CLEAR: Composition of Likelihoods for Evolve And Resequence Experiments

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

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

S1 Text Choosing Window Size

In genome-wide scans for detecting selection, we apply the CLEAR statistic on sliding windows of length Lbp. The single locus statistic values within the window are averaged to get the composite statistic. While the statistic is robust to variation in window-size, choosing a very large window where LD has decayed will weaken the composite signal, and choosing a small window will decrease the power of composite likelihoods. Here, we use a systematic calculation to choose L as the distance where the LD between the favored mutation and a site $L/2$bp away remains strong.

Consider a segregating site l bp away from the favored allele in a selective sweep. Let ρ_τ be the LD between the favored allele and the site, τ generations after the onset of selection. Then, we have (see Eqs. 30-31 in [2]):

$$\rho_\tau = \alpha_\tau \beta_\tau \rho_0 = e^{-r\tau l} \left(\frac{K^{(\tau)}}{K^{(0)}} \right) \rho_0, \tag{S1}$$

where $K^{(\tau)} = 2\nu_\tau(1-\nu_\tau)$ is the heterozygosity at the selected site, r is the recombination rate (crossovers/bp/gen). The 'decay factor', $\alpha_\tau = e^{-r\tau l}$, and 'growth factor', β_τ, are due to recombination and selection, respectively. Under regular parameter settings, linkage to the favored allele is expected to increase after onset of selection and then decreases due to crossover events (See Figure S15-A). While ρ_0 is unknown in pool-seq E&R experiments, we compute the value of l so that

$$\alpha_\tau \beta_\tau = 1. \tag{S2}$$

In E&R scenarios, we let τ be the time of the last sampling. For given s, we aim to compute the smallest window size L over all possible starting frequencies. Specifically,

$$L = 2 \min_{\nu_0} \left\{ \frac{1}{r\tau} \log \left(\frac{\hat{\nu}_\tau(1-\hat{\nu}_\tau)}{\nu_0(1-\nu_0)} \right) \right\}, \tag{S3}$$

where the term $\hat{\nu}_\tau$ depends on initial frequency ν_0 and selection strength s (Eq. 9).

We used *D. melanogaster* dataset parameters, $N = 250, r = 2 \times 10^{-8}$ and $\tau = 59$ to compute the optimal window size for different values of Ns, ranging from weak selection to strong selection: $Ns \in \{20, 100, 200, 500\}$, or $s \in \{0.08, 0.4, 0.8, 2\}$. We set $L = 30$Kbp (See Figure S15-B) to provide good resolution for detecting weak selection.

Generative Process 1: The Generative Process for Neutral Wright-Fisher Time-series Pool-seq Data.

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

Figure S1: **The Generative Process for Neutral Wright-Fisher Time-series Pool-seq Data.**

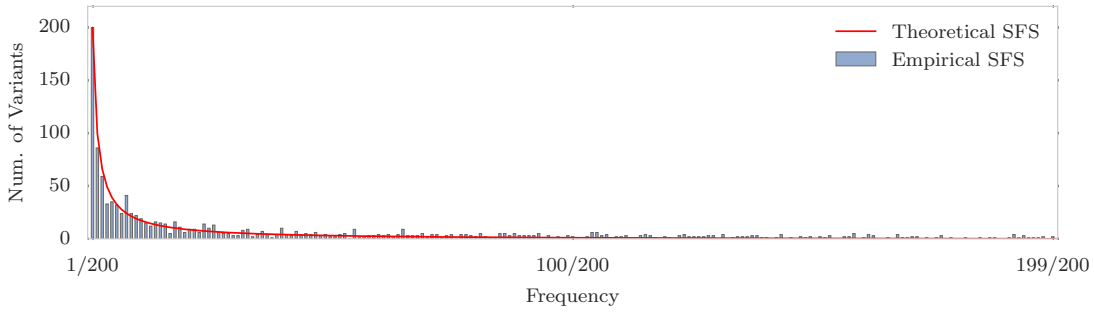

Figure S2: **Site Frequency Spectrum.**
Theoretical and Empirical SFS in a 50Kbp region for a neutral population of 200 individuals when $N_e = 10^6$ and $\mu = 10^{-9}$. The x-axis corresponds to site frequency, and the y-axis to the number of variants with a specific frequency. In a neural population,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

