Inferring epistasis from genetic time-series data

Muhammad Saqib Sohail[1], Raymond H.Y. Louie[2], Zhenchen Hong[3], John P. Barton[3,4,*] and Matthew R. McKay[1,5,6,7,*]

[1]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]The Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia
[3]Department of Physics and Astronomy, University of California, Riverside, CA, USA
[4]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[5]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China
[6]Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Victoria, Australia
[7]Department of Microbiology and Immunology, University of Melbourne, at The Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia

*Corresponding authors: matthew.mckay@unimelb.edu.au and jpbarton@pitt.edu.

Abstract

Epistasis refers to fitness or functional effects of mutations that depend on the sequence background in which these mutations arise. Epistasis is prevalent in nature, including populations of viruses, bacteria, and cancers, and can contribute to the evolution of drug resistance and immune escape. However, it is difficult to directly estimate epistatic effects from sampled observations of a population. At present, there are very few methods that can disentangle the effects of selection (including epistasis), mutation, recombination, genetic drift, and genetic linkage in evolving populations. Here we develop a method to infer epistasis, along with the fitness effects of individual mutations, from observed evolutionary histories. Simulations show that we can accurately infer pairwise epistatic interactions provided that there is sufficient genetic diversity in the data. Our method also allows us to identify which fitness parameters can be reliably inferred from a particular data set and which ones are unidentifiable. Our approach therefore allows for the inference of more complex models of selection from time series genetic data, while also quantifying uncertainty in the inferred parameters.

Keywords: Bayesian inference; selection; epistasis; linkage; path integral; diffusion; time-series data; longitudinal data

Epistasis refers to fitness effect of mutant alleles that differ from the sum of the fitness effects of each individual mutant (Carlborg and Haley 2004; Phillips 2008; Lehner 2011; de Visser *et al.* 2011). Epistasis therefore causes the fitness effect of a mutation to depend on the genetic background on which it arises. Theoretical and experimental studies have shown that epistasis can play a role in speciation (Gavrilets 2004; Wade 2002) and adaptation (Chou *et al.* 2011; Hansen 2013), and that it is intertwined with the evolutionary advantages of recombination (Kouyos *et al.* 2007; de Visser and Elena 2007). Epistasis is not uncommon in nature, and signatures of strong epistasis have been observed in laboratory evolution and site-directed mutagenesis experiments (Bershtein *et al.* 2006; Salverda *et al.* 2011; Khan *et al.* 2011; Gong *et al.* 2013).

Epistasis makes fitness landscapes more complex, shaping evolution (Phillips 2008; de Visser and Krug 2014). For example, epistasis may make certain mutational pathways more difficult to traverse while others become more readily accessible, depending on the sequence background (Weinreich *et al.* 2005, 2006; Phillips 2008; Salverda *et al.* 2011; de Visser and Krug 2014; Pedruzzi *et al.* 2018). A better understanding of epistasis could therefore help to characterize the evolutionary dynamics of novel viral strains capable of evading immune responses (Illingworth *et al.* 2014), pathogens that develop drug resistance (Hughes and Andersson 2015; Zhang *et al.* 2020) and tumor growth in cancers (Yates and Campbell 2012; Wang *et al.* 2014), as well as the adaptation of populations under lab settings (Domínguez-García *et al.* 2019).

Advances in sequencing technologies over the past decades have made it possible to obtain detailed, time-resolved population-level sequence data, enabling the study of evolving populations in fine detail. Examples of such data include those obtained from evolving populations in vitro (Barrick *et al.* 2009), ones sampled from naturally-infected hosts (Murcia *et al.* 2012; Zanini *et al.* 2015; Illingworth 2015; Xue *et al.* 2017), and time-resolved global influenza evolutionary records (Bao *et al.* 2008). These evolving populations contain multiple polymorphic loci, making the epistasis between

mutant alleles a potential factor in their evolution.

A complicating factor in inferring epistasis from such time-series data is the presence of linkage effects. Genetic linkage can arise by chance as a consequence of shared inheritance, or for functional reasons due to epistatic interactions between linked loci. Linkage can be especially strong when recombination is low, selection is strong, and novel mutations frequently appear and compete in a population (Desai and Fisher 2007; Neher and Shraiman 2009; Sniegowski and Gerrish 2010). The ability to distinguish the effects of epistasis from linkage due to chance is therefore important for the reliable inference of fitness from genetic time-series data.

The large majority of existing methods for inferring the fitness effects of mutations from genetic data ignore epistasis in their modeling. Hence they do not estimate epistasis, nor do they account for epistatic effects when estimating the fitness advantage of an allele. Most existing methods are based on single-locus models which assume independent evolution of loci (Bollback *et al.* 2008; Malaspinas *et al.* 2012; Mathieson and McVean 2013; Feder *et al.* 2014; Lacerda and Seoighe 2014; Steinrücken *et al.* 2014; Foll *et al.* 2015; Topa *et al.* 2015; Ferrer-Admetlla *et al.* 2016; Schraiber *et al.* 2016; Gompert 2016; Iranmehr *et al.* 2017; Taus *et al.* 2017; Zinger *et al.* 2019), thus they are unable to directly account for genetic linkage or epistasis. A few methods (Illingworth and Mustonen 2011; Terhorst *et al.* 2015; Sohail *et al.* 2021) have been developed that consider the joint evolution of multiple loci, but these assume additive fitness models. Hence, while they account for genetic linkage, they do not consider epistasis. A notable exception are the methods that use an extension of the multi-locus approach of Illingworth and Mustonen (2011) to account for epistatic interactions (Illingworth *et al.* 2014; Illingworth 2015). These fit a deterministic evolutionary model based on observed genotype frequencies, and while presenting an important advance, they require the use of computationally intensive numerical optimization methods.

New Approaches

Here we present a novel method that provides a closed-form, analytical solution for estimates of selection coefficients and pairwise epistatic interactions from genetic time-series data. Due to its analytical form, our approach is straightforward to implement and computationally efficient for moderate numbers of loci. Our method is based on an extension of the marginal path likelihood (MPL) framework (Sohail *et al.* 2021) to account for epistasis. We use a path integral method derived from statistical physics (Risken 1989) to efficiently represent the likelihood of an observed trajectory of single and double mutant allele frequencies. We then apply Bayesian theory to estimate the fitness parameters that best explain an observed evolutionary trajectory.

We model a population evolving under the Wright-Fisher (WF) model with mutation, selection, and recombination. First, we define $\mathbf{x}(t)$ as the vector of single and double mutant allele frequencies observed at generation t. For a system with L alleles labeled by $i = 1, 2, \ldots, L$, the first L entries of $\mathbf{x}(t)$ represent mutant allele frequencies $x_i(t)$, and entries from $L+1$ to $R = L(L+1)/2$ represent the frequencies of individuals in the populations with mutant i and j alleles, denoted $x_{ij}(t)$. Under WF dynamics the probability of observing a trajectory or 'path' $(\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_K))$ conditioned on $\mathbf{x}(t_0)$ is given by

$$P\left((\mathbf{x}(t_k))_{k=1}^{K} | \mathbf{x}(t_0)\right) = \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)\right). \tag{1}$$

We approximate the probability in (1) with a path integral. The first step of this approach is to approximate the WF process by a diffusion process (Kimura 1964; Ewens 2012; Tataru *et al.* 2015; He *et al.* 2017; Tataru *et al.* 2017). Under this approximation, the transition probabilities that appear on the right-hand side of (1) can be approximated by the transition probability density, ϕ, of a diffusion process (Durrett 2008), multiplied by a constant scaling term. In principle, $P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)\right)$ can be approximated using numerical integration techniques to solve the diffusion equations (Bollback *et al.* 2008; Malaspinas *et al.* 2012; Ferrer-Admetlla *et al.* 2016). Such approaches, however, are computationally intensive and lead to expressions that are difficult to treat analytically, even at the single locus level. Instead, the path integral approach we take allows for efficient computation of (1) by discretizing the transition probability density for small time

steps.

Taking a Gaussian prior for the selection coefficients and epistasis parameters and applying the maximum a posteriori criterion, we obtain an analytical expression for the estimates of selection coefficients and epistasis terms (collected into a vector $\hat{\mathbf{s}}$) given the observed allele frequency trajectories (see Materials and Methods for details):

$$\hat{\mathbf{s}} = [\mathrm{C}_{\mathrm{int}} + \gamma I]^{-1} \times [\Delta \mathbf{x} - \mu\, \mathrm{v}_{\mathrm{int}} - r\, \eta_{\mathrm{int}}] \ . \tag{2}$$

Here $\mathrm{C}_{\mathrm{int}}$ is the covariance matrix of single and double mutant allele frequencies integrated over time, γ is a regularization parameter, and I is the identity matrix. The net change in single and double mutant allele frequencies over the trajectory is denoted by $\Delta \mathbf{x}$. Entries of the $\mathrm{v}_{\mathrm{int}}$ and η_{int} vectors, which describe the flux in mutant allele frequencies due to mutation and recombination, respectively, are given by

$$\mathrm{v}_e(\mathbf{x}(t_k)) = \begin{cases} 1 - 2x_i(t_k) & 1 \le e \le L \\ x_i(t_k) + x_j(t_k) - 4x_{ij}(t_k) & L < e \le R, \end{cases} \tag{3}$$

and

$$\eta_e(\mathbf{x}(t_k)) = \begin{cases} 0 & 1 \le e \le L \\ (i-j)\left(x_{ij}(t_k) - x_i(t_k)x_j(t_k)\right) & L < e \le R, \end{cases} \tag{4}$$

integrated over time. Here $e \mapsto i$ for $e \le L$ and $e \mapsto (i,j)$ for $L < e \le R$. We use μ for the mutation probability per locus per generation (assuming for simplicity that the probability to mutate from mutant to wild-type (WT) is the same as that to mutate from WT to mutant, though we relax this assumption in Supplementary Material), and r is the recombination probability per locus per generation.

Intuitively, (2) shows that excess changes in allele frequencies/correlations that cannot be explained by mutation or recombination are evidence for selection/epistasis. The first term in the 'numerator', $\Delta \mathbf{x}$, gives the observed change in the single and double mutant allele frequencies between the final and the initial time points. This raw difference is then adjusted by the expected cumulative mutational flows of single and double mutant frequencies over the trajectory, $\mu\, \mathrm{v}_{\mathrm{int}}$. Finally, the change in double mutant frequencies is adjusted to account for their expected shift due to recombination, $r\, \eta_{\mathrm{int}}$. Linkage effects are incorporated through the inverse of the regularized integrated covariance matrix. This matrix, which captures the magnitude of allele frequency changes expected due to chance alone, also sets the scale of the inferred selection coefficients and epistatic interactions. Frequency changes that are significantly larger than chance expectations are therefore evidence for strong selection. As an example, distant loci that remain strongly linked over long times despite frequent recombination (see (4)) would suggest a strong, positive epistatic interaction between these loci.

