Efficient epistasis inference via higher-order covariance matrix factorization

Kai S. Shimagaki[1,2] **and John P. Barton**[1,2,†]

[1]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, USA. [2] Department of Physics and Astronomy, University of Pittsburgh, USA. †Address correspondence to: jpbarton@pitt.edu.

Epistasis can profoundly influence evolutionary dynamics. Temporal genetic data, consisting of sequences sampled repeatedly from a population over time, provides a unique resource to understand how epistasis shapes evolution. However, detecting epistatic interactions from sequence data is technically challenging. Existing methods for identifying epistasis are computationally demanding, limiting their applicability to real-world data. Here, we present a novel computational method for inferring epistasis that significantly reduces computational costs without sacrificing accuracy. We validated our approach in simulations and applied it to study HIV-1 evolution over multiple years in a data set of 16 individuals. There we observed a strong excess of negative epistatic interactions between beneficial mutations, especially mutations involved in immune escape. Our method is general and could be used to characterize epistasis in other large data sets.

Introduction

Epistasis is common in nature and plays an important role in evolution [1,2]. In the presence of epistasis, the fitness effects of mutations are contingent on the genetic background in which they appear, making the relationship between sequence and function complex [3–5]. More accurate estimates of epistasis could improve our ability to predict evolution, both at the level of genetic sequences and phenotypes [6–8].

Enormous amounts of time-resolved sequence data have been generated in recent years, opening the possibility of inferring epistasis from observations of evolution. Naively, sets of mutations that improve fitness will likely be found together in the same genetic sequence more often than expected by chance, while sets of deleterious mutations will be observed less frequently. However, phenomena such as genetic hitchhiking [9] and clonal interference [10] can also generate correlations between mutations that are unrelated to function. At present, a few methods exist to estimate pairwise epistatic interactions from temporal data, but computational constraints limit their applicability to small numbers of loci [11–13].

Here we propose an efficient method for inferring epistatic fitness that extends and vastly improves the computational efficiency of an approach developed by Sohail and collaborators [13]. With this new approach, the required memory and computational complexity scale only quadratically with the number of loci. These improvements are due to an efficient higher-order covariance matrix factorization (HCMF) method, which allows us to analyze much larger data sets than in previous analyses.

After validating our method in simulations, we apply it to study epistasis in within-host human immunodeficiency virus (HIV)-1 evolution in a cohort of 16 individuals. Several past studies have highlighted the role of epistasis in viral evolution. Early experimental work found evidence for both synergistic [14] and negative [15] epistasis in different viruses. Epistasis has been observed in influenza and in SARS-CoV-2, especially in the context of immune evasion [16–21]. In HIV-1, epistasis has been observed between mutations involved in drug resistance [22–24] and immune escape [25,26]. Here we found a consistent pattern of negative epistasis in HIV-1, with an interaction strength that typically scales along with the fitness effects of the individual mutations. Overall, our HCMF method enables the estimation of epistasis in large data sets, and our analysis contributes to the quantification of epistasis in viral evolution.

Recent inference methods for temporal genetic data

The widespread adoption of high-throughput sequencing has generated vast amounts of temporal genetic data, creating new opportunities to understand evolutionary processes. These datasets, which track genetic changes in populations over time, offer unique insights into natural selection and other evolutionary features. However, it can be challenging to extract meaningful information from temporal sequence data, particularly when considering the interplay between multiple genetic loci. Here, we review recent approaches to analyzing temporal genetic data and discuss how our method fits within this context.

Due to the technical challenges inherent in multi-locus data, most prior efforts to infer selection from temporal genetic data have focused on single-locus systems or treated different loci as independent [27–34]. To further simplify analyses, mathematical approximations of evolutionary dynamics are also common. Bollback et al. used a Bayesian approach to numerically estimate effective population size and selection from allele frequency time series based on an underlying Wright-Fisher (WF) diffusion model [27]. In the diffusion approximation, a continuous-time evolutionary process emerges as a limit of one formulated in discrete time steps, usually under the assumption that the population size, N, is large (going to infinity in the diffusion limit) and selection coefficients, s, are small (formally, $\mathcal{O}(1/N)$) [35,36]. Several groups developed numerical approximations of Wright-Fisher diffusion to estimate selection [32] and allele age [29,37,38]. Feder et al. used related approaches to develop a rigorous likelihood ratio-based test for selection [31]. Lacerda and Seoighe developed a different method for numerically approximating evolutionary dynamics without assuming that

selection is weak, which they used to estimate effective population size and selection[28]. Starting from the same mathematical foundation, Mathieson and collaborators generalized models of selection, allowing them to estimate selection coefficients that vary in space[30] or time[33] (see also ref.[34]).