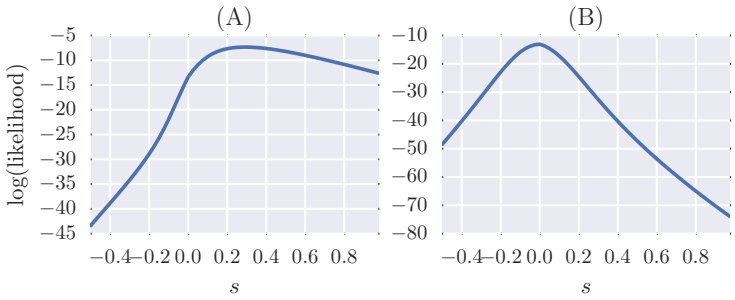

Figure S3: **Likelihoods of the parameter s.**
Likelihood of the parameter s in *D. melanogaster* data for a variant with $\widehat{s} = 0.2$ (A) and $\widehat{s} = 0$ (B).

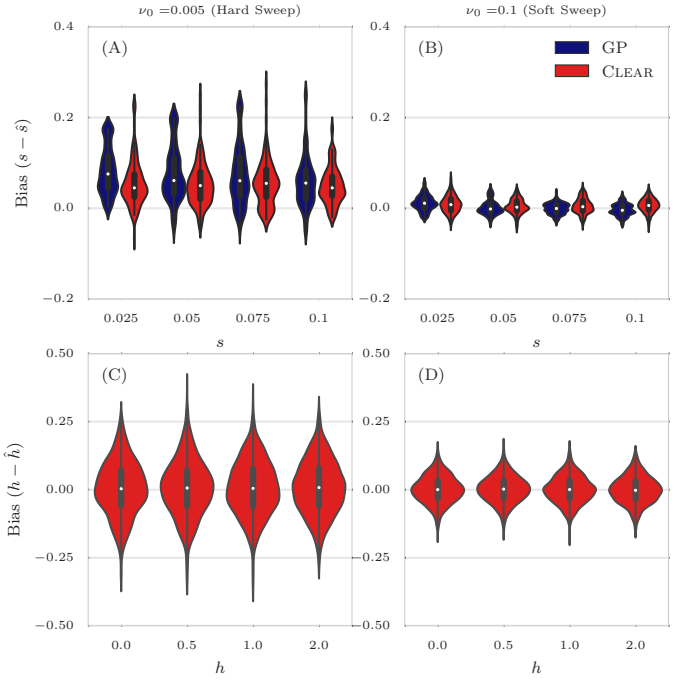

Figure S4: **Distribution of bias for 30× coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and CLEAR (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage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

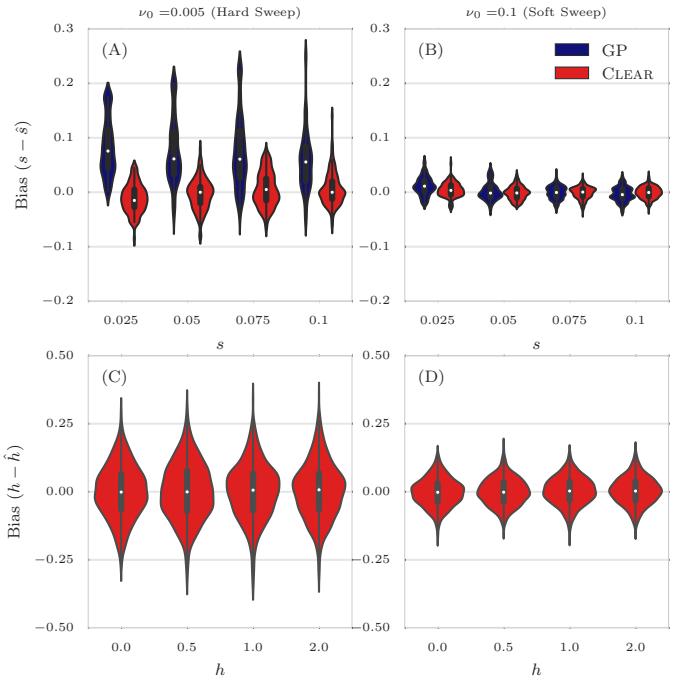

Figure S5: **Distribution of bias for 300× coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and CLEAR (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = \infty$ (Panels A,B). GP and CLEAR have similar variance in estimates of s for soft sweep, while CLEAR provides lower variance in hard sweep. Also see Table S2. Panels C,D show the variance in the estimation of h.

