

Detecting Selection in Experimental Evolution Experiments

Arya Iranmehr¹, Ali Akbari¹, and Vineet Bafna²

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

²Computer 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 wild and controlled environments. This paradigm of experiment was infeasible until recently when the whole-genome and whole-population was made possible by next-generation sequencing technologies. However, one of the primary constraints of the EE studies is the limited time for the experiment, which primarily depend on the organism’s generation time. This constraint impedes adaptation and optimization (evolvability) studies, where the population can only evolved and re-sequenced in a small number of generations, relative to the number of generations required for fixation of adaptive allele. Although a powerful library of tests-of-selection has already been developed, they are mainly designed for static data to identify adaptation when the sample is taken close enough (before/after) to the fixation of adaptive allele. In this article, we study the problem of identifying selective sweep in short-term experimental evolution of sexual organisms and propose Composite Of MArkovian Likelihoods for Experimental evolution (COMALE) statistic which computes its score by averaging likelihood ratios of polymorphisms for a genomic region. The likelihood of null (neutral) and alternative (selection) hypotheses calculated using the Wright-Fisher Markov chain model for each variant. Extensive simulation study shows that COMALE achieves higher detection power methods on both soft and hard sweep simulations for various selection strengths. Finally, we apply the COMALE statistic to the controlled experimental evolution of *D. melanogaster* to detect adaptive genes/alleles under alternating cold and hot temperatures.

1 Introduction

in have

Experimental Evolution. Recent advances of whole genome sequencing has enabled us to sequence populations at a reasonable cost to perform *longitudinal studies* and study different forces of evolution in real-time. Modern experimental evolution refers to the study of the evolutionary processes of a model organism at genomic level in a controlled [6, 8, 23, 31, 32, 41, 42] or natural [4, 7, 13, 14, 34, 46, 66] environment. Although constraints such as small population sizes, limited timescales and oversimplified laboratory environments limits interpreting experimental evolution results, they can be used been used to test different hypotheses [30] regarding mutation rate, inbreeding, environmental variability, sexual selection and conflict, kin selection and cooperation life history and sex allocation, sexual reproduction and mating systems, behavior and cognition, host-parasite interactions, speciation repeatability of evolution and make more accurate inferences than static data analysis [10, 16, 51]. In addition, dynamic data has been used to estimate model parameters including population size [43, 58, 62–64] strength of selection [9, 25, 26, 33, 36, 55, 58], allele age [33] recombination rate [58], mutation rate [5, 58] and test neutrality hypotheses [7, 11, 20, 58].

Among different types of evolution experiments [5, 52] in this paper we only focus on adaptive evolution of multicellular sexual organisms with continues culture, fixed population size, single

why only one

locus selection (only one causal mutation). For this setting, *D. melanogaster* is usually the model organism and it has been used to identify adaptive genes in longevity and aging [11, 47] (600 and generations respectively), courtship song [60] (100 generations), hypoxia tolerance [68] (200 generations), adaptation to new temperatures[41, 59] (59 generations), egg size [28] (40 generations), C virus resistance [35] (20 generations), and dark-fly [27] (49 generations) experiments.

Natural selection. Natural selection is one of the main forces of the evolutionary process and identifying it at the genomic level is one of the critical problems facing humanity. For example, drug resistance in HIV [21], cancer [22, 67], malaria [3, 38], pests [12] or antibiotic resistance [54] are instances of genetic adaptations that yet to be understood at genomic level, which could potentially return us to the pre-antibiotic age. Although, a wide range of computational methods [61] enable us to identify different regimes of genetic adaptations, Messer and Petrov [37] argued that “many, if not most, cases of adaptation are yet to be discovered”, false-negatives. In addition, current methods prone to “pathological false-positives” due to other confounding factors such as demography.

Selective sweep. Selective sweep [29, 53] is the model for describing directional single-locus selection, which takes into account of associations of the beneficial mutation with its surrounding loci. The extent of genetic loci that are in association with the adaptive allele depends on the amount of accumulated recombination events between adaptive allele and the rest of genome.

In the asexual populations, where no crossover occurs, the whole chromosome is perfectly linked to the adaptive allele the whole chromosome “hitchhikes” with the beneficial mutation in the sweep process. Also, when (beneficial) mutation rate is high, it is possible that more than one beneficial allele exist in the population at the same time and “clonal interference” [15, 32] best describes the adaptation process.

On the other hand, in sexual population, the favored mutation is only in linkage-disequilibrium (LD) with its nearby polymorphisms. Hence, methods for identifying selective sweep in sexual populations often analyze polymorphism data of a population of in a genomic region, rather than a single site.

Identifying Natural Selection. Adaptation leaves a variety of signatures in different kinds of genomic data, and methods for identifying natural selection are essentially *data-driven*. For instance, reduction in genetic diversity[19, 49, 57] in allele-frequency data, prevalence of long haplotypes [50, 61] in haplotype (phased) data, population differentiation [11, 24] in multiple-population data and rapid increase in allele frequencies [7] in the dynamic data are different signatures of selective sweep in the polymorphism data. In this paper, we restrict our attention to the experimental evolution experiments with pooled-sequencing, where dynamic allele frequency of the population is available.

The identification of an selection event can be done in different levels of detail. At the coarsest level, identification can be done by determining whether a region (e.g. a small region with no or low recombination) on genome is under selection. In the rest, we consider this task to be the task of *detection*. Then, finding the causal mutation/allele would be a more elaborate identification of selection, henceforth, *locating* selection. Finally, estimating model parameters such as strength of selection and overdominance at the site fully describes the selective sweep.

Static Data. Traditionally, given static allele frequency data, Site Frequency Spectrum (SFS) is computed to perform neutrality tests including Tajima’s D [57], Fay and Wu’s H [19], Composite Likelihood Ratio [40], SFSelect [49], in a genomic region¹. Despite their simplicity and clarity, it has been shown that SFS-based tests often fail to distinguish demographic changes from adaptation.

¹The extent of genomic region is mainly depend on the amount background linkage in the genome.

They are also prone to pathological false-positive/negatives due to low linkage of the adaptive allele to its surrounding variation and ascertainment bias [2, 37, 39, 44, 45].

SFS (Figure 1) shows the distribution of allele frequencies in a genomic region. Historically, SFS of static data has been extensively used to “detect” genetic adaption and demographic changes in a population by measuring the diversity in a genomic region. In general, reduction in diversity is a signal of selection, and detecting selection based on the reduction in genomic diversity is a subtle task, because

definition?

- (i) in soft sweeps the genomic diversity does not necessarily reduce.
- (ii) even in hard sweeps with no recombination, the reduction diversity is significant only when the SFS sample is taken close to fixation (not too far prior or after fixation)
- (iii) SFS change in the same way for both selection and demography changes.

positive and negative tests are mixed as also the methods.

Although conditions (hard sweep with no recombination, not far from fixation, and random-mating and constant size population) for detecting selection based on SFS are very restrictive, SFS-based tests are simple and inexpensive to use and often used in combination with other tests [2, 61].

In practice, (positive and sum-to-one) weighted linear combinations scaled SFS distribution [1] is used as different estimators of θ and pairwise discrepancy between them is used as a test statistic for detecting selection. Under neutrality, the discrepancies should be distributed around zero, and a simple t test can provide p-value for rejecting neutrality. For example, test statistics for Tajima’s D [57], Fay Wu’s H [19] and SFSelect [49] can be obtained by a dot product of the scaled SFS vector with their corresponding weight vector.

too harsh

Dynamic Data (ad-hoc). On the other hand, tests-of-selection for dynamic data is less studied, and often existing tests for static data are adopted for dynamic data in an ad-hoc manner. For example, Zhu et al. [68] used the ratio of the estimated population size of case and control populations to compute test statistic for each window. Burke et al. [11] applied Fisher exact test to the last observation of data on case and control populations. Bergland et al. [7] applied F_{st} to populations throughout time to signify their differentiation from ancestral as well as geographically different populations. Jha et al. [28] computed test statistic of generalized linear-mixed model (GLMM) directly from read counts.

Dynamic Data (general). To perform statistical test on time-series data directly, Bollback et al. [9] provided diffusion approximation to the continues Wright Fisher Markov process and estimated s numerically. Then, they tested likelihood ration on X^2 distribution.

Feder et al. [20] proposed Frequency Increment Test(FIT) for dynamic frequency data and Empirical Likelihood Ratio Test (ELRT) instead of X^2 test which performs poorly when the number of independent samples (replicates) are small. Specifically, FIT is a neutrality test which uses (continuous-time continuous-state) Brownian motion process for modeling variation of allele frequencies under genetic drift. Basically, given current allele frequency ν_t at a site, Brownian motion approximation assumes future-generations allele frequencies are drawn from the Gaussian distribution

$$\nu_{t+\tau} \sim \mathcal{N} \left(\nu_t, \frac{2\nu_t(1-\nu_t)}{N_e} \tau \right) \quad (1)$$

In other words, under neutrality, increments of the consecutive observations drawn from a Gaussian with zero mean and variance of (1), and p-value of and be readily computed via a Student’s t-test. More recently, Song et al. [58] computed LRT statistic by fitting parameters to a Gaussian process model to the time series frequency data.

Notation. Let $\mathbf{X} = (X_{ijk}) \in [0, 1]^{T \times M \times R}$ denote the population frequency where T is the number of samples in time, M is the number of segregating sites, and R is the number of replicates. Samples in time are taken such that $\tau_1 < \tau_2, \dots < \tau_T$. To simplify our notation, we define x_t as a allele frequency of a site in a replicate, and we denote it by $\nu_t(s, h)$ if such a site is directly under selection with the strength s and overdominance of h . For notation convenience, henceforth we omit dependence of ν_t to s and h . Thus, ν_t denotes the frequency of the carrier of beneficial allele in the population t generations, since the onset of selection.

Hard and Soft Sweep. Selective sweeps are classified by the amount of variation exist in the individuals carrying adaptive allele. By definition, hard sweep is the case when all the carriers coalesce after onset of selection and thus diversity between carriers is at its minimum. In general, excess of genetic variation in carriers of adaptive allele, makes it difficult to detect selection, i.e., soft sweep is difficult to detect. In the following, we conduct and evaluate our simulations for soft and hard sweep separately. Since here we do not take into account of de novo mutations, soft sweep can only happen in standing variation where the site under selection is at frequency is larger than $1/F$. In contrast, hard sweep experiments are those that their adaptive allele is at its minim frequency, $1/F$, at the onset of selection.

Challenges. The main constraint in the adaptive experimental evolution is the sampling-time-span (STS), the number of generations between the first and last sampled generations. Given a fixed amount of time for an study, the longer generation times of the organism, the smaller number of generations can be evolved and re-sequenced. This implies that, only quite strong selection pressures, STS will be of the same order of fixation time. As a result, the time series data usually only captures a “partial sweep”.

Moreover, in controlled experimental evolution experiments, populations are evolved and inbred. This scenario in which population size significantly drops from the large number of wild type (e.g. D. melanogaster $N_e \approx 10^6$) to a small number (typically F is between 100-1000) of founder lines for EE, resembles a severe population bottleneck. Such a intense reduction in effective population size increases the variance of binomial sampling in the Wright-Fisher model, and consequently makes genetic drift quite strong. Also, signal of selection in SFS is absorbed by such a strong bottleneck and makes it harder to identify selection.

2 Results

2.1 Tests for detecting selection on time series data.

FIT performs poorly when $\tau \neq 1$ or ν_0 is not close to 0.5. First we show that Brownian motion provides a poor approximation for the distribution of future allele frequencies ν_τ when the initial frequency ν_0 is not close to 0.5, or when sampling in time is sparse i.e., $\tau > 1$. To test this, we simulated 150K simulations for $\nu_0 = 0.005$ and $\tau \in \{1, 10, 100\}$ and 100K simulations for $\nu_0 = 0.1$ and $\tau \in \{1, 10, 100\}$. As shown in Figure 3, Brownian motion performs poorly when initial frequency is small or the sampling is sparse.

At the start of selection, ν_0 is at its minimum in hard sweep and is at low frequency in the soft sweep with standing variation due to SFS of a neutral population(Figure 1). Thus if the selecting site is chosen randomly with high probability its frequency will be far from 0.5 and FIT will be end up with high rates of false positives/negatives. Moreover, in experimental evolution experiments sampling in time is quite sparse e.g. $\tau_i - \tau_{i-1}$ is of the order of 10-100 generations in D melanogaster studies.

Markovian Likelihood for Neutral Evolution. Here we turn the problem of small population size in EE to our advantage to compute likelihood of time series observation exactly using Wright-Fisher Markov chain. Precisely, (discrete-time discrete-state-space) Wright-Fisher Markov Chain with transition matrix P for population of N diploid individuals [18]

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

where $P_{i,j}$ denotes probability of changing allele frequency of $\frac{i}{2N}$ to $\frac{j}{2N}$ in one generation, due to genetic drift. Transition probabilities for τ generations ahead can be computed by powering the transition matrix to τ . Figure 3 illustrates expected and observed distribution of frequencies after 1, 10 and 100 generations when starting frequency is 0.005 (top) and 0.1 (bottom) using Markov chain provides an accurate distribution of allele frequencies for future generations, under genetic drift.

Markovian Likelihood for Selection. We can also compute the Markov chain for single locus selection (see section 3), compute the conditional distribution $\Pr(\nu_{t+1} \mid \nu_t, s)$, and maximum likelihood estimate of s accordingly. Similar to the null model, we can compute distribution of t generation ahead by powering the transition matrix to t . We conducted 100K simulations with selection strength $s = 0.1$ on a site with frequency $\nu_0 = 0.005$ and computed empirical and predicted distribution of frequencies after 1, 10 and 100 generations by Markov chain (3). As shown in Figure 4 the distribution of empirical data follows that of Markov chain.

SFS based statistics in time series. We also considered SFS based tests including Tajima's D, Fay & Wu's H and SFSelect for detecting selection. As these tests work only on static data, we extend them for time series data for both null and alternative hypotheses. More precisely, we explicitly defined functionals D_t and H_t as function of initial carrier frequency ν_0 and strength of selection s (see section 3 for details). To show that the proposed models of D_t and H_t are valid models, we simulated 1000 populations and computed SFS based statistics every 10 generation and compared them with the proposed model. As shown in the Figures 1 proposed models are more consistent with data.

SFS based tests will fail when carrier frequency ν_t is low. Importantly, Figures 1 shows the increase in variance of the trajectories associated with SFS, compared to carrier frequency. Moreover, it is difficult to distinguish SFS trajectories of selection and neutral populations, when carrier frequency is small, e.g. first 50 generations of Figures 1. This, observation can be verified by examining the terms in the functional form of D_t . As shown in the Figure 5 right, in early generations of hard sweep where carrier frequency is low, D_t is either positive or close zero. In other words, the reduction in diversity become significant when carrier frequency is high enough. It can be shown that this argument holds for H_t in early generations of hard sweep, where carrier frequency is not high.