However, phenomena such as genetic hitchhiking and clonal interference, which can drive neutral alleles to high frequencies or result in the loss of beneficial alleles through competition, can cloud the association between allele frequency change and selection at a single locus. A variety of methods have therefore attempted to infer selection across multiple loci simultaneously[39–46]. The method developed by He et al. is similar to diffusion-based approaches described above, but it can be applied to systems of two loci[46]. Illingworth and collaborators used a deterministic model for allele frequency dynamics and fit selection coefficients by maximizing the likelihood of the data under finite sampling. Tataru et al. developed another approach that estimates allele frequency distributions and selection coefficients using Taylor expansions around deterministic trajectories[44,45]. The approach of Terhorst et al. is fully stochastic, using a Gaussian process approximation of the Wright-Fisher process to jointly estimate selection while considering linkage disequilibrium[43].

Distinct from these methods, Foll et al. used approximate Bayesian computation (ABC) to estimate effective population sizes and then inferred selection coefficients under the assumption of conditional independence[42]. Rather than attempting to compute or even approximate the likelihood, ABC works by performing many simulations with different parameter values, highlighting those that offer the best compromise between the Bayesian prior distributions for the parameters and how accurately they reproduce some summary statistics of the data.

One weakness of the multi-locus methods described above is that they are difficult to apply to data sets with many loci. For some approaches, this is simply due to modeling assumptions[46]. For others, computational constraints are the limiting factor[39–45]. Typically, applications to data sets with more than a few dozen loci quickly become computationally infeasible[47].

Another recent line of work by Buffalo and Coop has used temporal genetic data to measure the genome-wide effects of natural selection[48,49]. Through detailed theoretical analyses, they showed that linked selection leaves a signature in the frequency dynamics of neutral alleles. In short, neutral alleles that appear on the same genetic background with advantageous ones may steadily increase in frequency over time, creating positive covariation between allele frequency changes across time (and vice versa for neutral alleles on deleterious backgrounds). They used this information to estimate the net contribution of linked selection to allele frequency change, the additive genetic variation for fitness, and potential shifts in selection from data. Buffalo and Coop's framework differs from those described above in its focus on temporal covariation and its suitability for studying polygenic adaptation, which can occur through subtle frequency shifts at multiple loci.

In this study, we extend the approach originally developed by Sohail and collaborators[13,47]. Like some previous efforts, our framework is based on the WF diffusion approximation. However, our approach to inference is different: we compute analytical expressions for the parameters that best fit the data in a Bayesian sense, rather than attempting to fit the parameters numerically. We use an additive fitness model with both selection coefficients and pairwise epistatic interactions between loci, which is more complex than most alternative approaches. In prior work, one major drawback of this approach was that the computational time needed to estimate pairwise epistatic interactions scales roughly as L^6, where L is the number of loci to be analyzed. This poor computational scaling made it infeasible to analyze complex data sets featuring hundreds or thousands of loci. Here, we introduce a new numerical approach that dramatically decreases computational costs, enabling the estimation of epistasis in large data sets.

Epistasis inference framework

As in related work[13,47], our modeling framework is based on the Wright-Fisher (WF) model[36,50,51]. We consider the evolution of a population of N haploid individuals with genetic variation at L loci. For simplicity, we will use a binary model where the allele at each locus is either wild-type (WT) or mutant, but this can easily be extended to realistic sequence models (Methods). We write the fraction of individuals with genotype α at time t as $z_\alpha(t)$. The state of the population is then described by the set of all $M = 2^L$ genotype frequencies $\boldsymbol{z} = (z_1, z_2, \ldots, z_M) = (z_\alpha(t))_{\alpha=1}^{M}$.

To model the fitness effects of individual mutations and pairwise epistasis, we assume a fitness function

$$f^\alpha = f(\boldsymbol{g}^\alpha) = 1 + \sum_i s_i g_i^\alpha + \sum_{i<j} s_{ij} g_i^\alpha g_j^\alpha \,. \qquad \textbf{(1)}$$

Here, the g_i^α are indicator functions, with a value equal to one if genotype α has a mutant allele at locus i and zero otherwise. Each locus i has a corresponding selection coefficient s_i that quantifies the fitness effect of the mutant allele at that locus and epistatic interactions s_{ij} with mutant alleles at all other loci j. Ultimately, our goal will be to infer the underlying fitness parameters $(s_i)_{i=1}^{L}$ and $(s_{ij})_{i<j}$ from temporal genetic data.

Under the WF model, the probability of obtaining a certain distribution of genotype frequencies in the next generation $\boldsymbol{z}(t+1)$, given the current distribution $\boldsymbol{z}(t)$, is multinomial. In general, inferring fitness parameters directly from the multinomial likelihood is numerically challenging. However, when generation-to-generation changes in genotype frequencies are small, we can apply the simplified diffusion approximation of the WF model[35,52]. Through the diffusion approximation, we can obtain an expression for the probability of an entire evolutionary trajectory, which we refer to as the path likelihood[13,47] (Methods). This allows us to compute the fitness parameters (including individual selection coefficients s_i and pairwise epistatic interactions s_{ij}) that best fit a

data set of sequences collected over time.