In the following, we use simulations to demonstrate that our approach accurately infers fitness parameters using data from populations evolving under selection, mutation, recombination, epistasis, and nontrivial genetic linkage. We also show conditions under which reliable inference of selection and epistasis is possible. In cases where low genetic diversity precludes the accurate inference of some fitness parameters, MPL is still able to infer their collective fitness contributions.

Results

Accurate estimation of epistasis and selection coefficients

We first analyzed the performance of MPL on a two-locus bi-allelic system. We ran extensive simulations varying the selection strength, the composition of the initial population and different types of epistasis. The types of epistatic interactions we considered include positive epistasis, where the double mutant has a fitness higher than the sum of the individual fitness effects of each mutant allele; negative epistasis, where the fitness of the double mutant is lower than the sum of the individual fitness effects of each mutant allele; and sign epistasis, where the direction and the magnitude

of the fitness effect of epistasis is opposite to and larger than the sum of the individual fitness effect of the two mutant alleles.

We found that MPL is typically able to accurately infer underlying fitness parameters. In the simulation shown in Figure 1, the initial population consisted of only the WT genotype. MPL estimates (21) of selection coefficients were accurate in each simulated scenario. Estimates of the epistasis terms were better in scenarios where both the selection coefficients were beneficial (Figure 1A) compared with the scenarios where both were deleterious (Figure 1B), regardless of the type of epistasis. This is because double mutants tend to appear very rarely in cases where both single mutants were less fit than WT, as the single mutants are rapidly purged from the population. In such cases (Figure 1B), the double mutant genotype never exceeded 4% of the population in our simulations. A similar situation occurred in the positive sign epistasis scenario (Figure 1B *bottom* panel). Thus, genetic diversity constrains the accuracy of the epistasis estimates, which is also reflected in the uncertainty of the inferred parameters (Figure 2).

We further tested the ability of MPL to infer selection coefficients and epistasis terms under varying degrees of genetic diversity, on a two-locus bi-allelic system, by changing the composition of genotypes in the initial population. We found that the inference of these fitness parameters was quite accurate when all four genotypes appeared at high frequencies in the population, even when both single mutations were deleterious (*top left* panel of Figure 3). When some of the mutant genotypes are never present in the population, however, not all fitness parameters can be accurately inferred (e.g., the *top right* panel of Figure 3). These results show that genetic diversity in data limits which fitness parameters can be inferred.

Identifiability of fitness parameters

Based on patterns of genetic diversity in the time-series data, the estimated fitness parameters can be naturally classified into one of three categories: accessible, partially accessible, or inaccessible, by examining the structure of the integrated covariance matrix used as part of the MPL estimator. Accessible fitness parameters are ones that could be independently estimated in principle (vice-versa for the inaccessible parameters), whereas partially accessible fitness parameters can only be estimated as part of a sum. Specifically, this is done by reducing the integrated covariance matrix to its reduced row-echelon form and checking the linear dependencies of its rows. The fitness parameters whose corresponding rows of the integrated covariance matrix are linearly independent are denoted as accessible. These can be estimated meaningfully. The fitness parameters corresponding to linearly dependent rows are classed as partially accessible. While these parameters cannot be meaningfully estimated individually, we can still estimate their sum. Finally, fitness parameters corresponding to the rows of the integrated covariance matrix with all zero entries are referred to as inaccessible as there is insufficient data to provide a meaningful estimate, either individually or as part of a sum, of these parameters. As an example, we can consider a population with two loci labeled 1 and 2 where only two genotypes are ever observed, one with both WT and one with both mutant alleles. Then the individual coefficients s_1, s_2, s_{12} cannot be independently inferred, but their sum $s_1 + s_2 + s_{12}$ can be estimated.

When the population consisted of all but one of the single mutant genotypes (*right* and *left* panels of second row of Figure 3), one of the selection coefficients was accessible (and thus accurately inferred) while the remaining two fitness parameters were partially accessible. In scenarios where the double mutant was absent from the population (*left* panel of third row of Figure 3), the selection coefficients were accessible, however there was no data to make any meaningful inference of the epistasis term. When the data contained only the WT and the double mutant genotypes (*right* panel of third row of Figure 3), all three fitness parameters were partially accessible as their inferred sum was accurate even though neither the selection coefficients nor the epistasis terms could be accurately inferred individually. Finally, in scenarios where only one of the two loci was polymorphic, and thus accessible, it was not possible to make a meaningful inference about the selection coefficient at the non-polymorphic locus or the pairwise epistasis term (*bottom left* and *bottom right* panels of Figure 3).

Additional tests demonstrated that the performance of MPL was consistent across a variety of landscapes, comprising of beneficial and/or deleterious selection coefficients and various forms of epistasis like positive, negative, positive sign and negative sign epistasis (Supplementary Figure S1).

Analysis of a more complex five-locus epistatic fitness landscape

We ran further simulations on a more complex five-locus system to test the effects of genetic diversity on the inference of MPL. Genetic diversity in these simulations was controlled in two ways: (i) by specifying the number of unique genotypes in the initial population (Figure 4), and (ii) by combining data from multiple independent low genetic diversity replicates (Figure 5).

As expected, there was an increase in the fraction of accessible fitness parameters (Figures 4C, 4E, 5C and 5E) and better inference of the fitness landscape (Figures 4B and 5B) as the level of genetic diversity increased. Our results show that for a given level of genetic diversity, the fraction of accessible selection coefficients is higher than the fraction of accessible epistasis terms (Supplementary Figure S2), i.e., higher genetic diversity is required for inference of epistasis than that required for inference of selection coefficients alone. This is because, for an epistasis term to be accessible, both corresponding selection coefficients must also be accessible.

We used area under the receiver operating characteristic curve (AUROC) as a performance metric to quantify the ability of MPL to classify beneficial and deleterious fitness parameters. When computed over all selection coefficients (*left* panels of Figures 4D and 5D) and all pairwise epistasis terms (*left* panels of Figures 4F and 5F), the results showed higher detection performance with increasing genetic diversity. The poor performance at low genetic diversity was due to the large number of parameters that were either inaccessible or partially accessible, and thus cannot be meaningfully inferred due to lack of data. Computing the AUROC metric but restricted to *only* those selection coefficients classed as accessible revealed that the MPL estimator was able to correctly classify nearly all of such selection coefficients, under all scenarios considered (*right* panels of Figures 4D and 5D). The classification of accessible epistasis terms also showed good performance at moderate and high genetic diversity (*right* panels of Figures 4F and 5F). Although none of the epistasis terms were accessible at low genetic diversity, combining multiple replicates using (22) resulted in some epistasis terms becoming accessible (Figure 5E).

Similar results were obtained across a range of fitness landscapes differing in the degree of sparsity in their pairwise epistasis terms (Supplementary Figure S3). These tests demonstrate that MPL has a very good ability to detect those fitness parameters for which there is sufficient data to enable inference and classification.

Robustness to sampling parameters

The accuracy of the estimator depends on how well the underlying population dynamics is sampled. This includes how often the population is sampled in time, the number of samples measured at each time point, and the number of generations used for inference. Here we test the robustness of the MPL method with respect to these sampling parameters. In general, one would expect performance to degrade as samples are taken further apart in time (increasing time sampling step Δt) for a fixed number of generations used for inference, T, or as the number of generations used for inference is reduced for a fixed time sampling step, as less of the trajectory dynamics are captured in both these sampling scenarios. Moreover, taking limited samples at each time point would reduce the accuracy of the allele frequency estimates which may also compromise the accuracy of the MPL estimate.

To test the robustness of the estimator, we ran extensive simulations under various sampling conditions. These simulations demonstrated that MPL can accurately detect both accessible selection coefficients and accessible epistasis terms for a range of sampling parameters (Figure 6 and Supplementary Figure S4, respectively). MPL performed quite well even when the observed data consisted of a low number of samples, n_s, with only a few time samples (large time sampling step, Δt). For example, at $n_s = 50$ (from a population of $N = 1000$), the AUROC of detecting accessible beneficial selection coefficients (*top left* panel of Figure 6) varied from 0.94 to 0.9 when the time sampling step was increased from $\Delta t = 5$ to $\Delta t = 50$ (corresponding to 21 and 3 time samples respectively over $T = 100$ generations used for inference). Similarly, MPL performed well even when only a few time points that captured the evolutionary dynamics were used for inference. For example, the AUROC of detecting accessible beneficial selection coefficients (*bottom left* panel of Figure 6) varied from 0.91 to 0.95 when the number of generations used for inference was increased from $T = 30$ to $T = 140$ (corresponding to 7 and 29 time points respectively with $\Delta t = 5$). These results show that MPL estimator is

robust to reasonable limitations in sampling depth and frequency.

Comparison with other models

A model that does not account for epistasis: For fitness landscapes with epistasis, any inference model that does not explicitly account for epistasis will ascribe the effect of epistasis terms to individual selection coefficients, thereby over- or under-estimating them. To test this, we ran simulations to compare the performance of the MPL method, which accounts for both linkage and epistasis, with the one we proposed previously, which accounts only for linkage and considers a first order fitness model with no epistasis (Sohail *et al.* 2021). Here we term this variant as 'MPL (without epistasis)'. Simulations on simple two-locus systems with different fitness parameter settings showed that when the epistasis term was accessible (based on genetic diversity in the data), MPL estimates were more accurate than MPL (without epistasis) for scenarios where the fitness landscape had epistasis, particularly when any pair of fitness parameters had opposite signs (Supplementary Figure S5).

Next, we tested the classification performance of the two methods in a five-locus system. Initially, we chose relatively simple structures for the fitness landscapes; i.e., no epistasis links between loci with mutant allele selection coefficients of opposite signs, and all epistasis links having similar strengths and the same sign (*top* row panels of Figure 7A). We also varied the strength of the epistasis terms from strong (both selection coefficients and epistasis terms drawn from the same distribution) to weak (epistasis terms an order of magnitude weaker than the selection coefficients). Our results showed that when the genetic diversity in the data was low, the relative classification performance of the two methods was dependent on the underlying fitness landscape (*middle* row panels of Figure 7A), with MPL generally performing better than or as well as MPL (without epistasis). However, combining multiple low-diversity independent replicates using (22) resulted in MPL performing significantly better than MPL (without epistasis) in all scenarios tested, including weak, strong, positive, and negative epistasis (*bottom* row panels of Figure 7A).

