Supplementary Information

Evolutionary model

We assume a WF model consisting of N individuals evolving under mutation, selection, and recombination. Each individual is represented by a sequence of length L. The loci are assumed to be bi-allelic where the value of each locus is either 0 (wild-type (WT)) or 1 (mutant), thus resulting in $M = 2^L$ possible haplotypes. For clarity, we use $i, j, ...$ to refer to locus indices and $a, b, ...$ to refer to haplotype indices. The index is shown as a subscript when representing only one of the locus or haplotype indices. However, when both indices need to be shown simultaneously, the locus index is shown as a subscript while the haplotype index is shown as a superscript. Let $n_a(t)$ denote the number of individuals in the population that belong to haplotype a at generation t such that $\sum_{a=1}^{M} n_a(t) = N$. At generation t, denote $\boldsymbol{z}(t) = (z_1(t), ..., z_M(t))$ as the observed haplotype frequencies with $z_a(t) = \frac{n_a(t)}{N}$. The observed allele frequencies are correspondingly $\boldsymbol{x}(t) = (x_1(t), ..., x_L(t))$ and are related to haplotype frequencies by $x_i = \sum_a g_i^a x_a$ where g_i^a represents the allele (either 0 or 1) at the ith locus of the ath haplotype.

Here we assume that the fitness contribution from individual alleles is additive, such that the fitness f_a of the ath haplotype can be written

$$f_a = 1 + \sum_{i=1}^{L} s_i g_i^a \,.$$

Here the s_i denote the time-invariant selection coefficients for mutant alleles, which quantify the selective advantage of mutant allele i relative to wild-type (WT). This model is consistent with a multiplicative fitness model where the effects of individual mutations are small, or an exponential fitness function for an additive trait, where each mutation is assumed to make a small contribution to the trait value.

Under Wright-Fisher dynamics, the probability of observing haplotype frequencies $\boldsymbol{z}(t+1)$ at generation $t+1$, given haplotype frequencies of $\boldsymbol{z}(t)$ at generation t, is

$$P(\boldsymbol{z}(t+1)|\boldsymbol{z}(t)) = N! \prod_{a=1}^{M} \frac{p_a(\boldsymbol{z}(t))^{Nz_a(t+1)}(\boldsymbol{n})}{(Nz_a(t+1))!}\,, \tag{S1}$$

where $p_a(\boldsymbol{z}(t))$ is the probability of observing haplotype a at generation t. To derive this expression, we sum over contributions to generating haplotype a in the next generation, including the effects of mutation, recombination, and selection,

$$p_a(\boldsymbol{z}(t)) = \frac{y_a(t)f_a + \sum_{b|b\neq a}\left(\mu_{ba}y_b(t)f_b - \mu_{ab}y_a(t)f_a\right)}{\sum_{b=1}^{M} y_b(t)f_b}\,.$$

Here μ_{ba} is the probability for haplotype b to mutate to haplotype a in a generation. $y_a(t)$ is the frequency of haplotype a at generation t after recombination. Specifically, we write $y_a(t)$ as

$$y_a(t) = (1-r)^{L-1}z_a(t) + \left(1-(1-r)^{L-1}\right)\psi_a(\boldsymbol{z}(t))\,,$$

where r is the probability of recombination per locus per generation, and $\psi_a(t)$ is the probability of generating haplotype a through random recombination between two haplotypes in the population at time t.

We assume that mutations at different loci are independent with the same mutation rate μ, so that $\mu_{ba} = \mu^{d(b,a)}(1-\mu)^{L-d(b,a)}$ where $d(b,a)$ is the number of different alleles (Hamming distance) between haplotype b and haplotype a. Below, we will follow the assumption that the population size N is large, and that the selection coefficients s_i, mutation rate μ, and recombination rate r are small ($\mathcal{O}(1/N)$). For now, expanding the expression for $p_a((t))$ to eliminate terms of order μ^2 and higher, we have

$$p_a(\boldsymbol{z}(t)) = \frac{y_a(t)f_a + \mu\sum_{b|d(b,a)=1}(y_b(t)f_b - y_a(t)f_a)}{\sum_{b=1}^{M} y_b(t)f_b}\,. \tag{S2}$$

The probability that the haplotype frequency vector follows a particular evolutionary path $(\boldsymbol{z}(t_1), \boldsymbol{z}(t_2), ..., \boldsymbol{z}(t_K))$, conditioned on the initial state $\boldsymbol{z}(t_0)$, population size N, mutations rate μ and selection coefficients $\boldsymbol{s}$ is

$$P\left((\boldsymbol{z}(t_k))_{k=1}^{K}|\boldsymbol{z}(t_0), N, \mu, \boldsymbol{s}\right) = \prod_{k=0}^{K-1} P\left(\boldsymbol{z}(t+1)|\boldsymbol{z}(t)\right)\,.$$

Covariance can be estimated as $N\left\langle \Delta x_i \Delta x_j \right\rangle$

Recall that we work under the assumption that the population size N is large, selection coefficients s_i, mutation rate μ, and recombination rate r are $\mathcal{O}(1/N)$. We then expand to leading order in $1/N$. Following the multinomial distribution Equation (S1), we have

$$\left\langle \frac{\boldsymbol{n}(t+1) - \boldsymbol{n}(t)}{N} \right\rangle = \boldsymbol{p}(\boldsymbol{z}(t)) - \boldsymbol{z}(t)\,.$$

Expanding Equation (S2) yields

$$\begin{aligned} p_a(\boldsymbol{z}(t)) &= z_a(t)\left(f_a - \sum_b (f_b - 1) z_b(t)\right) + \mu \sum_{b|d(a,b)=1} (z_b(t) - z_a(t)) - r(L-1)\left(z_a(t) - \psi_a(t)\right) + O\left(\frac{1}{N^2}\right) \\ &= z_a(t) + O\left(\frac{1}{N}\right). \end{aligned}$$

Then the moment expansion can be derived as

$$\begin{aligned} \left\langle \left(\frac{n_a(t+1) - n_a(t)}{N}\right)^2 \right\rangle &= \frac{p_a(\boldsymbol{z}(t))\left(1 - p_a(\boldsymbol{z}(t))\right)}{N} + \left(p_a(\boldsymbol{z}(t)) - z_a(t)\right)^2 \\ &= \frac{z_a(t)\left(1 - z_a(t)\right)}{N} + O\left(\frac{1}{N^2}\right), \end{aligned}$$

$$\begin{aligned} \left\langle \frac{n_a(t+1) - n_a(t)}{N} \frac{n_b(t+1) - n_b(t)}{N} \right\rangle &= -\frac{p_a(\boldsymbol{z}(t)) p_b(\boldsymbol{z}(t))}{N} + \left(p_a(\boldsymbol{z}(t)) - z_a(t)\right)\left(p_b(\boldsymbol{z}(t)) - z_b(t)\right) \\ &= -\frac{z_a(t) z_b(t)}{N} + O\left(\frac{1}{N^2}\right). \end{aligned}$$

Then we arrive at the desired relation

$$\begin{aligned} \langle \Delta x_i(t) \Delta x_j(t) \rangle &= \left\langle \left(\sum_a g_i^a \Delta n_a(t)\right)\left(\sum_b g_j^b \Delta n_b(t)\right) \right\rangle \\ &= \left\langle \left(\sum_a g_i^a \frac{n_a(t+1) - n_a(t)}{N}\right)\left(\sum_b g_j^b \frac{n_b(t+1) - n_b(t)}{N}\right) \right\rangle \\ &= \sum_a g_i^a g_j^a \left\langle \left(\frac{n_a(t+1) - n_a(t)}{N}\right)^2 \right\rangle + \sum_{a \neq b} g_i^a g_j^b \left\langle \frac{n_a(t+1) - n_a(t)}{N} \frac{n_b(t+1) - n_b(t)}{N} \right\rangle \\ &= \sum_a g_i^a g_j^a \frac{z_a(t)(1 - z_a(t))}{N} - \sum_{a \neq b} g_i^a g_j^b \frac{z_a(t) z_b(t)}{N} + O\left(\frac{1}{N^2}\right) \\ &= \sum_a g_i^a g_j^a \frac{z_a(t)}{N} - \sum_{a,b} g_i^a g_j^b \frac{z_a(t) z_b(t)}{N} + O\left(\frac{1}{N^2}\right) \\ &= \frac{x_{ij}(t) - x_i(t) x_j(t)}{N} + O\left(\frac{1}{N^2}\right) \\ &= \frac{C_{ij}(\boldsymbol{z}(t))}{N} + O\left(\frac{1}{N^2}\right). \end{aligned}$$

where $x_{ij}(t)$ denotes the frequency of haplotypes in the population with mutant alleles at both sites i and j at time t. $C_{ij}(\boldsymbol{z}(t))$ are entries of the mutant allele frequency covariance matrix, which is equivalent to the linkage disequilibrium (LD) measure D. Note that for $i = j$, $x_{ij} = x_i$, and thus $\langle \Delta x_i(t) \Delta x_i(t) \rangle = x_i(t)\left(1 - x_i(t)\right)/N + O(1/N^2)$.