Comale is not even defined.

Power of detection. To evaluate detection power of COMALE statistic, we compare it with Gaussian process (GP) [58], FIT [20] statistics. For every experiment, (with a specific strength of selection, initial carrier frequency, sampling time span and number of replicates), we conduct 1000 simulations of which half are neutral and the rest are under selection. We define power of each method as average true-positive rate when false-positive rate is less than 0.05 over 1000 simulations.

First we compare all the methods under default parameter settings, i.e., $T = 5$, $\tau = 50$ and $R = 3$. As shown in the Figure 12, COMALE statistic provides a better detection power in both hard and soft sweep regimes for all values of selection strength.

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 beneficial allele.

For each setting of ν_0 and s , we conducted 100 simulations and computed the rank of the beneficial mutation in the sorted scores. Then for each configuration, Cumulative Distribution Function (CDF) of the distribution, of rank of the adaptive allele for 100 simulations is plotted in Figure 14.

COMALE works the best in hard sweep regime. For example, when $\nu_0 = 0.005$ and $s = 0.01$ (weakest selection), the beneficial allele is ranked first, in more than 60 experiments, and take rank of less than 5 in more than 90% of experiments of hard sweep. Accuracy of the locating the adaptive allele diminished as sweep become softer, i.e. larger values of ν_0 . Yet, in the worst case the beneficial allele is tanked among top 50 SNPs.

Strength of Selection Finally, estimating the model parameters such as strength of selection provides the finest detail of identification of a selective sweep process, since given the model parameters we can predict the state of population, e.g. fixation time, in future generations.

To estimate s , likelihood of data is computed via Wright-Fisher Markov chain (3), and maximum likelihood \hat{s} is readily computed. We computed bias, $s - \hat{s}$ for each experiment of COMALE , and GP. The distribution of the bias is presented in Figure 13 for different configurations. In general, both GP and COMALE are biased for weak selections, which genetic drift dominates. However, for stronger sweeps, e.g. $s = 0.1$, COMALE provides estimates with smaller bias and variance.

2.2 Analysis of Real Data

We finally apply COMALE method to the controlled experimental evolution experiment of [41], which evolves 5 replicates of a population of *Drosophila melanogaster* for 37 generations under alternating 12-hour cycles of hot (28°C) and cold (18°C) temperatures. Three replicates are sampled at the first generation, 2 replicates at generation 15, one replicate at generation 23, one replicate at generation 27 and three replicates at generation 59.

Heterogeneity of read depths. COMALE statistic is basically computed using site allele frequencies at different generations for all replicates. For real Pool-Seq data, however, allele frequencies is unknown and only read counts at each site measured. As shown in Figure ?? read depths are highly heterogeneous and filtering low coverage sites from data strictly reduces the number of SNPs. For example, by setting minimum read depth at each site (for all replicates and generations), the number of SNPs shrinks from 1,544,374 to 10,387.

Uncertainty in allele frequency. In addition to coverage heterogeneity, allele frequencies in principle are *hidden* variables for *observed* data, alleles read counts. We take into account of allele frequency uncertainty by computing the likelihood of data using Hidden Markov Model (HMM), see section ?? for details.

SNP-based analysis with ELRT statistic Manhattan plots in Figure 15 depicts the distribution of the top 2000 SNPs and corresponding 1961 genes. The list of genes is available in the same directory under name `genes_SNP.txt`.

The following genes are enriched with P-value of 0.0006725 using Fisher exact test and genes of 5.1.

FBgn0033216

FBgn0034662

FBgn0039611

FBgn0032297
 FBgn0041342
 FBgn0037788
 FBgn0039040
 FBgn0038236
 FBgn0039787
 FBgn0038701
 FBgn0038702
 FBgn0039760

Window-based analysis with COMALE statistic Manhattan plots in Figure 15 depicts the distribution of the top 100 50K regions and corresponding 421 genes. The list of genes is available in the same directory under name `genes_SNP.txt`.

3 Materials and Methods

Simulations For each experiment a diploid population is created and evolved as follows.

- I. **Creating initial founder line haplotypes** First using `msms` program, 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 and $μ = 2 * 10^{-9}$ is mutation rate and $L = 50K$ is the window size in base pairs which gives $θ = 2μN_eL = 200$ and $ρ = 2N_e r L = 2000$. For default parameter, the expected number of segregating sites in a window is

$$\mathbb{E}[M] = θ \sum_{i=1}^{F-1} \frac{1}{i} = 1175$$

- II. **Creating initial diploid population** To implement similar setting for experimental evolution of diploid organisms, initial haplotypes first cloned to create F diploid homozygotes. Then each diploid individual is cloned N/F times to yield diploid population of size N .
- III. **Forward Simulation** Given initial diploid population, position of the site under selection, selection strength s , number of replicates $R = 3$, recombination rate $r = 2 * 10^{-8}$ and sampling times $\mathcal{T} = \{10, 20, 30, 40, 50\}$, `simuPop` is used to perform forward simulation and compute allele frequencies for all of the R replicates. Also, to avoid spurious simulation samples, simulation results are constrained to those that the beneficial allele escapes stochastic loss of genetic drift and *establishes* in all the replicates.

Markov Chain In an diallelic diploid Wright-Fisher infinite size population with single locus selection with genotypes $0|0$, $0|1$ and $1|1$, and relative fitness $w_{00} = 1$, $w_{01} = 1 + hs$ and $w_{11} = 1 + s$ relative fitness respectively, then allele frequency of the adaptive allele change according to

$$ν_{t+1}(s, h) = \frac{w_{11}ν_t^2 + w_{01}ν_t(1 - ν_t)}{w_{11}ν_t^2 + 2w_{01}ν_t(1 - ν_t) + w_{00}(1 - ν_t)^2} = ν_t + \frac{s(h + (1 - 2h)ν_t)ν_t(1 - ν_t)}{1 + sν_t(2h + (1 - 2h)ν_t))} \quad (3)$$

where $s, h \in \mathbb{R}$ are selection strength and dominance parameters, respectively.

In a finite population size, however, the allele frequencies of adaptive allele are drawn from the Markov chain with transition probabilities ([18], eq 1.58-59)

$$Q_{i,j}(s, h) = \Pr \left(ν_{t+1} = \frac{j}{2N} \middle| ν_t = \frac{i}{2N} \right) = \binom{2N}{j} ν_{t+1}^j (1 - ν_{t+1})^{2N-j} \quad (4)$$

For a given \hat{s}, \hat{h} , the Markov chain 4 can be readily used to compute likelihood $\mathcal{L}(\hat{s}, \hat{h}|\mathbf{x})$ for the sequences of allele frequencies \mathbf{x} . Figure 4 shows that the predicted distributions of frequencies using Markov chain 3 after 10, 100 generations are consistent with simulated data.

Test statistic Likelihood of the trajectories under null and alternative hypothesis can be computed to perform likelihood ratio tests which provide predictors that are robust to pathological false-positives [20]. Specifically, for each site, we define the likelihood-ratio predictor

$$\lambda = s^* \log \left(\frac{\mathcal{L}(s^*, h^* | \mathbf{x})}{\mathcal{L}(0, 0 | \mathbf{x})} \right) \quad (5)$$

where \mathbf{x} is the vector of trajectories and \mathcal{L} is the likelihood function defined by the Markov chain with transition (3).

COMALE Test Given likelihood ratio (5) for each SNP we can further exclude false positives by combining likelihood ratio scores of the polymorphisms in a genomic region by computing Composite Likelihood Ratio(CLR)[40, 61, 65]. For each window, CLR is computed by taking average of top one percentile of the likelihood ratios.

Hidden Markov Model Let $\{(c_t, d_t)\}$ be the observed data for a locus of a replicate, where c_t is the read count of derived allele and d_t is the depth of the reads at that site, then for two time points we have

$$\begin{aligned} \Pr(\vec{x}_t; s, h) &= \sum_{\nu_t} \Pr(\vec{x}_t, \nu_t; s, h) = \sum_i (\vec{\alpha}_t)_i \\ \Pr(\vec{x}_t, \nu_t; s, h) &= \sum_{\nu_{t-1}} \underbrace{\Pr(x_t | \nu_t)}_{\vec{e}_{x_t}} \underbrace{\Pr(\nu_t | \nu_{t-1}; s, h)}_{\mathbf{Q}_{s,h}} \underbrace{\Pr(\vec{x}_{t-1}, \nu_{t-1}; s, h)}_{\vec{\alpha}_{t-1}} = (\vec{\alpha}_{t-1}^T \mathbf{Q}_{s,h}) \odot \vec{e}_{x_t} \end{aligned} \quad (6)$$

where Q is the state transition matrix defined in (4) and $\vec{e}_{x_t} \in [0, 1]^{2N}$ is the conditional distribution of the reads given allele frequency which is

$$\Pr(x|\nu) = \Pr(c, d|\nu) = \binom{c}{d} \nu^c (1-\nu)^{d-c} \quad (7)$$

Note that for a given s, h , the state transition $Q(s, h)$ is the same for all the sites, but observation-state distribution change from site to site. The complete likelihood is then computed

$$\mathcal{L}(s, h | \vec{x}_{1,T}, \dots, \vec{x}_{R,T}) = \prod_r \sum_i (\vec{\alpha}_{r,T})_i \quad (8)$$

Estimating parameters

$$s^*, h^* = \arg \max_{s,h} \sum_r^R \log (\mathcal{L}(s, h | \mathbf{x}_r)) \quad (9)$$

SFS based test on time series data To adapt SFS-based test for dynamic data, one can simply design statistical tests to examine whether the observations of D , H and SFSelect statistics are significantly deviated from the null distribution, i.e. constant values in time. On the other hand, we can model the dynamics of SFS as a function of selection strength to 1)have a model for estimating model parameter 2)obtain an insight regarding how does SFS statistics behave under different stages of selection.

As mentioned before, restricting population to F founder lines ($F \ll N_e$) resembles a severe bottleneck event, which confound SFS. Figure 6 demonstrates effect of controlled experimental evolution on different SFS statistics under neutral evolution for 1000 simulations. The mean of neutral simulations can be used to empirically filter out the effect of bottleneck in dynamic data.

I. **Tajima's D.** As shown in the Appendix 5.4, dynamic of Tajima's D in hard sweep is

$$D_t = D_0 - \log(1 - \nu_t) \frac{W_0}{\log(2N)} - \nu_t^2 \Pi_0 \quad (10)$$

where W_0 and Π_0 are Watterson and Tajima estimate of θ at the first generation. Given that ν_t defined in term of ν_0 and s in (16), we can use (10) to estimate the strength of selection by performing a nonlinear regression to find s . Then, a likelihood ratio, similar to (5), provides² a predictor for detecting selection in each window.

II. **Fay Wu's H** As shown in the Appendix 5.3, dynamic H statistic is directly related to average of Haplotype Allele Frequency (HAF) [48] and can be written

$$nH_t = \theta\nu_t \left(\frac{\nu_t + 1}{2} - \frac{1}{(1 - \nu_t)n + 1} \right) + \theta(1 - \nu_t) \left(\frac{n + 1}{2n} - \frac{1}{(1 - \nu_t)n + 1} \right) \quad (11)$$

which is a good approximation only in hard sweep with strong selection, Figures ?? and ?? (second row, left).

III. **SFSelect** statistic proposed by Ronen et al. [49] to predict selection by empirically learning the so-called *optimal* weights using Support Vector Machines.

Whether the dynamic of SFS in times-series can improve the detection power in the conditions (i-iii) is not well-studied yet. Evans et al. [17] developed diffusion equations for evolution of SFS in time series, they are difficult to solve. Hence, we analyze and visualize scalar test statistics D , H and SFSelect in the dynamic data under different selection regimes.

Dominance Note that, for $s > 0$ we have []

condition	comment
$h < 0$	underdominance
$h = 0$	recessive adaptive allele
$h = 0.5$	directional selection
$h = 1$	dominant adaptive allele
$h > 1$	overdominance

Software. 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 in second on a desktop computer with a Core i7 CPU and 16GB of RAM.

P value According to the Wilks theorem [] log odds ratios are asymptotically distributed according to \mathcal{X}^2 , yet [20] shown that in the empirical distribution is better when the number of independent samples (replicates) is small. Therefore we create 10^6 neutral unlinked loci simulations with the same sampling rate and the same number of replicates.

4 Discussion

5 Appendix

5.1 Allele frequencies under selective sweep

(12)

²The likelihood of the data to the model is the least-squares loss between model $D_t(\hat{s})$ and the observed D .

where $s \in \mathbb{R}$ is the selection coefficient and $o \in [0, 1]$ is the overdominance parameter which for $u = 0.5$ we have

$$x_{t+1} = x_t + \frac{sx_t(1-x_t)}{2+2sx_t}. \quad (13)$$

we also have

$$\frac{dx_t}{dt} = \frac{sx_t(1-x_t)}{2+2sx_t} \quad (14)$$

which is a differential equation that is difficult to solve. However if take the approximation $2+2sx_t \approx 2$, it becomes an ordinary differential equation that can be readily solved

$$\nu_t = \frac{1}{1 + \frac{1-x_0}{x_0} e^{-st/2}} = \sigma(st/2 + \eta(x_0)) \quad (15)$$

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

5.2 Logistic Model for Selection.

As maximum likelihood for estimating s using Markov chain is computationally expensive, here we use and logistic function for modeling allele frequencies undergoing selective sweep. In a pure selection process with no drift, i.e. infinite population size, dynamic of allele frequencies can be well approximated by the logistic function (see Appendix 5.1 for derivation)

$$\nu_t = \sigma(st + \eta(\nu_0)) \quad (16)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function, and $\eta(x) = \log(x)/\log(1 - x)$ is inverse of the logistic, aka logit, function. Figure 2 depicts the behavior of the logistic model for site allele frequencies and the default sampling time span(STS). Without genetic drift, soft sweep is easier to detect, because the logistic function happens to have steeper slope in th STS than those of hard sweeps, due to standing variation frequency. In addition, even under infinite population size regime, it is difficult differentiate between weak selections and genetic drift (a horizontal line), in short (e.g. 50 generations) experimental evolutions.

As shown in Figures ?? and ?? (first row), the approximate logistic model is consistent with simulated data and we use it to estimate the strength of selection for each site by solving a linear system of equations(see Section 3 for details).

5.3 Fay Wu's H

Lemma 1. *In any finite population size of n with m segregating sites, allele frequencies take discrete values, i.e., $x_j \in \{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}\}$, $\forall j \in 1, \dots, m$ and we can write*

$$\|\mathbf{x}\|^2 = \sum_{j=1}^m x_j^2 = \sum_{i=1}^{n-1} \left(\frac{i}{n}\right)^2 \xi_i = \frac{(n-1)}{2n} H \quad (17)$$

where ξ_i is the number of sites with frequency i/n and H is the Fay & Wu's estimate of θ .