To express the results, it's useful to define a new vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_L, s_{1,2}, s_{1,3}, \ldots, s_{L-1,L}) = (s_e)_e$. This vector combines both selection coefficients for individual mutations and pairwise epistatic interactions, with a generalized index e that runs over both single loci $(1,2,\ldots,L)$ and pairs of loci $((1,2),(1,3),\ldots,(L-1,L))$. Similarly, we can define a mutant allele frequency and correlation vector $\boldsymbol{x} = (x_1, \ldots, x_L, x_{1,2}, \ldots, x_{L-1,L})$, where

$$\begin{aligned} x_i &= \sum_\alpha g_i^\alpha z_\alpha \\ x_{i,j} &= \sum_\alpha g_i^\alpha g_j^\alpha z_\alpha \,. \end{aligned} \tag{2}$$

Higher order mutant allele frequencies (e.g., $x_{i,j,k}$) are defined similarly.

The fitness parameters $\hat{\boldsymbol{s}}$ that maximize the path likelihood, together with a Gaussian prior distribution (or equivalently, ridge regression penalty) for the selection coefficients and epistatic interactions, are then given by [13,47]

$$\hat{\boldsymbol{s}} = (C^{\text{int}} + \gamma I)^{-1} \left(\boldsymbol{\Delta x}^{\text{int}} - \boldsymbol{u}^{\text{int}} - \boldsymbol{v}^{\text{int}} \right) . \tag{3}$$

In this expression, $\boldsymbol{\Delta x}^{\text{int}}$ represents the observed net allele frequency change over the observation period,

$$\begin{aligned} \boldsymbol{\Delta x}^{\text{int}} &= \sum_{k=0}^{K-1} \boldsymbol{\Delta x}(t_k) = \sum_{k=0}^{K-1} \left(\boldsymbol{x}(t_{k+1}) - \boldsymbol{x}(t_k) \right) \\ &= \boldsymbol{x}(t_K) - \boldsymbol{x}(t_0) \,. \end{aligned}$$

Here, the $(t_k)_{k=0}^K$ denote a series of sample collection times. The $\boldsymbol{u}^{\text{int}}$ and $\boldsymbol{v}^{\text{int}}$ terms quantify the expected cumulative allele frequency changes due to mutation and recombination, respectively (for explicit expressions, see Methods). γ quantifies the width of the prior distribution for the selection coefficients and epistatic interactions. C^{int} is the allele frequency covariance matrix summed over time. Let $i,j,k,l \in \{1,2,\ldots,L\}$, then $C(t) = C(\boldsymbol{x}(t))$ is

$$C_{e,f}(t) = \begin{cases} x_{ij}(t) - x_i(t)\, x_j(t) & e = i, \;\; f = j \\ x_{ikl}(t) - x_i(t)\, x_{kl}(t) & e = i, \;\; f = (k,l) \\ x_{ijk}(t) - x_{ij}(t)\, x_k(t) & e = (i,j), \;\; f = k \\ x_{ijkl}(t) - x_{ij}(t)\, x_{kl}(t) & e = (i,j), \;\; f = (k,l) \,. \end{cases} \tag{4}$$

and

$$C^{\text{int}} = \sum_{k=0}^{K-1} \Delta t_k C(t_k) \,, \tag{5}$$

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

Although (3) is complex, it can be interpreted intuitively. Essentially, (3) states that net allele frequency change that is *not* explained by the forces of mutation or recombination is evidence of selection. The sign and magnitude of inferred selection depend on the total change in allele frequencies (how much, quantified by $\boldsymbol{\Delta x}^{\text{int}}$, and how fast, quantified by the diagonal part of C^{int}) as well as the effects of genetic background (quantified by the off-diagonal terms of C^{int}).

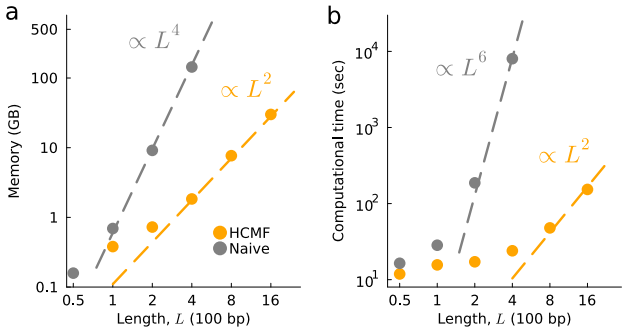

Fig. 1. HCMF substantially reduces the required memory size and computational time. a, Required memory size versus number of loci L (measured in 100 bp). The required memory size of the naive method scales as $\mathcal{O}(L^4)$, while our method reduces it to $\mathcal{O}(L^2)$. For the naive method, we did not consider $L > 400$ due to computational constraints. **b**, Required computational time (in seconds) versus number of loci. As anticipated, the computational time of the naive method and HCMF scale by $\mathcal{O}(L^6)$ and $\mathcal{O}(L^2)$, respectively.