Marginal path likelihood (MPL) inference

Sohail et al. presents a diffusion approximation and path integral expression for the stochastic haplotype frequency dynamics under the WF model setting (Sohail *et al.* 2021). The path integral expression for the probability of observing a trajectory of haplotype frequencies $(\boldsymbol{z}(t_1), \boldsymbol{z}(t_2), ..., \boldsymbol{z}(t_K))$ is given by

$$\begin{aligned} P\left((\boldsymbol{z}(t_k))_{k=1}^K \,|\boldsymbol{z}(t_0), N, \mu, \boldsymbol{s}\right) &= \prod_{k=0}^{K-1} P\left(\boldsymbol{z}(t+1)|\boldsymbol{z}(t)\right) \\ &\approx \left(\prod_{k=0}^{K-1} \left[\frac{1}{\sqrt{det\, C(\boldsymbol{z}(t_k))}} \left(\frac{N}{2\pi \Delta t_k}\right)^{M/2} d\boldsymbol{z}(t_{k+1})\right]\right) \exp\left(-\frac{N}{2} S\left((\boldsymbol{z}(t_k))_{k=0}^K\right)\right), \end{aligned} \tag{S3}$$

where $C(\boldsymbol{z}(t_k))$ is the haplotype covariance matrix, i.e.,

$$(C(\boldsymbol{z}(t)))_{ab} \coloneqq \begin{cases} z_a(t)(1-z_a(t)) & a=b \\ -z_a(t)z_b(t) & a\neq b \end{cases}.$$

and

$$S\left((\boldsymbol{z}(t_k))_{k=0}^{K}\right) = \sum_{k=0}^{K-1}\frac{1}{\Delta t_k}\sum_{a=1}^{M}\sum_{b=1}^{M}\left[z_a(t_{k+1})-z_a(t_k)-\Delta t_k d_a(\boldsymbol{z}(t_k))\right]\left(C^{-1}(z(t_k))\right)_{ab}\left[z_b(t_{k+1})-z_b(t_k)-\Delta t_k d_b(\boldsymbol{z}(t_k))\right]$$

where $d_a(\boldsymbol{z}(t_k))$ is the expected change in haplotype frequency z_a at time t_k.

The MAP estimate of the selection coefficients can be obtained by solving

$$\hat{\boldsymbol{s}} = \underset{s}{\operatorname{argmax}}\mathcal{L}\left(\boldsymbol{s}|N,\mu,(\boldsymbol{z}(t_k))_{k=1}^{K}\right)P^{prior}(\boldsymbol{s}),$$

where $P^{prior}(\boldsymbol{s})$ is an assumed prior

$$P^{prior}(\boldsymbol{s}) = \frac{1}{(2\pi\sigma^2)^{R/2}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{s}^T\boldsymbol{s}\right),$$

with mean zero and variance $\sigma^2>0$, and the likelihood function is given as

$$\mathcal{L}\left(\boldsymbol{s}|N,\mu,(\boldsymbol{z}(t_k))_{k=1}^{K}\right) = P\left((\boldsymbol{z}(t_k))_{k=1}^{K}|\boldsymbol{z}(t_0),N,\mu,\boldsymbol{s}\right).$$

The right-hand side is approximated by Equation (S3). Differentiating the expression with respect to $\boldsymbol{s}$ and equating to zero leading to the MAP estimator of selection coefficients s

$$\hat{\boldsymbol{s}} = \left[\sum_{k=0}^{K-1}\Delta t_k C\left(\boldsymbol{x}(t_k)\right)+\gamma I\right]^{-1}\times\left[\boldsymbol{x}(t_k)-\boldsymbol{x}(t_0)+\mu\sum_{k=0}^{K-1}\Delta t_k\left(2\boldsymbol{x}(t_k)-1\right)\right],$$

where $\gamma = 1/N\sigma^2$.

Simulation

To benchmark the performance of our method, we generated artificial time series sequence data by simulating evolution as a Wright-Fisher process. As in previous section, the loci were assumed to be bi-allelic. We used 10 different sets of selection coefficients (see Supplementary Figure S1) and simulate 20 replicates of data for each set, totalling 200 simulations. In each simulation, a population of $N = 1000$ sequences were simulated to evolve for $T = 700$ generations. At the first generation, the population consisted of four haplotypes that were randomly generated, and the individuals were randomly distributed over the four haplotypes. At a generation t, the population first goes through a multinomial sampling process to determine the number of sequences $z_a(t)$ for each existing haplotype a in the current generation, where the probability p_a for a haplotype a to be drawn is proportional to the product of its frequency in the last generation $z_a(t-1)$ and fitness f_a, i.e.,

$$p_a(\boldsymbol{z}(t)) = \frac{z_a(t-1)f_a}{\sum_{b=1}^{M}z_b(t-1)f_b}.$$

Mutations were then induced randomly by a preset mutation rate $\mu = 10^{-3}$, under the constraint that each sequence can have at most one mutation per generation.

In order to test effects of recombination, we performed another 200 simulations with the same setup as above, but with recombination enabled. In these simulations, after mutations were randomly generated, the population went through a recombination step at a rate of $r = 10^{-5}$, i.e., each pair of sequences had a probability of r to recombine at a random locus. With this setup each simulation had around 3.5×10^4 mutation events, and around 3.5×10^3 recombination events.

MPL implementation

Linear interpolation

When implementing MPL inference following Equation (4), with time-series allele frequencies and covariances (observed or estimated), we linearly interpolate allele frequencies $\{x_i\}, i = 1,...,L$ and pairwise allele frequencies $\{x_{ij}\}, i,j = 1,2,...,L$ between each two adjacent time points in order to compute the integrated covariance matrix, i.e.,

$$C_{integrated} = \sum_k\int_{t_k}^{t_{k+1}}C(t)dt = \sum_k\int_{t_k}^{t_{k+1}}x_{ij}(t)-x_i(t)x_j(t)dt,$$

where $x_i(t) = x_i(t_k)+\frac{t-t_k}{t_{k+1}-t_k}(x_i(t_{k+1})-x_i(t_k))$, and $x_{ij}(t) = x_{ij}(t_k)+\frac{t-t_k}{t_{k+1}-t_k}(x_{ij}(t_{k+1})-x_{ij}(t_k))$ for $t\in[t_k,t_{k+1}]$. The linear interpolation is adopted as a standard procedure for MPL inference, and has little influences for results presented in this paper.

The only exception is for unnormalized results, where the linear interpolation interferes with unnormalized estimates, leading to poor inference of selection coefficients. Therefore for unnormalized results the integrated covariance matrix is computed as sum of covariance matrix at each time point, times the average time interval around it, i.e.,

$$C_{integrated} = \sum_k C(t_k) \frac{t_{k+1} - t_{k-1}}{2} .$$

Supplementary results

Choices of the time window

Here we show how different choices of time window affects estimation of covariance matrix and inference of selection coefficients in Supplementary Figure S4. Performance improves with larger time windows, and saturates at certain values. The window needed for optimal performance gets larger when data gets more limited. Further increasing the window size from the saturation point does not show harm to performance in our testing.

Effect of normalization

In Equation (3), we normalize the initial estimate E with observed variances. Here in Supplementary Figure S3 we compare its performance with or without the normalization step. The normalization step reliably decreases error of the estimated covariance matrix, and subsequently improves inference of selection coefficients.

Choices of the shrinker and η

In Supplementary Figure S6 we show how values of the regularization strength η and choices of shrinkers affect performance of nonlinear shrinkage regularization method in various degrees of finite sampling effects. We tested 10 optimal shrinkers that are respectively derived using loss functions for the Frobenius norm or nuclear norm of $\hat{R} - R$, $\hat{R}^{-1} - R^{-1}$, $R^{-1}\hat{R} - I$, $\hat{R}^{-1}R - I$, $\hat{R}^{-1/2}R\hat{R}^{-1/2} - I$, where Frobenius norm and nuclear norm for a $m \times n$ matrix A are defined as

$$\begin{aligned} ||A||_{Frobenius} &= \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2} , \\ ||A||_{nuclear} &= \sum_i \sigma_i(A) . \end{aligned} \tag{S4}$$

where $\sigma_i(A)$ is the ith sigular value of matrix A. We find that shrinkers corresponding to Frobenius loss functions typically outperform those corresponding to nuclear loss functions. And the shrinker corresponding to the loss function of Frobenius norm of $\hat{R}^{-1}R - I$ has best performance across almost all sampling variations. As for choices of the regularization strength η, we find that a small strength ($\eta = 1 \times 10^{-5}$) is best for most cases when sampling time interval is smaller than 10, with which all shrinkers collapse to the same performance. In fact with a vanishing η, these optimal shrinkers for different loss functions approximately collapse to one same shrinker, which merely sets eigenvalues smaller than $\lambda_+(\eta)$ to 1 and largely preserves larger ones' original values, where $\lambda_+(\eta)$ is the right bound of eigenvalue distribution of a high-dimensional random covariance matrix given by the Marchenko–Pastur law (Marčenko and Pastur 1967). Overall we find that the shrinker corresponding to the loss function of Frobenius norm of $\hat{R}^{-1}R - I$, combined with a small regularization strength $\eta = 1 \times 10^{-5}$ yields near-optimal performance in almost all cases.

Performance comparison of all methods

Here we include side-by-side comparisons for performances of all tested methods under various limited sampling effects in both measures, e.g., Spearman's ρ and MAE, in Supplementary Figure S7-S8. On average, linear shrinkage tends to perform very slightly better than nonlinear methods when the time interval between samples is small. However, the regularization strength for the linear method needs to be tuned for optimal performance. For large sampling intervals, the linear regularization strength needed to achieve optimal rank correlation between the true and inferred selection coefficients increases in proportion to Δg, which results in extremely small magnitudes for inferred selection coefficients. For these reasons, nonlinear regularization (with strength $\eta = 1 \times 10^{-5}$ and using loss function of Frobenius norm of $\hat{R}^{-1}R - I$) is likely the best choice for arbitrary inference problems.

Performance on simulated data with recombination

We perform another 200 simulations with the same setup as described in , except that the recombination is now enabled with a rate of 1×10^{-5}. In each simulation, a population of $N = 1000$ sequences with $L = 50$ loci evolves for $T = 700$ generations. The mutation rate of $\mu = 1e^{-3}$ expects to generate around 3.5×10^4 mutation events. The recombination rate of $r = 1 \times 10^{-5}$ expects to generate around 3.5×10^3 recombination events. Performances of various methods on selection inference are shown in Supplementary Figure S9-S10. Performances are consistent with those evaluated on simulated data without recombination shown in . As expected from our derivation, the Est method and its variations apply to evolution with recombination.