We also compared the performance of the two methods on more complicated fitness landscapes with both positive and negative epistasis terms, and on fitness landscapes of varying density of non-zero epistasis terms (i.e., different epistasis sparsity levels). We generated several such fitness landscapes, with similar magnitudes of selection coefficients and pairwise epistasis terms as the fitness landscape in Figure 4A, but differing in the density of epistasis terms, ranging from a purely additive landscape (no epistasis terms) to a highly epistatic landscape (with all pairwise epistasis terms being non-zero). We grouped these landscapes on the basis of number of non-zero pairwise epistasis terms. Our results demonstrated that in high genetic diversity scenarios, MPL had better performance than MPL (without epistasis) regardless of the density of epistasis terms in the fitness landscape (Figure 7B). In scenarios where the genetic diversity in the data was low, the two methods had similar performance when the underlying fitness landscape was additive or had low density of pairwise epistasis terms, while MPL had superior performance when the fitness landscape was highly epistatic (Figure 7B). Interestingly, our simulations showed that even in scenarios where none of the epistasis terms were accessible (Supplementary Figure S2), MPL still showed a marked improvement in performance over MPL (without epistasis) in classifying accessible selection coefficients (Supplementary Figure S6). Overall, our approach enabled us to disentangle the confounding effects of linkage and epistasis from data, resulting in more accurate inference of fitness parameters.

A model that accounts for epistasis: We further compared our method with that of Illingworth *et al.* (2014) (see Supplementary Material), a state-of-the-art method that also accounts for epistasis, but unlike MPL, is based a deterministic evolutionary model, while requiring the use of numerical optimization algorithms. Our simulations demonstrated that MPL showed better classification performance and was considerably faster (Supplementary Figure S7). The performance improvement of MPL was particularly evident for scenarios with deleterious selection coefficients and with negative sign epistasis (Supplementary Figure S8).

Scaling to larger systems

For a bi-allelic system, the number of selection coefficients grow linearly with the number of loci in the system, L, while the number of epistasis terms grow quadratically as $L(L-1)/2$. To test the effect this increase in the number of fitness parameters has on their accessibility and the performance of MPL, we simulated systems of different sizes (10-, 20-, and 30-locus systems). Our results showed that the fraction of accessible parameters and the AUROC classification performance tended to reduce as the number of loci increases. However, by increasing the genetic diversity through multiple replicate combining, all fitness parameters eventually became accessible for all three system sizes (*top* row of Supplementary Figure S9). Simulations also showed that the accessibility of fitness parameters was robust to the level of sparsity in the underlying fitness landscape (Supplementary Figure S10). Consistent with earlier results, increased genetic diversity for a given system size resulted in improved classification performance (Supplementary Figure S9).

Computational complexity

The closed-form nature of the MPL estimate (21) makes it potentially computationally efficient. The two most computationally-intensive steps in the algorithm are (i) calculating the triple and quadruple mutant allele frequencies from the data, and (ii) inversion of the regularized integrated covariance matrix. The number of triple and quadruple mutant frequencies required for computing the inverse term in (21) increases as L^4, where L is the number of loci. However, this number can be reduced based on the genetic diversity in the data. For instance, for any locus-pair (i, j) whose double mutant frequency is zero, it follows that any three tuple (i, j, k) involving the same pair will have a triple mutant frequency of zero and hence its calculation can be avoided. Similarly the number of quadruple mutant frequencies that need to be computed can also be reduced. The computations required for computing the inverse term can also be reduced by considering only the polymorphic loci $L_p < L$, instead of the whole sequence, leading to $R_p = L_p(L_p+1)/2$ parameters to be estimated. The inverse would then require $\mathcal{O}(R_p^3)$ computations, with $R_p \ll R$ in practice for realistic data sets.

Discussion

Epistasis is a pervasive phenomenon that can strongly shape the evolution. Genetic time-series data provides an opportunity to detect and estimate epistatic contributions to fitness. However, developing methods that can efficiently yield accurate inferences has remained a challenge. Here we proposed a method to address this challenge. Our approach is a physics-based approach that builds upon a framework that we recently introduced for non-epistatic models (Sohail *et al.* 2021). Through simulations, we demonstrated that our method can accurately infer both pairwise epistasis effects and individual selection coefficients, provided sufficient variation exists in the data. Moreover, the method systematically reveals necessary conditions on genetic variation in the data in order for accurate inferences to be possible, and for the separate contributions of epistasis and allele selection coefficients to be inferrable.

MPL uses a path integral to approximate the likelihood of a set evolutionary parameters (including epistasis), given an observed time-series of allele frequencies and their correlations. This framework can also be adapted for different evolutionary scenarios. In recent work, it was applied together with epidemiological models to infer the transmission effects of mutations from genomic surveillance data, and to study the evolution of SARS-CoV-2 (Lee *et al.* unpublished data).

The data input to MPL, under a fitness model with pairwise epistasis terms, consists of the single, double, triple and quadruple mutant allele frequencies. While these are readily available from long-read sequencing data, the double and higher mutant allele frequencies cannot be computed extensively for short-read data. More work is required to develop methods that can accurately detect or infer selection and epistasis for such data sets. However, the trend toward longer read lengths in third-generation sequencing technologies (Pollard *et al.* 2018) suggests that higher order mutant frequencies will be more readily available in future data sets. While fitness models with higher-order epistasis involving more than two mutant alleles are also possible (Weinreich *et al.* 2013), here we restricted our analysis to a fitness model with pairwise epistasis terms. In principle, the MPL framework can be extended to account for higher-order epistasis

terms by explicitly modeling the evolution of higher-order mutant allele frequencies. However, the contribution of epistasis terms to fitness typically decline with order (Weinreich *et al.* 2018) and, at least in some scenarios, the gain achieved by modeling higher-order epistasis beyond pairwise terms appear to be minimal (Lozovsky *et al.* 2021).

MPL, like all inference methods, requires sufficient diversity to enable parameter inference. For a fitness model with pairwise epistasis terms, the number of model parameters to be inferred increases quadratically with the sequence length. As such, data with insufficient variation may lead to a situation where most of the model parameters are partially accessible or inaccessible (Supplementary Figure S2). This is not intrinsically a limitation of our specific method, but rather of a lack of exploratory power in the data. However, in scenarios where multiple of such low-diversity independent replicates are available, MPL offers a solution to overcome this limitation by providing a systematic way to combine low-diversity replicates.

The current approach infers a fitness landscape with epistasis terms between every pair of mutant alleles, in contrast to an additive fitness landscape inferred in Sohail *et al.* (2021). One can also consider selecting the most likely fitness model, given the data, from a reduced set of models with different densities of epistasis terms using a model selection approach. However, it may only be feasible to pursue selection approaches for moderate sized systems due to the exponential increase in the number of possible models with increasing system size. An alternative approach can be to apply a sparsity constraint on the epistasis terms. Future work on this problem can leverage sparsity inducing techniques such as the least absolute shrinkage and selection operator (LASSO) regression family of methods (Tibshirani 1996; Yuan and Lin 2006), to come up with a computationally efficient algorithm suitable for systems with hundreds or thousands of segregating mutations.

Materials and Methods

Model

We consider a population of N individuals evolving under a WF model with selection, mutation and recombination. Each individual is represented by a sequence of length L. The loci are assumed to be bi-allelic where each locus is either 0 (wild-type (WT)) or 1 (mutant), thus resulting in $M = 2^L$ genotypes. We consider a fitness model that accounts for epistasis arising due to pairwise interactions between alleles at different loci. The Wrightian fitness f_a of the ath genotype can then be written as

$$f_a = 1 + \sum_{i=1}^{L} s_i g_i^a + \sum_{i=1}^{L} \sum_{j=i+1}^{L} s_{ij} g_i^a g_j^a, \tag{5}$$

where s_i and s_{ij} denote the time-invariant selection coefficients and pairwise epistasis terms respectively, and g_i^a represents the allele (either 0 or 1) at the ith locus of the ath genotype. The population is completely specified by the $M \times 1$ genotype frequency vector $\mathbf{z}(t) = (z_1(t), \cdots, z_M(t))$, where $z_a(t) = n_a(t)/N$ and $n_a(t)$ denotes the number of individuals in the population that belong to genotype a at generation t.

Under WF dynamics, the probability of observing genotype frequencies $\mathbf{z}(t+1)$ at generation $t+1$, given genotype frequencies of $\mathbf{z}(t)$ at generation t is

$$P\left(\mathbf{z}(t+1)\middle|\mathbf{z}(t)\right) = N! \prod_{a=1}^{M} \frac{\left(p_a(\mathbf{z}(t))\right)^{Nz_a(t+1)}}{(Nz_a(t+1))!} \tag{6}$$

with

$$p_a(\mathbf{z}(t)) = \frac{y_a(t) f_a + \sum_{b \neq a} \left(\mu_{ba} y_b(t) f_b - \mu_{ab} y_a(t) f_a\right)}{\sum_{b=1}^{M} y_b(t) f_b}. \tag{7}$$

Here μ_{ba} is the probability of genotype b mutating to genotype a, and $y_a(t)$ is the frequency of genotype a after

recombination

$$y_a(t) = (1-r)^{L-1} z_a(t) + \left(1 - (1-r)^{L-1}\right) \psi_a(\mathbf{z}(t)), \tag{8}$$

where r is the recombination probability per locus per generation and $\psi_a(\mathbf{z}(t))$ is the probability that a recombination of two individuals results in an individual of genotype a (see Supplementary Material for details).

We assume the genotype frequencies are observed at non-consecutive generations t_k, with $k \in \{0, 1, \ldots, K\}$. Then, the probability that the genotype frequency vector follows a particular evolutionary path $(\mathbf{z}(t_1), \mathbf{z}(t_2), \cdots, \mathbf{z}(t_K))$, conditioned on the initial state $\mathbf{z}(t_0)$, is

$$P\left((\mathbf{z}(t_k))_{k=1}^{K} | \mathbf{z}(t_0)\right) = \prod_{k=0}^{K-1} P\left(\mathbf{z}(t_{k+1}) | \mathbf{z}(t_k)\right). \tag{9}$$

This expression can be used to infer evolutionary parameters. However, the inference problem is difficult due to the intractability of the fractional form of (7). Following the approach used in Sohail *et al.* (2021), we simplify the inference problem using a path integral. This allows us to obtain closed-form estimates of selection coefficients and epistasis terms. Even though the WF dynamics is defined at the genotype level (9), here we develop its simplified allele-level version for transparency. We show later in this section that both the genotype and allele-level analyses lead to the same expression for the estimate of fitness parameters. For ease of exposition, we assume here that the probability of mutating from a WT to mutant allele is the same as that from mutant allele to WT, which we denote by μ. However, this assumption can be easily relaxed (see Supplementary Material for details where we derive the estimator with asymmetrical mutation probabilities).

Linear mapping between genotype and allele frequencies: The allele frequencies can be described by taking a linear combination of genotype frequencies. Specifically,

$$\begin{aligned} x_i(t) &= \sum_{a=1}^{M} g_i^a z_a(t), & x_{ij}(t) &= \sum_{a=1}^{M} g_i^a g_j^a z_a(t), \\ x_{ijk}(t) &= \sum_{a=1}^{M} g_i^a g_j^a g_k^a z_a(t), & x_{ijkl}(t) &= \sum_{a=1}^{M} g_i^a g_j^a g_k^a g_l^a z_a(t), \end{aligned} \tag{10}$$

where $x_i(t)$, $x_{ij}(t)$, $x_{ijk}(t)$ and $x_{ijkl}(t)$ are the single, double, triple, and quadruple mutant allele frequencies at locus i, locus-pair (i,j), locus-triplet (i,j,k) and locus-quartet (i,j,k,l) respectively at generation t.

