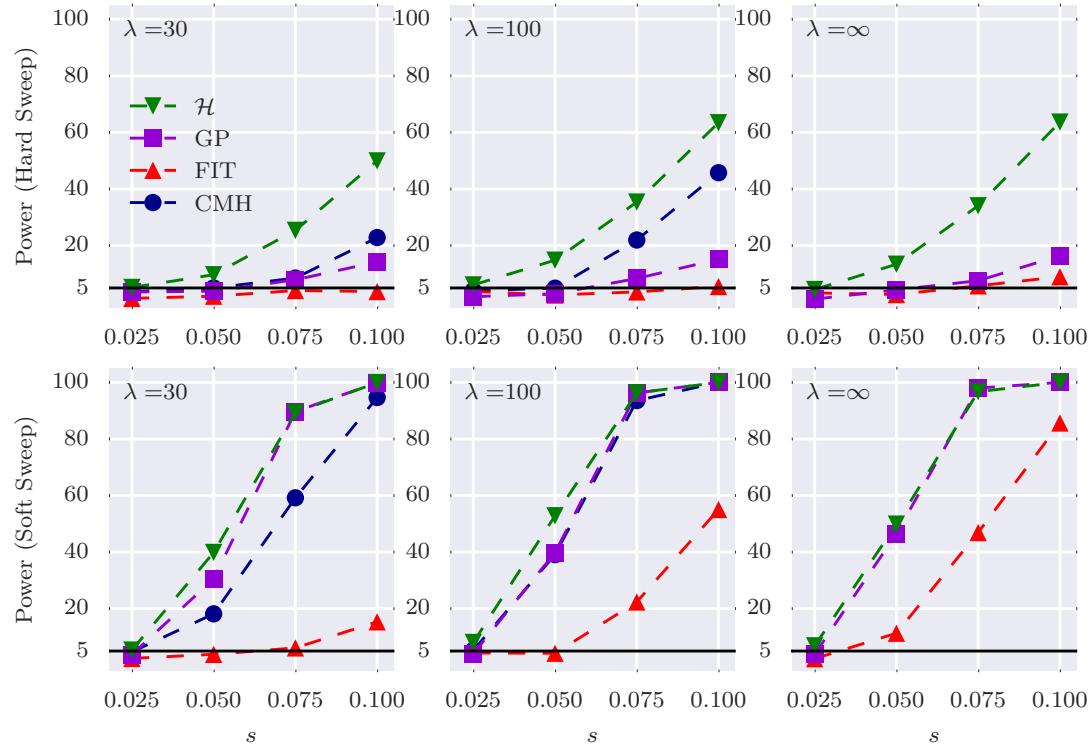


Fig. 4: Power calculations for detection of selection. Detection power for COMALE(\mathcal{H}), Frequency Increment Test (FIT), Gaussian Process, and CMH under hard (top) and soft sweep (bottom) scenarios. λ , s denote the mean coverage and selection coefficient, respectively. The y -axis measures power – sensitivity with false positive rate $FPR \leq 0.05$ – for 1,000 simulations of 50Kbp regions.

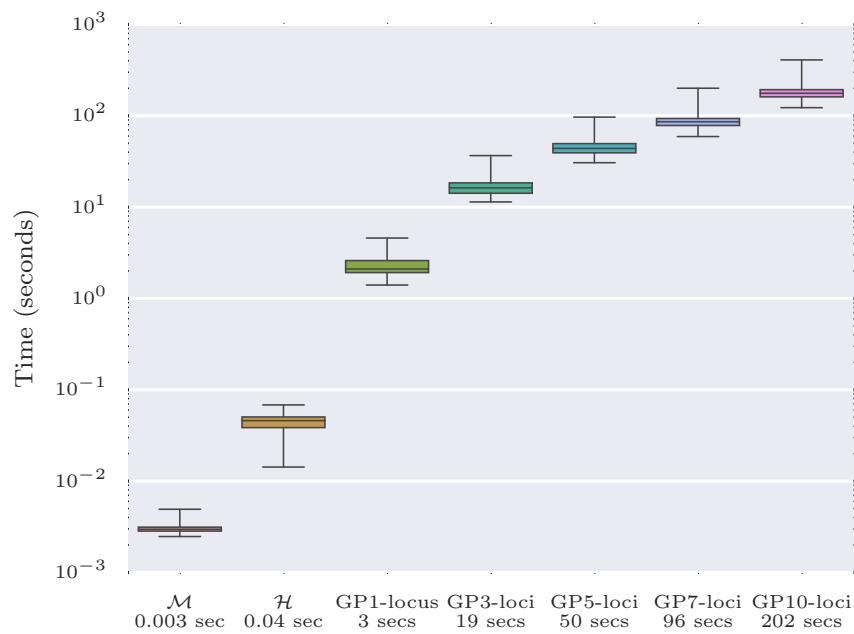


Fig. 5: Running time. Box plot of running times (cpu-secs.) of COMALE, HMM, GP with single, 3, 5, 7, and 10 loci over 1000 simulations conducted on a workstation with 4th Generation Intel Core i7 processor. The average running time for each method is shown on the x-axis.

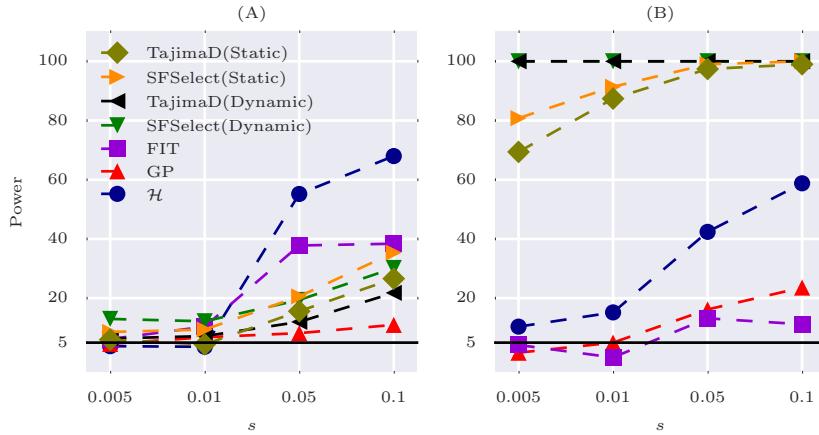


Fig. 6: Power of SFS based statistics. Power of detecting selection for Frequency Increment Test (FIT), Gaussian Process (GP), COMALE (\mathcal{H}) on hard-sweep natural experimental evolution with $N_e = 10^4$ and depth $\lambda = \infty$. The measurements are conducted for a range of selection coefficients, s . Each point represents the mean of 200 simulations. For each simulation, sampling starts at a randomly chosen time, and subsequently 5 replicate samples are acquired every 10 generations. (A) Start of sampling is chosen randomly throughout the sweep $\tau_1 \sim U[1, t_{\nu=1}(s, N_e)]$, where $t_{\nu=x}(s, N_e)$ denotes the expected time to reach carrier frequency x in a hard sweep and $U[a, b]$ is discrete uniform distribution. (B) The start of sampling is chosen near fixation of the favored allele, i.e. $\tau_1 \sim U[t_{\nu=0.9}(s, N_e), t_{\nu=1}(s, N_e)]$.

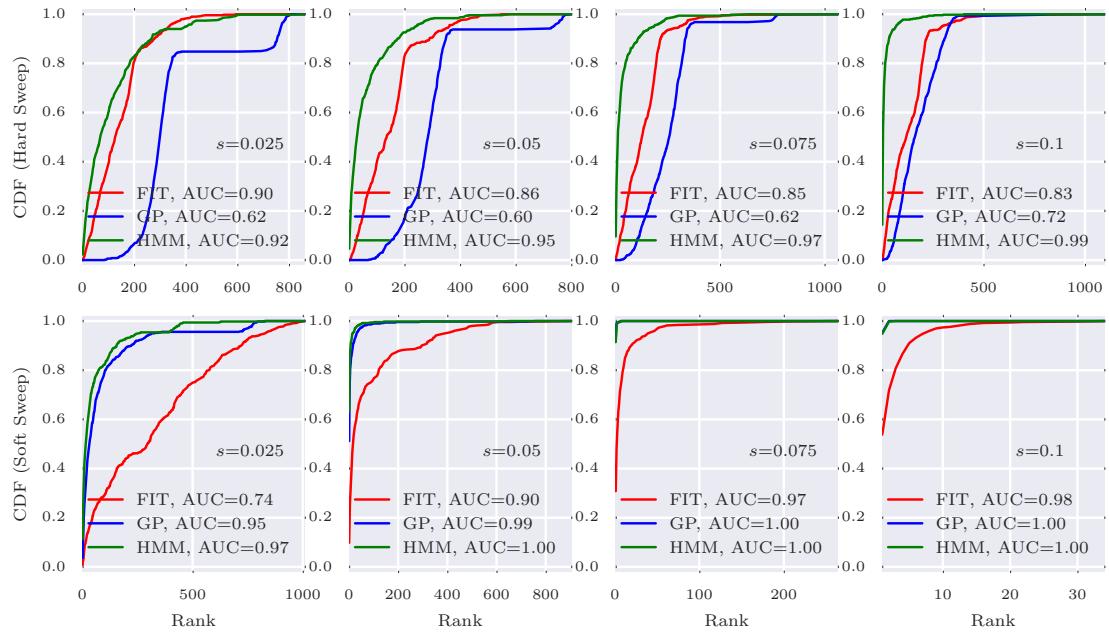


Fig. 7: Ranking performance for 100X coverage. Cumulative Distribution Function (CDF) of the distribution of the rank of the adaptive allele in 500 simulations for Hidden Markov Model (HMM), Gaussian Process (GP), CMH, and Frequency Increment Test (FIT), for different values of selection coefficient s and initial carrier frequency. Area Under Curve (AUC) is computed as a quantitative measure ranking performance of methods for each configuration.