Performance combining multiple replicate data with different founder haplotypes

We generate 20 different initial haplotype distributions, each with four random founder haplotypes. Individuals in the initial population are randomly distributed across founder haplotypes. We perform a WF simulation for each combination of the 20 initial distributions and 10 sets of selection coefficients, yielding 200 simulations in total. We then investigate the effects of combining 20 replicate data with the same selection but different founder haplotypes. As shown in Supplementary Figure S11, performances on selection inference of all methods are greatly improved.

Literature cited

Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C, O'sullivan KM, DeSouza I, Feeney ME, Eldridge RL, Maier EL *et al*. 2005. Selective escape from CD8+ t-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. J. Virol.. 79:13239–13249.

Anand S, Mangano E, Barizzone N, Bordoni R, Sorosina M, Clarelli F, Corrado L, Boneschi FM, D'Alfonso S, De Bellis G. 2016. Next generation sequencing of pooled samples: Guideline for variants' filtering.

Badran AH, Guzov VM, Huai Q, Kemp MM, Vishwanath P, Kain W, Nance AM, Evdokimov A, Moshiri F, Turner KH *et al*. 2016. Continuous evolution of bacillus thuringiensis toxins overcomes insect resistance. Nature. 533:58–63.

Bansal V, Bafna V. 2008. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics. 24:i153–9.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. 2008. The influenza virus resource at the national center for biotechnology information. J. Virol.. 82:596–601.

Barghi N, Tobler R, Nolte V, Jakšić AM, Mallard F, Otte KA, Dolezal M, Taus T, Kofler R, Schlötterer C. 2019. Genetic redundancy fuels polygenic adaptation in drosophila. PLoS Biol.. 17:e3000128.

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. 2009. Genome evolution and adaptation in a long-term experiment with escherichia coli. Nature. 461:1243–1247.

Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol.. 3:329.

Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C *et al*. 2010. Signatures of mutation and selection in the cancer genome. Nature. 463:893–898.

Buffalo V, Coop G. 2019. The linked selection signature of rapid adaptation in temporal genomic data. Genetics. 213:1007–1045.

Buffalo V, Coop G. 2020. Estimating the genome-wide contribution of selection to temporal allele frequency change. Proc. Natl. Acad. Sci. U. S. A.. 117:20672–20680.

Burrell RA, McGranahan N, Bartek J, Swanton C. 2013. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 501:338–345.

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res.. 63:213–227.

Donoho DL, Gavish M, Johnstone IM. 2018. Optimal shrinkage of eigenvalues in the spiked covariance model. Ann. Stat.. 46:1742–1778.

Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N. 2008. Viral population estimation using pyrosequencing. PLoS Comput. Biol.. 4:e1000074.

Esvelt KM, Carlson JC, Liu DR. 2011. A system for the continuous directed evolution of biomolecules. Nature. 472:499–503.

Ewens WJ. 2012. *Mathematical Population Genetics 1: Theoretical Introduction*. Springer Science & Business Media.

Feder AF, Petrov DA, Bergland AO. 2012. LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data. PLoS One. 7:e48588.

Franssen SU, Barton NH, Schlötterer C. 2017. Reconstruction of Haplotype-Blocks selected during experimental evolution. Mol. Biol. Evol.. 34:174–184.

Franssen SU, Nolte V, Tobler R, Schlötterer C. 2015. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental drosophila melanogaster populations. Molecular biology and evolution. 32:495–509.

Gerrish PJ, Lenski RE. 1998. The fate of competing beneficial mutations in an asexual population. Genetica. 102-103:127–144.

Hedrick PW. 1987. Gametic disequilibrium measures: proceed with caution. Genetics. 117:331–341.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet.. 38:226–231.

Illingworth CJR, Fischer A, Mustonen V. 2014. Identifying selection in the within-host evolution of influenza using viral sequence data. PLoS Comput. Biol.. 10:e1003755.

Illingworth CJR, Mustonen V. 2011. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. Genetics. 189:989–1000.

Johnstone IM. 2001. On the distribution of the largest eigenvalue in principal components analysis. Ann. Stat.. 29:295–327.

Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L *et al*. 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 152:714–726.

Ledoit O, Wolf M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal.. 88:365–411.

Ledoit O, Wolf M. 2020. The power of (Non-)Linear shrinking: A review and guide to covariance matrix estimation. Journal of Financial Econometrics. .

Lee B, Sohail MS, Finney E, Ahmed SF, Quadeer AA, McKay MR, Barton JP. 2022. Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. medRxiv. pp. 2021–12.

Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, Bloom JD. 2018. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc. Natl. Acad. Sci. U. S. A.. 115:E8276–E8285.

Long A, Liti G, Luptak A, Tenaillon O. 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat. Rev. Genet.. 16:567–582.
Luksza M, Lässig M. 2014. A predictive fitness model for influenza. Nature. 507:57–61.
Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, Rizvi NA, Merghoub T, Levine AJ, Chan TA *et al*. 2017. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature. 551:517–520.
Lynch M, Bost D, Wilson S, Maruki T, Harrison S. 2014. Population-genetic inference from pooled-sequencing data. Genome Biol. Evol.. 6:1210–1218.
Marčenko VA, Pastur LA. 1967. DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES.
McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. 2010. The immune response during acute HIV-1 infection: clues for vaccine development. Nat. Rev. Immunol.. 10:11–23.
Metzker ML. 2010. Sequencing technologies - the next generation. Nat. Rev. Genet.. 11:31–46.
Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. 2016. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc. Natl. Acad. Sci. U. S. A.. 113:E1701–9.
Pelizzola M, Behr M, Li H, Munk A, Futschik A. 2021. Multiple haplotype reconstruction from allele frequency data. Nature Computational Science. 1:262–271.
Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, Elvin JG, Rothbard JA, Bangham CR, Rizza CR. 1991. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature. 354:453–459.
Rambaut A, Posada D, Crandall KA, Holmes EC. 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet.. 5:52–61.
Shen MW, Zhao KT, Liu DR. 2021. Reconstruction of evolving gene variants and fitness from short sequencing reads. Nat. Chem. Biol.. 17:1188–1198.
Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res.. 23:23–35.
Sohail MS, Louie RHY, Hong Z, Barton JP, McKay MR. 2022. Inferring epistasis from genetic time-series data. Mol. Biol. Evol.. 39.
Sohail MS, Louie RHY, McKay MR, Barton JP. 2021. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. Nat. Biotechnol.. 39:472–479.
Terhorst J, Schlötterer C, Song YS. 2015. Multi-locus analysis of genomic time series data from experimental evolution. PLoS Genet.. 11:e1005069.
Wu NC, Wilson IA. 2017. A perspective on the structural and functional constraints for immune evasion: Insights from influenza virus. J. Mol. Biol.. 429:2694–2709.
Xue KS, Stevens-Ayers T, Campbell AP, Englund JA, Pergam SA, Boeckh M, Bloom JD. 2017. Parallel evolution of influenza across multiple spatiotemporal scales. Elife. 6.
Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 12:119.
Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA. 2015. Population genomics of intrapatient HIV-1 evolution. Elife. 4.