We will explicitly model the evolution of the single and double mutant allele frequencies, which we represent by the single vector,

$$\mathbf{x}(t) = \left(x_1(t), \cdots, x_L(t), x_{12}(t), x_{13}(t), \cdots, x_{(L-1)L}(t)\right). \tag{11}$$

For notational convenience (to facilitate sequential indexing), we equivalently write

$$\mathbf{x}(t) = \left(\mathrm{x}_1(t), \cdots, \mathrm{x}_L(t), \mathrm{x}_{L+1}(t), \cdots, \mathrm{x}_R(t)\right), \tag{12}$$

where $R = L(L+1)/2$. Here, and in the following, we differentiate between non-italic and italic scalar notation. From (11) and (12), we have $\mathrm{x}_e(t) = x_i(t)$ for $e \leq L$, and $\mathrm{x}_e(t) = x_{ij}(t)$ for $L < e \leq R$. We will explicitly denote the index mapping as $e \mapsto i$ for the former case, and $e \mapsto (i,j)$ for the latter.

Path integral: We model the evolution of both the single and the double mutant allele frequencies. In the allele-level path integral, these are required to obtain estimates of the selection coefficients and the pairwise epistasis terms (see Supplementary Material for the genotype-level path integral formulation).

The probability of observing a path of allele frequencies $(\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_K))$ conditioned on $\mathbf{x}(t_0)$ is given by

$$P\left((\mathbf{x}(t_k))_{k=1}^{K} | \mathbf{x}(t_0)\right) = \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)\right). \tag{13}$$

We use a path integral to approximate this probability, as described in New Approaches. This gives the following closed-form approximation of the transition probability (see Supplementary Material for details)

$$P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)\right) \approx \phi(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)) \prod_{e=1}^{R} \mathrm{d}x_e(t_{k+1}), \tag{14}$$

where

$$\phi(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k)) = \left(\frac{N}{2\pi\Delta t_k}\right)^{R/2} \frac{\exp(-\frac{N}{2}\Theta(\mathbf{x}(t_{k+1}), \mathbf{x}(t_k)))}{\sqrt{\det C(\mathbf{x}(t_k))}}$$

with $\Delta t_k = t_{k+1} - t_k$ and

$$\begin{aligned}
&\Theta(\mathbf{x}(t_{k+1}), \mathbf{x}(t_k)) \\
&= \frac{1}{\Delta t_k} \sum_{e=1}^{R} \sum_{f=1}^{R} \left[\mathrm{x}_e(t_{k+1}) - \left(\mathrm{x}_e(t_k) + \mathrm{d}_e(\mathbf{x}(t_k))\Delta t_k\right)\right] \\
&\quad \times \left(\mathrm{C}^{-1}(\mathbf{x}(t_k))\right)_{ef} \left[\mathrm{x}_f(t_{k+1}) - \left(\mathrm{x}_f(t_k) + \mathrm{d}_f(\mathbf{x}(t_k))\Delta t_k\right)\right].
\end{aligned}$$

Here $\mathrm{d}_e(\mathbf{x}(t_k))$ describes the expected change in allele frequencies (either single mutant or pairwise, depending on e) from generation t_k to t_{k+1}, given by

$$\begin{aligned}
\mathrm{d}_e(\mathbf{x}(t_k)) &= \mathrm{x}_e(t_k)\left(1 - \mathrm{x}_e(t_k)\right)\mathrm{s}_e + \sum_{f \neq e} \mathrm{C}_{ef}(\mathbf{x}(t_k))\mathrm{s}_f \\
&\quad + \mu \mathrm{v}_e(\mathbf{x}(t_k)) + r\, \eta_e(\mathbf{x}(t_k)).
\end{aligned} \tag{15}$$

Above, $\mathbf{s} = \left(s_1, \cdots, s_L, s_{12}, s_{13}, \cdots, s_{(L-1)L}\right)$ is the vector of selection coefficients and pairwise epistasis terms, with terms having the corresponding non-italic representation s_e, defined analogously to (12). We have also defined

$$\mathrm{v}_e(\mathbf{x}(t_k)) = \begin{cases} 1 - 2x_i(t_k) & 1 \leq e \leq L \\ x_i(t_k) + x_j(t_k) - 4x_{ij}(t_k) & L < e \leq R, \end{cases}$$

and

$$\eta_e(\mathbf{x}(t_k)) = \begin{cases} 0 & 1 \leq e \leq L \\ (i-j)\left(x_{ij}(t_k) - x_i(t_k)x_j(t_k)\right) & L < e \leq R, \end{cases}$$

where $e \mapsto i$ for $e \leq L$ and $e \mapsto (i, j)$ for $L < e \leq R$. The first term in (15) represents the expected change in allele frequencies (either single mutant or pairwise, depending on e) due to e-th fitness parameter, the second term represents the change due to all but the e-th fitness parameter, the third term in (15) represents the contribution due to net mutational flow, while the fourth term represents the contributions to the expected change in allele frequencies due to recombination.

The matrix $C(\mathbf{x}(t_k))$ is a symmetric $R \times R$ matrix describing the covariances of the allele frequencies at generation t_k. This can be partitioned into four sub-matrices, each with an intuitive interpretation (details in Supplementary Material).

Briefly, for $e \leq L$ and $f \leq L$, with mapping $e \mapsto i$ and $f \mapsto j$, the elements

$$\mathrm{C}_{ef}(\mathbf{x}(t_k)) = x_{ij}(t_k) - x_i(t_k)x_j(t_k) \tag{16}$$

are the covariance between mutants at loci i and j; for $e \leq L$ and $L < f \leq R$, with mapping $e \mapsto i$ and $f \mapsto (j,k)$, the elements

$$\mathrm{C}_{ef}(\mathbf{x}(t_k)) = x_{ijk}(t_k) - x_i(t_k)x_{jk}(t_k) \tag{17}$$

are the covariance between mutant at locus i and double mutant at locus-pair (j,k); the elements of $C(\mathbf{x}(t_k))$ for $L < e \leq R$ and $f \leq L$ are the same as (17) due to the symmetric nature of the covariance matrix; while for $L < e \leq R$ and $L < f \leq R$, with mapping $e \mapsto (i,j)$ and $f \mapsto (k,l)$, the elements

$$\mathrm{C}_{ef}(\mathbf{x}(t_k)) = x_{ijkl}(t_k) - x_{ij}(t_k)x_{kl}(t_k) \tag{18}$$

are the covariance between the double mutants at locus-pair (i,j) and double mutants at locus-pair (k,l).

Substituting (14) in (13) gives an approximation for the probability of the single and pairwise mutant allele frequencies following the evolutionary path $\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_K)$, conditioned on $\mathbf{x}(t_0)$.

Marginal path likelihood (MPL) estimator with epistasis

The MPL parameter estimates are obtained by adopting a Bayesian approach. Specifically, we use the maximum a posteriori (MAP) criterion to find the most likely selection coefficients and epistasis terms given the measured single, double, triple and quadruple mutant frequencies at each sampling time point, along with knowledge of evolutionary parameters N, μ and r. For the purpose of developing an efficient inference approach, we assume that the observed frequencies are equal to the true frequencies in the population. The MPL estimate of the selection coefficients and epistasis terms can thus be obtained by solving

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \mathfrak{L}\left(\mathbf{s}; N, r, \mu, (\mathbf{x}(t_k))_{k=0}^{K}\right) P^{\text{prior}}(\mathbf{s}), \tag{19}$$

where $P^{\text{prior}}(\mathbf{s})$ is the assumed (conjugate) prior

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

with mean zero and variance $\sigma^2 > 0$, and the likelihood of the selection coefficients and epistasis terms, $\mathbf{s}$, given the observed data can be expressed as

$$\begin{aligned}\mathfrak{L}\left(\mathbf{s}; N, r, \mu, (\mathbf{x}(t_k))_{k=0}^{K}\right) &= P\left((\mathbf{x}(t_k))_{k=1}^{K} | \mathbf{x}(t_0), N, r, \mu, \mathbf{s}\right) \\ &= \prod_{k=0}^{K-1} P\left(\mathbf{x}(t_{k+1}) | \mathbf{x}(t_k), N, r, \mu, \mathbf{s}\right).\end{aligned} \tag{20}$$

While it is challenging to calculate the likelihood (20) exactly, the task is simplified by using the path integral approach outlined in the previous section with some modifications (see Supplementary Material for details) to account for time-samples drawn from non-unit time intervals, $\Delta t_k = t_{k+1} - t_k$. Following this approach, the MAP solution is evaluated

as

$$\hat{\mathrm{s}}_e = \sum_{f=1}^{R} \left[\sum_{k=0}^{K-1} \Delta t_k \mathrm{C}(\mathbf{x}(t_k)) + \gamma I \right]^{-1}_{ef} \times \left[\mathrm{x}_f(t_K) - \mathrm{x}_f(t_0) - \mu \sum_{k=0}^{K-1} \Delta t_k \mathrm{v}_f(\mathbf{x}(t_k)) - r \sum_{k=0}^{K-1} \Delta t_k \eta_f(\mathbf{x}(t_k)) \right], \tag{21}$$

for $e = 1, \ldots, R$. The inverse term consists of the covariance matrix of single and double mutant allele frequencies (a function of single, double, triple, and quadruple mutant allele frequencies) integrated over time, which we refer to as the *integrated covariance matrix*, plus a regularization term, where $\gamma = 1/N\sigma^2$ and I is the identity matrix.

The MPL estimator (21) has an intuitive interpretation. It computes the observed change in the single and double mutant allele frequencies between the final and the initial time points, adjusts it by accounting for the (inward and outward) mutational flows of single and double mutant frequencies over time, and then applies a correction to the double mutant frequencies to account for the effect of recombination. Finally, it accounts for linkage effects through the inverse of the (regularized) integrated covariance matrix.

As shown in (21), significant changes in mutant frequencies – that is, ones that are substantially larger than those expected due to finite population size alone – that cannot readily be explained by mutation, recombination, or the effects of background mutations provide evidence of selection or epistatic interactions. For example, mutant alleles that are separated by a long distance on the genome and which remain strongly linked despite recombination would be evidence of a positive epistatic interaction.