Results

Factorization of higher-order integrated covariance matrix and efficient inference framework

While (3) provides a powerful expression to simultaneously estimate the fitness effects of mutations and pairwise epistasis from temporal genetic data, it faces a serious computational limitation. The covariance matrix C^{int} is $D = qL(q(L-1)/2+1)$-dimensional, where q is the number of alleles at each locus. The computational complexity of inverting the covariance matrix thus scales as $\mathcal{O}\left((qL)^6\right)$, with memory costs related to the storage of covariance matrix entries scaling as $\mathcal{O}\left((qL)^4\right)$. For data sets with hundreds or thousands of loci, even storing the covariance matrix in memory becomes challenging.

We developed an efficient and generic method to resolve the major computational bottleneck hindering the application of this approach to larger data sets. The key idea of our approach is to exploit the regular structure of the covariance matrix, allowing us to factorize the matrix and perform calculations in a lower-dimensional space without any loss of information. Specifically, we can write the integrated covariance matrix in terms of a rectangular matrix with dimensions $D \times d$, which depends on the number of unique sequences in the data set. Writing the number of unique sequences in the data set at time t_k as d_k, we have $d = \sum_{k=0}^K d_k + K - 1$. As we will show below, d may be multiple orders of magnitude smaller than D for relevant data sets of interest, allowing this factorization to dramatically speed up analyses.

Practically, we linearly interpolate allele frequencies between sampled time points and replace the sum in (5) with an integral. This interpolation method mitigates periods of sparse sampling and consistently improves inference accuracy [13,47,53]. The expression for C^{int} with linearly interpolated allele frequencies can be factorized as follows (see

Methods for details):

$$
\begin{aligned}
C^{\mathrm{int}} &= \sum_{k=0}^{K-1} \Delta t_k \frac{C(t_k)+C(t_{k+1})}{2} \\
&\quad + \sum_{k=1}^{K-1} \Delta t_k \frac{\boldsymbol{\Delta x}(t_k)\boldsymbol{\Delta x}(t_k)^\top}{6} \\
&= \sum_{k=0}^{K} \sum_{\alpha} \boldsymbol{\xi}_\alpha(t_k)\boldsymbol{\xi}_\alpha(t_k)^\top + \sum_{k=1}^{K-1} \boldsymbol{\xi}(t_k)\boldsymbol{\xi}(t_k)^\top \\
&=: \Xi\Xi^\top .
\end{aligned}
\tag{6}
$$

The $\boldsymbol{\xi}$ vectors are defined as

$$
\begin{aligned}
\boldsymbol{\xi}_\alpha(t_k) &= \begin{cases} \sqrt{\frac{z_\alpha(t_0)\Delta t_0}{2}}\Big(\boldsymbol{\sigma}^\alpha - \boldsymbol{x}(t_0)\Big) & k=0 \\ \sqrt{\frac{z_\alpha(t_k)(\Delta t_k+\Delta t_{k-1})}{2}}\Big(\boldsymbol{\sigma}^\alpha - \boldsymbol{x}(t_k)\Big) & 0<k<K \\ \sqrt{\frac{z_\alpha(t_K)\Delta t_{K-1}}{2}}\Big(\boldsymbol{\sigma}^\alpha - \boldsymbol{x}(t_K)\Big) & k=K\,, \end{cases} \\
\boldsymbol{\xi}(t_k) &= \sqrt{\frac{\Delta t_k}{6}}\boldsymbol{\Delta x}(t_k)\,,
\end{aligned}
\tag{7}
$$

where $z_\alpha(t_k)$ is the frequency of genotype α in the data at time t_k, and $\boldsymbol{\sigma}^\alpha$ is a D-dimensional vector with entries

$$
\sigma_e^\alpha = \begin{cases} g_i^\alpha & \text{for } e=i \\ g_i^\alpha g_j^\alpha & \text{for } e=(i,j)\,. \end{cases}
\tag{8}
$$

Using the factorized Ξ matrix (6), we can rewrite the equation for the estimated selection coefficients and epistatic interactions (3) as

$$
\hat{\boldsymbol{s}} = \gamma^{-1}(\boldsymbol{\Delta}\tilde{\boldsymbol{x}}^{\mathrm{int}} - \Xi\boldsymbol{\Delta\eta})\,,
\tag{9}
$$

with

$$
\begin{aligned}
\boldsymbol{\Delta}\tilde{\boldsymbol{x}}^{\mathrm{int}} &= \boldsymbol{\Delta x}^{\mathrm{int}} - \boldsymbol{u}^{\mathrm{int}} - \boldsymbol{v}^{\mathrm{int}}\,, \\
\boldsymbol{\Delta\eta} &= \Big(\Xi^\top\Xi + \gamma I\Big)^{-1}\ \Big(\Xi^\top\boldsymbol{\Delta}\tilde{\boldsymbol{x}}^{\mathrm{int}}\Big)\,.
\end{aligned}
\tag{10}
$$