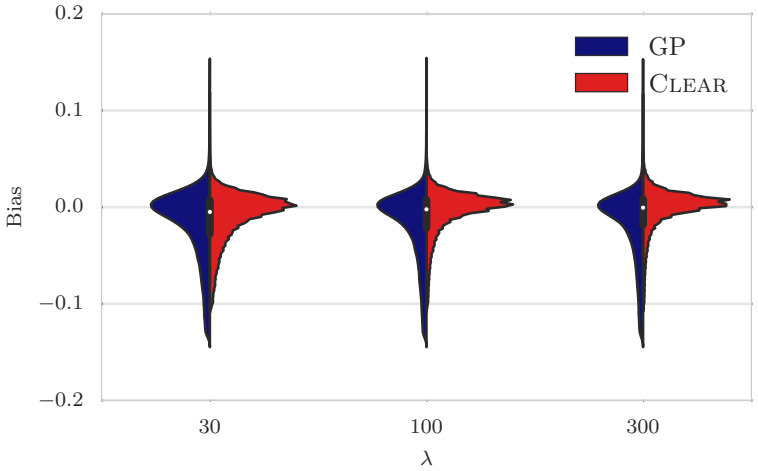

Figure S6: Distribution of bias for null simulations with coverage $\lambda \in \{30, 100, 300\}$.

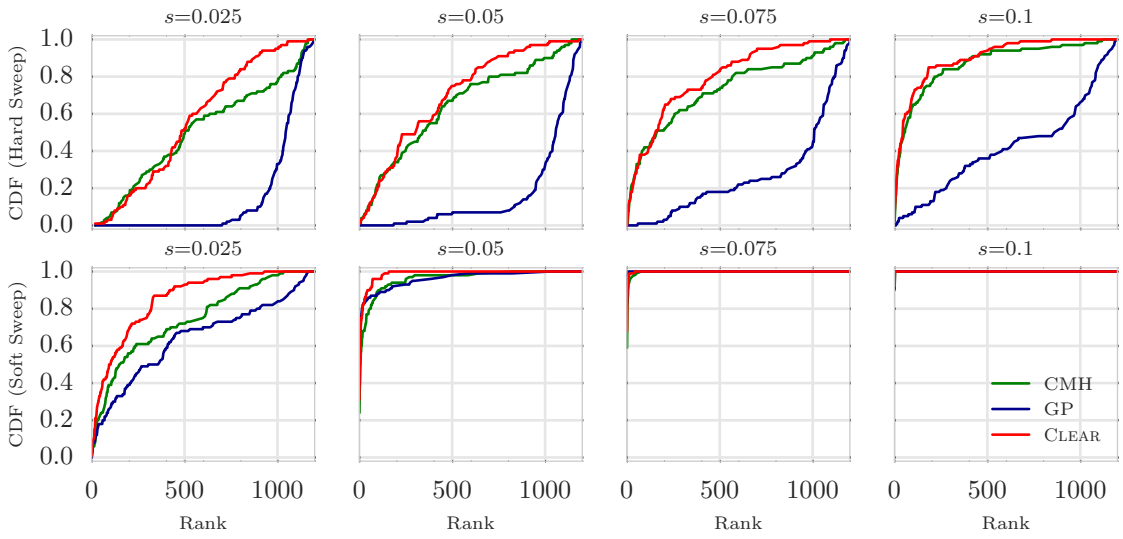

Figure S7: **Ranking performance for 30× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (H score), Gaussian Process (GP), and Cochran Mantel Haenszel (CMH), for different values of selection coefficient s and initial carrier frequency.