Combining multiple independent observations

The approach naturally lends itself to incorporating data from multiple replicates. These replicates may represent independent evolutionary paths with possibly distinct sampling parameters and starting conditions. Let $t_1^q, \ldots, t_{K_q}^q$ be the sampling times of the qth replicate and $x_i^q(t_k^q)$, $x_{ij}^q(t_k^q)$ be the single and double mutant allele frequencies at the ith locus and the (i,j)th locus-pair respectively at generation t_k^q. The observed trajectory of the single and double mutant allele frequencies of the qth replicate is thus denoted as $\mathbf{x}^q(t_k^q) = \left(x_1^q(t_k^q), \cdots, x_L^q(t_k^q), x_{12}^q(t_k^q), x_{13}^q(t_k^q), \cdots, x_{(L-1)L}^q(t_k^q) \right)$. The MPL estimate in this case is given as (see Supplementary Material for details)

$$\begin{aligned} \hat{\mathrm{s}}_e = \sum_{f=1}^{R} \left[\sum_{q=1}^{Q} \sum_{k=0}^{K_q-1} \Delta t_k^q \mathrm{C}(\mathbf{x}^q(t_k^q)) + \gamma I \right]^{-1}_{ef} \\ \times \sum_{q=1}^{Q} \left(\mathrm{x}_f^q(t_{K_q}^q) - \mathrm{x}_f^q(t_0^q) - \mu \sum_{k=0}^{K_q-1} \Delta t_k^q \mathrm{v}_f(\mathbf{x}^q(t_k^q)) \right. \\ \left. - r \sum_{k=0}^{K_q-1} \Delta t_k^q \eta_f(\mathbf{x}^q(t_k^q)) \right), \end{aligned} \tag{22}$$

where Q is the number of replicates being combined, $\Delta t_k^q = t_{k+1}^q - t_k^q$, $\gamma = 1/N\sigma^2$ as before, and $C(\mathbf{x}^q(t_k^q))$ is the covariance matrix of the mutant allele frequencies at generation t_k^q for the qth replicate.

Equivalence with the genotype estimate

The MPL estimate above was derived using a path integral for the mutant allele frequencies even though the WF evolutionary process is defined at the genotype level. One may ask if working at the level of allele frequencies leads to some loss in optimality? To check this, we derive the MPL estimate of the selection coefficients and epistasis terms directly from the genotype path-likelihood (see Supplementary Material for details), in contrast to the mutant allele path-likelihood (20) as was done above. For the observed path of the genotype frequencies $(\mathbf{z}(t_0), \mathbf{z}(t_1), \cdots, \mathbf{z}(t_K))$, the

MPL estimate is obtained by solving

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \mathfrak{L}\left(\mathbf{s}; N, \mu, (\mathbf{z}(t_k))_{k=0}^{K}\right) P^{\text{prior}}(\mathbf{s}), \tag{23}$$

where

$$\begin{aligned}\mathfrak{L}\left(\mathbf{s}; N, \mu, (\mathbf{z}(t_k))_{k=0}^{K}\right) &= P\left((\mathbf{z}(t_k))_{k=1}^{K} | \mathbf{z}(t_0), N, \mu, \mathbf{s}\right) \\ &= \prod_{k=0}^{K-1} P\left(\mathbf{z}(t_{k+1}) | \mathbf{z}(t_k), N, \mu, \mathbf{s}\right).\end{aligned} \tag{24}$$

We obtain the same expression for the MPL estimate (21) by solving (23) as shown in the Supplementary Material, i.e., there is no loss in optimality by working with the marginal allele frequencies. This implies that knowledge of up to fourth-order allele frequencies is sufficient to estimate selection coefficients and pairwise epistatic interactions. At least within the diffusion approximation, higher order frequencies do not carry additional information needed to estimate the fitness effects of individual mutations or pairwise epistasis.

Simulation setup

We generated evolutionary histories by running WF simulations, with selection, mutation and recombination, consisting of a population of N bi-allelic sequences evolving for T generations. We then randomly sampled n_s sequences every Δt generations, and used these sampled trajectories for inference of fitness parameters. The specific values of these parameters used in simulations are specified in the figure captions.

In simulations where it was required to control genetic diversity in a population, we specified the number and the frequencies of the unique genotypes in the initial population, and disallowed mutations and recombination. We refer to the all-zero genotype as the WT genotype. In simulations where the initial population contained more than one unique genotype, one of these was always the WT while the others were chosen from the set of remaining $2^L - 1$ possible genotypes at random, without replacement. All simulation results were computed over 1000 Monte Carlo runs. Unless stated otherwise, the initial frequency of each non-WT genotype was set to 5% of the population size, the sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, $T = 100$ generation were used for inference, and the regularization parameter, γ, was set to one.

Data availability

Simulation data and a MATLAB (version R2017a) implementation of MPL used for reproducing results in the paper are freely available at https://github.com/mssohail/epistasis-inference.

Acknowledgements

M.S.S. and M.R.M. were supported by the Hong Kong Research Grants Council (grant numbers 16204121 and 16201620). The work of Z.H. and J.P.B. reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM138233. M.R.M. is the recipient of an Australian Research Council Future Fellowship (project number FT200100928) funded by the Australian Government.

Conflicts of interest

The authors declare no conflict of interest.

Literature cited

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. Journal of Virology. 82:596–601.

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.

Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. 2006. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 444:929–932.

Bollback JP, York TL, Nielsen R. 2008. Estimation of $2N_e s$ from temporal allele frequency data. Genetics. 179:497–502.

Carlborg Ö, Haley CS. 2004. Epistasis: Too often neglected in complex trait studies? Nature Reviews Genetics. 5:618–625.

Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ. 2011. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science. 332:1190–1192.

de Visser JAG, Cooper TF, Elena SF. 2011. The causes of epistasis. Proceedings of the Royal Society B: Biological Sciences. 278:3617–3624.

de Visser JAG, Elena SF. 2007. The evolution of sex: Empirical insights into the roles of epistasis and drift. Nature Reviews Genetics. 8:139–149.

de Visser JAGM, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics. 15:480–490.

Desai MM, Fisher DS. 2007. Beneficial mutation–selection balance and the effect of linkage on positive selection. Genetics. 176:1759–1798.

Domínguez-García S, García C, Quesada H, Caballero A. 2019. Accelerated inbreeding depression suggests synergistic epistasis for deleterious mutations in Drosophila melanogaster. Heredity. 123:709–722.

Durrett R. 2008. *Probability Models for DNA Sequence Evolution*. Springer Science & Business Media.

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

Feder AF, Kryazhimskiy S, Plotkin JB. 2014. Identifying signatures of selection in genetic time series. Genetics. 196:509–522.

Ferrer-Admetlla A, Leuenberger C, Jensen JD, Wegmann D. 2016. An approximate Markov model for the Wright–Fisher diffusion and its application to time series data. Genetics. 203:831–846.

Foll M, Shim H, Jensen JD. 2015. WFABC: a Wright–Fisher ABC–based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular Ecology Resources. 15:87–98.

Gavrilets S. 2004. *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press.

Gompert Z. 2016. Bayesian inference of selection in a heterogeneous environment from genetic time-series data. Molecular Ecology. 25:121–134.

Gong LI, Suchard MA, Bloom JD. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. eLife. 2:e00631.

Hansen TF. 2013. Why epistasis is important for selection and adaptation. Evolution. 67:3501–3511.

He Z, Beaumont M, Yu F. 2017. Effects of the ordering of natural selection and population regulation mechanisms on Wright-Fisher models. G3: Genes, Genomes, Genetics. 7:2095–2106.

Hughes D, Andersson DI. 2015. Evolutionary consequences of drug resistance: Shared principles across diverse targets and organisms. Nature Reviews Genetics. 16:459–471.

Illingworth CJ. 2015. Fitness inference from short-read data: Within-host evolution of a reassortant H5N1 influenza virus. Molecular Biology and Evolution. 32:3012–3026.

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

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

Iranmehr A, Akbari A, Schlötterer C, Bafna V. 2017. CLEAR: Composition of likelihoods for evolve and resequence experiments. Genetics. 206:1011–1023.

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 332:1193–1196.

Kimura M. 1964. Diffusion models in population genetics. Journal of Applied Probability. 1:177–232.

Kouyos RD, Silander OK, Bonhoeffer S. 2007. Epistasis between deleterious mutations and the evolution of recombination. Trends in Ecology & Evolution. 22:308–315.

Lacerda M, Seoighe C. 2014. Population genetics inference for longitudinally-sampled mutants under strong selection. Genetics. 198:1237–1250.

Lee B, Sohail MS, Finney E, Ahmed SF, Quadeer AA, McKay MR, Barton JP. unpublished data. Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data, url: https://www.medrxiv.org/content/early/2022/01/01/2021. 12.31.21268591.full.pdf. medRxiv. .

Lehner B. 2011. Molecular mechanisms of epistasis within and between genes. Trends in Genetics. 27:323–331.

Lozovsky ER, Daniels RF, Heffernan GD, Jacobus DP, Hartl DL. 2021. Relevance of higher-order epistasis in drug resistance. Molecular Biology and Evolution. 38:142–151.

Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. 2012. Estimating allele age and selection coefficient from time-serial data. Genetics. 192:599–607.

Mathieson I, McVean G. 2013. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics. 193:973–984.

Murcia PR, Hughes J, Battista P, Lloyd L, Baillie GJ, Ramirez-Gonzalez RH, Ormond D, Oliver K, Elton D, Mumford JA *et al*. 2012. Evolution of an eurasian avian-like influenza virus in naive and vaccinated pigs. PLoS Pathogens. 8:e1002730.

Neher RA, Shraiman BI. 2009. Competition between recombination and epistasis can cause a transition from allele to genotype selection. Proceedings of the National Academy of Sciences. 106:6866–6871.

Pedruzzi G, Barlukova A, Rouzine IM. 2018. Evolutionary footprint of epistasis. PLoS Computational Biology. 14:e1006426.

Phillips PC. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics. 9:855–867.

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: Their purpose and place. Human Molecular Genetics. 27:R234–R241.

Risken H. 1989. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer-Verlag. second edition.

Salverda ML, Dellus E, Gorter FA, Debets AJ, Van Der Oost J, Hoekstra RF, Tawfik DS, de Visser JAG. 2011. Initial mutations direct alternative pathways of protein evolution. PLoS Genetics. 7:e1001321.

Schraiber JG, Evans SN, Slatkin M. 2016. Bayesian inference of natural selection from allele frequency time series. Genetics. 203:493–511.

Sniegowski PD, Gerrish PJ. 2010. Beneficial mutations and the dynamics of adaptation in asexual populations. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 365:1255–1263.

Sohail MS, Louie RH, McKay MR, Barton JP. 2021. MPL resolves genetic linkage in fitness inference from complex evolutionary histories. Nature Biotechnology. 39:472–479.

Steinrücken M, Bhaskar A, Song YS. 2014. A novel spectral method for inferring general diploid selection from time series genetic data. The Annals of Applied Statistics. 8:2203–2222.

Tataru P, Bataillon T, Hobolth A. 2015. Inference under a Wright-Fisher model using an accurate beta approximation. Genetics. 201:1133–1141.