Critically, computing (9) is far less computationally intensive than (3) when $D \gg d$, as the matrix to be inverted in (9) is only $d \times d$. In total, the computational complexities of calculations in (9) are: matrix-vector products of $\Xi\boldsymbol{\Delta\eta}$ and $\Xi^\top\boldsymbol{\Delta}\tilde{\boldsymbol{x}}^{\mathrm{int}}$ take $\mathcal{O}(dD)$; matrix-matrix product of $\Xi^\top\Xi$ requires $\mathcal{O}(d^2D)$; solving the equation for $\boldsymbol{\Delta\eta}$ without directly solving its inverse is smaller than $\mathcal{O}(d^{2+\omega})$, with ω a small positive number $0<\omega\leq 1$, depending on linear optimization solvers.

This substantial computational reduction was achieved by implicitly computing the integrated covariance matrix without ever storing the covariance matrix itself. Therefore, our epistasis inference scheme is more efficient and scalable as the computational complexity scales only linearly with D (and thus quadratically with L). In comparison, even naive selection inference without epistasis scales as $\mathcal{O}(D^3)$. The expression for the selection coefficients in (9) uses no approximations. Thus, its solution is exact in the diffusion limit[13].

For simplicity, we initially assumed the same regularization γ for selection and epistasis. However, we have also generalized our approach so that the regularization values γ_e can differ and implemented this in our code (Methods). While our analysis considers only pairwise epistatic interactions, one could further extend the fitness function to consider even higher-order interactions. For p-way epistatic interactions with $p>1$, the computational complexity would become $\mathcal{O}(dD)$ with $D=\sum_{l=1}^{p} q^l \binom{L}{l}$.

The discussions so far are general and can be applied to a standard covariance matrix that depends on low-order moments of the sequence distribution. While the explicit expression of the $\boldsymbol{\xi}$ would need to be adjusted, this covariance factorization method is applicable to other interpolation frameworks and estimation for more complex fitness models. For example, piecewise constant interpolation is readily applicable, and more complex non-linear interpolation, such as spline curves, can also be used, effectively introducing additional basis vectors, $\boldsymbol{\xi}$ (ref.[53]).

HCMF substantially reduces computational costs

To assess the efficiency of HCMF, we simulated population evolution under the WF model using different numbers of loci, ranging from $L=50$ to 1600. We used a constant population size of $N=10^3$ and a symmetric mutation rate of $\mu=10^{-3}$, which gives the probability that per site per generation of mutating from the wild-type allele to the mutant (or vice versa). We chose a recombination rate of $r=10^{-4}$ per site per generation. Our simulations ranged over 2000 generations, with virtual samples collected for inference every 10 generations. We used a fitness landscape in which 25% of mutations were beneficial ($s_i=0.03$), 25% were deleterious ($s_i=-0.03$), and 50% were neutral ($s_i=0$). Similarly, 25% of all pairs of sites were randomly selected to have positive/negative epistatic interactions ($s_{ij}=0.03$ or -0.03, respectively), with the remaining 50% of the possible epistatic interactions set to zero. To ensure sufficient sampling to measure typical results, we performed 500 simulations for each condition.

Since the size of the covariance matrix increases quadratically with sequence length, the required memory size of the naive approach increases as $\mathcal{O}(L^4)$. However, memory requirements only scale as $\mathcal{O}(L^2)$ for the HCMF method (**Fig. 1a**). HCMF also dramatically reduces the run time of the inference, scaling as $\mathcal{O}(L^2)$ compared to $\mathcal{O}(L^6)$ for the naive approach (**Fig. 1b**). For example, for $L=400$, HCMF is 10^4 times faster than the naive approach. This computational advantage should further increase for larger sequence lengths. As noted above, since the HCMF approach involves no approximations, the selection coefficients and epistatic interactions inferred by this approach match the ones from the naive method exactly within machine precision.