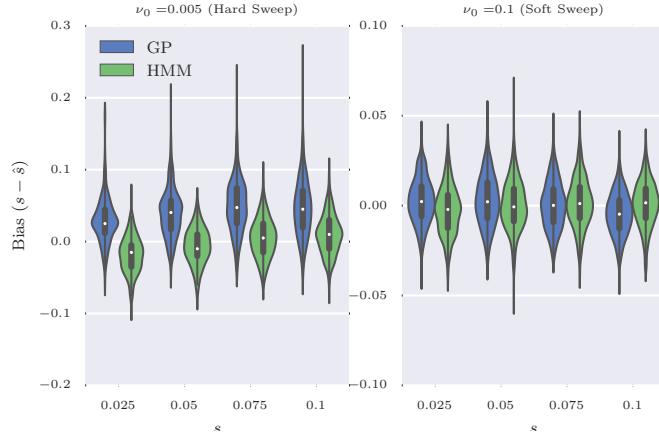


Fig. 8: Distribution of bias for 100X coverage. The distribution of bias ($s - \hat{s}$) in estimating selection coefficient over 500 simulations using Gaussian Process (GP) and Hidden Markov Model (HMM) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0 , when coverage is 100. The estimation performance of GP and HMM are very similar in soft sweep, while HMM provides lower variance in hard sweep (overall standard deviation is 0.036 for GO and is 0.028 for HMM).

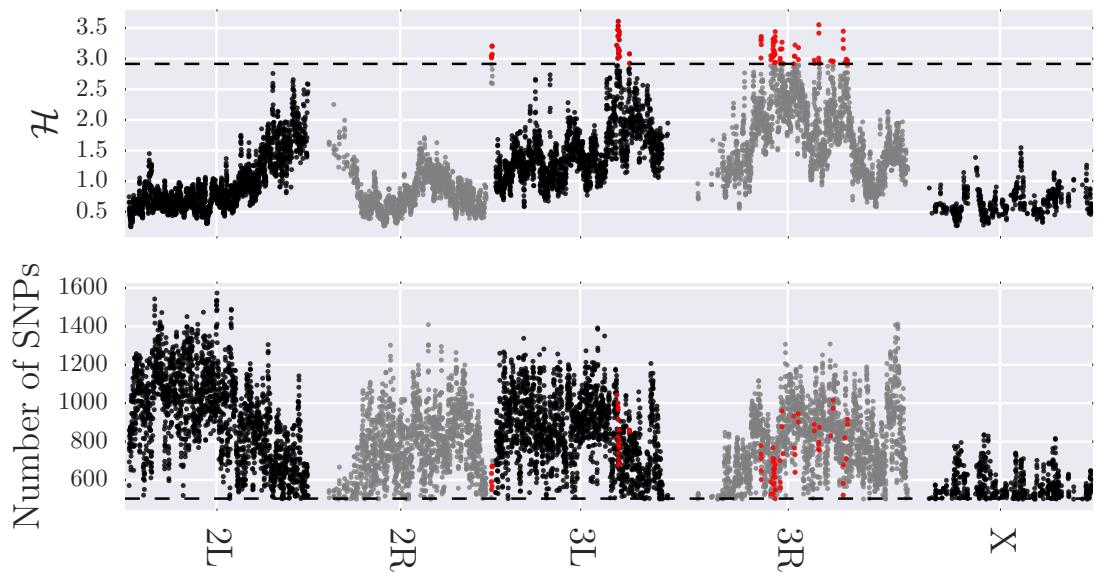


Fig. 9: COMALE analysis of a *Drosophila* EE experiment. Manhattan plot of the composite \mathcal{H} statistic (top) and the number of SNPs (bottom) in 50Kbp sliding window with steps of 10Kbp, excluding windows with less than 500 SNPs.

What is the dotted line? Are these all chromosomes. We need to cite the drosophila experiment here

Tables

Table is spilling over

Rank	GO ID	GO Term	-log(<i>p</i> -value)	Hits	Num of Genes
1	GO:0004046	aminoacylase activity	9.7	5	7
3	GO:0015101	organic cation transmembrane transporter activity	6.4	3	5
4	GO:0007501	mesodermal cell fate specification	6.2	4	11
5	GO:0004601	peroxidase activity	5.1	4	17
6	GO:0006979	response to oxidative stress	5.0	8	79
11	GO:0009312	oligosaccharide biosynthetic process	4.2	3	13
12	GO:0004653	polypeptide N-acetylgalactosaminyltransferase activity	4.1	3	14
19	GO:0040014	regulation of multicellular organism growth	3.6	3	18
20	GO:0016485	protein processing	3.5	3	19
21	GO:0006030	chitin metabolic process	3.5	7	99
22	GO:0020037	heme binding	3.4	8	127
35	GO:0008061	chitin binding	3.1	7	113
36	GO:0004702	receptor signaling protein serine/threonine kinase activity	3.0	3	26

Table 1: Gene set enrichment of 243 genes within 25 intervals under selection using Fisher exact test.
No detail?

References new page

- [1] Guillaume Achaz. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258, 2009. [Is this a correct style?](#)
- [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] 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.
- [5] 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.
- [6] 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.
- [7] Jeffrey E Barrick and Richard E Lenski. Genome dynamics during experimental evolution. *Nat Rev Genet*, 14(12):827–839, dec 2013.
- [8] 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.

- [9] 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.
- [10] 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.
- [11] 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.
- [12] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of 2Nes from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] Michael M Desai and Joshua B Plotkin. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, 180(4):2175–2191, 2008.
- [19] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.
- [20] Steven N Evans, Yelena Shvets, and Montgomery Slatkin. Non-equilibrium theory of the allele frequency spectrum. *Theoretical population biology*, 71(1):109–119, 2007.
- [21] Warren J Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media, 2012.
- [22] 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.

- [23] Justin C Fay and Chung-I Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413, 2000.
- [24] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. Identifying signatures of selection in genetic time series. *Genetics*, 196(2):509–522, 2014.
- [25] 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.
- [26] Yun-Xin Fu. Statistical properties of segregating sites. *Theoretical population biology*, 48(2):172–197, 1995.
- [27] Michael M Gottesman. Mechanisms of cancer drug resistance. *Annual review of medicine*, 53(1):615–627, 2002.
- [28] Hatice Gulen and Atilla Eris. Effect of heat stress on peroxidase activity and total protein content in strawberry plants. *Plant Science*, 166(3):739–744, 2004.
- [29] Matthew Hegerenness, Noam Shresh, Daniel Hartl, and Roy Kishony. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311(5767):1615–1617, 2006.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] Robert Kofler and Christian Schlötterer. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics*, 28(15):2084–2085, 2012.
- [37] Ourania A Konstandi, Issidora S Papassideri, Dimitrios J Stravopodis, Christos A Kenoutis, Zulfiqar Hasan, Theodoros Katsorchis, Ron Wever, and Lukas H Margaritis. The enzymatic component of *Drosophila melanogaster* chorion is the Pxd peroxidase. *Insect biochemistry and molecular biology*, 35(9):1043–1057, 2005.

- [38] 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.
- [39] 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.
- [40] Junsuo S Li and Jianyong Li. Major chorion proteins and their crosslinking during chorion hardening in *Aedes aegypti* mosquitoes. *Insect biochemistry and molecular biology*, 36(12):954–964, 2006.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] PABLO OROZCO-terWENGEL, 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.
- [50] 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.

- [51] Edward Pollak. A new method for estimating the effective population size from allele frequency changes. *Genetics*, 104(3):531–548, 1983.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] Roy Ronen, Glenn Tesler, Ali Akbari, Shay Zakov, Noah A Rosenberg, and Vineet Bafna. Predicting Carriers of Ongoing Selective Sweeps Without Knowledge of the Favored Allele. *PLoS Genet*, 11(9):e1005527, 2015.
- [57] Roy Ronen, Nitin Udpa, Eran Halperin, and Vineet Bafna. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–193, 2013.
- [58] 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.
- [59] Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Steven A McCarroll, Rachelle Gaudet, and Others. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918, 2007.
- [60] Stanley A Sawyer and Daniel L Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, 1992.
- [61] 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.
- [62] 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.
- [63] 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.
- [64] 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.

- [65] Wolfgang Stephan, Yun S Song, and Charles H Langley. The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci. *Genetics*, 172(4):2647–2663, apr 2006.
- [66] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.
- [67] 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.
- [68] Ray Tobler, Susanne U Franssen, Robert Kofler, Pablo Orozco-terWengel, Viola Nolte, Joachim Hermissen, 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.
- [69] 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.
- [70] 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.
- [71] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting natural selection in genomic data. *Annual review of genetics*, 47:97–120, 2013.
- [72] Jinliang Wang. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical research*, 78(03):243–257, 2001.
- [73] Robin S Waples. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391, 1989.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] Xin Yi, Yu Liang, Emilia Huerta-Sánchez, 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.
- [78] Hiba Zahreddine and K L Borden. Mechanisms and insights into drug resistance in cancer. *Front Pharmacol*, 4(28.10):3389, 2013.
- [79] Dan Zhou, Nitin Udupa, Merril Gersten, DeeAnn W Visk, Ali Bashir, Jin Xue, Kelly A Frazer, James W Posakony, Shankar Subramaniam, Vineet Bafna, and Others. Experimental selection of hypoxia-tolerant *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 108(6):2349–2354, 2011.