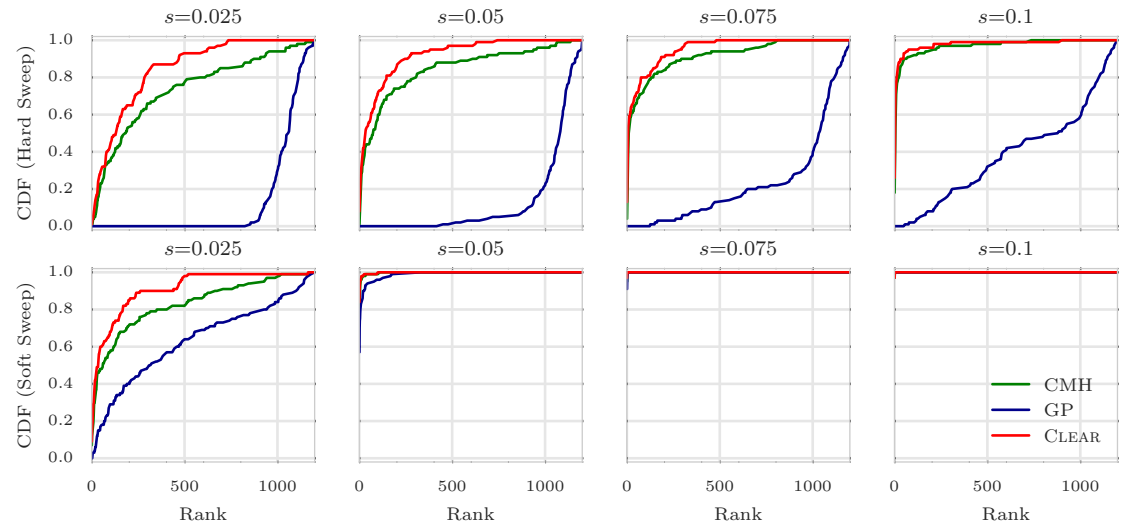

Figure S8: **Ranking performance for 300× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (H score), Gaussian Process (GP), and Cochran Mantel Haenszel (CMH), for different values of selection coefficient s and initial carrier frequency.

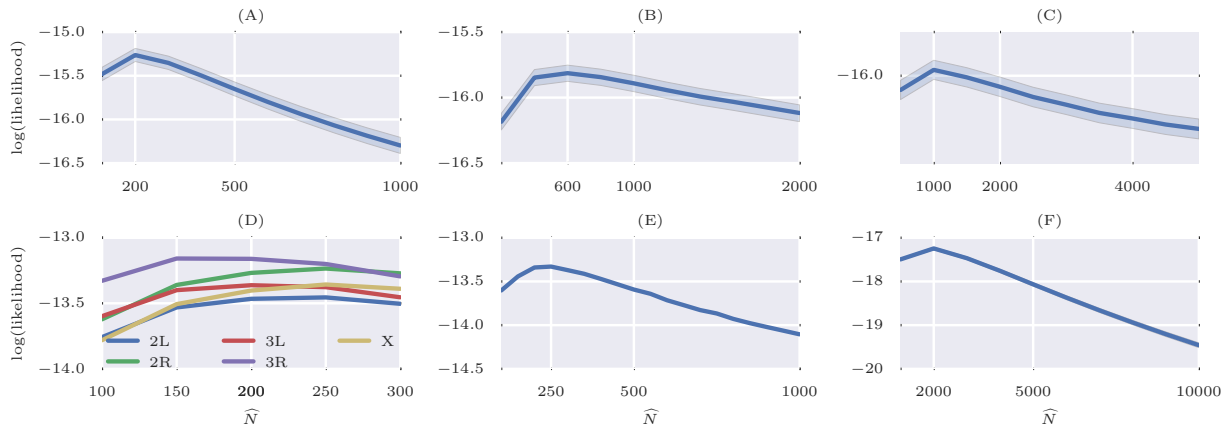

Figure S9: **Maximum likelihood Estimates of N.** Mean and 95% confidence interval of likelihoods of N on simulated data with $N = 200$ (A), $N = 600$(B), and $N = 1000$ individuals, over 1000 simulations. Chromosome-wise (D) and genome-wide (E) likelihood of population size for data from a study of *D. melanogaster* adaptation to alternating temperatures. Likelihood of the Chromosome 3R is attained at 150, while genome-wide maximum likelihood estimate for population size is 250. (F) Likelihood of the population size with respect to all the variants in the yeast dataset. Despite large census population size ($10^6 - 10^7$ [1]), this dataset exhibits much smaller effective population size ($\widehat{N} = 2000$).