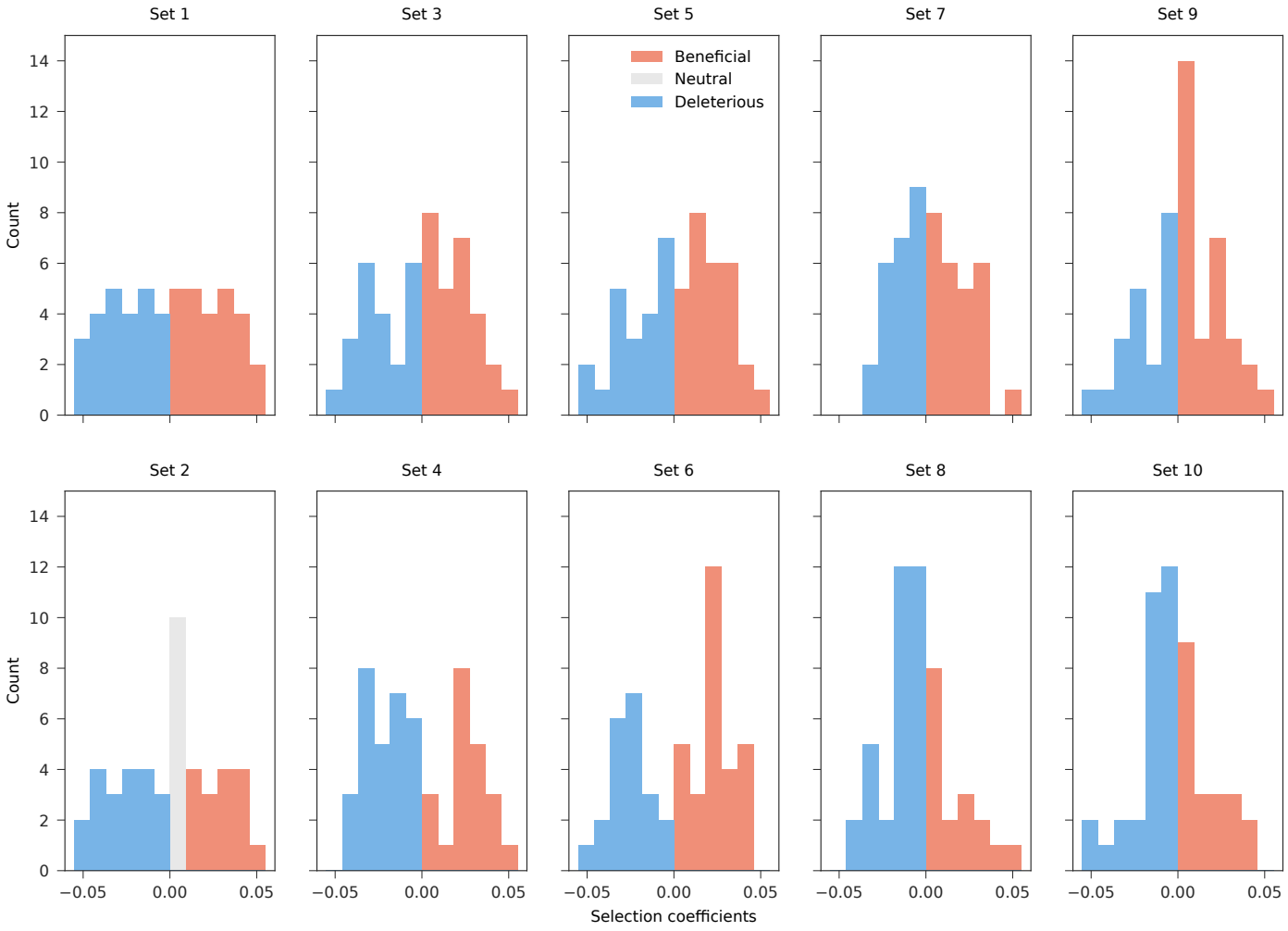

Figure S1 Uniform, trimodal and Gaussian distributions are used to generate 10 sets of selection coefficients used in simulations. The first two sets (Set 1-2) of selection coefficients are drawn from uniform distribution in range [-0.05, 0.05], with 0 and 10 neutral alleles, respectively. The next four sets (Set 3-6) are drawn from a trimodal distribution, a combination of three Gaussian distributions with standard deviation of 0.01, centered at -0.03, 0, 0.03 respectively. Another four sets (Set 7-10) are drawn from a Gaussian distribution with standard deviation of 0.02, centered at 0.

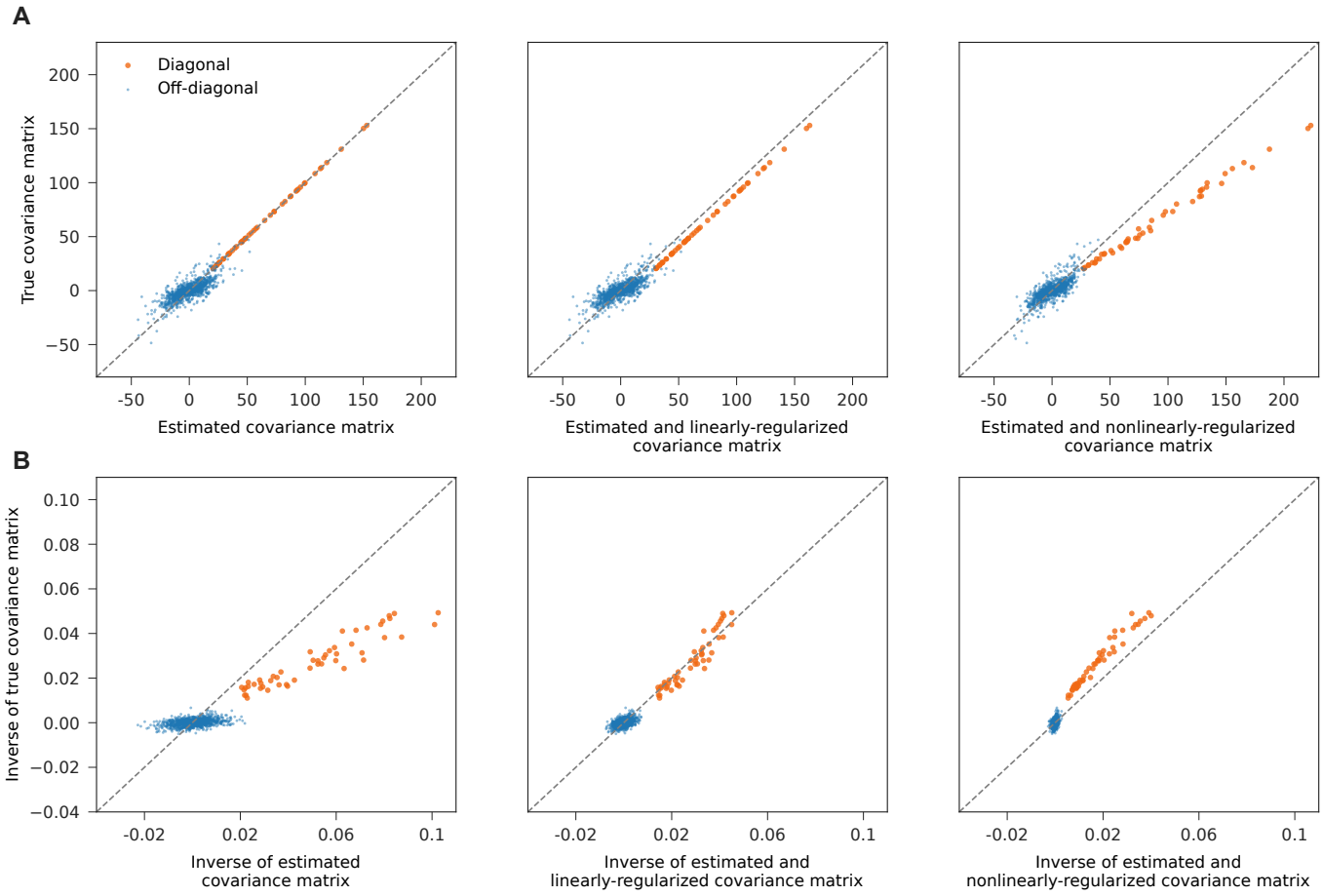

Figure S2 Regularization restrains large terms in the inverse of estimated covariance matrix. (A) Terms in true covariance matrix are compared with estimated covariance matrix (left), estimated and linearly-regularized covariance matrix (middle), and estimated and nonlinearly-regularized covariance matrix (right). Both regularization methods decrease magnitudes of diagonal terms, but have little influence on off-diagonal terms. (B) Terms in the inverse of true covariance matrix are compared with inverse of estimated covariance matrix (left), inverse of estimated and linearly-regularized covariance matrix (middle), and inverse of estimated and nonlinearly-regularized covariance matrix (right). Both regularization methods greatly restrain magnitudes of both diagonal and off-diagonal terms in the inverse of covariance matrix. The covariance matrices compared here are all integrated throughout the simulation shown in Figure 1A, using data sampled every 10 generations with 100 sequences per sample. The linear shrinkage regularization adopts a strength of 10, and the nonlinear shrinkage regularization adopts the optimal shrinker corresponding to the loss function of the Frobenius norm of $\hat{R}^{-1}R - I$ combined with a regularization strength $\eta = 1 \times 10^{-5}$.