Importance of higher-order covariance information for inferring epistasis

One of the main barriers to inferring epistatic interactions via (3) is computing and inverting the integrated covariance ma-

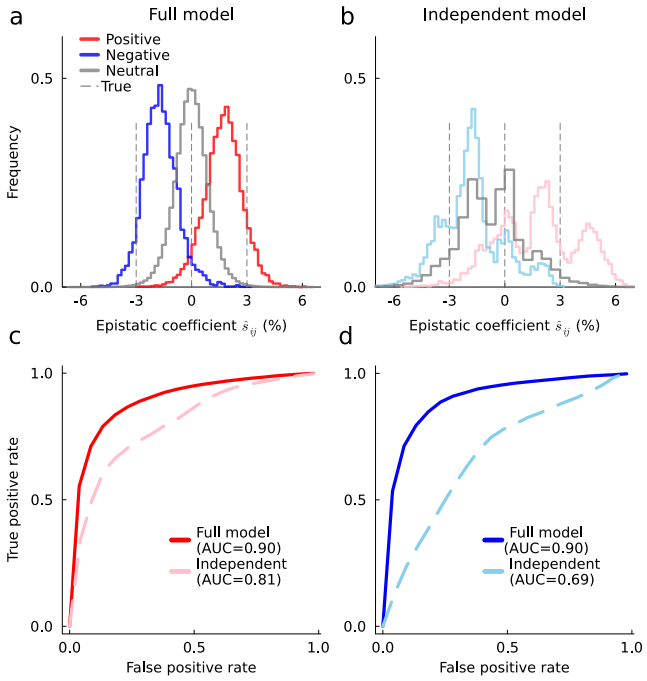

Fig. 2. Higher-order covariance information improves the inference of epistasis. Distribution of inferred epistasis using the full model inferred via HCMF (**a**), which includes higher-order covariance information, and using the independent model (**b**), in which off-diagonal terms of the covariance matrix are set to zero. **c**, Receiver operating characteristic (ROC) curve for identifying positive epistasis. The area under the curve (AUC) value is 0.90 for the full model, while it drops to 0.81 for the independent model. **d**, Analogous ROC and AUC values for identifying negative epistasis. The AUC values are 0.90 and 0.69 for full and independent models, respectively.

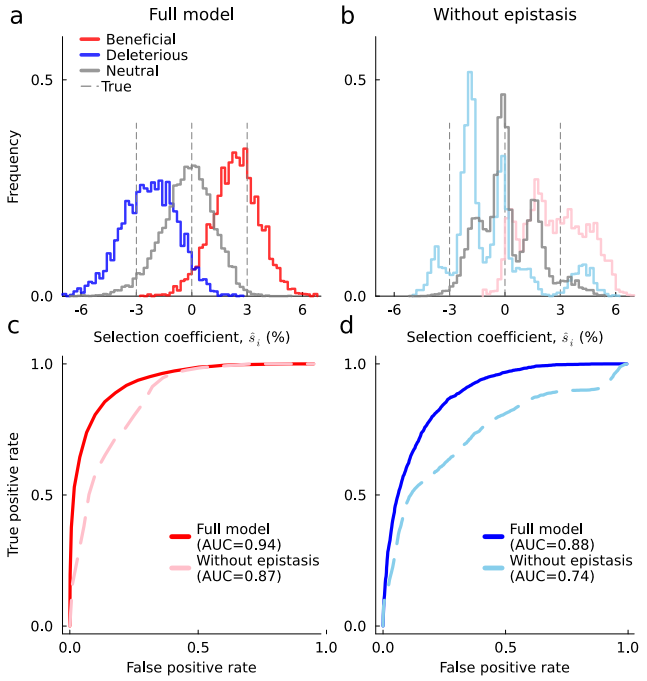

Fig. 3. Modeling epistasis improves the inference of additive selection coefficients. Distribution of selection coefficients inferred with the full model via HCMF (**a**) and a simpler model with no epistatic interactions (**b**). When epistasis is present, including it in the model also improves estimates of selection coefficients. **c**, ROC curves and their AUC values for identifying positive selection coefficients. The AUC value of the full model is 0.94, while the AUC value of the model without epistasis drops to 0.87. **d**, Analogous ROC and AUC values for identifying deleterious selection coefficients. The AUC values are 0.88 and 0.74 for the full model and the model without epistasis, respectively. Simulation parameters are the same as in **Fig. 2**.

trix. The HCMF approach we have developed offers one solution to this problem. However, one could also simplify (3) by neglecting the off-diagonal terms of the covariance matrix. This greatly reduces the computational burden of the problem, but neglects important information about linkage disequilibrium that could inform the inference of selection coefficients and epistatic interactions. We refer to this approximation as the independent model (analogous to the single locus model of ref. [47]).

We performed additional simulations to compare the accuracy of the HCMF method, which includes higher-order covariance information, and the independent model, which does not. These simulations were performed with the same parameters as in the previous section, using $L = 50$ loci and sparser epistatic interactions. Here we chose a random set of $L/2 = 25$ pairs of sites to have positive epistatic interactions ($s_{ij} = 0.03$), $L/2$ pairs with negative interactions ($s_{ij} = -0.03$), and set the remaining epistatic interactions to zero.