Recently, Ronen et al. [48] devised the 1-HAF statistic for identifying selection on static data, which has the expected value related to the $\|\mathbf{x}\|^2$:

$$\mathbb{E}[1\text{-HAF}(t)] = n\|\mathbf{x}_t\|^2 \approx ng(\nu_t) \quad (18)$$

where

$$g(\nu_t) = \theta\nu_t \left(\frac{\nu_t + 1}{2} - \frac{1}{(1 - \nu_t)n + 1} \right) + \theta(1 - \nu_t) \left(\frac{n + 1}{2n} - \frac{1}{(1 - \nu_t)n + 1} \right) \quad (19)$$

which easily follows that

$$\theta_H(t) = \frac{n-1}{2}g(\nu_t) \quad (20)$$

5.4 Tajima's D

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 :

$$\Pi_t = (1 - \nu_t^2)\Pi_0 \quad (21)$$

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

$$\frac{W_t}{W_0} = \frac{\frac{m_t}{S}}{\frac{m_0}{S}} \Rightarrow W_t = \frac{m_t}{m_0}W_0 \quad (22)$$

where m_t to be interpreted as the expected number of segregating sites at time t , under neutral evolution. At time t , the number of individuals that undergone neutral evolution is $(1 - \nu_t)n + 1$, which leads to

$$\frac{m_t}{m_0} = \frac{\log((1 - \nu_t)n + 1)\theta}{\log(n)\theta} \approx \frac{\log((1 - \nu_t)n)}{\log(n)} = \frac{\log(1 - \nu_t) + \log(n)}{\log(n)} = 1 + \frac{\log(1 - \nu_t)}{\log(n)} \quad (23)$$

putting all together

$$D_t = (1 - \nu_t^2)\Pi_0 - (1 + \frac{\log(1 - \nu_t)}{\log(n)})W_0 = D_0 - \log(1 - \nu_t)\frac{W_0}{\log(n)} - \nu_t^2\Pi_0 \quad (24)$$

5.5 Linkage Disequilibrium

Nonrandom associations between polymorphisms are established in the substitution process according to the phylogeny, broken by recombination events and reinforced by selection. Although in EE the experiments with pooled sequencing, LD can not be measured throughout evolution, it is still worthwhile to examine the behavior of LD as a result of the interaction between recombination and natural selection, to take into account of some of EE implicit constraints.

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

$$\rho_t = \alpha_t \beta_t \rho_0 = e^{-rlt} \left(\frac{H_t}{H_0} \right) \rho_0 \quad (25)$$

where $H_T = 2\nu_0(1 - \nu_0)$ is the heterozygosity at the selected site, r is the recombination rate/bp/gen. The decay factor $\alpha_t = e^{-rlt}$ is the product of recombination and growth factor β_t (eq. 30-31 in [56]) is the outcome of selection. For $s = 0.01$, $l = 100Kbp$, the log of decay, growth and product

of both is depicted in Figure 8. It is evident that, for these parameters LD does not start to decay until generation 1000, which would be problematic when ρ_0 . For example, in the case of hard sweep, the selection is imposed on the site with minimum AF, which is at perfect linkage ($|D'| = 1$) with all the other loci.³ We this phenomenon is shown in the Figures 9, 10 where at generation zero the site at position 500K is at perfect linkage with all the other sites, and linkage of the middle site with all the genome is depicted for both genetic drift and natural selection, in different generations. Also, a window of 50Kbp around the selected site is shaded in Figure 9 to demonstrate the value of LD in the window under drift and hard sweep. This implies that the precision of locating the selection on the genome is tightly dependent on a set of parameters including, recombination rate, selection strength, initial carrier frequency, and the initial linkage.

6 Figures

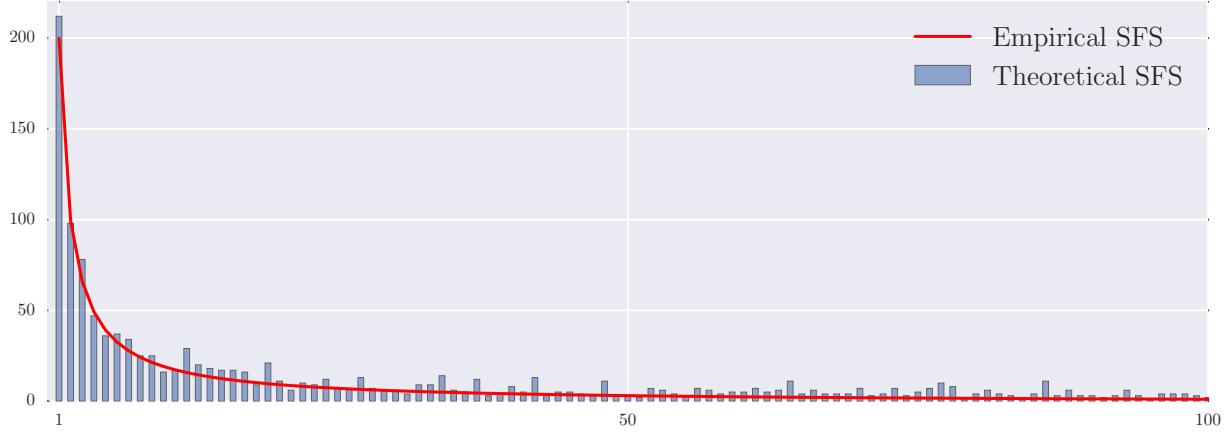


Figure 1: Theoretical and Empirical SFS for a neutral population of 200 individuals.

³This is because, between the selected site and all the other sites frequency of one gamete is zero.

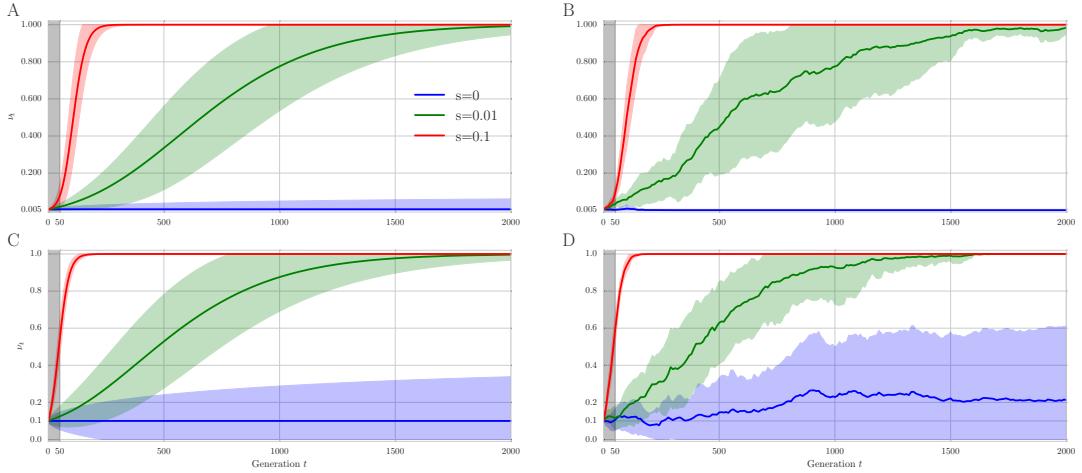


Figure 2: Logistic model for different selection strengths for soft (left) and hard (right) sweep as a function of time in generations. The first 50 generations, which observations are sampled is shaded.

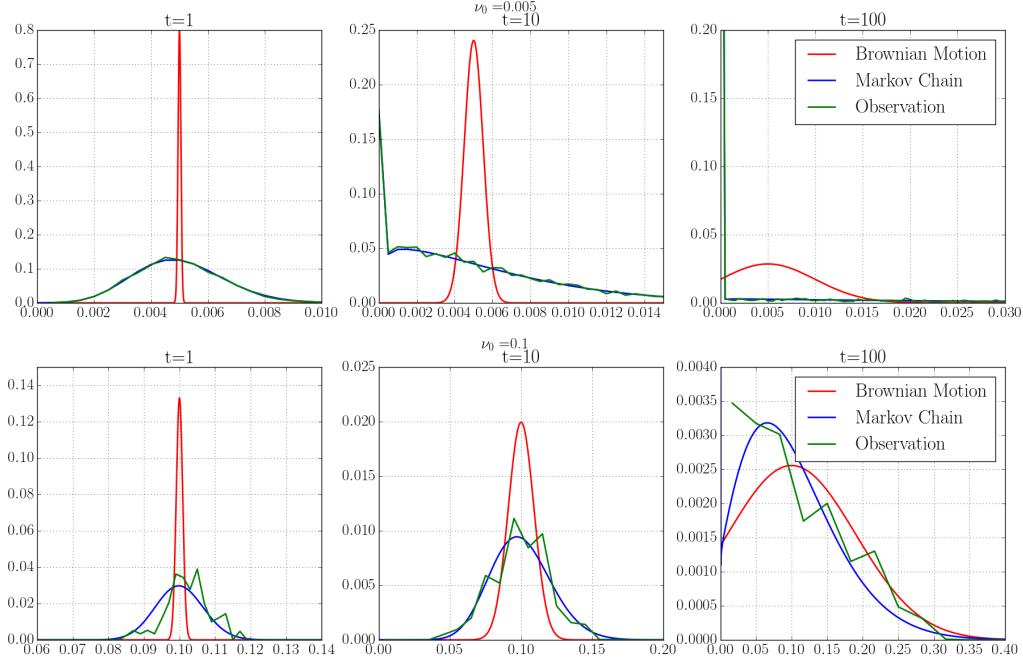


Figure 3: Predicted distributions of allele frequencies of sites with initial frequency $\nu_0 = 0.005$ (top) and $\nu_0 = 0.1$ (bottom), after $t = \{1, 10, 100\}$ generations, under neutral evolution, by Brownian motion and Markov chain. Observed distribution of 143900 sites with $\nu_0 = 0.005$ and 47500 with variants $\nu_0 = 0.1$ computed from neutrally evolving simulations is depicted in green lines.

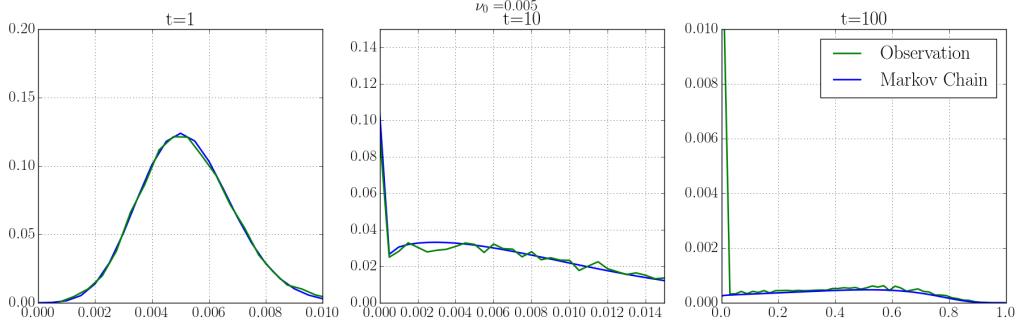


Figure 4: Predicted distributions of allele frequencies of sites with initial frequency $\nu_0 = 0.005$ and selection strength $s = 0.1$, after $t = \{1, 10, 100\}$ generations, by Markov chain. Observed distribution of the adaptive allele (with $\nu_0 = 0.005$) in 10000 simulation is shown in green.

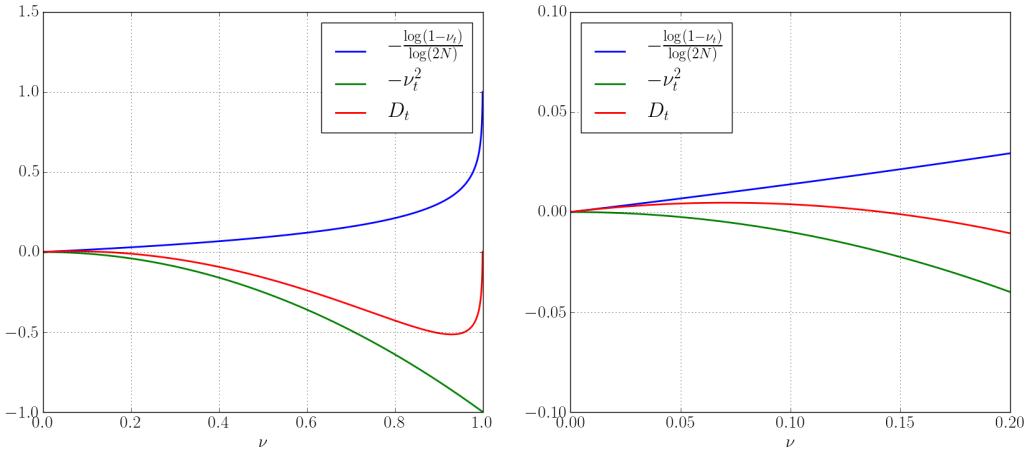


Figure 5: Interactions of two terms in D . W.l.o.g when $D_0 = 0$ and $\Pi_0 = W_0 = 1$, D_t is sum of the logarithmic $-\frac{\log(1-\nu_t)}{\log(2N)}$ and the squared term ν_t^2 .

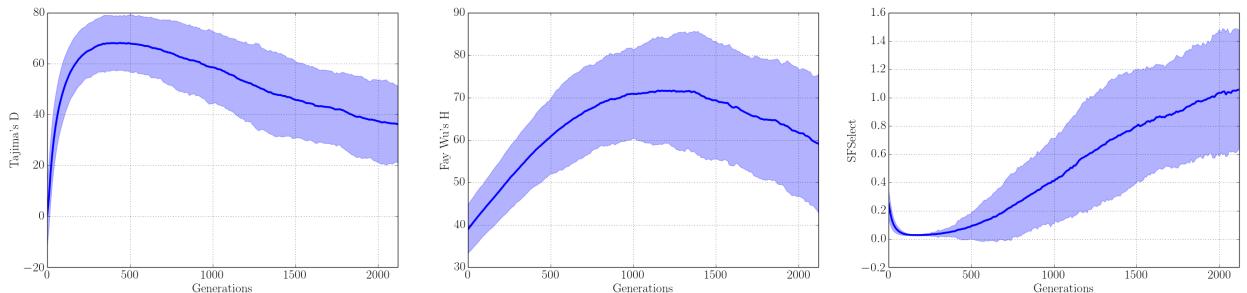


Figure 6: Effect of bottle neck in a typical experimental evolution experiment where a restricted number of founder lines (here $F = 200$) is selected out of a larger population size ($N_e = 10^{-6}$). Tajima's D (left), Fay Wu's H (middle) and SFSelect is computed for 1000 neutral simulations and mean and 95% confidence interval is plotted.