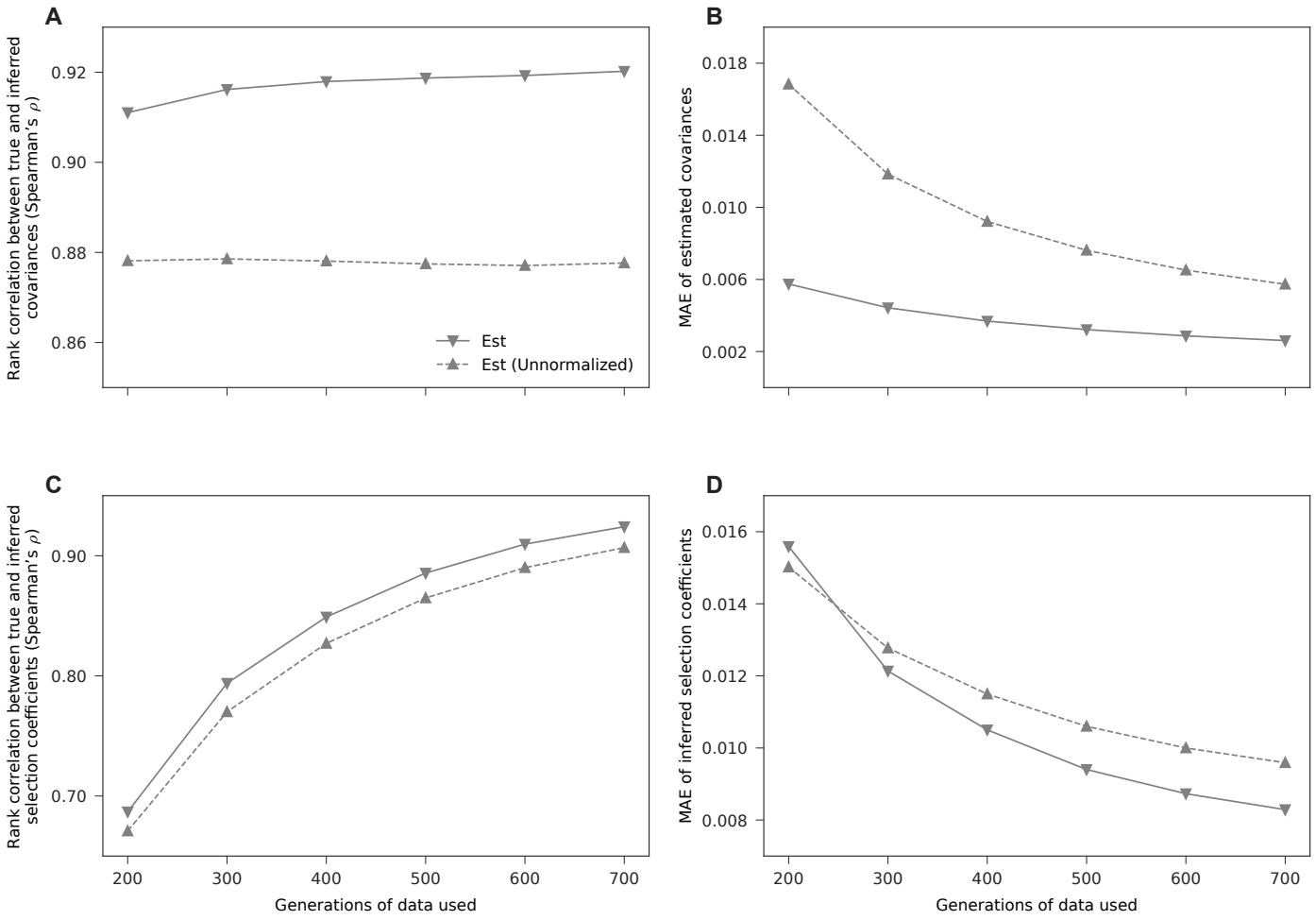

Figure S3 Normalization improves performance of both estimating covariance and inferring selection. (A) Spearman's rank correlation coefficients between estimated and true integrated covariance matrix entries, (B) MAE of estimated integrated covariance matrix normalized by number of generations used, (C) Spearman's rank correlation coefficients between inferred and true selection coefficients, (D) MAE of inferred selection coefficients, with or without normalization, averaged over 200 simulations with same setup as shown in Figure 1A, are compared here. The normalization step reliably decreases error of estimation of the integrated covariance matrix, and improves accuracy of inference of selection coefficients.

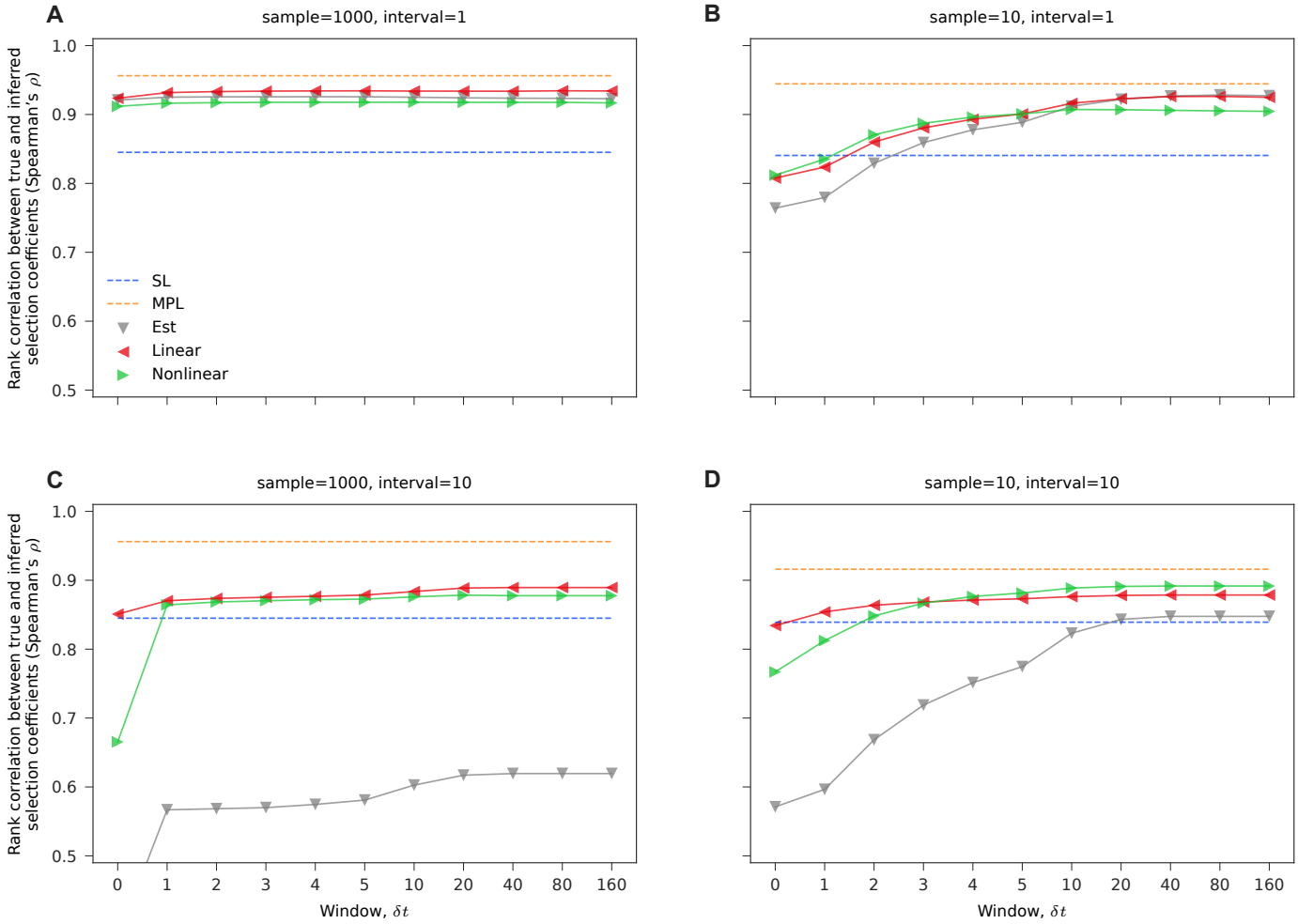

Figure S4 Performance improves with larger time windows. Spearman's rank correlation coefficients between inferred and true selection coefficients, averaged over 200 simulations with same setup as shown in Figure 1A, are shown for four cases (A) when data is ample, (B) when only 10 samples are drawn at each time point, (C) when the sampling time interval is 10 generations, (D) when only 10 samples are drawn at each time point and the sampling time interval is 10 generations. In general the performance gets better with larger time window, and saturates at certain values, e.g., (A) at window=1, (B)(C)(D) around window=20. Overall a window of 20 is sufficient for optimal performance for these four cases, while further increasing the window does not show harm to the performance.

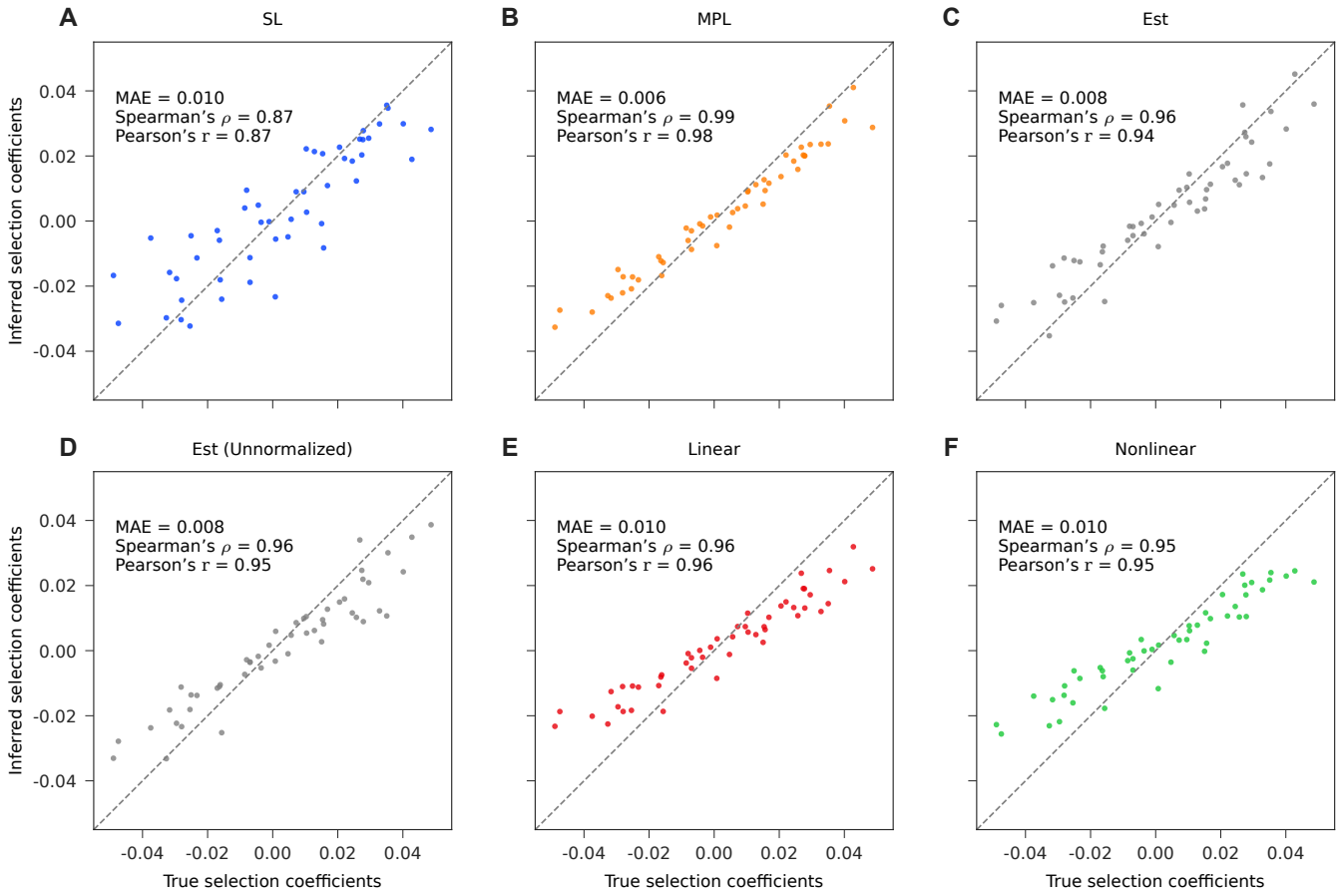

Figure S5 Inferred selection coefficients for an example simulation. Selection coefficients inferred by various methods with ample data are compared with true values for the example simulation plotted in Figure 1A. Each dot shows the true versus inferred selection coefficients for a locus in a single simulation. When sampling is complete, the naive Est method is sufficient to recover both ordering and magnitudes of the underlying selection coefficients and has similar performance as the MPL method which uses true covariance information. The regularized methods preserve the performance on Spearmanr's ρ, while having smaller magnitudes.