6 Appendix

6.1 An approximate logistic function for allele frequency dynamics

Assume that a site is evolving under selection constraints $s, h \in \mathbb{R}$, where s and h denote selection strength and dominance, respectively. Let ν_t denotes the frequency of the site at time $\tau_t \in \mathcal{T}$. Then, ν_{t+} , the frequency at time $\tau_t + 1$ can be estimated (See Eq. 6.1) using:

$$\hat{\nu}_{t+} = \nu_t + \frac{s(h + (1 - 2h)\nu_t)\nu_t(1 - \nu_t)}{1 + s\nu_t(2h + (1 - 2h)\nu_t)}.$$

We can show that the dynamic of the beneficial allele can be modeled via a logistic function, in the case of directional selection ($h = 0.5$). Taking derivatives of Eq. 6.1, we have

$$\frac{d\nu_t}{dt} = \frac{s\nu_t(1 - \nu_t)}{2 + 2s\nu_t} \quad (\text{S1})$$

To, solve the differential equation, note that for small s , $2 + 2s\nu_t \approx 2$. Substituting,

$$\nu_t = \frac{1}{1 + \frac{1 - \nu_0}{\nu_0} e^{-st/2}} = \sigma(st/2 + \eta(\nu_0)) \quad (\text{S2})$$

where $\sigma(\cdot)$ is the logistic function and $\eta(\cdot)$ is logit function (inverse of the logistic function).

6.2 Dynamic of Tajima's D

In this part we derive dynamic of Tajima's D statistic in *hard sweep* as function of its value at the onset of selection, D_0 , selection strength and the frequency of the beneficial allele at the onset of selection. Let D_0, Π_0, W_0 , be Tajima's D, Tajima's estimate of θ , and Watterson's estimate of θ at time zero and $D_0 = \Pi_0 - W_0$. In order to compute, $D_t = \Pi_t - W_t$ we compute Π_t and W_t separately as follows. Let P be the $n \times n$ matrix of pairwise heterozygosity if individuals, then $\Pi = \frac{1}{n^2} \sum P_{ij}$. So, if the population consist of νn identical carrier haplotype (due to lack of recombination), their pairwise hamming distance is zero and should be subtracted from the total Π_t :

$$\Pi_t = (1 - \nu_t^2)\Pi_0 \quad (\text{S3})$$

To compute W_t , first remember that $W_t = \frac{m_t}{S_n}$ where m_t is the number of segregating sites at time t and $S_n = \sum_i^n 1/i \approx \log(n)$. Also we have

$$\frac{W_t}{W_0} = \frac{\frac{m_t}{S}}{\frac{m_0}{S}} \Rightarrow W_t = \frac{m_t}{m_0} W_0 \quad (\text{S4})$$

Because of hard sweep and lack of recombination assumption, the population at time t consist of $(1 - \nu_t)n$ non-carrier haplotypes and $\nu_t n$ identical carrier haplotypes. While not strictly correct, we assume that the $(1 - \nu_t)n + 1$ individuals are evolving neutrally. Using this assumption, we have

$$\frac{m_t}{m_0} = \frac{\log((1 - \nu_t)n + 1)\theta}{\log(n)\theta} \approx \frac{\log((1 - \nu_t)n)}{\log(n)} = \frac{\log(1 - \nu_t) + \log(n)}{\log(n)} = 1 + \frac{\log(1 - \nu_t)}{\log(n)}. \quad (\text{S5})$$

Finally, by putting Eqs. S3, S4, S5 together, we can explicitly write the dynamics of D statistic as

$$\begin{aligned} D_t &= (1 - \nu_t^2)\Pi_0 - \left(1 + \frac{\log(1 - \nu_t)}{\log(n)}\right)W_0 \\ &= D_0 - \log(1 - \nu_t) \frac{W_0}{\log(n)} - \nu_t^2 \Pi_0 \\ &\approx D_0 - \log(1 - \sigma(st/2 + \eta(\nu_0))) \frac{W_0}{\log(n)} - \sigma(st/2 + \eta(\nu_0))^2 \Pi_0. \end{aligned} \quad (\text{S6})$$

where σ and η are logistic and logit functions.

6.3 Dynamics of Fay and Wu's H

In any finite population size of n with m segregating sites, allele frequencies take discrete values, i.e., $x_j \in \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}\}$, $\forall j \in 1, \dots, m$. We have the following:

$$\|\mathbf{x}\|^2 = \sum_{j=1}^m x_j^2 = \sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^2 \xi_i = \frac{(n-1)}{2n} H,$$

where ξ_i is the number of sites with frequency i/n and H is the Fay & Wu's estimate of θ and $\mathbf{x} \in (0, 1)^m$ is the vector of allele frequency of a region with m segregating sites. Recently, Ronen *et al.* [56] devised the 1-HAF statistic for identifying selection on static data, and showed that the expected value of 1-HAF statistic is given by:

$$\mathbb{E}[1\text{-HAF}(t)] = n\|\mathbf{x}_t\|^2 \approx ng(\nu_t) \quad (\text{S7})$$

where

$$g(\nu_t) = \theta\nu_t \left(\frac{\nu_t + 1}{2} - \frac{1}{(1 - \nu_t)n + 1} \right) + \theta(1 - \nu_t) \left(\frac{n + 1}{2n} - \frac{1}{(1 - \nu_t)n + 1} \right) \quad (\text{S8})$$

The dynamics of Fay & Wu's estimate are given by

$$H_t = \frac{n-1}{2} g(\nu_t) \quad (\text{S9})$$

6.4 Greedy computation of time-series SFS-based statistics

As discussed in Section 2.2, modeling dynamic of Tajima's D (and Fay&Wu's H) requires knowledge of initial carrier frequency ν_0 and the value of D (and H) statistic at the onset of selection, which are often unknown. As these statistics are monotonically decreasing (or increasing for SFSelect) under no demographic changes, we chose to greedily aggregate statistics throughout time. For example, for Tajima's D , we have

$$\mathcal{D} = \sum_{t \in \mathcal{T}} D_t \quad (\text{S10})$$

where the same procedure applies to Fay&Wu's H and SFSelect.

6.5 Linkage Disequilibrium

Nonrandom associations, Linkage Disequilibrium (LD), between polymorphisms are established in the substitution process, broken by recombination events and reinforced by selection. Although LD can not be measured in pooled sequencing data (phased haplotype data is required), it is still worthwhile to examine the behavior of LD as a result of the interaction between recombination and natural selection. In this part we theoretically overview expected LD in short EEs.

Let ρ_0 be the LD at time zero between the beneficial allele and a segregating site l base-pairs away, then under natural selection we have

$$\rho_t = \alpha_t \beta_t \rho_0 = e^{-r t l} \left(\frac{H_t}{H_0} \right) \rho_0 \quad (\text{S11})$$

where $H_t = 2\nu_t(1-\nu_t)$ is the heterozygosity at the selected site, r is the recombination rate/bp/gen. The *decay factor*, $\alpha_t = e^{-r t l}$, and *growth factor*, β_t (see Eq. 30-31 in [65]), are result of recombination and selection, respectively. Fig. S9 presents the expected theoretical value of LD when $\rho_0 = 0.5$ between beneficial allele (site at position 500K) and the rest of genome, and $\nu_0 = 0.1$. For

neutral evolution (top), LD decays exponentially through space and time, while in natural selection (bottom), LD increases and then decreases. Interestingly, LD increases to its maximum value, 1, for the nearby region (the plateau in the Fig. S9 bottom) of the beneficial allele.

In principle, LD increases after the onset of selection, until $\log(\alpha_t) + \log(\beta_t) > 0$, see Eq. S11. Specifically, log of decay term is linear and, using Eq. S2, we write growth factor in term of initial frequency ν_0 and selection strength s . Fig. S10 depicts interaction of decay and growth factors for weak and strong selection and soft and hard sweeps. In all the case, LD of the beneficial allele with a segregating site 50Kbp away, increases in the first 50 generations, which give rise to increasing number of *hitchhikers*.

Increase of LD in a large (100Kbp) region is particularly advantageous to the task of identifying the region under selection, if the composite statistics is used. As a result, \mathcal{H} statistic outperforms existing (single-loci) tools in identifying selection. In contrast, augmentation of LD, increases the number of candidates for the beneficial allele, which makes it difficult to localize the beneficial allele.

[New Page](#)

7 Supplemental Figs. and Tables

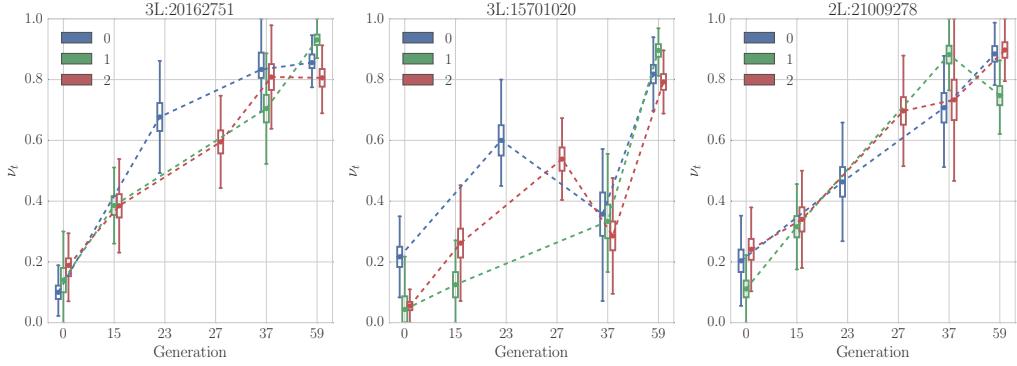


Fig. S1: Trajectory of pool-sequenced variants. Trajectory of three different variants that are increasing in frequency over time. Note that for read count data, the true allele frequency is not known. Here we draw the posterior distribution of the allele frequency at each time point using box plot. The median of each distribution is denoted by dots. The variance of each box is seen to be inversely related to the depth of the measurement.