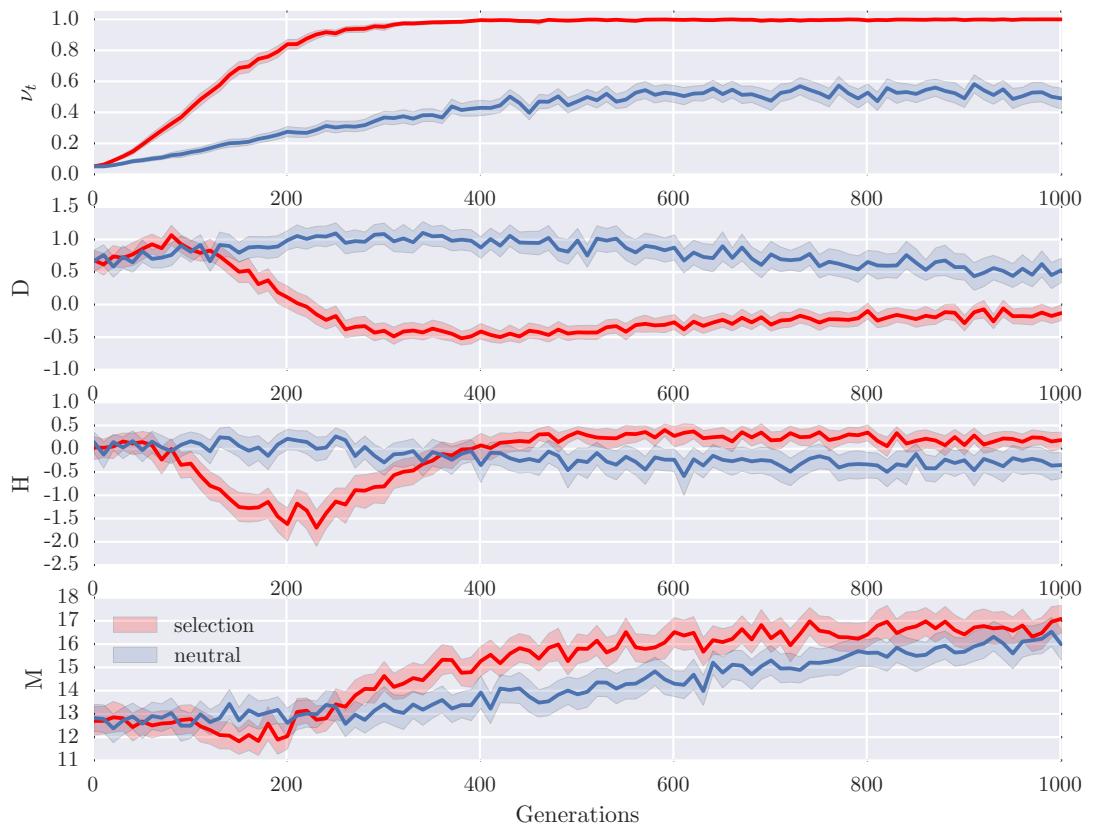


Figure 7: Mean and 95% CI of 1000 simulations for neutral (blue trajectories) selection with $s = 0.1$ (red trajectories).

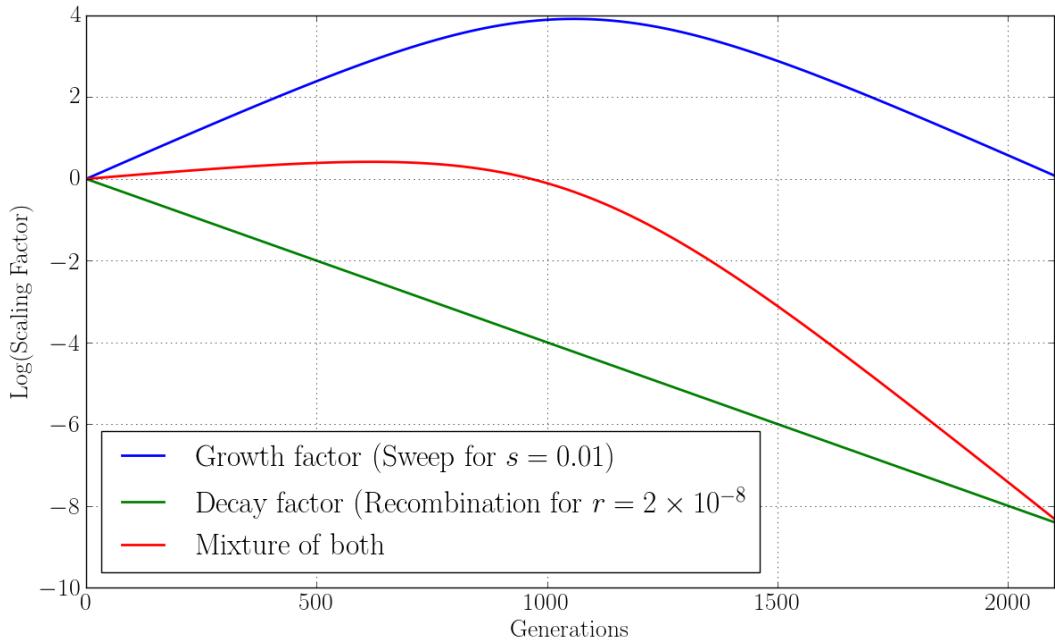


Figure 8: Interaction between productive factors of LD under natural selection for weak selection ($s=0.01$) and a distance of 100Kb between sites. In this setting, after about 1000 generations LD start to decay (red curve).

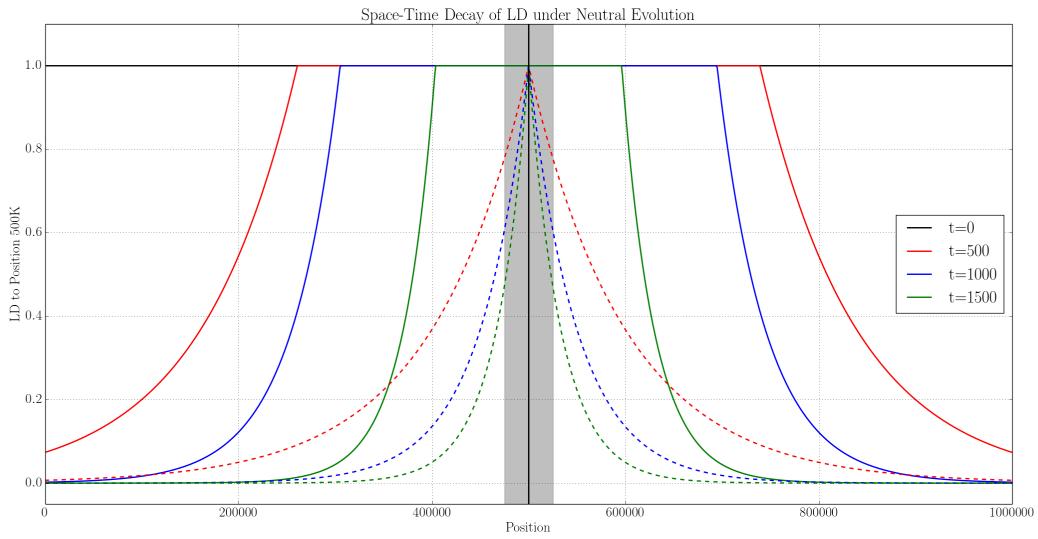


Figure 9: Decay of LD ($|D'|$ measure) of the minimum AF site at position 500K with the rest of genome when $s = 0.01$ and $r = 2 \times 10^{-8}$. A window of 50Kb is shaded at the center of genome to illustrate high values of linkage in both selection and drift.

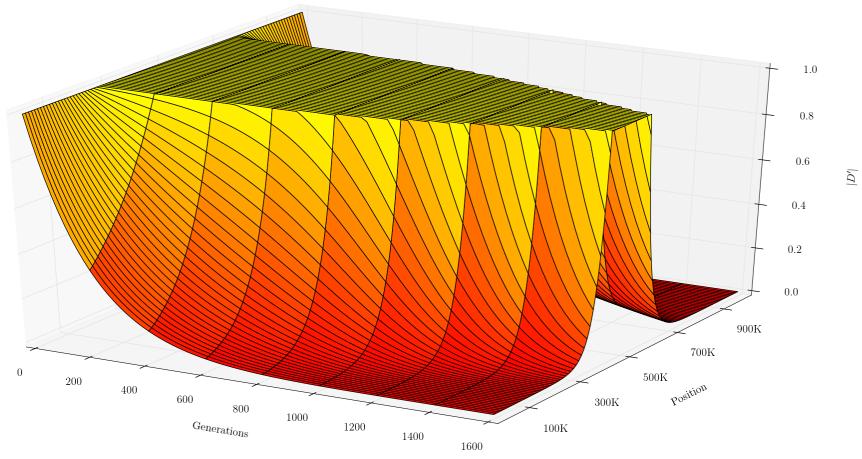
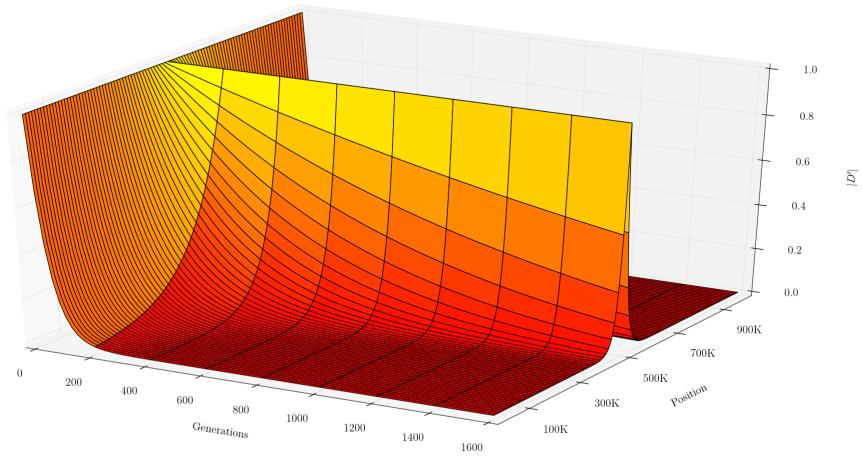


Figure 10: ld

Figure 11: 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).

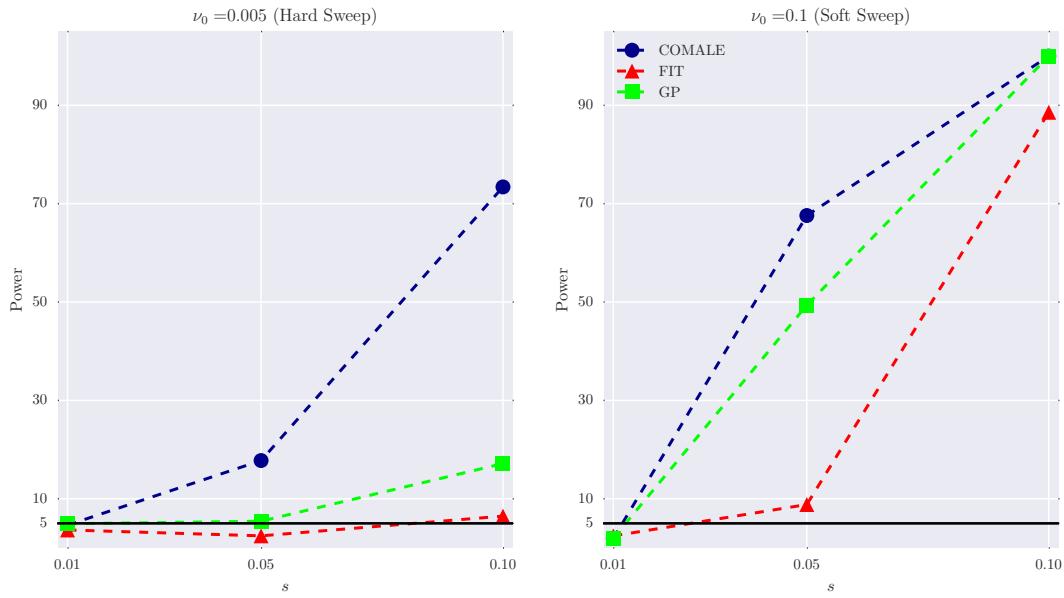


Figure 12: Predictive performance of different method is evaluated on 200 simulations for different values of selection strength s and initial carrier frequency ν_0 .

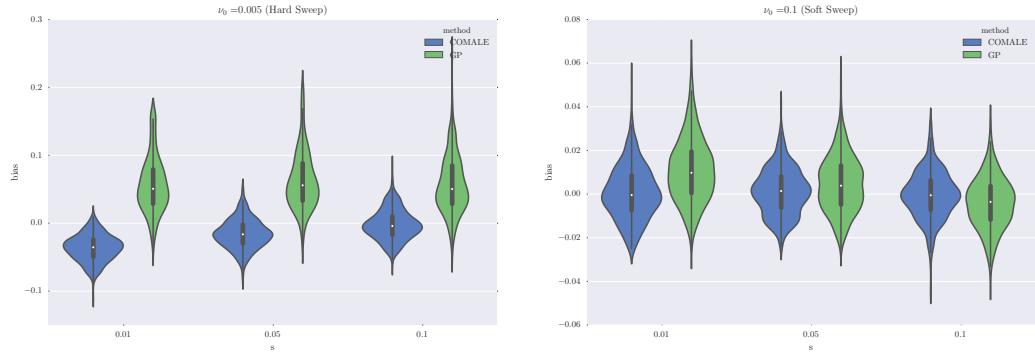


Figure 13: .

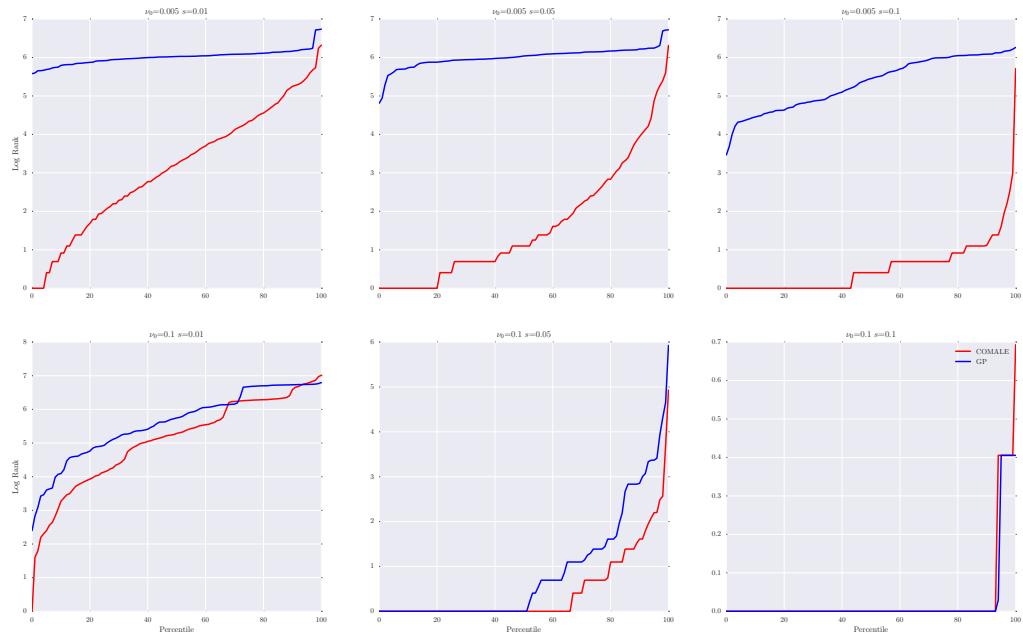


Figure 14: CDF of the rank of the adaptive allele in 100 simulations.