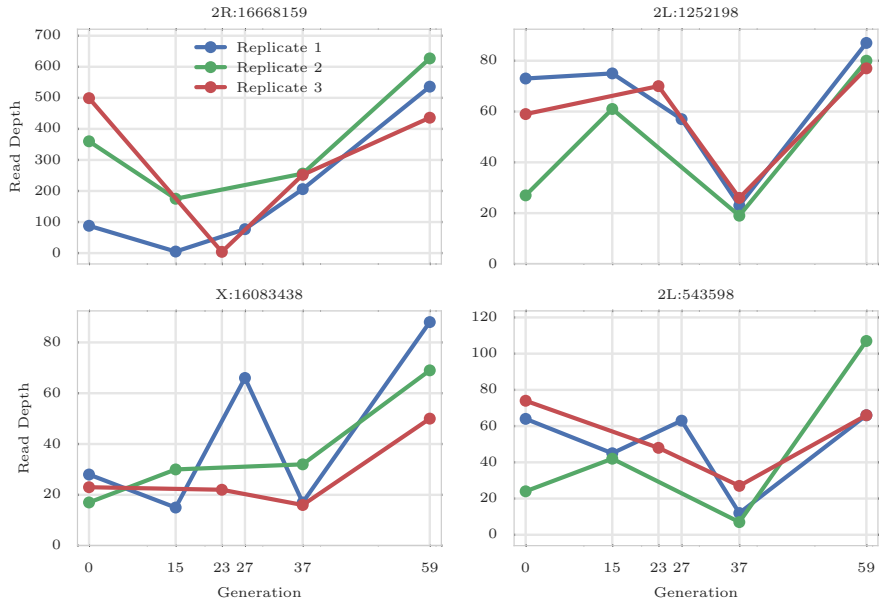

Figure S10: **Coverage heterogeneity in time series data.**
Each panel shows the read depth for 3 replicates of the data from a study of *D. melanogaster* adaptation to alternating temperatures data (see section). Heterogeneity in depth of coverage is seen between replicates, and also at different time points, in all 4 variants. None of these sites pass the the hard filtering with minimum depth of 30.

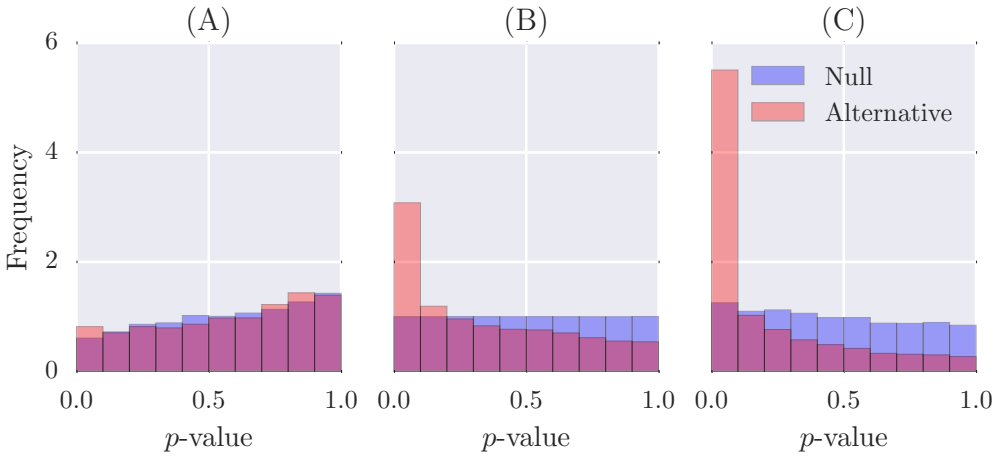

Figure S11: **Distribution of p-values.** Distribution of p-values of CLEAR in null simulations and experimental data when $N = 250$. Panel (A),(C) shows the effect of under estimations ($\widehat{N} = 100$) and over-estimation ($\widehat{N} = 500$) of population size in computing p-values, and panel (B) shows the distribution of $p-$values when unbiased estimate is used to create simulations. .