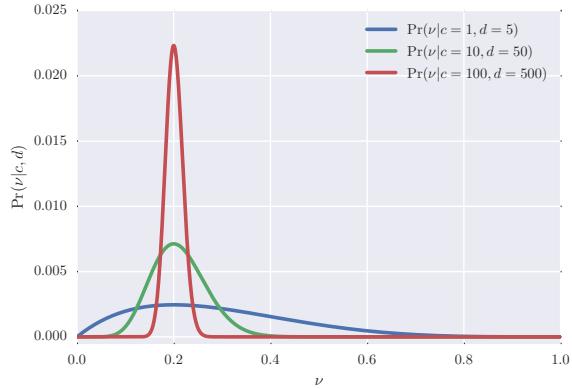


Fig. S2: Posterior distribution of allele frequency. Distribution of allele hidden frequency for different values of depth $d = \{5, 50, 500\}$. In all cases, the true frequency is 0.5, while the estimated frequency varies around it, with variance higher for low read-depths.

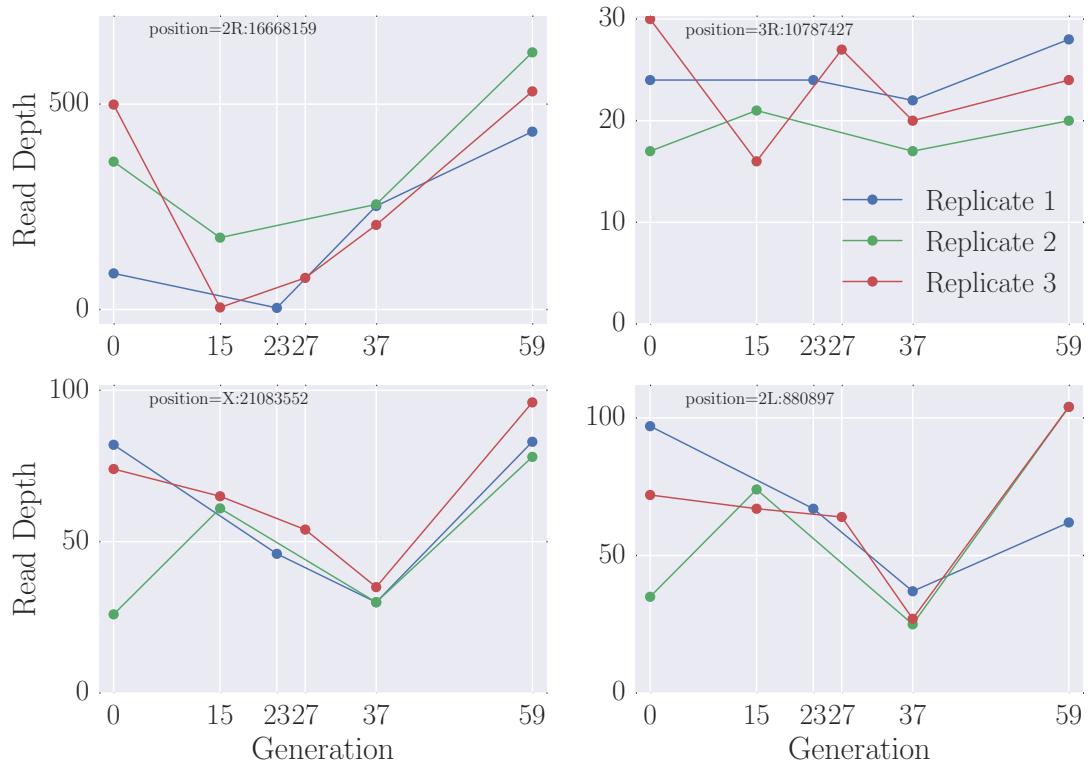


Fig. S3: Heterogeneity in time series data. Each panel shows the read depth for a different variant using 3 replicates of the *D. melanogaster* adaptation to alternating temperatures experimental evolution data (see section 3.1). Heterogeneity in depth of coverage is seen between replicates, and also at different time points, in all 4 variants.

Data source.

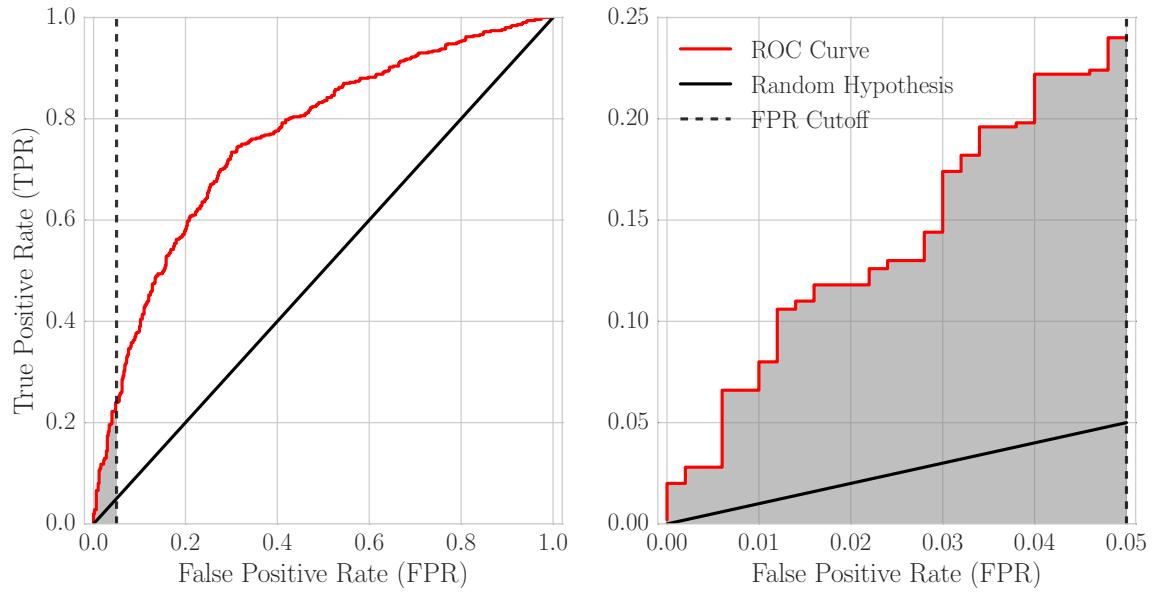


Fig. S4: Schematic for computing power of detecting selection. Standard Receiver operator curve (ROC) (left) is shown for illustration of the procedure for computing power of detection from 1,000 simulations (500 selection and 500 neutral). The Area Under the Curve (AUC) represents overall performance. The diagonal black line represents performance of a random hypothesis which achieves Area under the curve (AUC) of 0.5. To avoid computing AUC for the regions where FPR is unacceptably high, we restrict ROC curve to the region where $\text{FPR} \leq 0.05$ (right). In this case, we define power to be the (scaled) AUC of the restricted region.

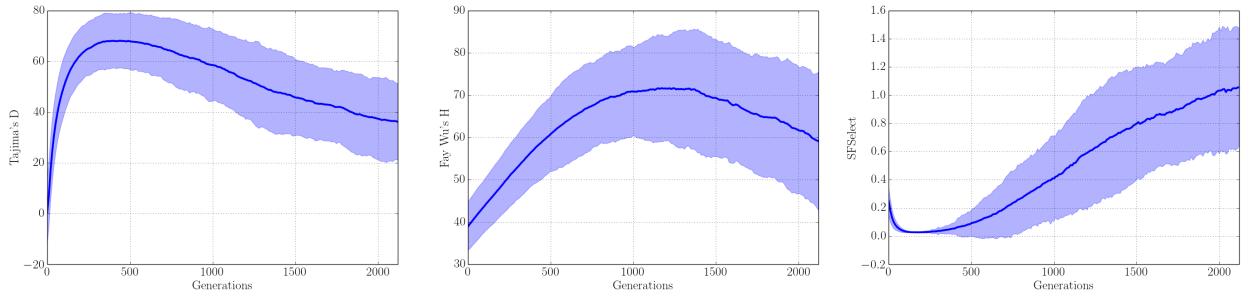


Fig. S5: Effect of bottleneck in a typical experimental evolution experiment with restricted number of founder lines. For the experiment, $F = 200$ founders were selected from a larger population size ($N_e = 10^{-6}$), and evolved under neutral scenario ($s = h = 0$). The statistics for Tajima's D (left), Fay Wu's H (middle) and SFSelect were computed for 1000 neutral simulations and the mean and 95% confidence interval plotted. Under neutral evolution, all the statistics are expected to vary around a fixed mean through time. However, under selective constraint, D and H take negative values, while SFSelect take positive values. In experimental evolution, bottleneck effect will suppress the signal of selection, especially in early generations.