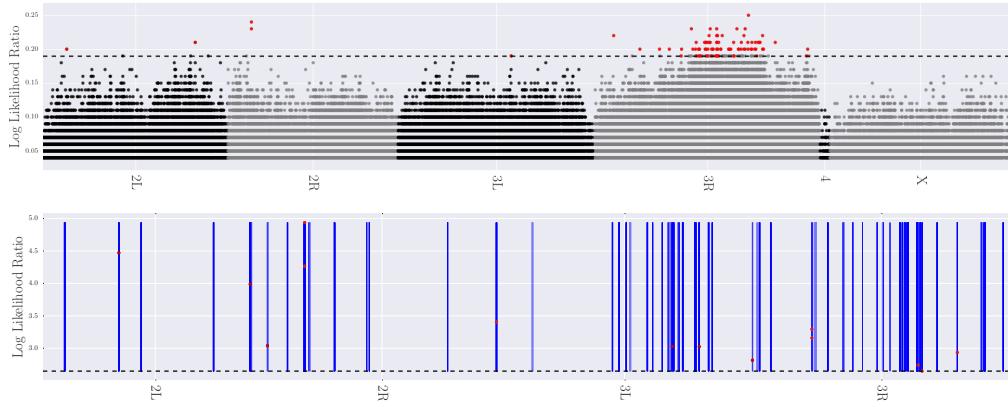


Figure 15: Manhattan plots shows the distribution of candidate (top 2000) SNPs and the associated (1961) genes are denoted with blue line in the bottom plot.

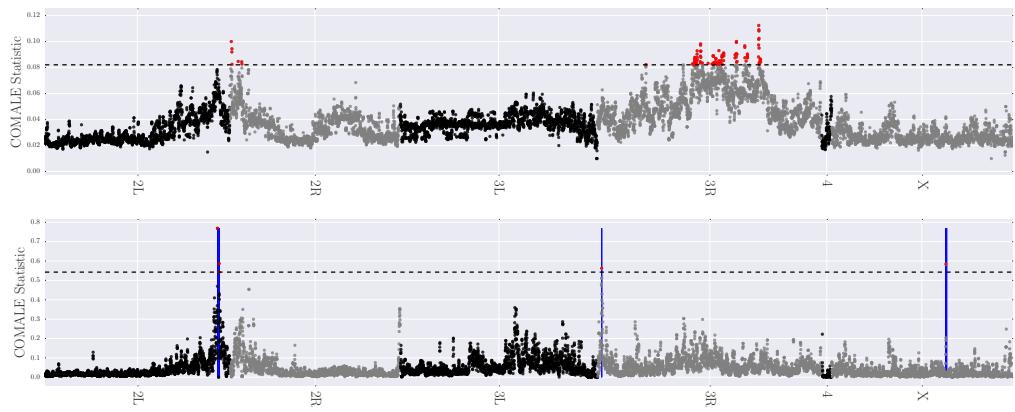


Figure 16: Manhattan plots shows the distribution of candidate regions using COMALE statistic which give rise to (421) genes are denoted with blue line in the bottom plot.

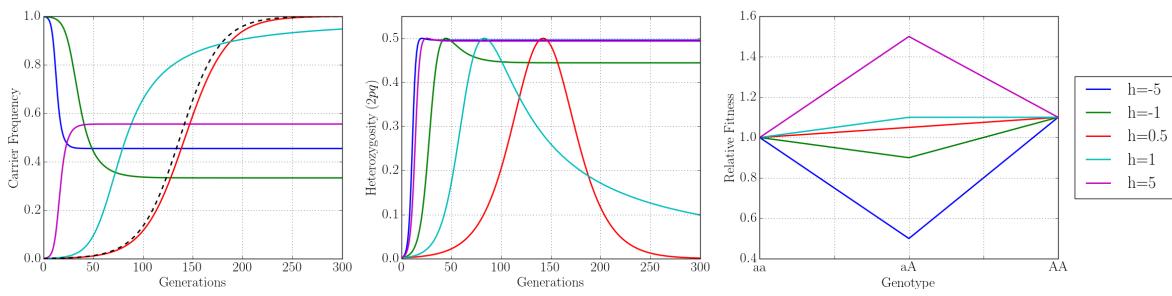


Figure 17: Dominance for $s = 0.1$.

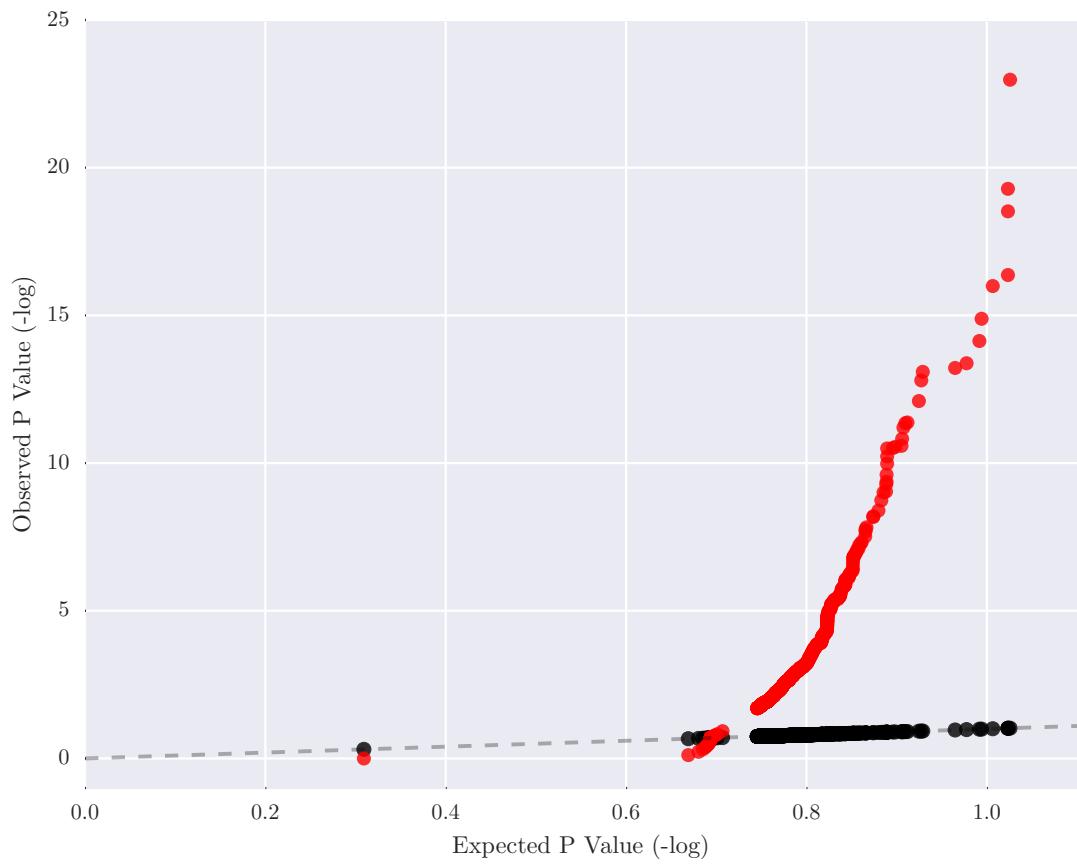


Figure 18: QQ plot for a simulation.

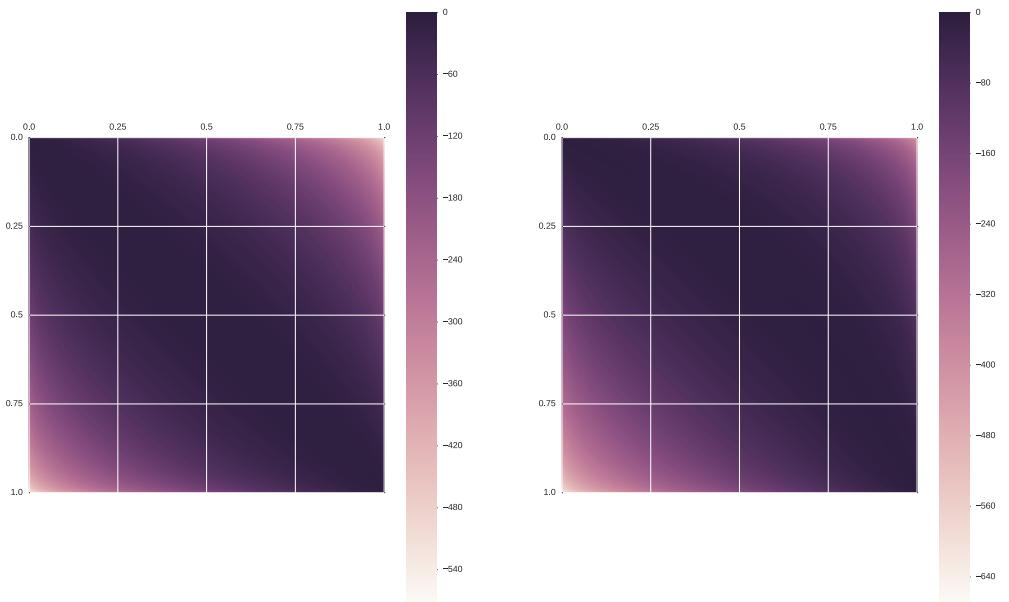


Figure 19: 100^{th} power of the transition matrix for neutral (left) and selection with $s = 0.1$ (right).

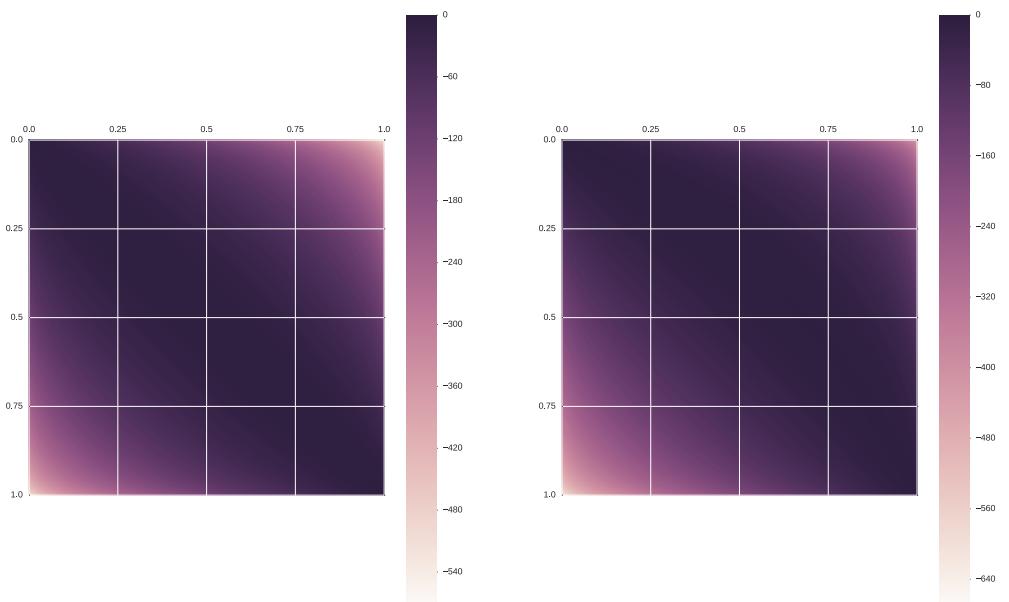


Figure 20: 100^{th} power of the transition matrix for neutral (left) and selection with $s = 0.1$ (right).

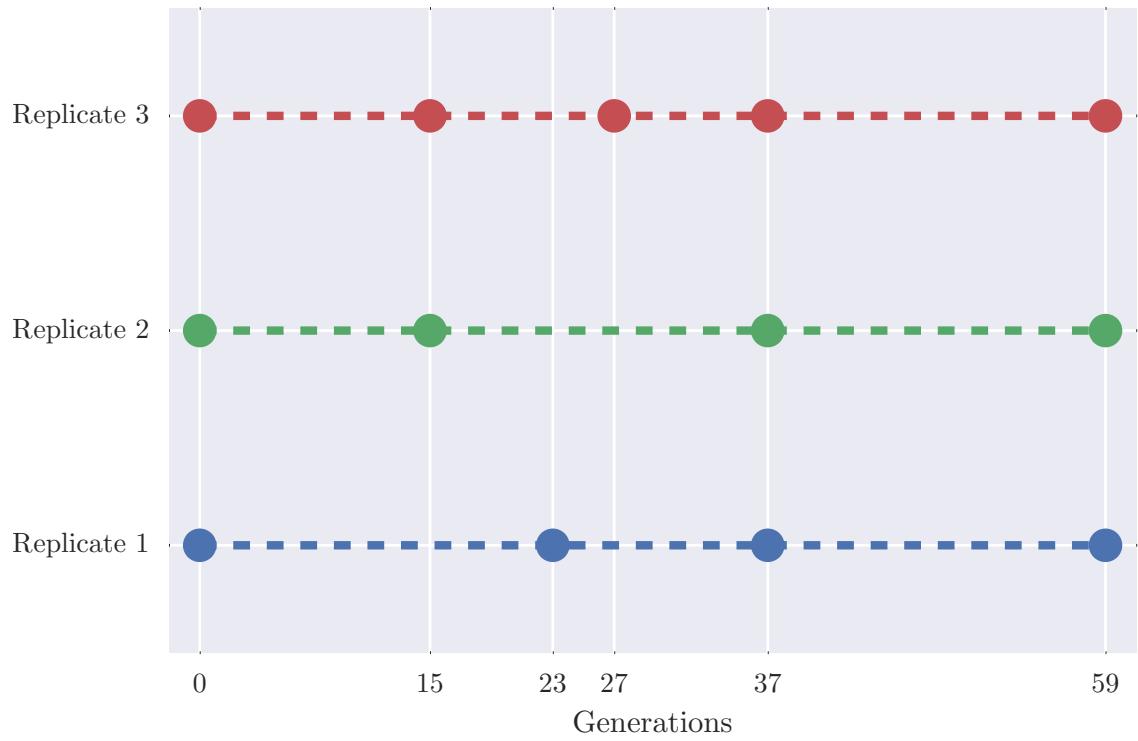


Figure 21: sampling times.