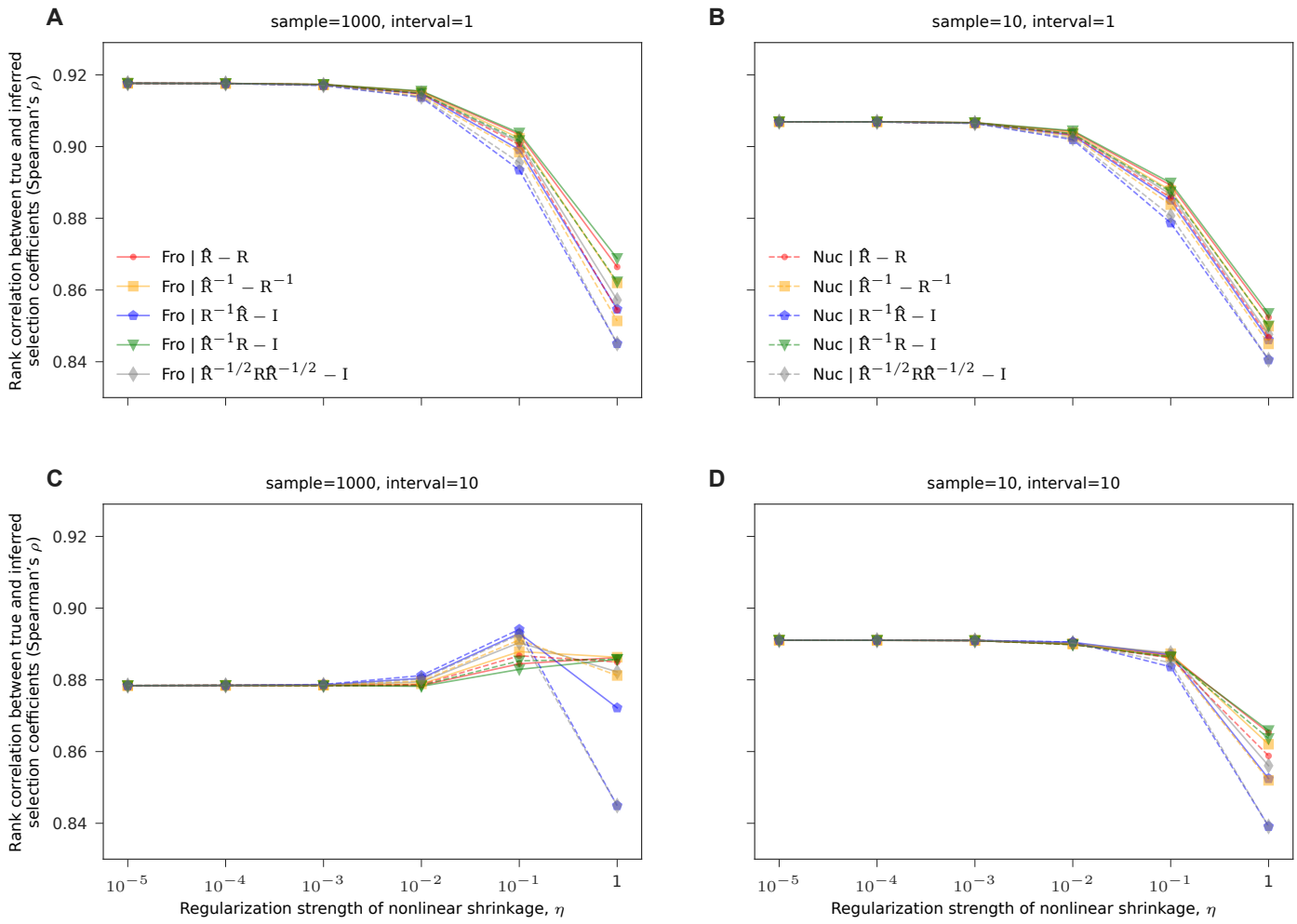

Figure S6 Nonlinear shrinkage performs best with a small regularization strength η. Spearman's rank correlation coefficients between inferred and true selection coefficients, averaged over 200 simulations with same setup as shown in Figure 1A, are shown for different values of gamma using 10 different shrinkers, (A) when data is ample, (B) when only 10 samples are drawn at each time point, (C) when sampling time interval is 10 generations, (D) when only 10 samples are drawn at each time point and sampling time interval is 10 generations. Shrinkers optimal for Frobenius loss are plotted as solid lines, while those for nuclear loss are plotted as dashed lines.

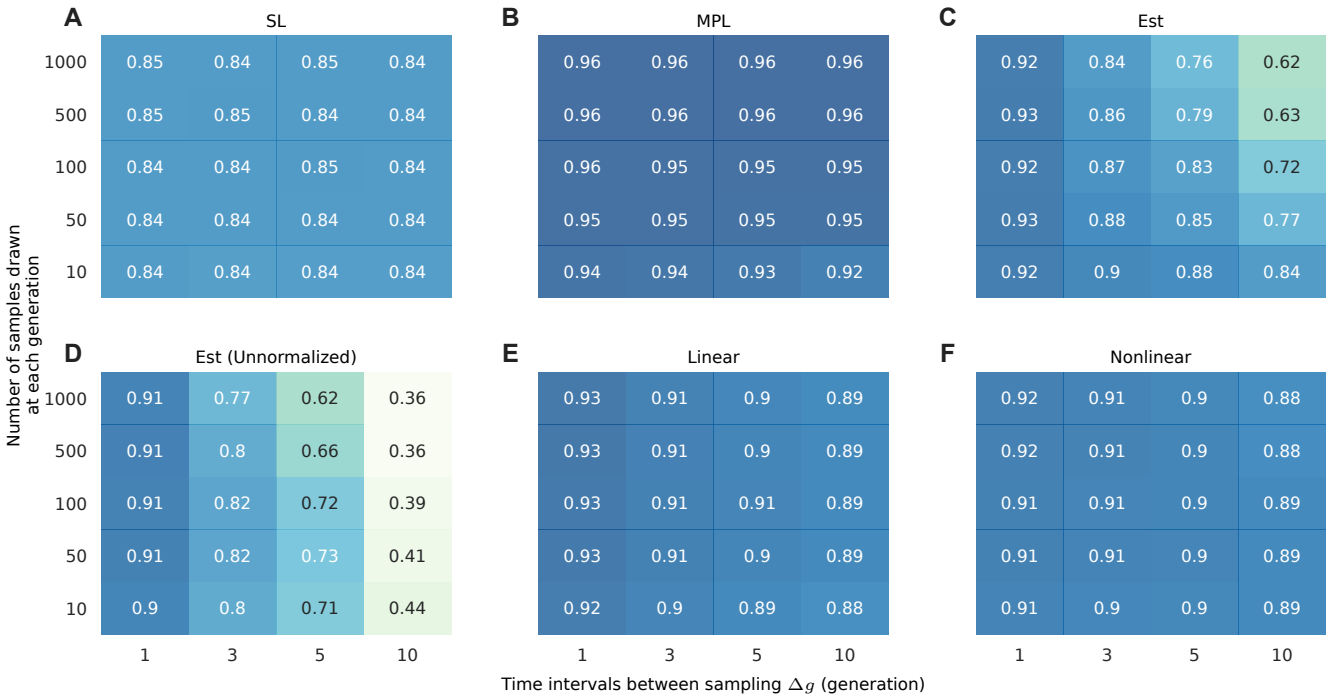

Figure S7 Spearmanr's ρ of all tested methods. Spearman's rank correlation coefficients between inferred and true selection coefficients, averaged over 200 simulations with same setup as shown in Figure 1A, are shown for different limited sampling effects using (A) SL method which ignores genetic linkage, (B) MPL method which uses true sample covariance, (C) Est method with normalized estimate of covariance, (D) Est method with unnormalized estimate of covariance, (E) Est method with normalized estimate of covariance, regularized by linear shrinkage of strength $\gamma = 10\Delta g$, (F) Est method with normalized estimate of covariance, regularized by nonlinear shrinkage, with strength $\eta = 1 \times 10^{-5}$ and using the optimal shrinker derived for the loss function of Frobenius norm of $\hat{R}^{-1}R - I$.