The inferred epistatic interactions using the full model with HCMF are much closer to the true ones than those inferred with the independent model (**Fig. 2a-b**). For the full model, the distribution of inferred positive/neutral/negative epistatic interactions is roughly normal, with peaks that can easily be distinguished from one another. In contrast, the epistatic coefficients inferred using the independent model are distributed much more broadly and irregularly. While positive and negative interactions can, on average, still be distinguished from one another using the independent model, it is more difficult to do so. To quantify this difference, we computed the receiver operating characteristic (ROC) curve and area under the curve (AUC) for identifying positive (**Fig. 2c**) and negative (**Fig. 2d**) epistatic interactions using the full and independent models. Thus, by all metrics we find that the inclusion of higher-order covariance information improves the ability to identify epistatic interactions from data.

Modeling epistasis improves the inference of selection coefficients

An alternative approach to reducing the computational costs of (3) is to use a simpler fitness landscape, such as a purely additive one with all $s_{ij} = 0$. This assumption may be especially appropriate for analyzing highly similar sequences. However, in models with substantial, strong epistatic interactions, omitting epistasis could skew fitness estimates (**Fig. 3a-b**). In these conditions, inference using models with epistasis yields more accurate estimates of individual selection coefficients and improves the detection of beneficial and deleterious mutations (**Fig. 3c-d**).

Joint inference from multiple replicates

Some evolution experiments, such as deep mutational scanning studies [54,55], have multiple independent replicates collected under the same conditions. Using our approach, we

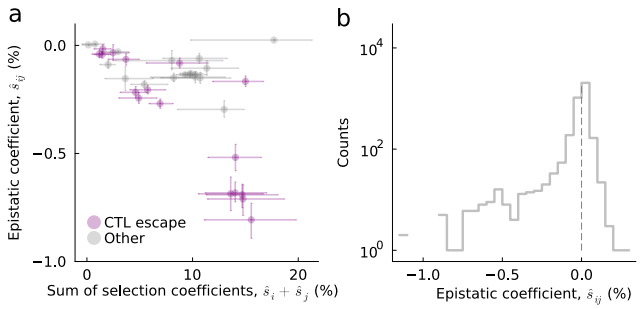

Fig. 4. Predominance of negative epistasis between beneficial HIV-1 mutations. a, Comparison of inferred epistatic interactions, $\hat{s}_{ij}$, and the corresponding sum of individual selection coefficients, $\hat{s}_i + \hat{s}_j$, in a typical case (700010077-3; see **Supplementary Fig. 2** for all individuals). Generically, we find negative correlations between inferred epistatic interactions and selection coefficients. CTL escape mutations are typically found to be both strongly beneficial and to have negative epistatic interactions with other escape mutations. **b**, Distribution of inferred epistatic interactions across all individuals. Most terms are near zero, but a few epistatic interactions are significantly negative.

can estimate selection coefficients and epistatic interactions that best explain the data across all replicates, as shown in prior work[13,56,57]. To demonstrate this phenomenon in a challenging setting for inference, we increased the density of epistatic interactions to 50%, with half set to be positive ($s_{ij} = 0.03$) and half negative ($s_{ij} = -0.03$). In this case, epistatic effects dominate the fitness function. With a single replicate, the AUC values were 0.82 (0.81) for identifying beneficial (deleterious) selection coefficients and 0.74 (0.72) for positive (negative) epistasis. Combining data from two replicates raised the AUC to 0.93 (0.89) for selection coefficients and 0.86 (0.85) for epistasis, with further improvements as the number of replicates increases (**Supplementary Fig. 1**).

Epistasis in intrahost HIV-1 evolution

As a practical application of our approach, we studied within-host HIV-1 evolution in 16 individuals who were not treated with antiretroviral drugs during the sampling time. This data set included individuals enrolled in the CHAVI 001 and CAPRISA 002 studies in the United States, Malawi, and South Africa[58,59]. Each individual was identified shortly after HIV-1 infection, and the viral population within each individual was sampled frequently for several months to years afterward. For most individuals, the 3′ and 5′ halves of the HIV-1 genome were sequenced separately using single genome amplification methods, preserving information about linkage disequilibrium between mutations even at long distances. For two individuals, denoted CH505 and CAP256, only the HIV-1 surface protein Env was sequenced. Most data sets consisted of around 50-100 HIV-1 sequences in total for each sequencing region, collected over 5-8 time points, with several hundred polymorphic loci (see **Supplementary Table 1**). However, the viral population was also sequenced more deeply in a few individuals, featuring as many as 1205 HIV-1 sequences collected at 31 time points over roughly 5 years.

Using this data, we inferred selection coefficients and epistatic interactions between HIV-1 mutations for each individual and sampling region. We used prior estimates to set the mutation[60] and effective recombination rates[61] in (9). By convention, we set the selection coefficients and epistatic interactions for the transmitted/founder (TF) sequence, the natural analog of WT, to zero (Methods). Thus, fitness effects are expressed relative to the strain of the virus that originally infected each individual. In general, the ability to transform the model parameters (i.e., selection coefficients and epistatic interactions) without affecting the dynamics of the model is referred to as a gauge freedom. Choosing a specific convention for the parameters is important for comparing fitness effects in different contexts and for improving the interpretability of the model[57,62–64].