Fonts are too small

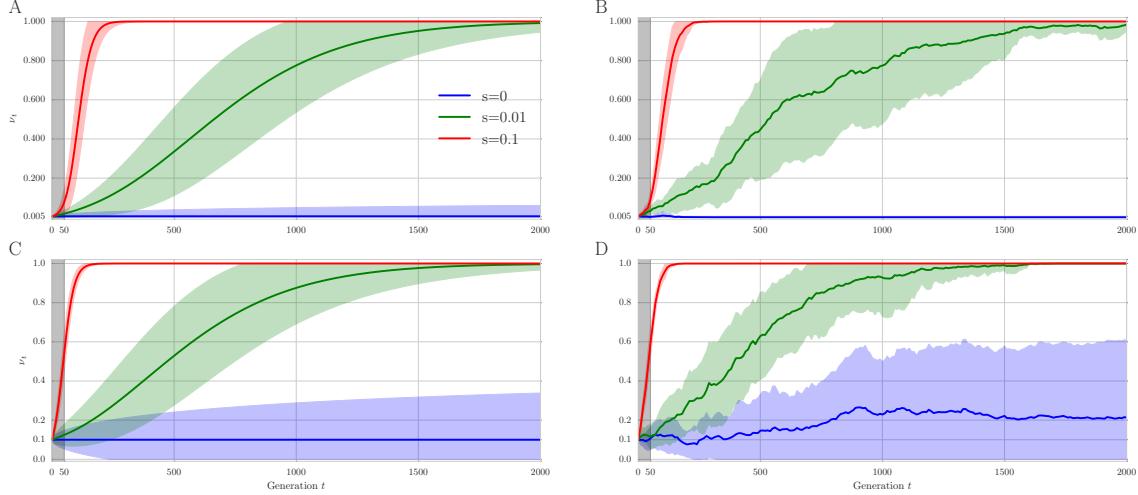


Fig. S6: Theoretical (Markov Chain) and Empirical trajectories of favored allele for hard and soft sweep scenarios. Theoretical and empirical trajectories of the frequency of the favored allele are computed for 1,000 EE simulations after choosing the diploid population size as 1000. Each curve shows the mean and the 95% confidence interval. Panels A and C represent theoretical markov chain based calculations of the favored allele frequency under hard (ν_0 is small), and soft sweep due to standing variation (higher ν_0), for a range of values of s . Note that $s = 0$ corresponds to neutral evolution. Similarly, panels B and D show the empirical forward simulations of populations under the same selection regimes, and hard/soft-sweep scenarios. The first 50 generations are shaded in gray to represent the typical sampling span of EE experiments. The plot illustrates the difficulty of EE experiments in having to predict selection at a very early stage of the sweep. The signal is slightly stronger under standing variation scenario. The theoretical and empirical simulations are in close correspondence.

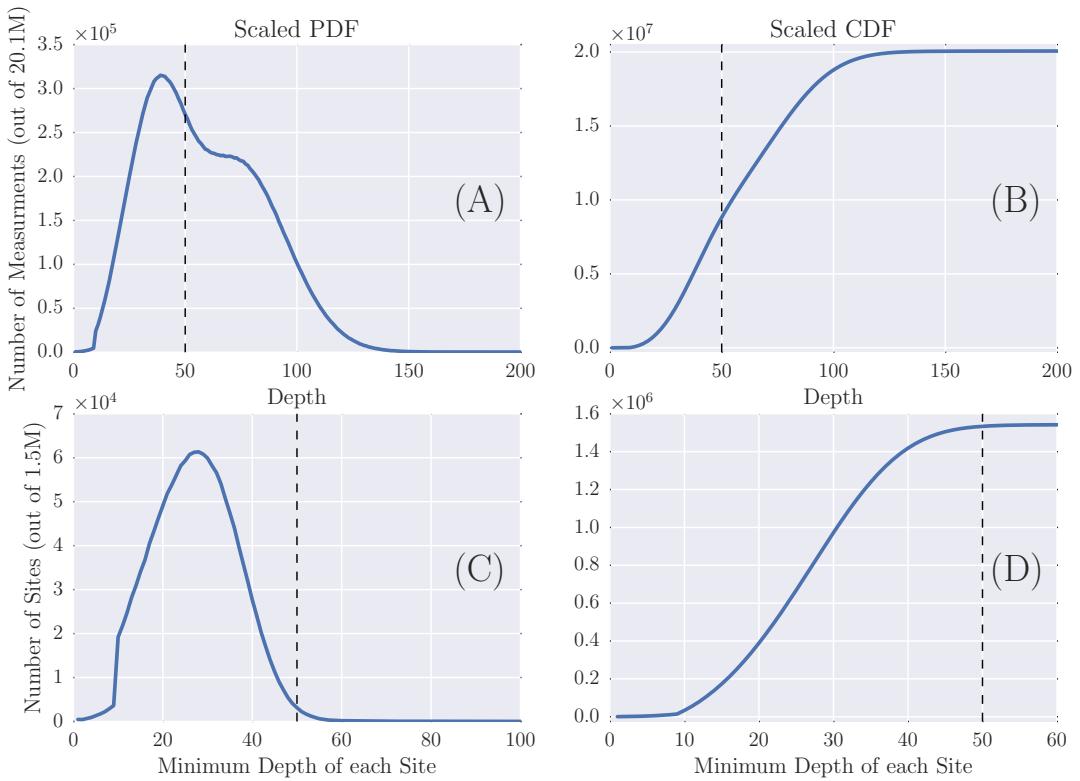


Fig. S7: Distribution of depth in the real data. Scaled PDF (A) and CDF (B) of the read depths of all ($\approx 20.1M$) measurements, i.e., all replicates and time points of the all ($\approx 1.5M$) variants. Scaled PDF (C) and CDF (D) of the minimum depth of sites. Although most ($\approx 11.5M$) of the measurements have depth of 50 or greater (dashed line in (A),(B)), only a small fraction ($\approx 10K$) of variants (dashed line in (C),(D)) pass the filter of having minimum depth of 50.

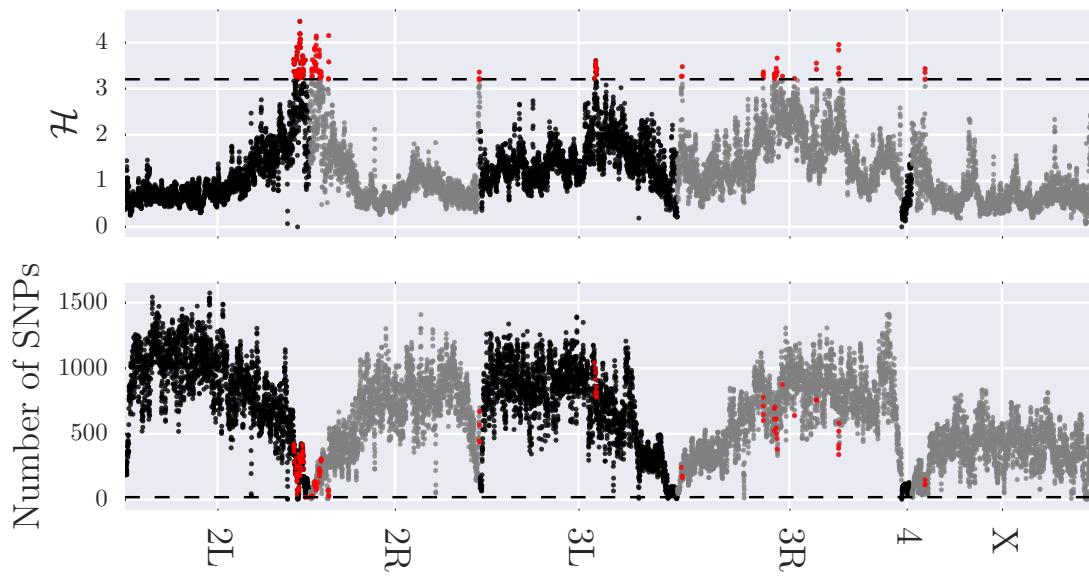


Fig. S8: Manhattan plot. Distribution of the composite \mathcal{H} statistic (top) and the number of SNPs (bottom) for 50Kbp sliding window with steps of 10Kbp across genome. Regions in which their \mathcal{H} statistic falls in top one percentile distribution is denoted with red color. Pearson correlation between the number of SNPs and \mathcal{H} of all windows is -0.03 and is -0.27 when restricting to the candidate regions is -0.27. This nine-fold increase in correlation implies that the \mathcal{H} statistic takes more extreme values as the number of SNPs in the window is less than expected (≈ 1100). [Manhattan plot of what?](#)