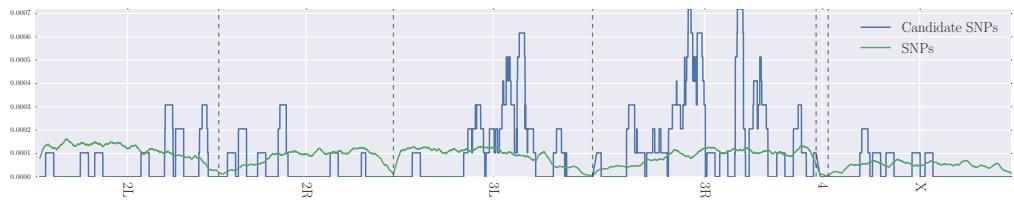


Figure 22: Density.

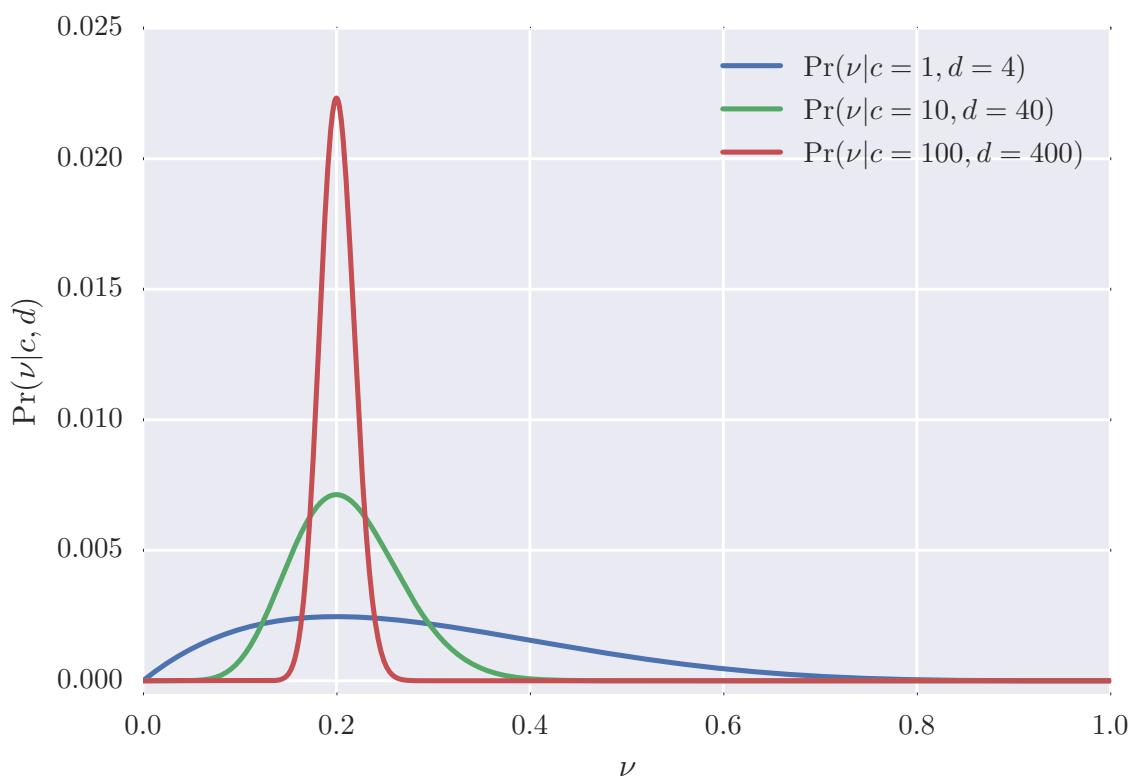


Figure 23: observation state transition.

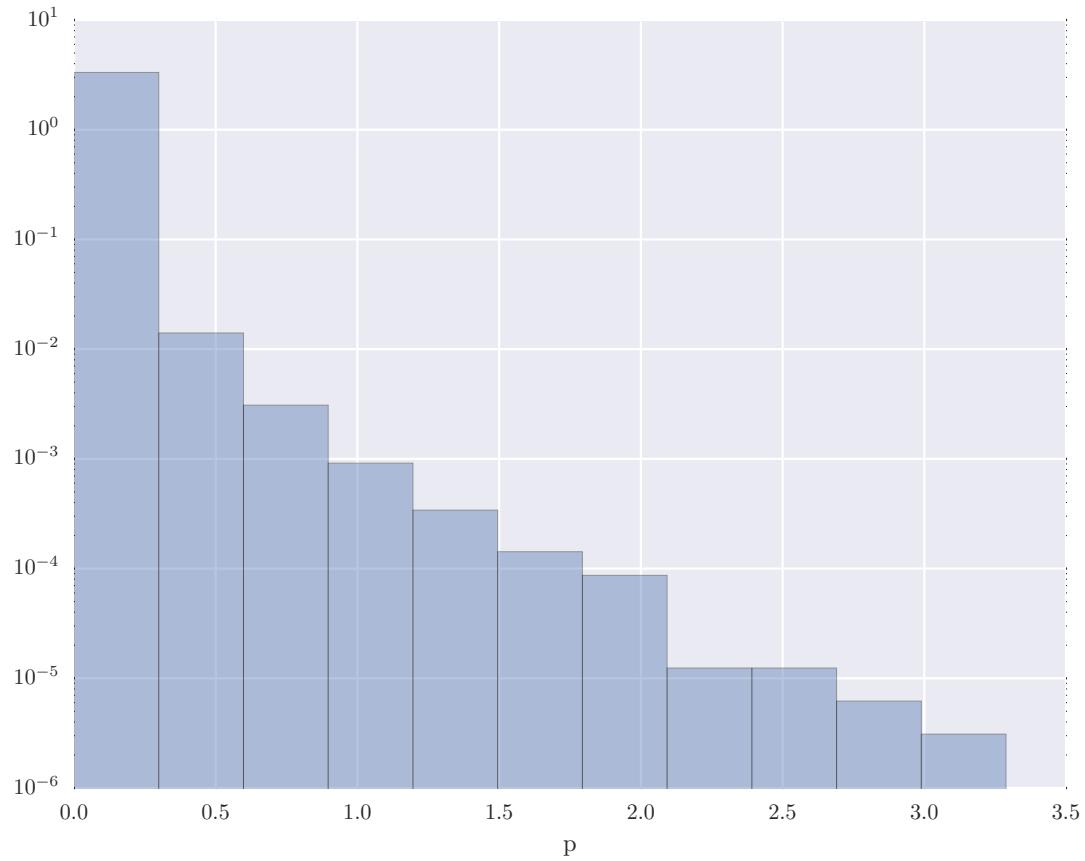


Figure 24: Empirical null distribution of test statistic on 1M simulations.

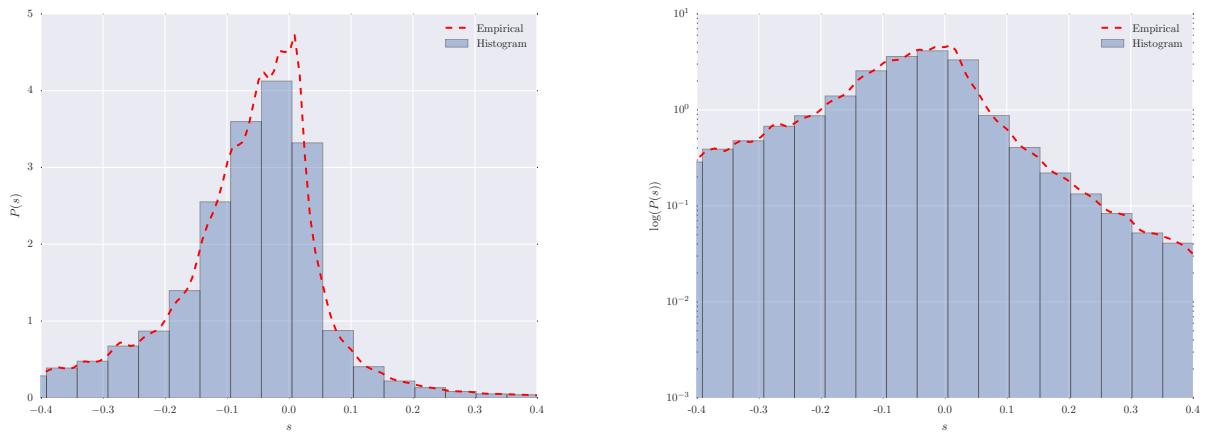


Figure 25: Empirical null distribution of s statistic on 1M simulations.

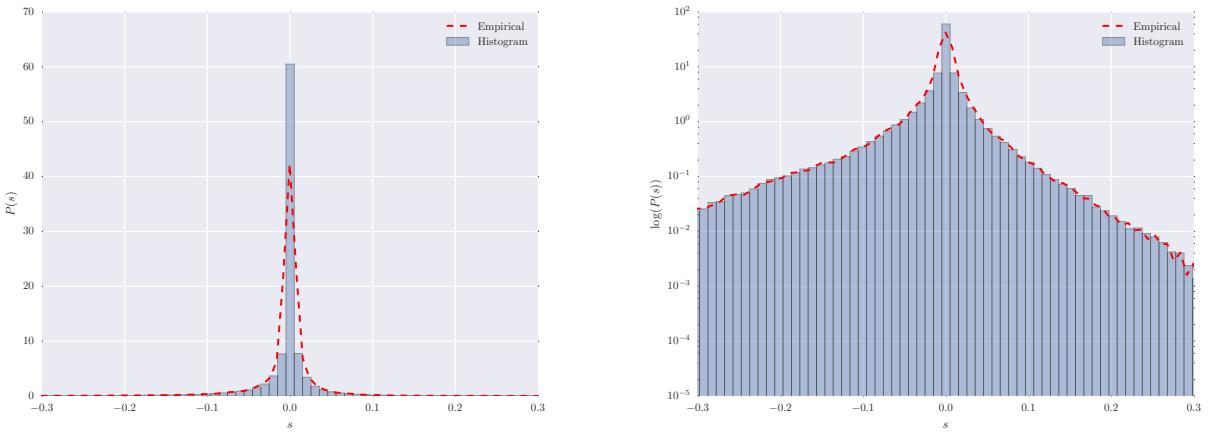


Figure 26: Empirical null distribution of regularized s statistic on 1M simulations.

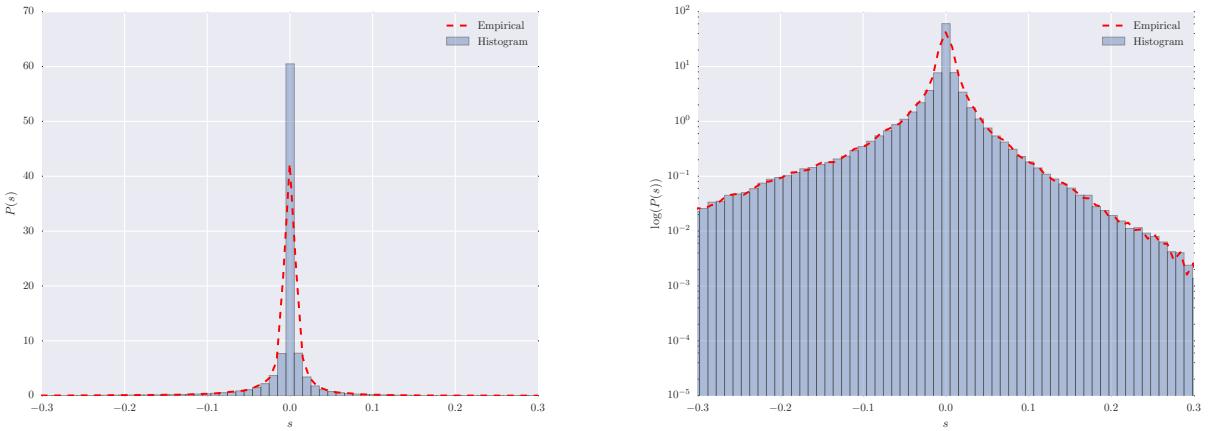
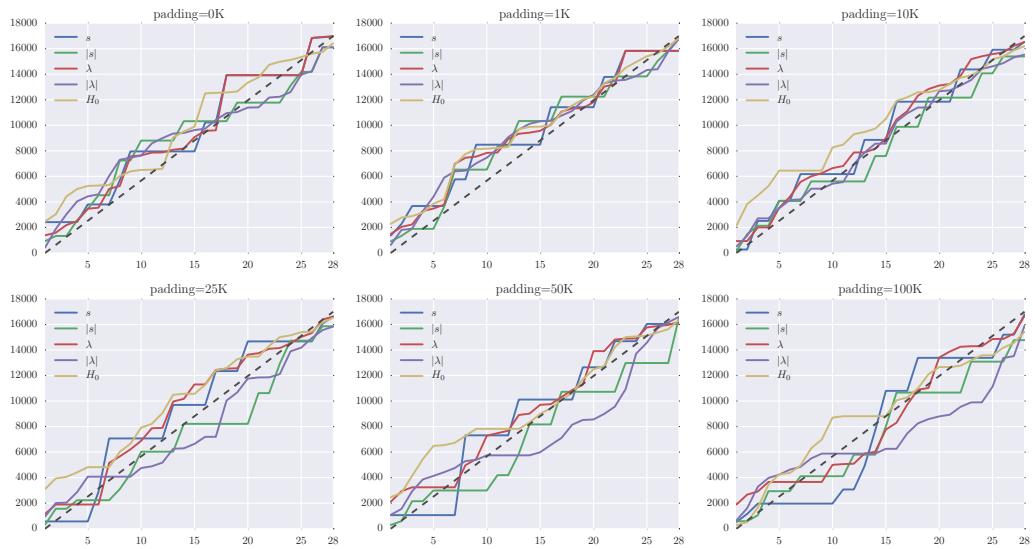
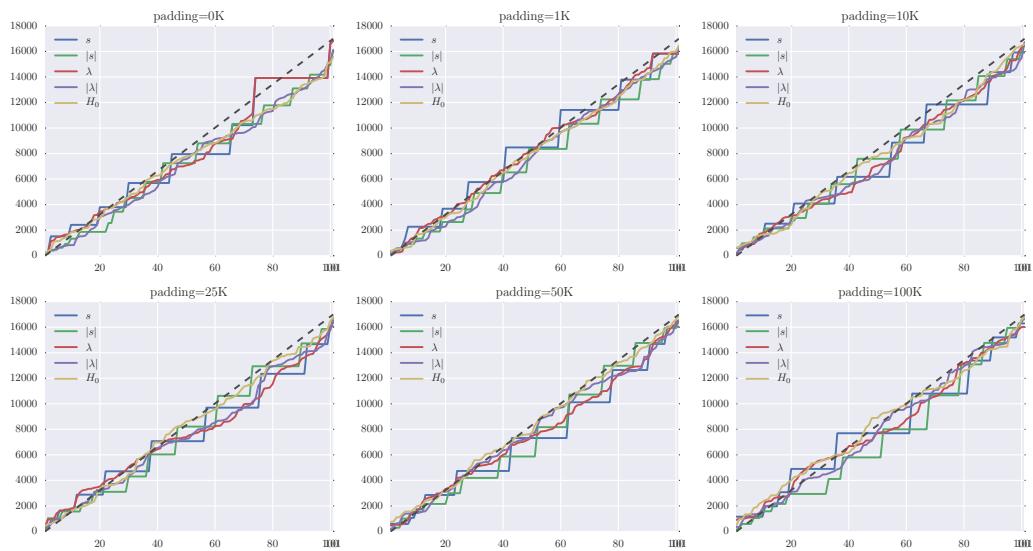


Figure 27: Empirical null distribution of regularized s statistic on 1M simulations.