Tataru P, Simonsen M, Bataillon T, Hobolth A. 2017. Statistical inference in the Wright-Fisher model using allele frequency data. Systematic Biology. 66:e30–e46.

Taus T, Futschik A, Schlötterer C. 2017. Quantifying selection with pool-seq time series data. Molecular Biology and Evolution. 34:3023–3034.

Terhorst J, Schlötterer C, Song YS. 2015. Multi-locus analysis of genomic time series data from experimental evolution. PLoS Genetics. 11:e1005069.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 58:267–288.

Topa H, Jónás Á, Kofler R, Kosiol C, Honkela A. 2015. Gaussian process test for high-throughput sequencing time series:

application to experimental evolution. Bioinformatics. 31:1762–1770.
Wade MJ. 2002. A gene's eye view of epistasis, selection and speciation. Journal of Evolutionary Biology. 15:337–346.
Wang X, Fu AQ, McNerney ME, White KP. 2014. Widespread genetic epistasis among cancer genes. Nature Communications. 5:1–10.
Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science. 312:111–114.
Weinreich DM, Lan Y, Jaffe J, Heckendorn RB. 2018. The influence of higher-order epistasis on biological fitness landscape topography. Journal of Statistical Physics. 172:208–225.
Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. 2013. Should evolutionary geneticists worry about higher-order epistasis? Current Opinion in Genetics & Development. 23:700–707.
Weinreich DM, Watson RA, Chao L. 2005. Perspective: Sign epistasis and genetic costraint on evolutionary trajectories. Evolution. 59:1165–1174.
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:e26875.
Yates LR, Campbell PJ. 2012. Evolution of the cancer genome. Nature Reviews Genetics. 13:795–806.
Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 68:49–67.
Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA. 2015. Population genomics of intrapatient HIV-1 evolution. eLife. 4:e11282.
Zhang Th, Dai L, Barton JP, Du Y, Tan Y, Pang W, Chakraborty AK, Lloyd-Smith JO, Sun R. 2020. Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. PLoS Genetics. 16:e1009009.
Zinger T, Gelbart M, Miller D, Pennings PS, Stern A. 2019. Inferring population genetics parameters of evolving viruses using time-series data. Virus Evolution. 5:vez011.

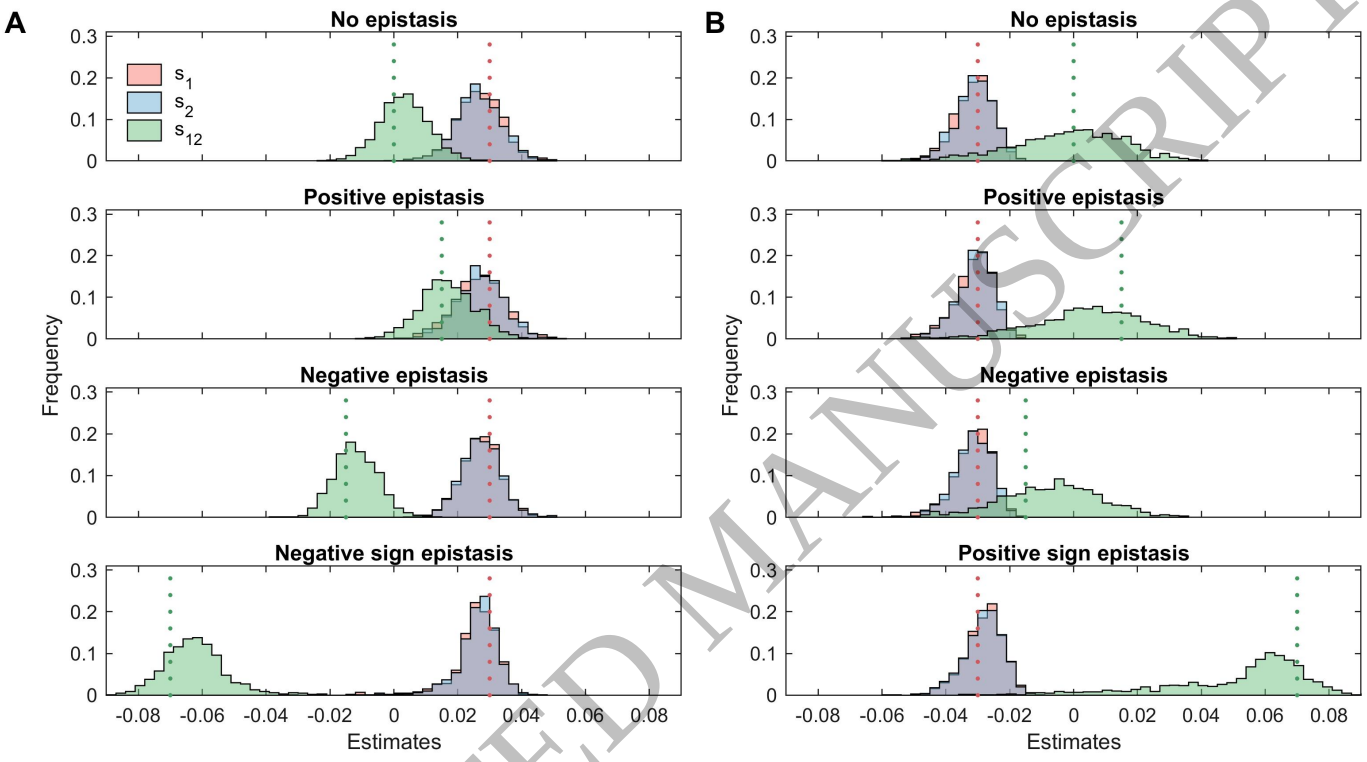

Figure 1 MPL can accurately infer selection coefficients and pairwise epistasis terms. Results were obtained for a two-locus system with selection, mutation, and recombination. (A) shows distribution of inferred selection coefficients and pairwise epistasis terms for various forms of epistasis when both selection coefficients are positive ($s_1 = s_2 = 0.03$), while (B) shows the same for the case when both selection coefficients are negative ($s_1 = s_2 = -0.03$). The pairwise epistasis term s_{12} was set to $\{0, 0.015, -0.015, 0.07, -0.07\}$ to simulate the scenarios of no epistasis, positive epistasis, negative epistasis, positive sign epistasis, and negative sign epistasis respectively. Other simulation parameters included per locus mutation probability $\mu = 10^{-3}$, per locus recombination probability $r = 10^{-3}$, and population size $N = 1000$. The initial population consisted of only the WT genotype (00). The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 1000$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs. The dashed lines represent the true selection coefficients (s_1 and s_2) and epistasis term (s_{12}). In these simulations, $s_1 = s_2$, hence the histograms of the estimates of the two have a significant overlap shown in grey color.

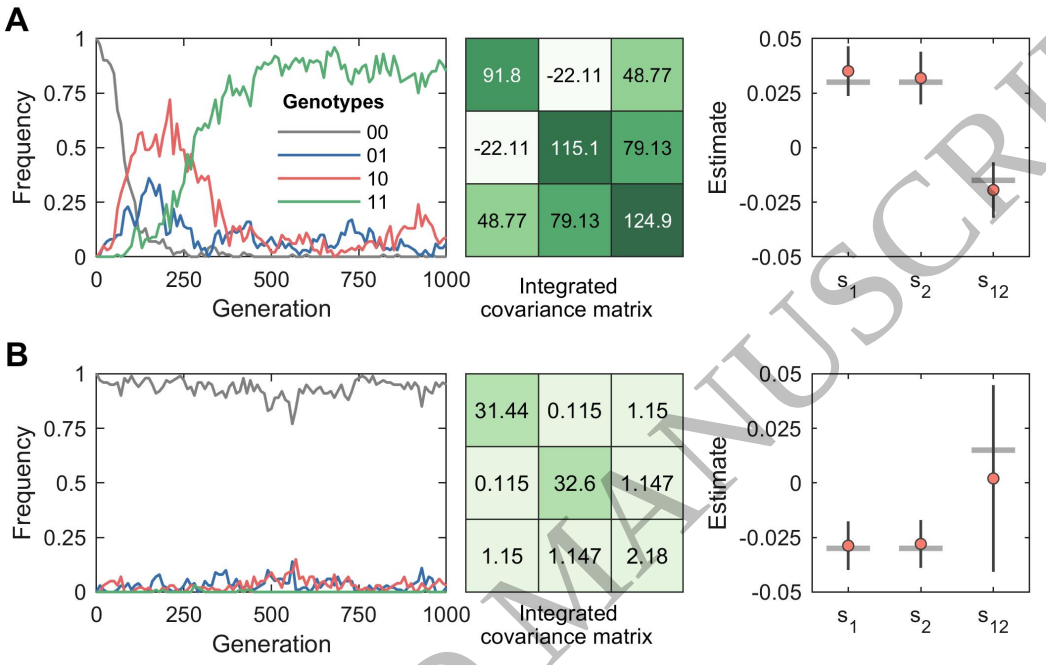

Figure 2 Higher genetic diversity leads to more accurate inference. (A) shows a sample run (*left* panel) of a two-locus system (negative epistasis scenario) where all genotypes are well represented in the data, as indicated by the magnitude of the diagonal entries of the integrated covariance matrix (*center* panel). This leads to accurate estimation of the epistasis term and the selection coefficients (*right* panel). The vertical bars in the *right* panel indicate the 95% confidence intervals while the horizontal bars indicate the true selection coefficients and epistasis terms. (B) shows a sample run (*left* panel) of a two-locus system (positive epistasis scenario) where the double mutant genotype has limited diversity, as indicated by the magnitude of the bottom right entry of the integrated covariance matrix (*center* panel). This leads to low accuracy in the estimate of the epistasis term. The selection coefficient estimates are still accurate because the single mutant genotypes, although present at low frequencies, are well represented in the data as indicated by the first two entries of the diagonal of the integrated covariance matrix (*center* panel). The results were obtained for a two-locus system with selection, mutation, and recombination. We set the selection coefficients s_1, s_2 and epistasis term s_{12} to $\{0.03, 0.03, -0.015\}$ and $\{-0.03, -0.03, 0.015\}$ in (A) and (B), respectively. Other system parameters included per locus mutation probability $\mu = 10^{-3}$, per locus recombination probability $r = 10^{-3}$, and population size $N = 1000$. The initial population consisted of only the WT genotype. The sampling parameters, in both simulations, were set to $n_s = 100$, $\Delta t = 10$, and $T = 1000$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference.