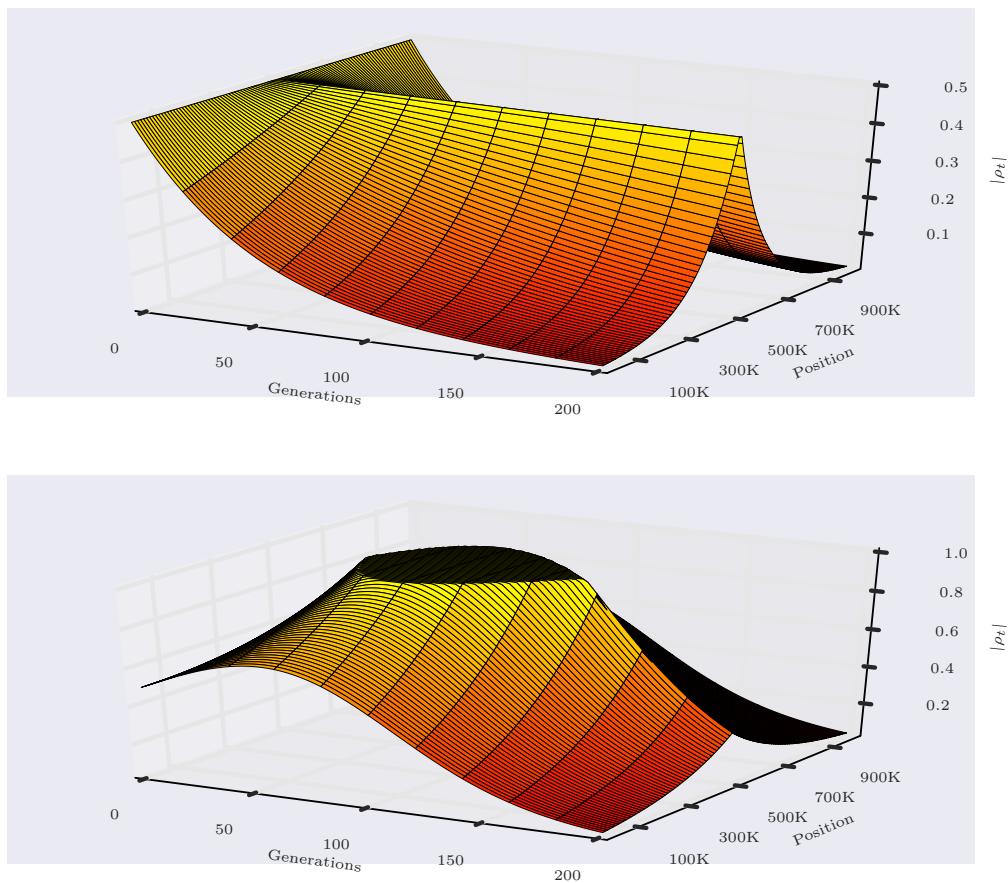
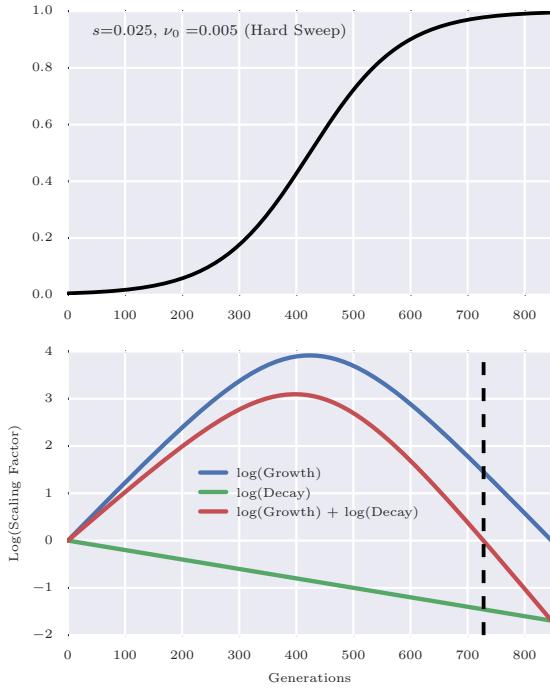
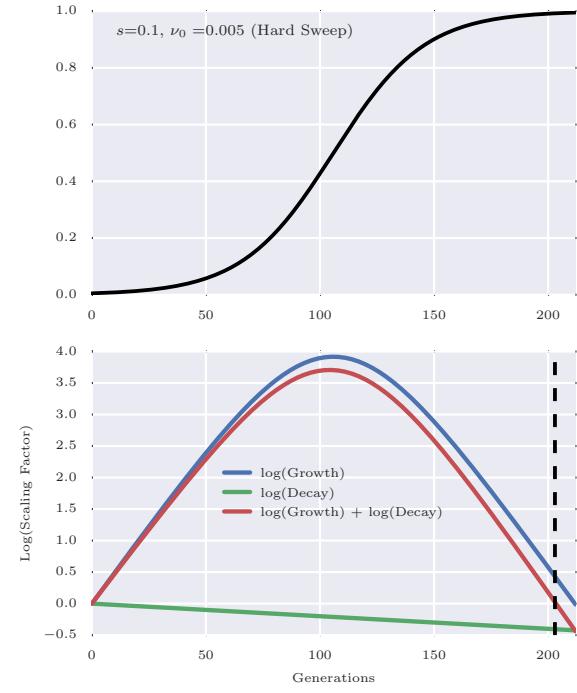


Fig. S9: Dynamic of LD. Decay of LD ($|D'|$ measure) of the minimum AF site at position 500K with the rest of genome in genetic drift with $r = 2 \times 10^{-8}$ (top) and hard sweep with $s = 0.01$ (bottom).

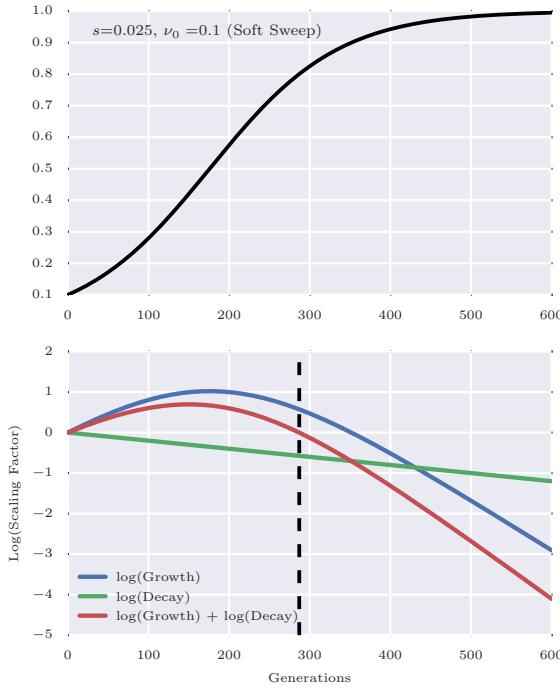
(A)



(B)



(C)



(D)

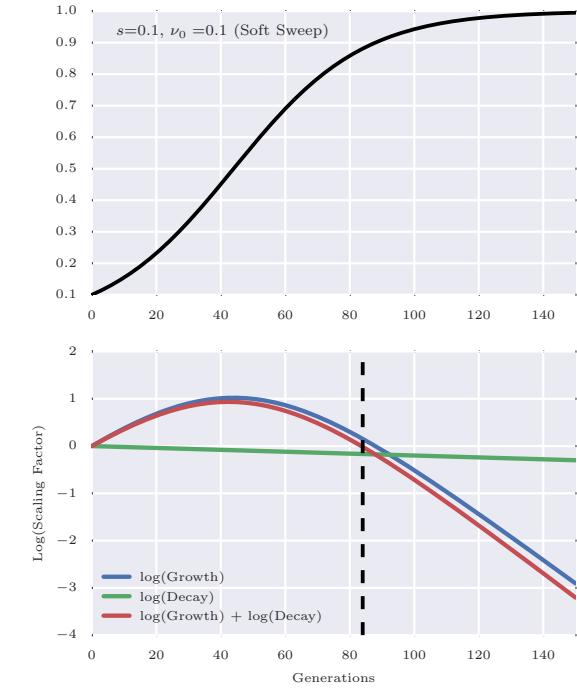


Fig. S10: Interaction between growth and decay factors of LD. Expected evolution of LD under natural selection for weak selection ($s=0.01$) and a distance of 100Kb between sites. In this setting, after about 1000 generations LD start to decay (red curve).

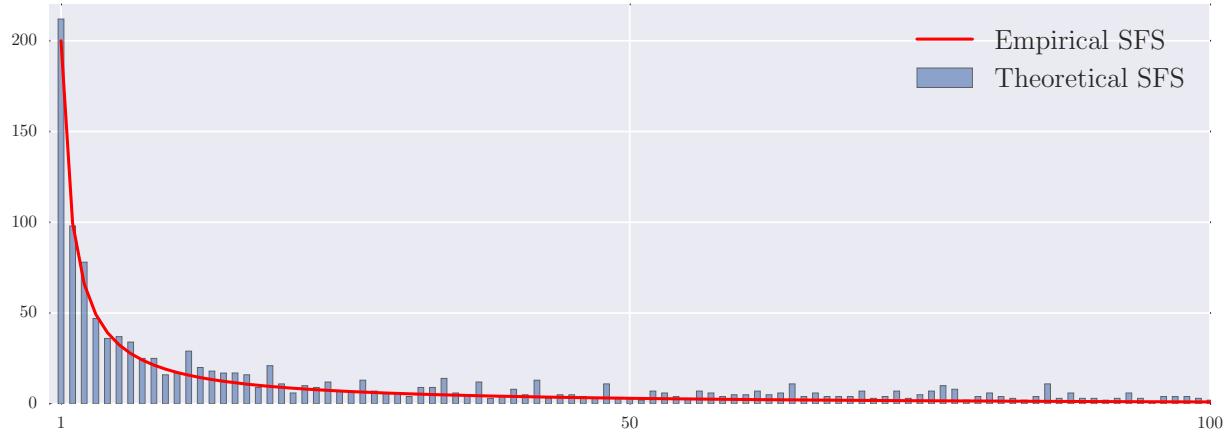


Fig. S11: Site Frequency Spectrum. Theoretical and Empirical SFS for a neutral population of 200 individuals.