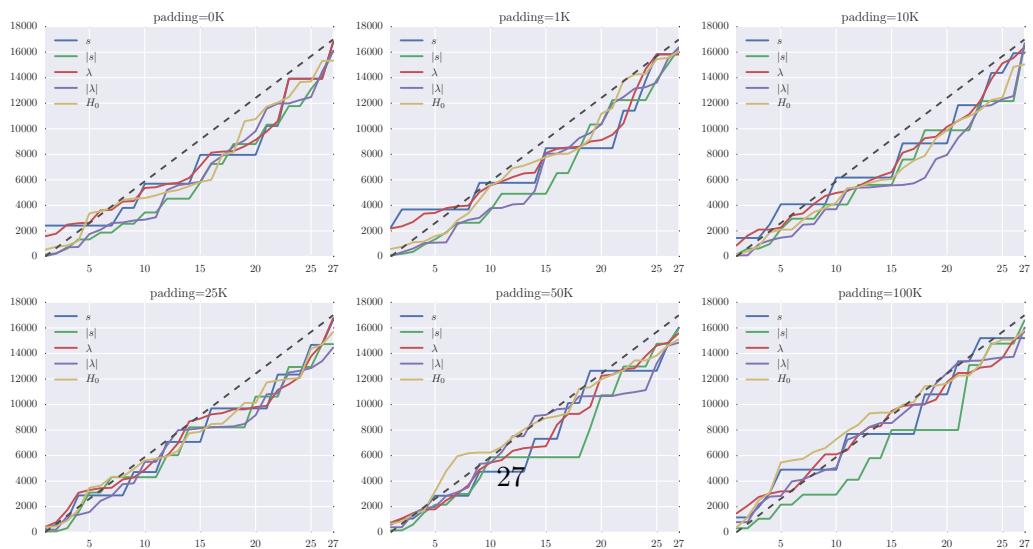
Rank of earlyUp Genes



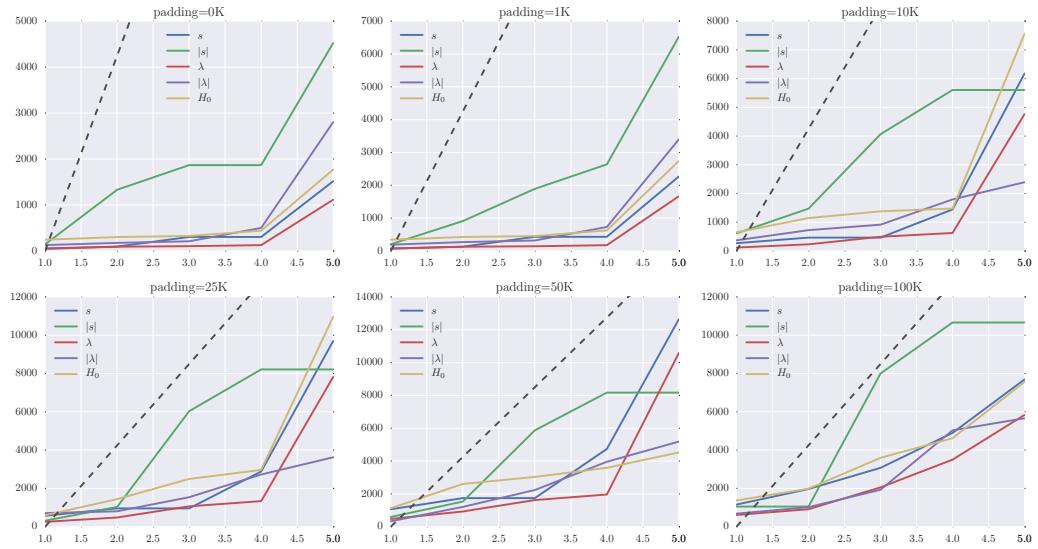
Rank of earlyDown Genes



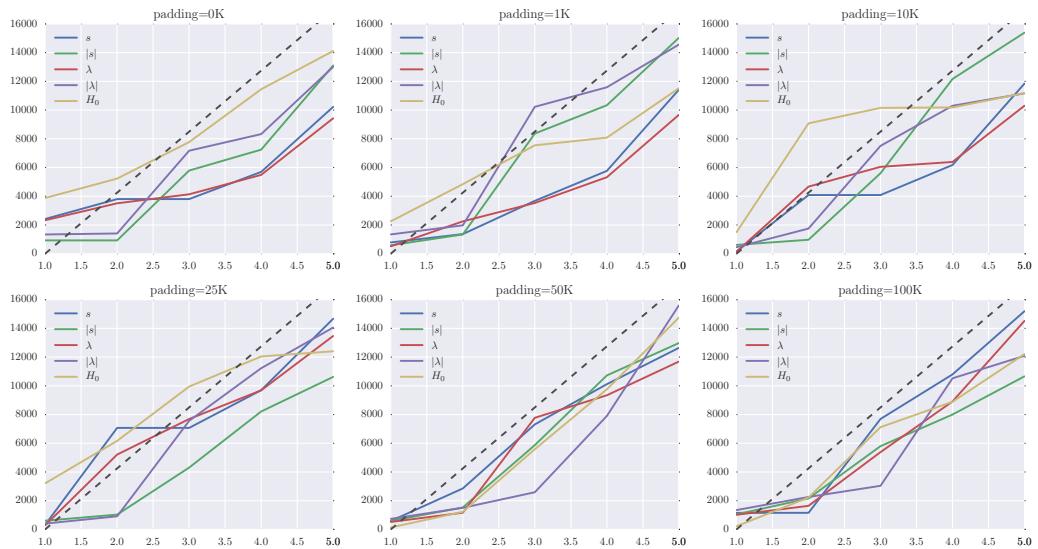
Rank of lateUp Genes



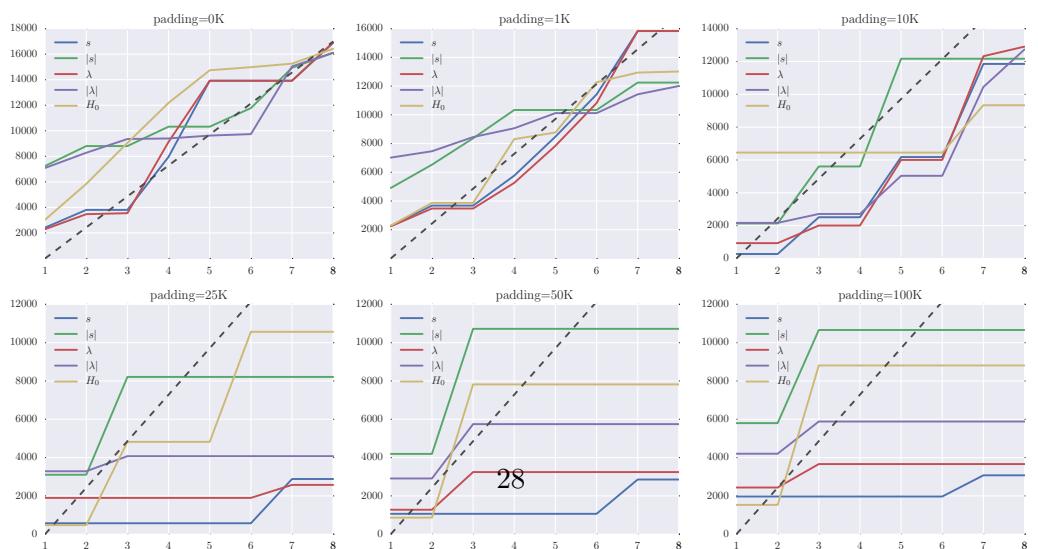
Rank of Song Genes



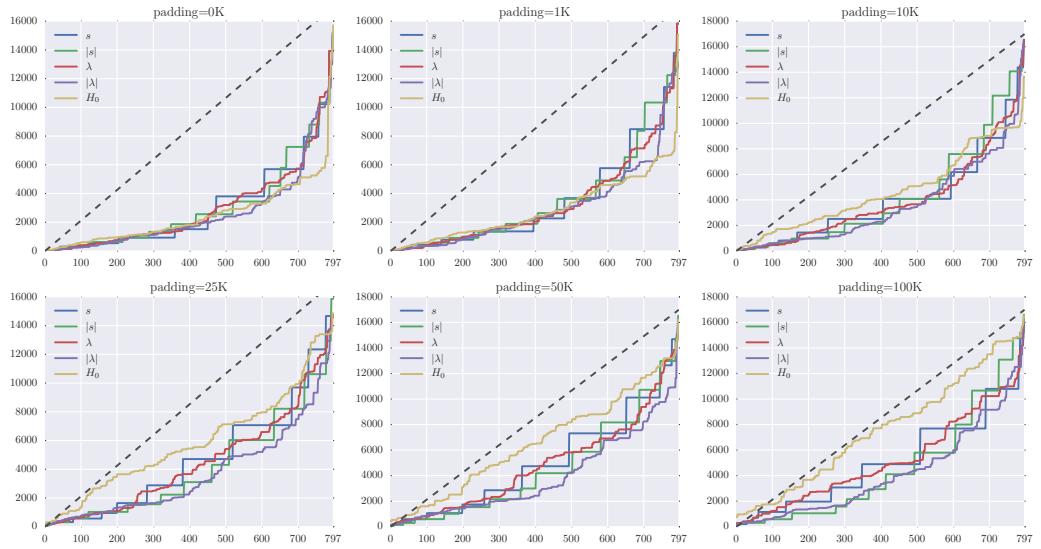
Rank of HSF Genes



Rank of HSP Genes



Rank of BM Genes



Rank of BF Genes

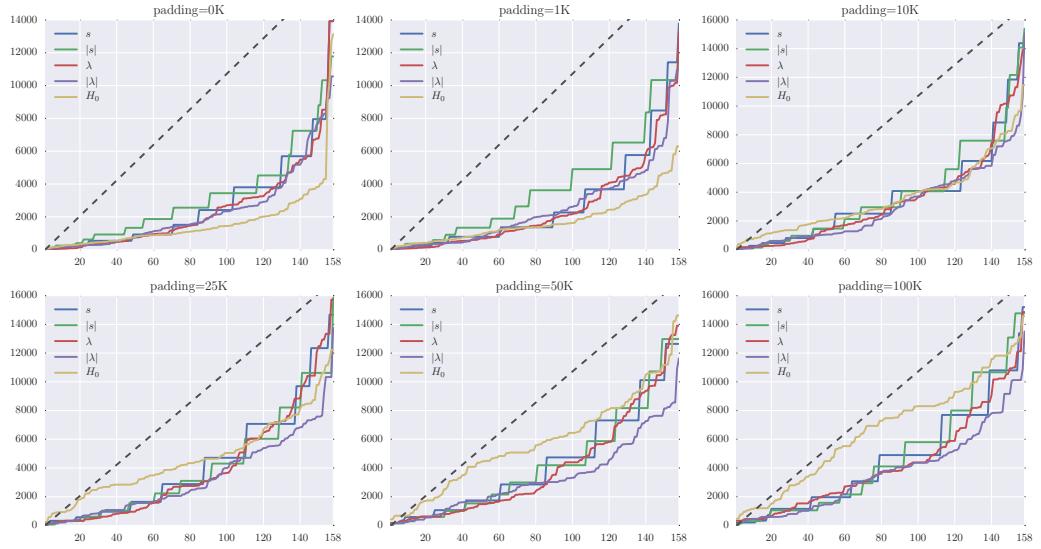


Figure 30: Song, HSF, HSP.

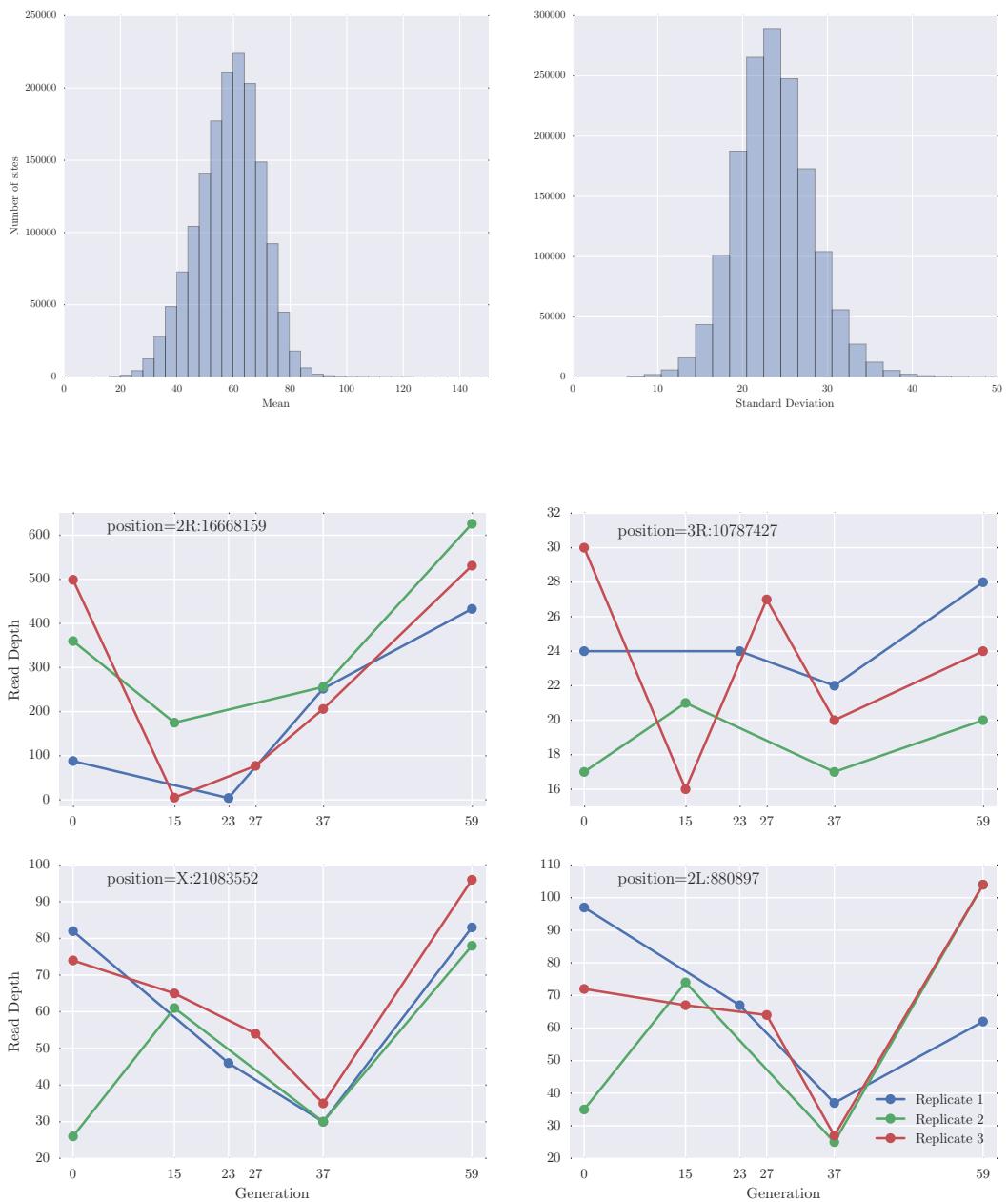


Figure 31: Song, HSF, HSP.