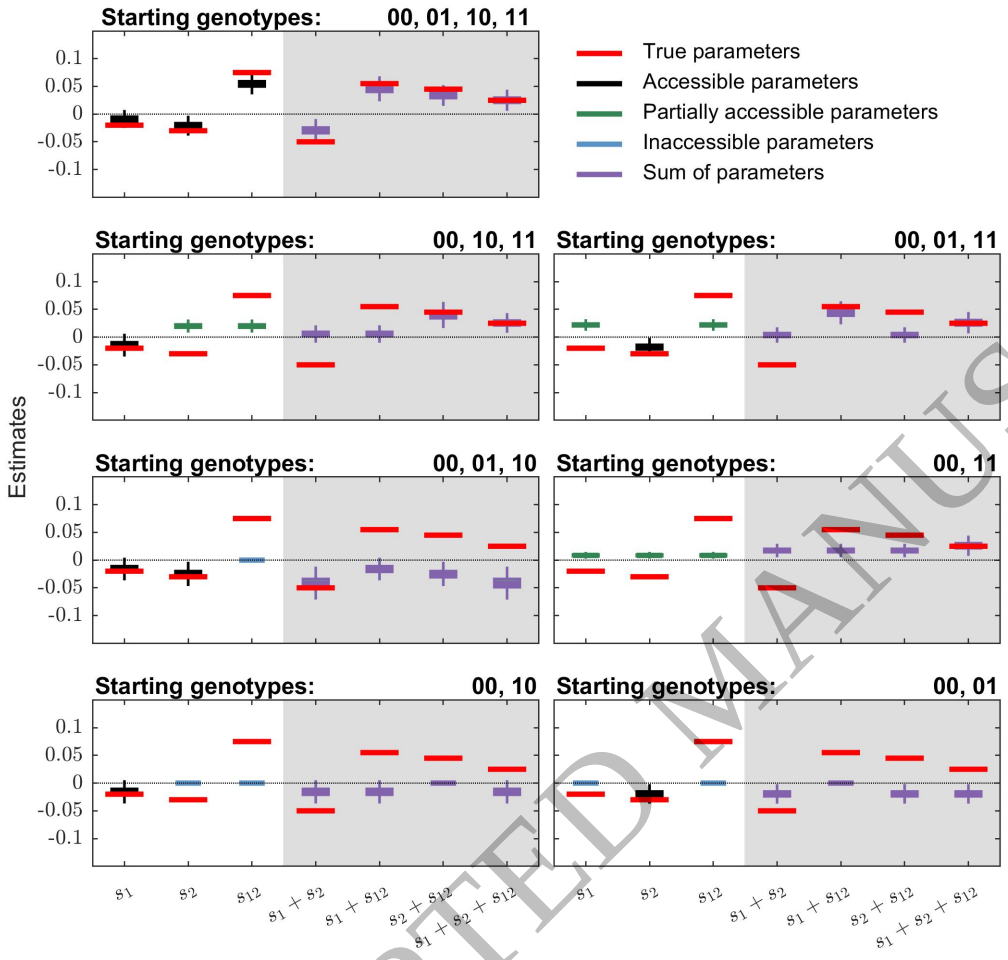

Figure 3 MPL can accurately estimate individual fitness parameters (selection coefficients and epistasis terms) and/or their sums depending on the genetic diversity present in the population. The results are for a two-locus system with positive sign epistasis (selection coefficients $s_1 = -0.02$, $s_2 = -0.03$ and pairwise epistasis term $s_{12} = 0.075$). All simulation results were computed over 1000 Monte Carlo runs. The boxplots of inferred selection coefficients and epistasis terms are shown on white background in each panel, while those of their sums are shown on grey background. The red bars indicate the true values of the respective terms. The boxplots show the standard data summary (first quartile, median, third quartile) with the whiskers showing 1.5 times the interquartile range. In order to control genetic diversity, both the per locus mutation probability and the per locus recombination probability were set to zero. The population size N was set to 1000. The panels depict scenarios with different starting populations. The genotypes contained in the starting population of each simulated scenario are mentioned on top of each panel. The frequency of each non-WT genotype in the initial population was set to 10% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 150$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference.

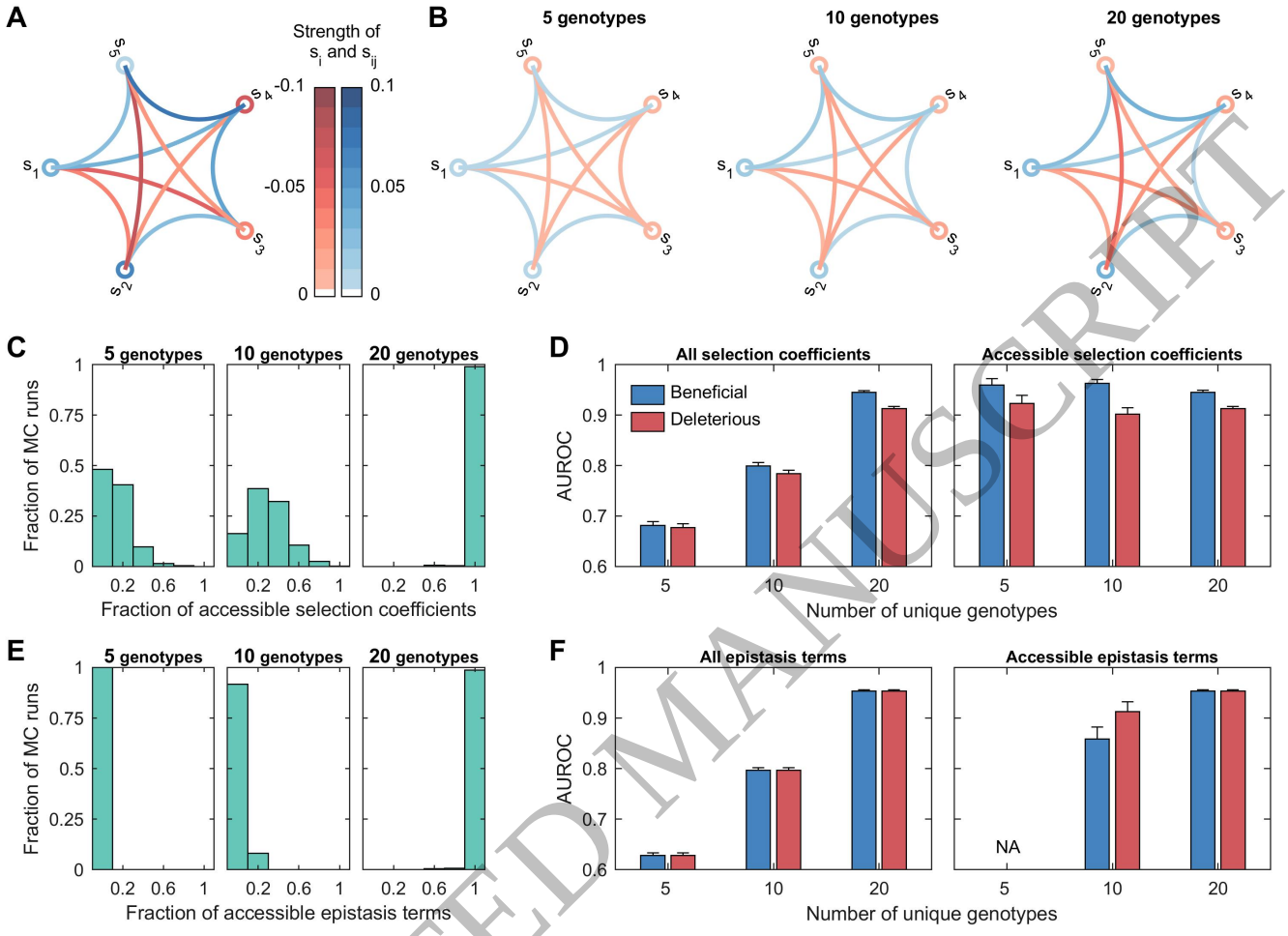

Figure 4 The fraction of selection coefficients and epistasis terms that are accessible depends on the genetic diversity in data. (A) shows the true fitness parameters of a five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). Specifically, we set selection coefficients as $s_1 = 0.0385$, $s_2 = 0.0605$, $s_3 = -0.0318$, $s_4 = -0.0632$, $s_5 = 0.002$, and epistasis terms as $s_{12} = -0.0361$, $s_{13} = -0.052$, $s_{14} = 0.0341$, $s_{15} = 0.0262$, $s_{23} = 0.0293$, $s_{24} = -0.0278$, $s_{25} = -0.075$, $s_{34} = 0.0498$, $s_{35} = -0.0283$, $s_{45} = 0.0721$. The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of genetic diversity (controlled by varying the number of unique genotypes in the initial population to either 5, 10 or 20). (C) shows the fraction of accessible selection coefficients as a function of genetic diversity. The *left* and *right* panels of (D) show the mean classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of genetic diversity. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. The population size N was set 1000. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control genetic diversity. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs.