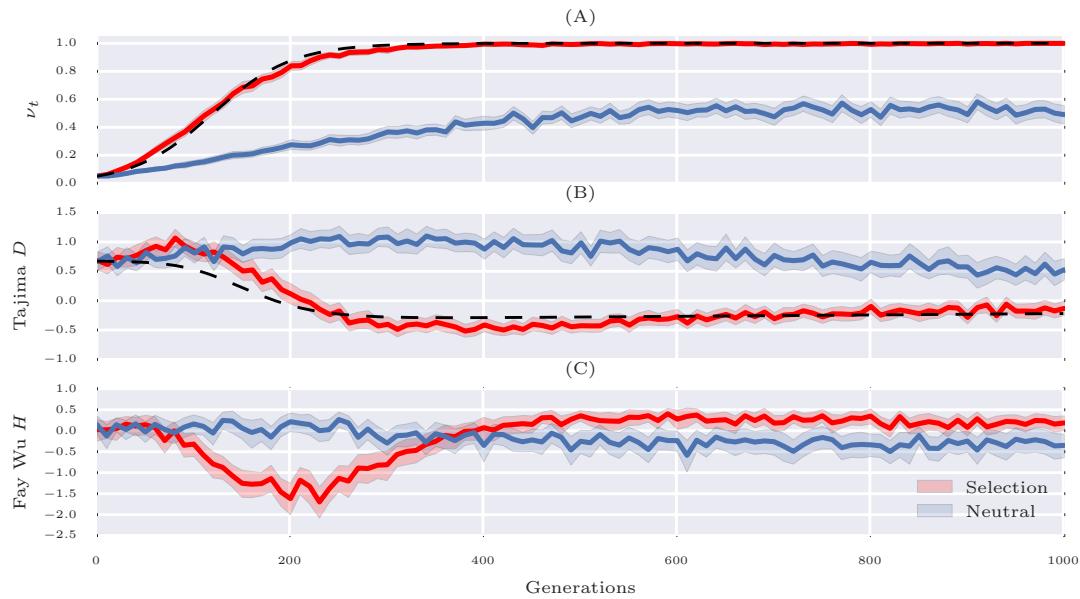


Fig. S12: Dynamic SFS-base statistics. Mean and 95% CI of 100 simulations for neutral (blue trajectories) selection with $s = 0.1$ (red trajectories).

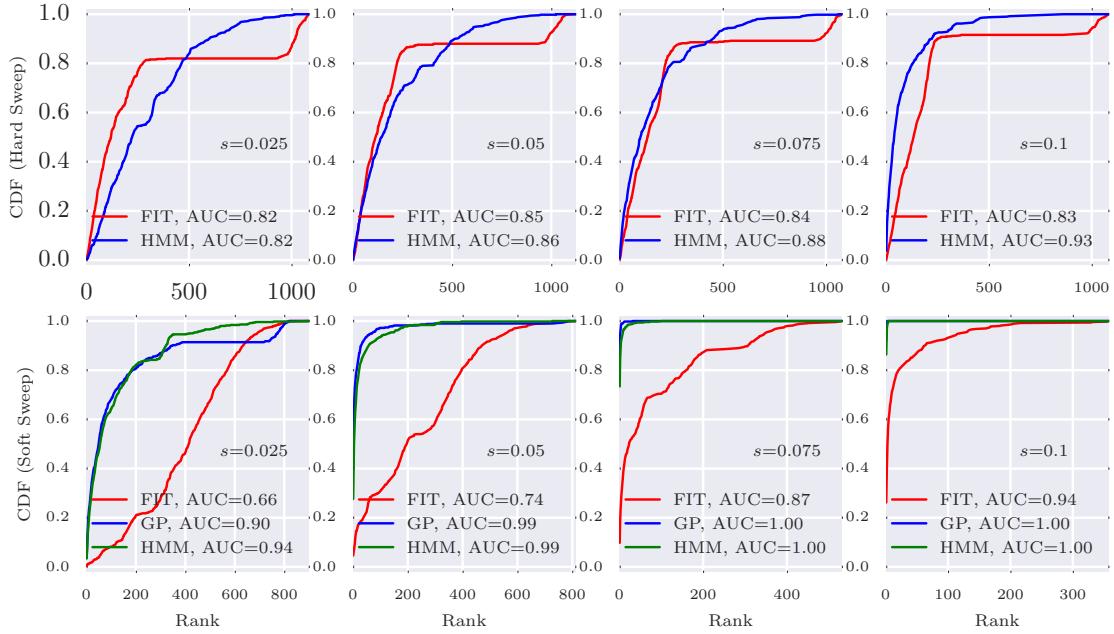


Fig. S13: Ranking performance for 30X coverage. Cumulative Distribution Function (CDF) of the distribution of the rank of the adaptive allele in 500 simulations for Markov Chain (\mathcal{M}), Gaussian Process (GP) and Frequency Increment Test (FIT), for different values of selection strength s and initial carrier frequency, when coverage is 30. Area Under Curve (AUC) is computed as a quantitative measure ranking performance of methods for each configuration.

Why do we need this figure? CMH? Ranking what performance?

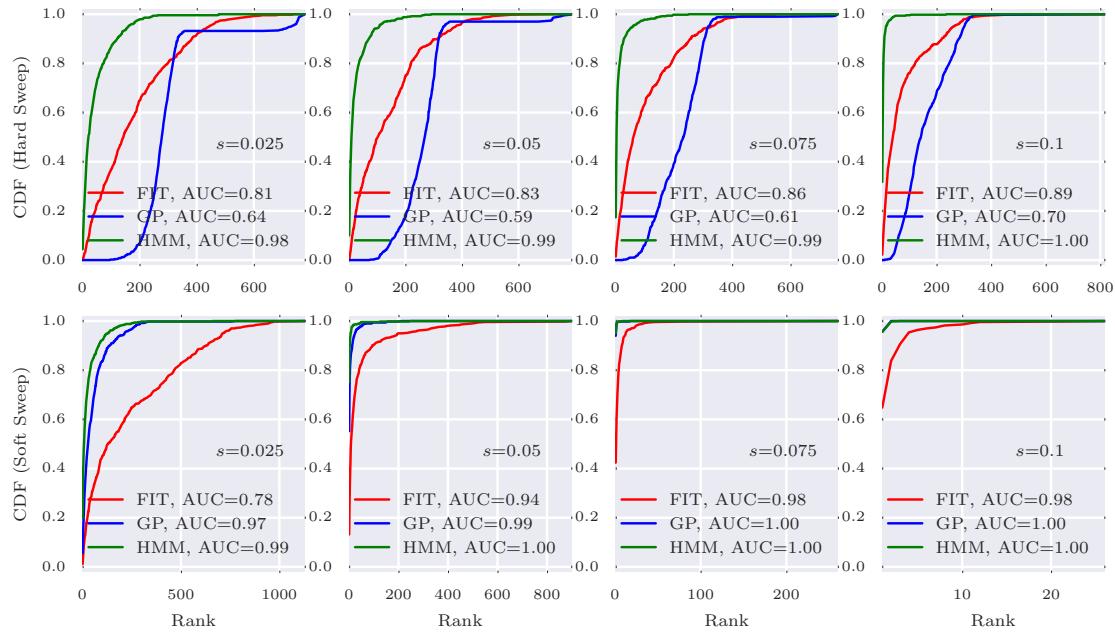


Fig. S14: Ranking performance for 300X coverage. Cumulative Distribution Function (CDF) of the distribution of the rank of the adaptive allele in 500 simulations for Markov Chain (\mathcal{M}), Gaussian Process (GP) and Frequency Increment Test (FIT), for different values of selection strength s and initial carrier frequency, when coverage is 300. Area Under Curve (AUC) is computed as a quantitative measure ranking performance of methods for each configuration.

[Why markov chain and not HMM?](#)

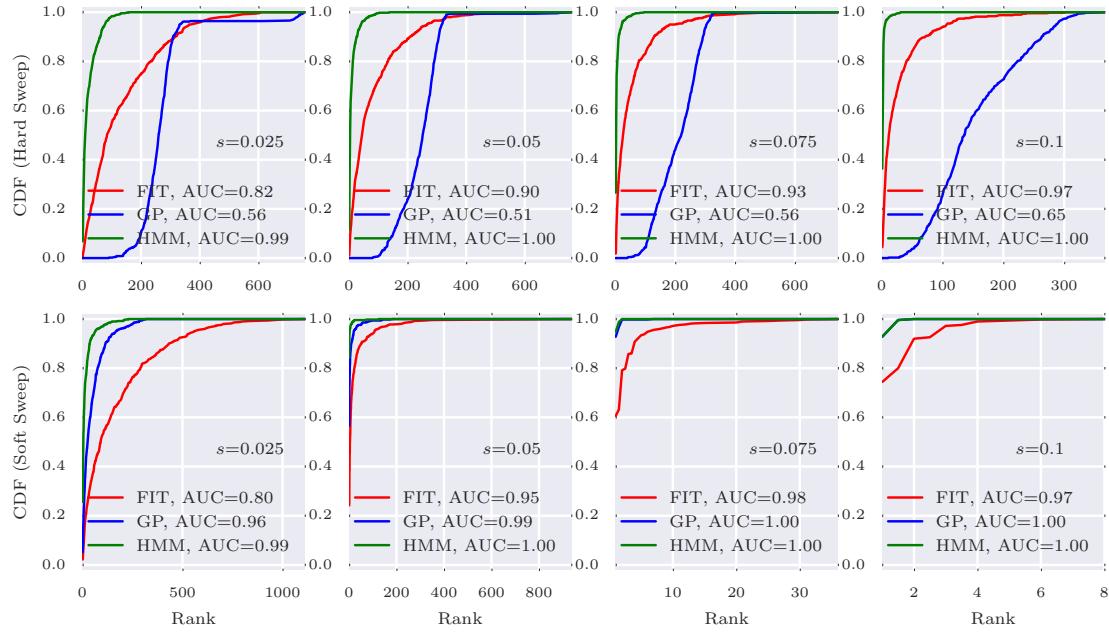


Fig. S15: Ranking performance for infinite coverage. Cumulative Distribution Function (CDF) of the distribution of the rank of the adaptive allele in 500 simulations for Markov Chain (\mathcal{M}), Gaussian Process (GP) and Frequency Increment Test (FIT), for different values of selection strength s and initial carrier frequency, when coverage is infinite. Area Under Curve (AUC) is computed as a quantitative measure ranking performance of methods for each configuration.