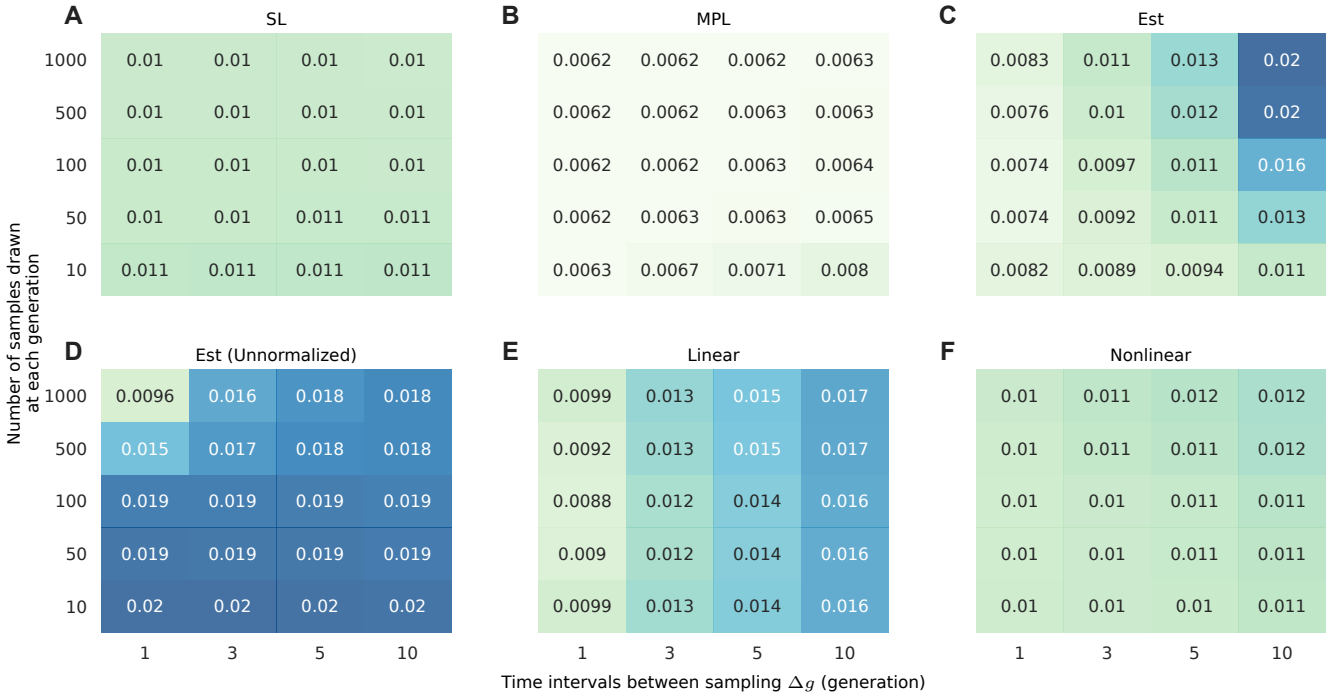

Figure S8 MAE of inferred selection coefficients for all tested methods. MAE of inferred selection coefficients, averaged over 200 simulations with same setup as shown in Figure 1A, are shown for different limited sampling effects using (A) SL method which ignores genetic linkage, (B) MPL method which uses true sample covariance, (C) Est method with normalized estimate of covariance, (D) Est method with unnormalized estimate of covariance, (E) Est method with normalized estimate of covariance, regularized by linear shrinkage of strength $\gamma = 10\Delta g$, (F) Est method with normalized estimate of covariance, regularized by nonlinear shrinkage, with strength $\eta = 1 \times 10^{-5}$ and using the optimal shrinker derived for the loss function of Frobenius norm of $\hat{R}^{-1}R - I$. Note that the mean absolute value of true selection coefficients is about 0.015. We find that when sampling time interval is large ($\Delta g = 10$), the linear shrinkage regularization method has high MAE, which is because it infers selection coefficients with vanishing magnitudes.

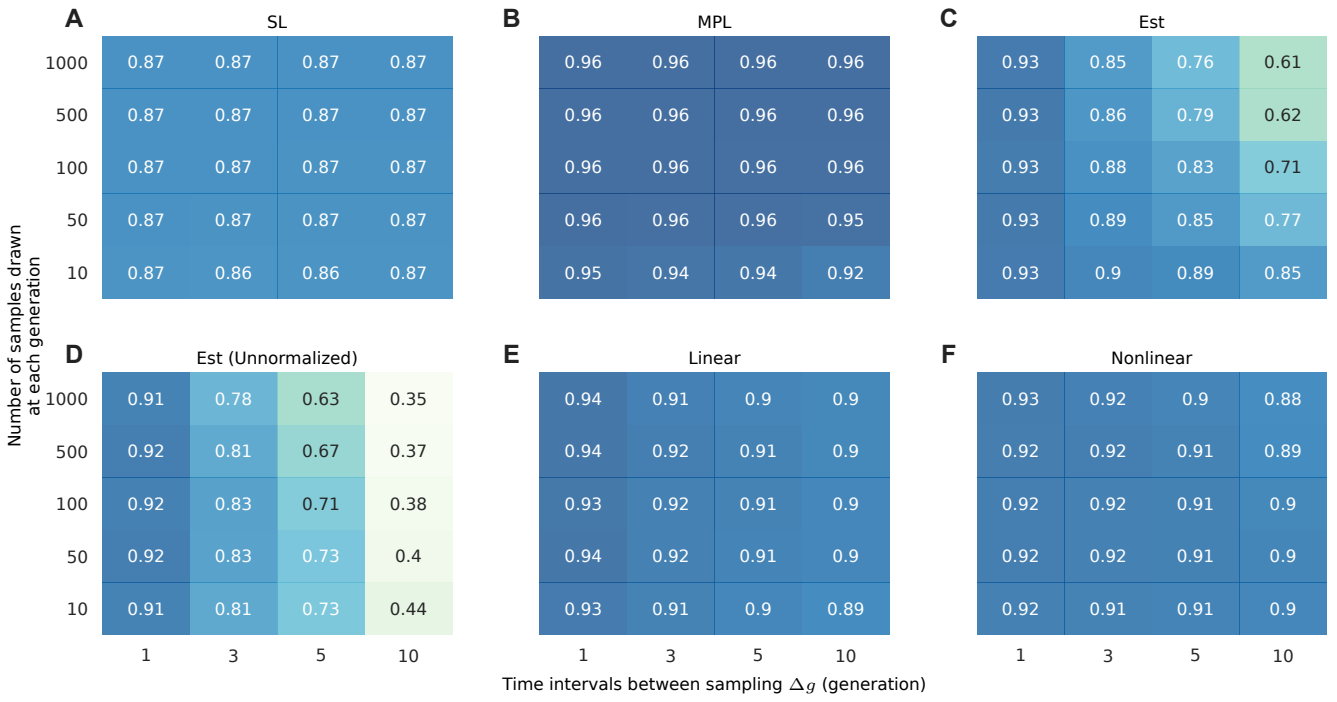

Figure S9 Spearmanr's ρ of all tested methods on simulated data with recombination. Performance of various methods on selection inference are evaluated and averaged over 200 simulations. The recombination rate is set as 10^{-5} per locus per generation. Other parameters are the same as the example shown in Figure 1A. Spearman's rank correlation coefficients between inferred and true selection coefficients are shown for different limited sampling effects using (A) SL, which ignores linkage disequilibrium, (B) MPL, which uses the true population covariance, (C) Est with normalized estimates of covariance, (D) Est with unnormalized estimates of covariance, (E) Est with normalized estimates of covariance, regularized by linear shrinkage with strength $\gamma = 10\Delta g$, (F) Est with normalized estimates of covariance, regularized by nonlinear shrinkage, with strength $\eta = 1 \times 10^{-5}$ and using the optimal shrinker derived for the loss function of Frobenius norm of $\hat{R}^{-1}R - I$.

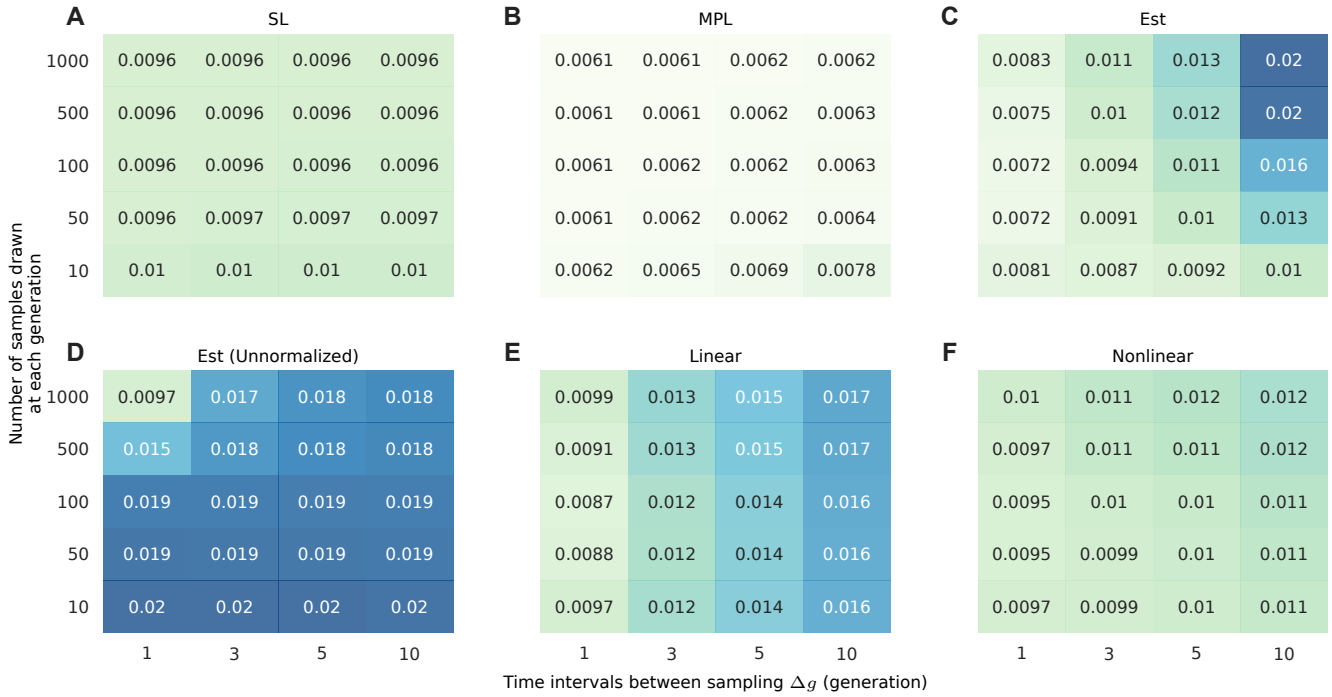

Figure S10 MAE of inferred selection coefficients for all tested methods on simulated data with recombination. The performance of various methods on selection inference are evaluated and averaged over 200 simulations. The recombination rate is set as 10^{-5} per locus per generation. Other parameters are the same as the example shown in Figure 1A. MAE of inferred selection coefficients are shown for different limited sampling effects using (A) SL method which ignores genetic linkage, (B) MPL method which uses true sample covariance, (C) Est method with normalized estimate of covariance, (D) Est method with unnormalized estimate of covariance, (E) Est method with normalized estimate of covariance, regularized by linear shrinkage of strength $\gamma = 10\Delta g$, (F) Est method with normalized estimate of covariance, regularized by nonlinear shrinkage, with strength $\eta = 1 \times 10^{-5}$ and using the optimal shrinker derived for the loss function of Frobenius norm of $\hat{R}^{-1}R - I$. Note that the mean absolute value of true selection coefficients is about 0.015. We find that when the sampling time interval is large ($\Delta g = 10$), the linear shrinkage regularization method has high MAE, which is because it infers selection coefficients with vanishing magnitudes.