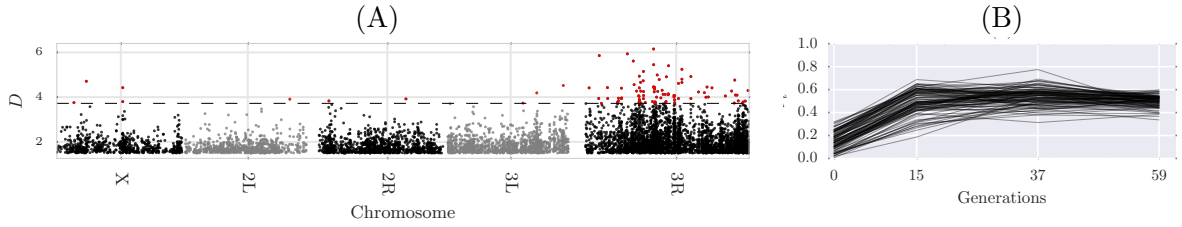

Figure S12: **Single locus analysis of the data from a study of *D. melanogaster* adaptation to alternating temperatures.**
Manhattan plot of scan for testing dominant selection (A). Significant variants with FDR ≤ 0.01 are denoted in red, and their trajectories are depicted in panel (B).

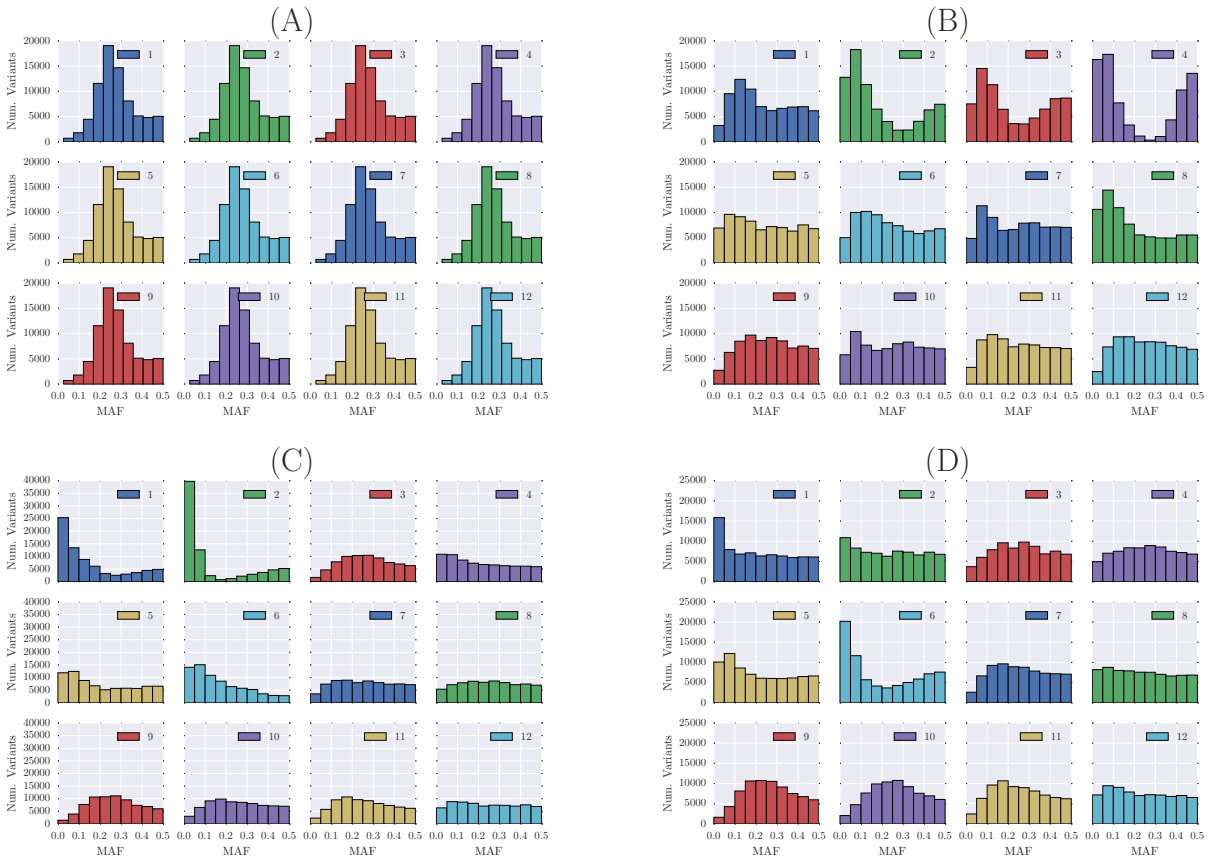