Why Markov chain?

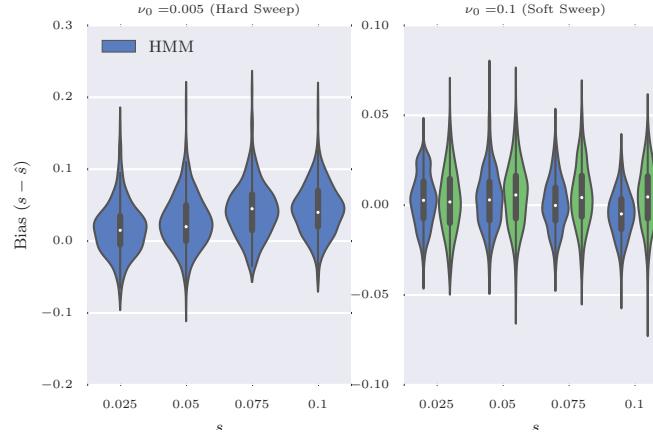


Fig. S16: Distribution of bias for 30X coverage. The distribution of bias ($s - \hat{s}$) in estimating selection coefficient over 500 simulations is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0 , when coverage is 30.

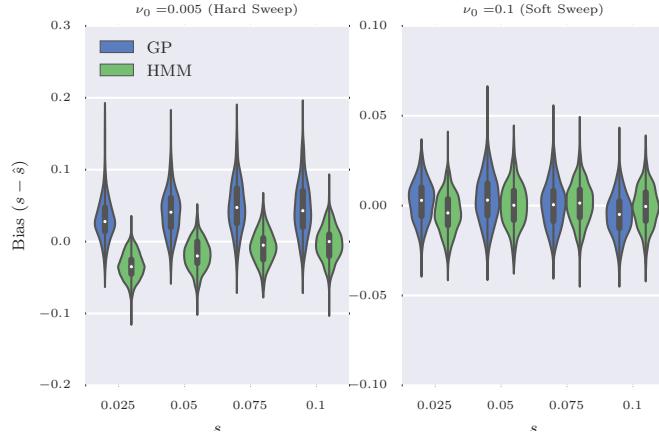


Fig. S17: Distribution of bias for 300X coverage. The distribution of bias ($s - \hat{s}$) in estimating selection coefficient over 500 simulations is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0 , when coverage is 300.

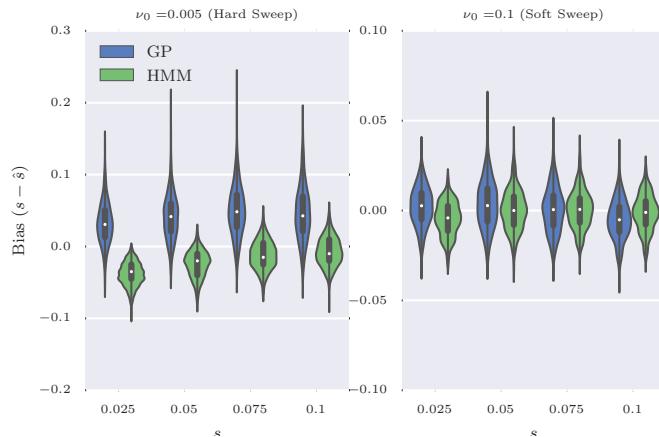


Fig. S18: Distribution of bias for infinite coverage. The distribution of bias ($s - \hat{s}$) in estimating selection coefficient over 500 simulations is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0 , when coverage is infinity.

Hard Sweep			Soft Sweep		
λ	Method	Avg Power	λ	Method	Avg Power
100	\mathcal{H}	30	100	\mathcal{H}	64
∞	\mathcal{H}	29	∞	\mathcal{H}	63
30	\mathcal{H}	23	∞	GP	62
100	CMH	19	100	GP	60
30	CMH	10	100	CMH	59
∞	GP	7	30	\mathcal{H}	59
100	GP	7	30	GP	56
30	GP	7	30	CMH	44
∞	FIT	5	∞	FIT	37
100	FIT	4	100	FIT	21
30	FIT	3	30	FIT	7

Table S1: Average Power – average true-positive-rate when FDR ≤ 0.05 – for detecting selection in a 50Kbp region for Frequency Increment Test (FIT), Gaussian Process (GP), Markov Chain (\mathcal{M}), Hidden Markov Model (\mathcal{H}) on 1,000 simulations for each of selection strength s and initial carrier frequency ν_0 . The power is averaged over all selection coefficients to provide a single value for comparison.

Better to group and sort by coverage, rather than method

Hard Sweep					Soft Sweep				
λ	π	LR	Method	Avg Power	λ	π	LR	Method	Avg Power
∞	99	+	\mathcal{M}	31	∞	100		\mathcal{M}	73
∞	99		\mathcal{M}	31	∞	100	+	\mathcal{M}	72
∞	100		\mathcal{M}	26	∞	99	+	\mathcal{M}	70
∞	50		\mathcal{M}	26	∞	99		\mathcal{M}	69
∞	100	+	\mathcal{M}	24	∞	50		\mathcal{M}	64
∞	50	+	\mathcal{M}	24	∞	50	+	\mathcal{M}	63
∞	0	+	\mathcal{M}	24	∞	0	+	\mathcal{M}	63
∞	0		\mathcal{M}	20	∞	0		\mathcal{M}	46
100	50	+	\mathcal{H}	24	100	99	+	\mathcal{H}	65
100	0	+	\mathcal{H}	24	100	99		\mathcal{H}	65
100	50		\mathcal{H}	24	100	99		\mathcal{M}	65
100	50	+	\mathcal{M}	22	100	50	+	\mathcal{H}	64
100	0	+	\mathcal{M}	22	100	0	+	\mathcal{H}	64
100	99		\mathcal{M}	22	100	50		\mathcal{H}	64
100	50		\mathcal{M}	22	100	50	+	\mathcal{M}	62
100	99		\mathcal{H}	21	100	0	+	\mathcal{M}	62
100	99	+	\mathcal{H}	20	100	100		\mathcal{H}	62
100	0		\mathcal{H}	16	100	50		\mathcal{M}	62
100	99	+	\mathcal{M}	14	100	100	+	\mathcal{H}	61
100	100		\mathcal{M}	12	100	100		\mathcal{M}	61
100	100	+	\mathcal{H}	11	100	99	+	\mathcal{M}	58
100	100		\mathcal{H}	11	100	100	+	\mathcal{M}	54
100	0		\mathcal{M}	8	100	0		\mathcal{H}	42
100	100	+	\mathcal{M}	7	100	0		\mathcal{M}	25
30	50	+	\mathcal{H}	18	30	50	+	\mathcal{H}	58
30	0	+	\mathcal{H}	18	30	0	+	\mathcal{H}	58
30	50		\mathcal{H}	17	30	50		\mathcal{H}	57
30	50	+	\mathcal{M}	12	30	99		\mathcal{H}	53
30	0	+	\mathcal{M}	12	30	99	+	\mathcal{H}	51
30	99		\mathcal{H}	12	30	99		\mathcal{M}	48
30	50		\mathcal{M}	12	30	50	+	\mathcal{M}	47
30	99	+	\mathcal{H}	10	30	0	+	\mathcal{M}	47
30	99		\mathcal{M}	10	30	100		\mathcal{H}	47
30	0		\mathcal{H}	6	30	50		\mathcal{M}	47
30	100	+	\mathcal{H}	5	30	100	+	\mathcal{H}	46
30	99	+	\mathcal{M}	5	30	100		\mathcal{M}	38
30	100		\mathcal{H}	5	30	99	+	\mathcal{M}	34
30	100	+	\mathcal{M}	4	30	100	+	\mathcal{M}	25
30	100		\mathcal{M}	3	30	0		\mathcal{H}	14
30	0		\mathcal{M}	3	30	0		\mathcal{M}	6

Table S2: Average power, average true-positive-rate when FDR ≤ 0.05 , for detecting selection in a 50Kbp region for Frequency Increment Test (FIT), Gaussian Process (GP), Markov Chain (\mathcal{M}), Hidden Markov Model (\mathcal{H}) on 1000 simulations for each of selection strength s and initial carrier frequency ν_0 . To incorporate ascertainment bias, each setting is evaluated such that depth of each SNP is identically distributed from Poisson(λ) for $\lambda \in \{30, 100, \infty\}$.

	CHROM	start	end
1	3L	14330000	14420000
2	3L	14490000	14540000
3	3R	10640000	10730000
4	3R	11770000	11850000
5	3R	11980000	12240000
6	3R	12250000	12330000
7	3R	12360000	12410000
8	3R	12780000	12900000
9	3R	13040000	13100000
10	3R	13160000	13220000
11	3R	14380000	14460000
12	3R	14470000	14520000
13	3R	14540000	14590000
14	3R	14700000	14760000
15	3R	14920000	14980000
16	3R	17240000	17390000
17	3R	18300000	18360000
18	3R	18630000	18690000
19	3R	20050000	20130000
20	3R	20230000	20280000
21	3R	20400000	20450000

Table S3: Coordinates of the enriched intervals of the *D. melanogaster* adaptation to alternating temperatures experimental evolution data (see part 3.1), assembly 5.7, found by COMALE statistic.