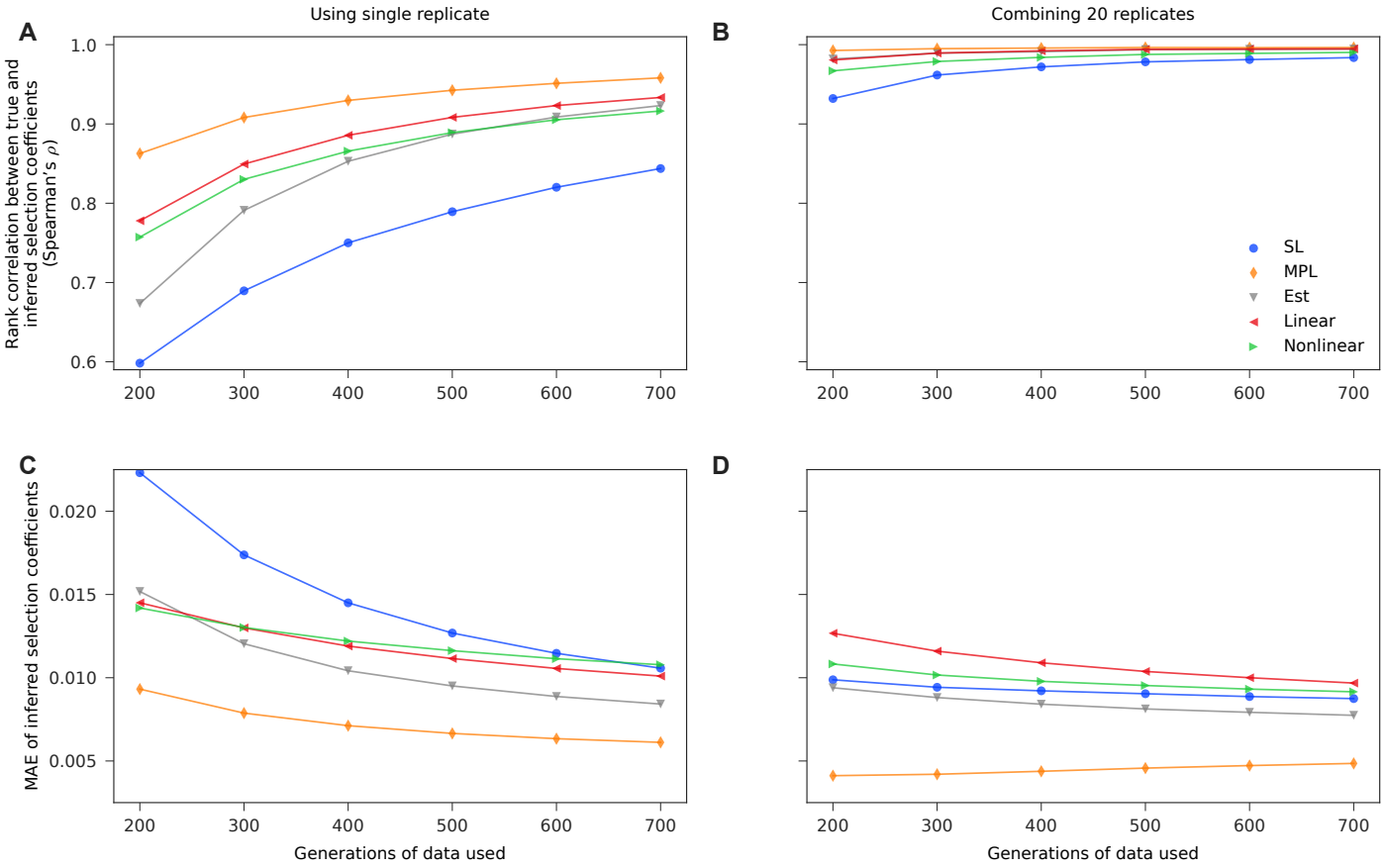

Figure S11 Performance can be greatly improved by combining data from multiple replicates with different founder haplotypes. Spearman's rank correlation coefficients between inferred and true selection coefficients, averaged over 200 simulations generated from 10 sets of selection coefficients and 20 sets of initial populations (each initial population has individuals randomly distributed over four randomly generated founder haplotypes), are shown (A) when using single replicate, and (B) when combining 20 replicates with different initial haplotype distributions. The performance on selection inference for all methods is greatly improved.

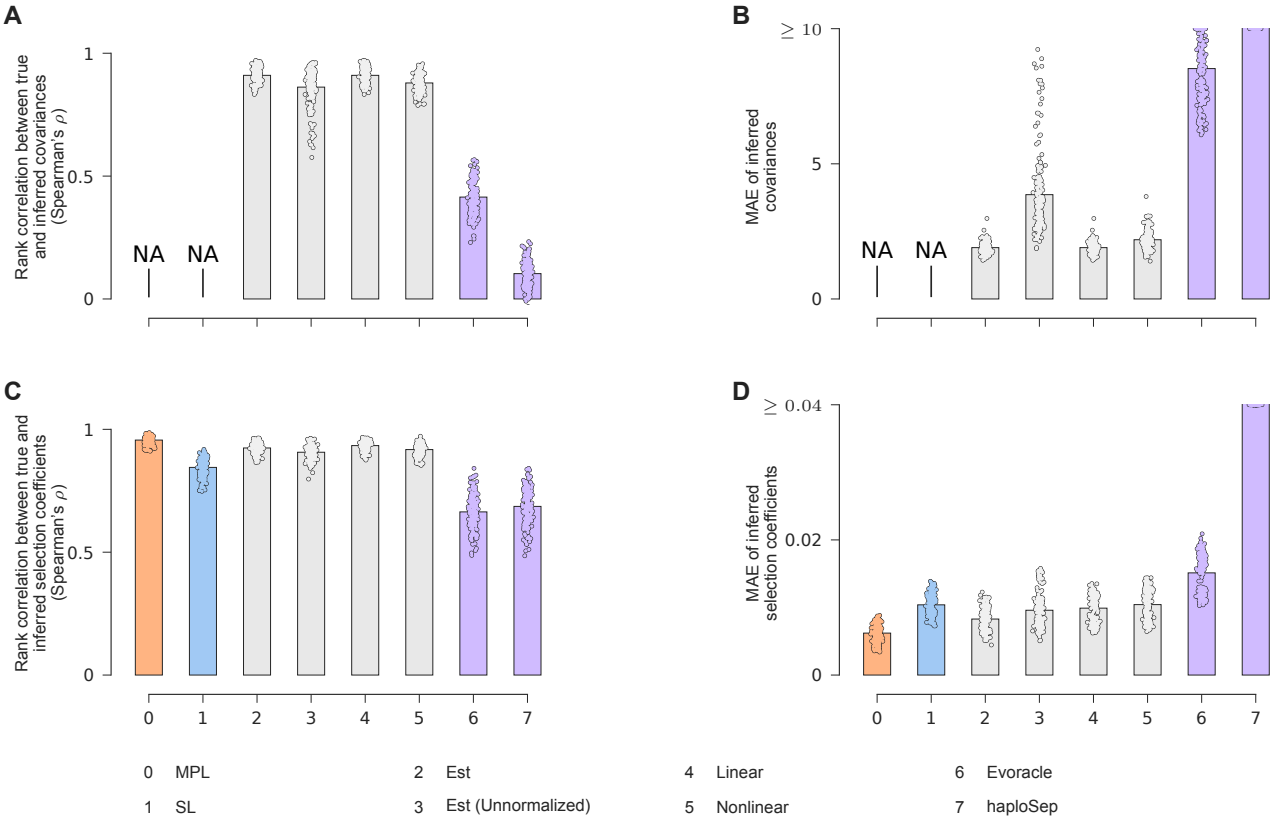

Figure S12 Comparative performance of haplotype reconstruction methods on simulated data with complex dynamics. Performance of various methods, including two haplotype reconstruction methods (*haploSep* and *Evoracle*), on selection inference are evaluated for 200 simulations with the same setup as the example shown in Figure 1A. Each dot corresponds to one of 200 simulations, and each bar denotes the mean value of all corresponding dots. The data used in this comparison is ample without finite sampling. We compare their performance on (A) Spearman's rank correlation coefficients and (B) MAE between inferred and true covariances (not including variances), (C) Spearman's rank correlation coefficients, and (D) MAE between inferred and true selection coefficients. Note that we use the true covariance information for MPL and only variances for SL, hence no covariance terms are estimated and correlation/error terms for MPL and SL in parts A and B are not applicable. With ample sampling, our (naive) method *Est* and its variations have similar performance, with all variations except for the unnormalized option clearly superior to SL. Methods that aim to reconstruct haplotypes (*haploSep* and *Evoracle*) do not show performance improvements over SL for these simulations.