Figure S13: **Site frequency spectrum of the Yeast dataset.** Whole-genome site frequency spectrum of the Yeast dataset at generations 0 (A), 180 (B), 360 (C) and 540 (D). Some replicates, e.g. replicate 2, undergoing severe demographic events.

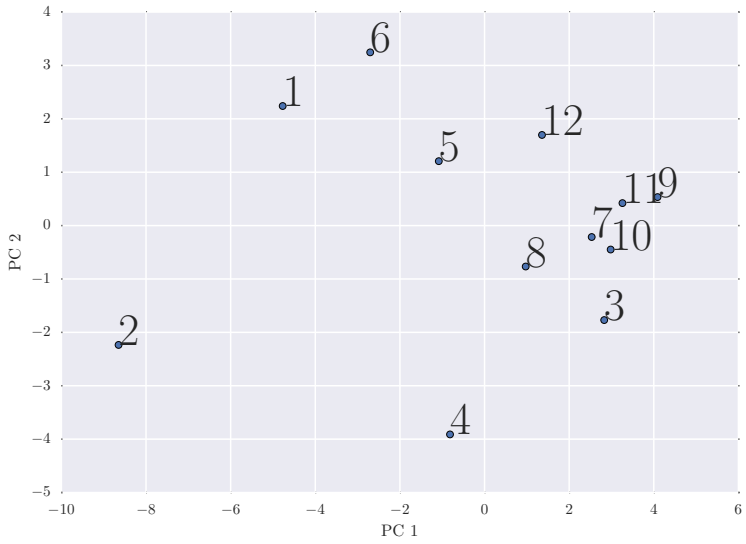

Figure S14: **Population similarity.** Principle component analysis of the 12 replicates throughout the experiment, showing that some populations exhibiting distinct frequency spectra.