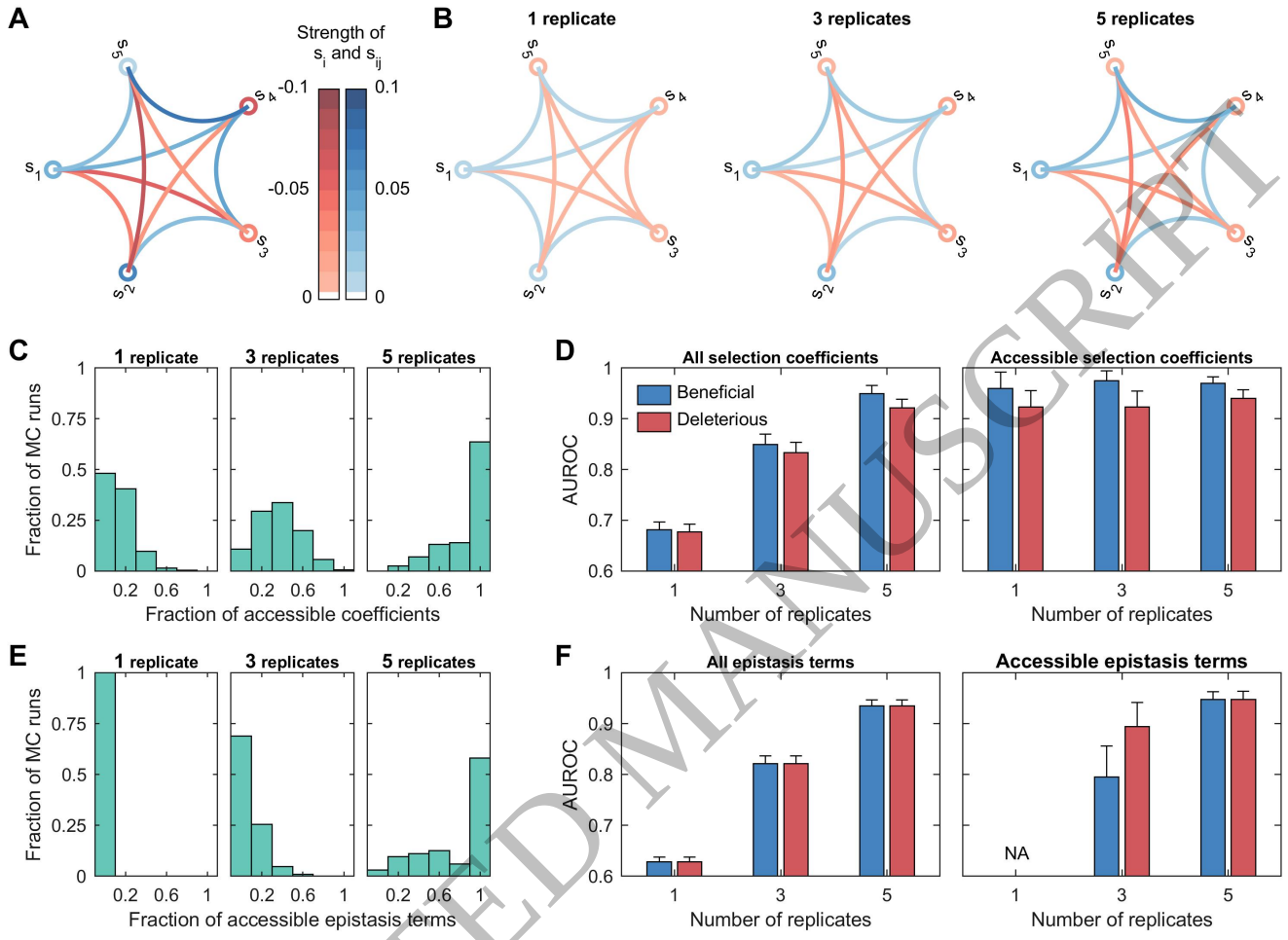

Figure 5 The fraction of selection coefficients and epistasis terms accessible in low genetic diversity scenarios can be increased by combining multiple independent replicates. (A) shows the true fitness parameters of a five-locus system, where the selection coefficients at loci are shown by circles and pairwise epistasis terms by chords between loci (*blue*: beneficial and *red*: deleterious). The underlying fitness landscape was the same as in Figure 4A. The *left*, *center*, and *right* panels of (B) show the average inferred fitness parameters obtained for different levels of genetic diversity (controlled by using either 1, 3 or 5 replicates for inference). (C) shows the fraction of accessible selection coefficients increases with the increase in genetic diversity. The *left* and *right* panels of (D) show the mean classification performance computed over all selection coefficients and over only the accessible selection coefficients respectively. The error bars indicate the standard error of the mean. (E) shows the fraction of accessible epistasis terms as a function of genetic diversity. The *left* and *right* panels of (F) show the classification performance computed over all and only the accessible epistasis terms respectively. 'NA' indicates the metric was not computed due to lack of data. Both the per locus mutation probability and the per locus recombination probability were set to zero in this simulation to control genetic diversity. The population size N was set 1000, and the initial population contained five unique genotypes. The frequency of each non-WT genotype in the initial population was set to 5% of the population size. The sampling parameters were set to $n_s = 100$, $\Delta t = 10$, and $T = 100$, where n_s is the number of samples, Δt is the time sampling step and T is the number of generations used for inference. All simulation results were computed over 1000 Monte Carlo runs.

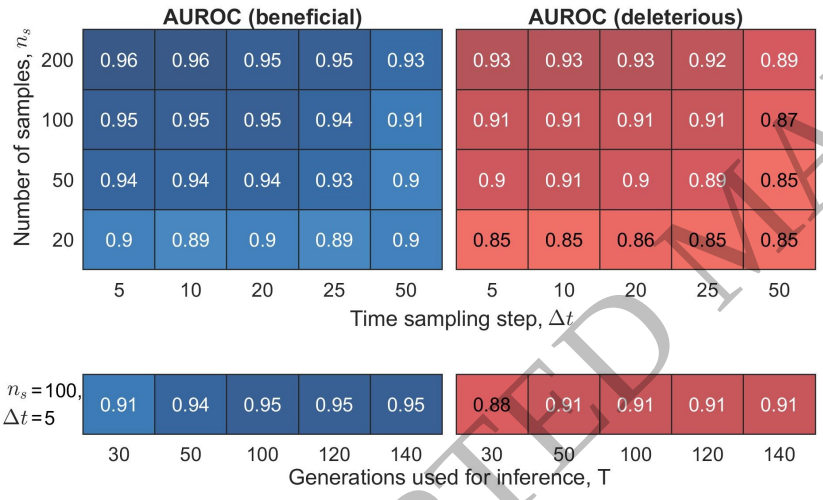

Figure 6 MPL is robust to variation in sampling parameters. The *top left* and *top right* panels show the mean AUROC performance of detecting accessible beneficial and deleterious selection coefficients, respectively. The *top* panels show mean AUROC performance for a range of values of number of samples, n_s, and time sampling step, Δt, with a fixed value of number of generations used for inference, $T = 100$, while the *bottom* panels show the performance for a range of values of T with $n_s = 100$ and $\Delta t = 5$. Results are for a five-locus system with the fitness landscape shown in Figure 4A. The population size N was set to 1000 and the initial population contained twenty unique genotypes. Other simulation parameters included per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$. All results were averaged over 1000 Monte Carlo runs.

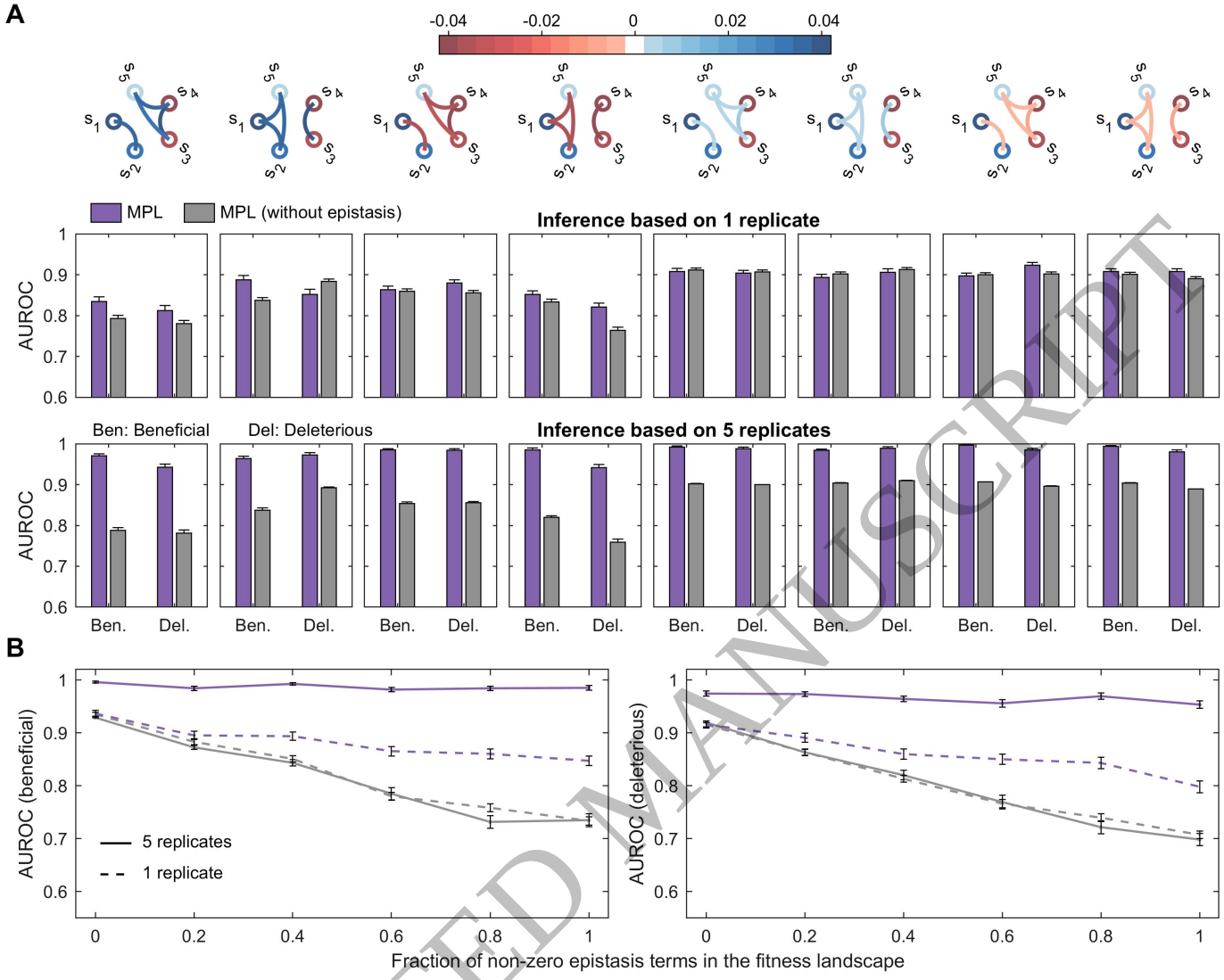

Figure 7 Ability of MPL to accurately identify selection coefficients is robust to the density and the strength of non-zero epistasis terms in the fitness landscape. (A) panels in the *top* row show simple fitness landscapes, i.e., no epistasis links between loci with mutant allele selection coefficients of opposite signs and all epistasis links of similar strengths and the same sign, while panels in the *center* and *bottom* rows show the classification performance of MPL and MPL (without epistasis) on data consisting of a single low genetic diversity replicate and that where five low-diversity replicates are combined, respectively. (B) shows the AUROC performance of both methods for varying density of epistasis terms in the fitness landscape under high and low genetic diversity scenarios. Error bars indicate the standard error of the mean. All fitness landscapes had two beneficial, two deleterious, and one neutral selection coefficients. Both the selection coefficients and epistasis terms, in the fitness landscapes with strong epistasis in (A), were randomly drawn from uniform distributions over the ranges $[0.03, 0.04]$ and $[-0.03, -0.04]$ for beneficial and deleterious fitness parameters respectively. While, the selection coefficients of the fitness landscapes with weak epistasis in (A) were drawn from the same distributions as before but the epistasis terms, positive and negative, were drawn from uniform distributions over the ranges $[0.003, 0.004]$ and $[-0.003, -0.004]$ respectively. For the fully connected fitness landscape in (B), we used the same fitness landscape as in Figure 4A. Half of the epistasis terms in this fitness landscape were positive while the other half were negative. To obtain a fitness landscape with a desired sparsity level, we set a randomly selected set of epistasis terms to zero. For a given sparsity level, we averaged the performance results over 10 randomly selected landscapes, except for the fully connected and the purely additive (all epistasis terms set to zero) landscape cases where the results are for a single landscape. Numerical values of all fitness landscapes used in these simulations are provided in Supplementary Table S1. The initial population contained 5 unique genotypes, with per locus mutation probability $\mu = 10^{-4}$ and per locus recombination probability $r = 10^{-4}$, and population size $N = 1000$. The sampling parameters were set to $n_s = 100$ and $\Delta t = 10$, with $T = 100$ generation used for inference. All simulation results were computed over 1000 Monte Carlo runs.