References

- [1] ACHAZ, G. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183, 1 (2009), 249–258.
- [2] AKEY, J. M. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome research* 19, 5 (2009), 711–722.
- [3] ARIEY, F., WITKOWSKI, B., AMARATUNGA, C., BEGHAIN, J., LANGLOIS, A.-C., KHIM, N., KIM, S., DURU, V., BOUCHIER, C., MA, L., AND OTHERS. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature* 505, 7481 (2014), 50–55.
- [4] BARRETT, R. D. H., ROGERS, S. M., AND SCHLUTER, D. Natural selection on a major armor gene in threespine stickleback. *Science* 322, 5899 (2008), 255–257.
- [5] BARRICK, J. E., AND LENSKI, R. E. Genome dynamics during experimental evolution. *Nat Rev Genet* 14, 12 (dec 2013), 827–839.
- [6] BARRICK, J. E., YU, D. S., YOON, S. H., JEONG, H., OH, T. K., SCHNEIDER, D., LENSKI, R. E., AND KIM, J. F. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* 461, 7268 (2009), 1243–1247.
- [7] BERGLAND, A. O., BEHRMAN, E. L., O'BRIEN, K. R., SCHMIDT, P. S., AND PETROV, D. A. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS Genet* 10, 11 (2014), e1004775.
- [8] BOLBACK, J. P., AND HUELSENBECK, J. P. Clonal interference is alleviated by high mutation rates in large populations. *Molecular biology and evolution* 24, 6 (2007), 1397–1406.
- [9] BOLBACK, J. P., YORK, T. L., AND NIELSEN, R. Estimation of 2Nes from temporal allele frequency data. *Genetics* 179, 1 (2008), 497–502.
- [10] BOYKO, A. R., WILLIAMSON, S. H., INDAP, A. R., DEGENHARDT, J. D., HERNANDEZ, R. D., LOHMUELLER, K. E., ADAMS, M. D., SCHMIDT, S., SNINSKY, J. J., SUNYAEV, S. R., AND OTHERS. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4, 5 (2008), e1000083.
- [11] BURKE, M. K., DUNHAM, J. P., SHAHRESTANI, P., THORNTON, K. R., ROSE, M. R., AND LONG, A. D. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* 467, 7315 (2010), 587–590.
- [12] DABORN, P., BOUNDY, S., YEN, J., PITTEENDRIGH, B., 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 (2001), 556–563.
- [13] DANIELS, R., CHANG, H.-H., SÉNE, P. D., PARK, D. C., NEAFSEY, D. E., SCHAFFNER, S. F., HAMILTON, E. J., LUKENS, A. K., VAN TYNE, D., MBOUP, S., AND OTHERS. Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* 8, 4 (2013), e60780.
- [14] DENEF, V. J., AND BANFIELD, J. F. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* 336, 6080 (2012), 462–466.

- [15] DESAI, M. M., AND FISHER, D. S. Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics* 176, 3 (2007), 1759–1798.
- [16] DESAI, M. M., AND PLOTKIN, J. B. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics* 180, 4 (2008), 2175–2191.
- [17] EVANS, S. N., SHVETS, Y., AND SLATKIN, M. Non-equilibrium theory of the allele frequency spectrum. *Theoretical population biology* 71, 1 (2007), 109–119.
- [18] EWENS, W. J. *Mathematical Population Genetics 1: Theoretical Introduction*, vol. 27. Springer Science & Business Media, 2012.
- [19] FAY, J. C., AND WU, C.-I. Hitchhiking under positive Darwinian selection. *Genetics* 155, 3 (2000), 1405–1413.
- [20] FEDER, A. F., KRYAZHIMSKIY, S., AND PLOTKIN, J. B. Identifying signatures of selection in genetic time series. *Genetics* 196, 2 (2014), 509–522.
- [21] FEDER, A. F., RHEE, S.-Y., HOLMES, S. P., SHAFER, R. W., PETROV, D. A., AND PENNINGS, P. S. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *eLife* 5 (jan 2016).
- [22] GOTTESMAN, M. M. Mechanisms of cancer drug resistance. *Annual review of medicine* 53, 1 (2002), 615–627.
- [23] HEGRENESS, M., SHORESH, N., HARTL, D., AND KISHONY, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* 311, 5767 (2006), 1615–1617.
- [24] HOLSINGER, K. E., AND WEIR, B. S. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics* 10, 9 (2009), 639–650.
- [25] ILLINGWORTH, C. J. R., AND MUSTONEN, V. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics* 189, 3 (2011), 989–1000.
- [26] ILLINGWORTH, C. J. R., PARTS, L., SCHIFFELS, S., LITI, G., AND MUSTONEN, V. Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular biology and evolution* 29, 4 (2012), 1187–1197.
- [27] IZUTSU, M., TOYODA, A., FUJIYAMA, A., AGATA, K., AND FUSE, N. Dynamics of Dark-Fly Genome Under Environmental Selections. *G3: Genes—Genomes—Genetics* (2015), g3—115.
- [28] JHA, A. R., MILES, C. M., LIPPERT, N. R., BROWN, C. D., WHITE, K. P., AND KREITMAN, M. Whole-genome resequencing of experimental populations reveals polygenic basis of egg-size variation in *Drosophila melanogaster*. *Molecular biology and evolution* 32, 10 (2015), 2616–2632.
- [29] KAPLAN, N. L., HUDSON, R. R., AND LANGLEY, C. H. The “ hitchhiking effect” revisited. *Genetics* 123, 4 (1989), 887–899.
- [30] KAWEKKI, T. J., LENSKI, R. E., EBERT, D., HOLLIS, B., OLIVIERI, I., AND WHITLOCK, M. C. Experimental evolution. *Trends in ecology & evolution* 27, 10 (2012), 547–560.

- [31] LANG, G. I., BOTSTEIN, D., AND DESAI, M. M. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* 188, 3 (2011), 647–661.
- [32] LANG, G. I., RICE, D. P., HICKMAN, M. J., SODERGREN, E., WEINSTOCK, G. M., BOTSTEIN, D., AND DESAI, M. M. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500, 7464 (2013), 571–574.
- [33] MALASPINAS, A.-S., MALASPINAS, O., EVANS, S. N., AND SLATKIN, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* 192, 2 (2012), 599–607.
- [34] MALDARELLI, F., KEARNEY, M., PALMER, S., STEPHENS, R., MICAN, J., POLIS, M. A., DAVEY, R. T., KOVACS, J., SHAO, W., ROCK-KRESS, D., AND OTHERS. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of virology* 87, 18 (2013), 10313–10323.
- [35] MARTINS, N. E., FARIA, V. G., NOLTE, V., SCHLÖTTERER, C., TEIXEIRA, L., SUCENA, É., AND MAGALHÃES, S. Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences* 111, 16 (2014), 5938–5943.
- [36] MATHIESON, I., AND MCVEAN, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* 193, 3 (2013), 973–984.
- [37] MESSER, P. W., AND PETROV, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution* 28, 11 (2013), 659–669.
- [38] NAIR, S., NASH, D., SUDIMACK, D., JAIDEE, A., BARENDS, M., UHLEMANN, A.-C., KRISHNA, S., NOSTEN, F., AND ANDERSON, T. J. C. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Molecular Biology and Evolution* 24, 2 (2007), 562–573.
- [39] NIELSEN, R., AND SIGNOROVITCH, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical population biology* 63, 3 (2003), 245–255.
- [40] NIELSEN, R., WILLIAMSON, S., KIM, Y., HUBISZ, M. J., CLARK, A. G., AND BUSTAMANTE, C. Genomic scans for selective sweeps using SNP data. *Genome research* 15, 11 (2005), 1566–1575.
- [41] OROZCO-TERWENGEL, P., KAPUN, M., NOLTE, V., KOFLER, R., FLATT, T., AND SCHLÖTTERER, C. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular ecology* 21, 20 (2012), 4931–4941.
- [42] OZ, T., GUVENEK, A., YILDIZ, S., KARABOGA, E., TAMER, Y. T., MUMCUYAN, N., OZAN, V. B., SENTURK, G. H., COKOL, M., YEH, P., AND OTHERS. Strength of selection pressure is an important parameter contributing to the complexity of antibiotic resistance evolution. *Molecular biology and evolution* (2014), msu191.
- [43] POLLAK, E. A new method for estimating the effective population size from allele frequency changes. *Genetics* 104, 3 (1983), 531–548.

- [44] PTAK, S. E., AND PRZEWORSKI, M. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics* 18, 11 (2002), 559–563.
- [45] RAMOS-ONSINS, S. E., AND ROZAS, J. Statistical properties of new neutrality tests against population growth. *Molecular biology and evolution* 19, 12 (2002), 2092–2100.
- [46] REID, B. J., KOSTADINOV, R., AND MALEY, C. C. New strategies in Barrett's esophagus: integrating clonal evolutionary theory with clinical management. *Clinical Cancer Research* 17, 11 (2011), 3512–3519.
- [47] REMOLINA, S. C., CHANG, P. L., LEIPS, J., NUZHDIN, S. V., AND HUGHES, K. A. Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. *Evolution* 66, 11 (2012), 3390–3403.
- [48] RONEN, R., TESLER, G., AKBARI, A., ZAKOV, S., ROSENBERG, N. A., AND BAFNA, V. Predicting Carriers of Ongoing Selective Sweeps Without Knowledge of the Favored Allele. *PLoS Genet* 11, 9 (2015), e1005527.
- [49] RONEN, R., UDPA, N., HALPERIN, E., AND BAFNA, V. Learning natural selection from the site frequency spectrum. *Genetics* 195, 1 (2013), 181–193.
- [50] SABETI, P. C., SCHAFFNER, S. F., FRY, B., LOHMUELLER, J., VARILLY, P., SHAMOVSKY, O., PALMA, A., MIKKELSEN, T. S., ALTSCHULER, D., AND LANDER, E. S. Positive natural selection in the human lineage. *science* 312, 5780 (2006), 1614–1620.
- [51] SAWYER, S. A., AND HARTL, D. L. Population genetics of polymorphism and divergence. *Genetics* 132, 4 (1992), 1161–1176.
- [52] SCHLÖTTERER, C., KOFLER, R., VERSACE, E., TOBLER, R., AND FRANSSEN, S. U. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* 114, 5 (2015), 431–440.
- [53] SMITH, J. M., AND HAIGH, J. The hitch-hiking effect of a favourable gene. *Genetical research* 23, 01 (1974), 23–35.
- [54] SPELLBERG, B., GUIDOS, R., GILBERT, D., BRADLEY, J., BOUCHER, H. W., SCHELD, W. M., BARTLETT, J. G., EDWARDS, J., OF AMERICA, I. D. S., 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 (2008), 155–164.
- [55] STEINRÜCKEN, M., BHASKAR, A., AND SONG, Y. S. A novel spectral method for inferring general diploid selection from time series genetic data. *The annals of applied statistics* 8, 4 (2014), 2203.
- [56] STEPHAN, W., SONG, Y. S., AND LANGLEY, C. H. The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci. *Genetics* 172, 4 (apr 2006), 2647–2663.
- [57] TAJIMA, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 3 (1989), 585–595.
- [58] TERHORST, J., SCHLÖTTERER, C., AND SONG, Y. S. Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. *PLoS Genet* 11, 4 (2015), e1005069.

- [59] TOBLER, R., FRANSSEN, S. U., KOFLER, R., OROZCO-TERWENGEL, P., NOLTE, V., HERMISSON, J., AND SCHLÖTTERER, C. Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. *Molecular biology and evolution* 31, 2 (2014), 364–375.
- [60] TURNER, T. L., STEWART, A. D., FIELDS, A. T., RICE, W. R., AND TARONE, A. M. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet* 7, 3 (2011), e1001336.
- [61] VITTI, J. J., GROSSMAN, S. R., AND SABETI, P. C. Detecting natural selection in genomic data. *Annual review of genetics* 47 (2013), 97–120.
- [62] WANG, J. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical research* 78, 03 (2001), 243–257.
- [63] WAPLES, R. S. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121, 2 (1989), 379–391.
- [64] WILLIAMSON, E. G., AND SLATKIN, M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* 152, 2 (1999), 755–761.
- [65] WILLIAMSON, S. H., HUBISZ, M. J., CLARK, A. G., PAYSEUR, B. A., BUSTAMANTE, C. D., AND NIELSEN, R. Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3, 6 (2007), e90.
- [66] WINTERS, M. A., LLOYD JR, R. M., SHAFER, R. W., KOZAL, M. J., MILLER, M. D., AND HOLODNIY, M. Development of elvitegravir resistance and linkage of integrase inhibitor mutations with protease and reverse transcriptase resistance mutations. *PloS one* 7, 7 (2012), e40514.
- [67] ZAHREDDINE, H., AND BORDEN, K. L. Mechanisms and insights into drug resistance in cancer. *Front Pharmacol* 4, 28.10 (2013), 3389.
- [68] ZHOU, D., UDPA, N., GERSTEN, M., VISK, D. W., BASHIR, A., XUE, J., FRAZER, K. A., POSAKONY, J. W., SUBRAMANIAM, S., BAFNA, V., AND OTHERS. Experimental selection of hypoxia-tolerant *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences* 108, 6 (2011), 2349–2354.