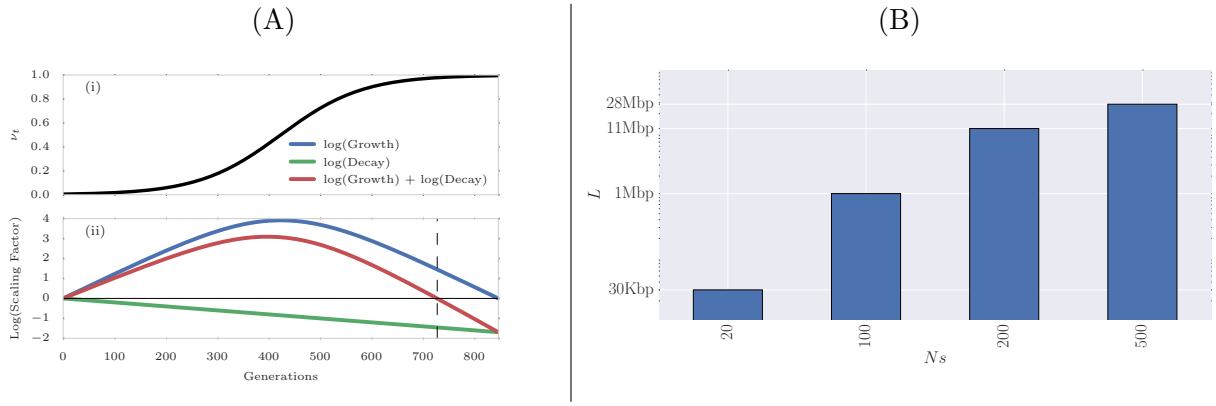

Figure S15: **Choosing window size for** CLEAR **statistic.** (A) Expected dynamics of LD between favored allele ($s = 0.025$) and a variant 50Kbp away, with initial frequency $\nu_0 = 0.01$. (A-i) depicts the dynamic of the favored allele during the selective sweep. (A-ii) illustrates interaction of the growth and decay factors introduced in Eq. S1, with the red line describing overall effect of selection and recombination on LD. The vertical dashed line points to the time when the LD value starts to decrease below original LD. (B) Alternatively, we can fix time, and find the window-size at which LD decays below the original LD (Eq. S3). The plot shows the window size as a function of Ns (20,100,200,500), after fixing other model parameters to match *D. melanogaster* E&R experiments ($N = 250$, $r = 2 \times 10^8, \tau = 59$).

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

Hard Sweep

λ	Method	Avg Power
300	Clear	34
300	Clear($L=1$)	21
300	CMH	12
300	FIT	2
300	GP	0
100	Clear	31
100	Clear($L=1$)	10
100	CMH	4
100	FIT	2
100	GP	0
30	Clear	20
30	Clear($L=1$)	3
30	FIT	2
30	CMH	0
30	GP	0

Soft Sweep

λ	Method	Avg Power
300	Clear	69
300	CMH	69
300	Clear($L=1$)	68
300	GP	61
300	FIT	8
100	Clear	67
100	CMH	60
100	Clear($L=1$)	60
100	GP	59
100	FIT	1
30	Clear	57
30	GP	53
30	CMH	39
30	Clear($L=1$)	39
30	FIT	3

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

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

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

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

Interval	Position	FBgn	Gene Name	GO Function
I1	X:1.567-1.824M	FBgn0023531	CG32809	NA
		FBgn0023130	a6	embryonic development via the syncytial blastoderm
		FBgn0025378	CG3795	serine-type endopeptidase activity
		FBgn0025391	Scgdelta	heart contraction, mesoderm development
		FBgn0261548	CG42666	NA
		FBgn0026086	Adar	RNA editing
		FBgn0026090	CG14812	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process
		FBgn0023522	CG11596	NA
I2	X:7.175-7.241M	FBgn0029941	CG1677	NA
		FBgn0029944	Dok	stress activated protein kinase signaling
		FBgn0029946	CG15034	NA
I3	2L:16.878-16.993M	FBgn0052832	CG32832	mitochondrial pyruvate transport
		FBgn0032618	CG31743	sulfotransferase activity
		FBgn0085342	CG34313	NA
		FBgn0040985	CG6115	NA
		FBgn0261671	tweek	synaptic vesicle endocytosis
		FBgn0026150	ApepP	metalloaminopeptidase activity
		FBgn0262355	CR43053	NA
		FBgn0053179	beat-IIIb	NA
I4	2R:2.725-2.810M	FBgn0040674	CG9445	NA
		FBgn0265935	coro	adult somatic muscle development
		FBgn0033110	CG9447	NA
		FBgn0033113	Spn42Dc	Inhibitory Serpins
		FBgn0028988	Spn42Dd	Inhibitory Serpins
		FBgn0033115	Spn42De	Inhibitory Serpins
		FBgn0050158	CG30158	small GTPase mediated signal transduction
I5	3L:14.362-14.514M	FBgn0036421	CG13481	ubiquitin-protein transferase activity
		FBgn0262580	CG43120	NA
		FBgn0036422	CG3868	NA
		FBgn0087007	bbg	PDZ domain
		FBgn0036426	CG9592	NA
		FBgn0036427	CG4613	serine-type endopeptidase activity

References

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

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