Here we focused specifically on epistatic interactions between nearby sites (separated by <50 bp), with distant epistatic interactions suppressed by strong regularization. There were two reasons for our focus on short-range interactions. First, due to the high effective recombination rate in HIV-1, the size of the sequencing region, and some large time gaps between samples, the expected change in correlations between mutant alleles due to recombination may violate the mathematical assumptions of the diffusion approximation, biasing our inferences for these sites. Second, short-range epistatic interactions may be of particular biological interest in HIV-1 evolution.

The accumulation of mutations within cytotoxic T lymphocyte (CTL) epitopes – linear peptides roughly 10 amino acids in length that are recognized by cytotoxic T cells – allows mutant viruses to escape from the immune system. Past work has shown that T cells are especially important in controlling HIV-1 replication[65], and that the virus faces significant selective pressure to escape from CTLs[47,59,65–67]. However, because the recognition of CTL epitopes is highly specific, even one nonsynonymous mutation within the epitope can be sufficient to confer escape[68–70]. We anticipate that this phenomenon could lead to negative epistasis between CTL escape mutations, as the fitness benefit of multiple mutations within the epitope should be lower than expected based on the beneficial effect of each individual escape mutation.

While most of the epistatic interactions we inferred were very close to zero, a few were significantly negative (**Fig. 4**). Consistent with the biological intuition above, we found that negative epistatic interactions were more common between beneficial mutations, especially CTL escape mutations (**Fig. 4**). This pattern of negative epistasis between beneficial mutations, including CTL escape mutations, was robustly observed across all individuals and sequencing regions that we studied (**Supplementary Fig. 2**).

Consistency with prior estimates of selection in HIV-1

Past work has studied HIV-1 evolution in part of this data set with different modeling choices, including a model with purely additive selection[47] and one that includes specific terms for CTL escape[71]. Neither of these models includes pairwise epistatic interactions. Thus, we compared the selection coefficients inferred in our analysis with those from previous models to understand how the inclusion of pairwise

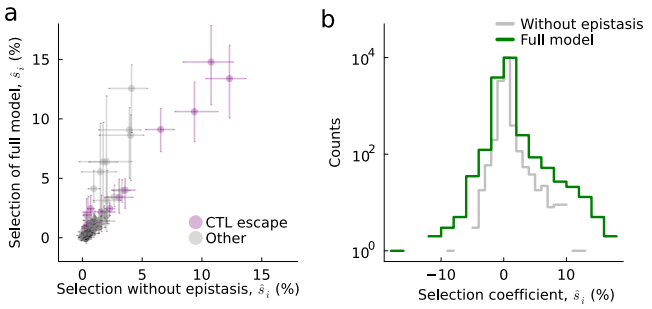

Fig. 5. Consistency of inferred selection coefficients in models with and without epistasis. **a**, Comparison of inferred selection coefficients in models with and without epistasis in a typical case (700010077-3; see **Supplementary Fig. 3** for all individuals). While the exact values differ, there is excellent general agreement between the mutations that are inferred to strongly affect fitness and those that are inferred to be nearly neutral. **b**, Distribution of selection coefficients across all individuals. Both distributions are peaked near zero, but the tails of the distributions in the full model are longer.

epistasis affects the interpretation of the fitness effects of individual mutations.

Figure 5a shows a typical example of the inferred selection coefficients, $\hat{s}_i$, with and without the inclusion of epistasis. Overall, we find that the inferred selection coefficients are similar to those in past models (mean Pearson's $R = 0.94$). In particular, all models find very strong selection for CTL escape mutations[47,71]. As in previous work, the great majority of inferred selection coefficients are very close to zero (**Fig. 5b**). However, the model without epistasis also features heavier tails in the distribution of inferred selection coefficients, with more mutations inferred to have either very beneficial or very deleterious individual effects.

Discussion

Epistasis is prevalent in nature, and has been observed to influence viral evolution[1,15,72–74]. However, inferring epistasis from data is technically and computationally challenging. Here, we developed a new approach for the path likelihood inference framework[13,47,57,71,75] that greatly reduces computational costs for many data sets of interest, especially for inferring epistasis. Our key innovation was the efficient factorization of the higher-order covariance matrix, which allows us to analytically estimate selection coefficients and epistatic interactions from data without ever explicitly computing the covariance matrix or its inverse. For this reason, we referred to our method as higher-order matrix factorization (HCMF). The HCMF approach is general and can be applied under different assumptions about the structure of the fitness landscape. HCMF does not introduce any new approximations, so it suffers no loss in accuracy compared to prior approaches.

After validating our approach in simulations, we applied HCMF to study HIV-1 evolution within 16 individuals. The fitness effects of mutations that we inferred were consistent with past computational results[47,56,67,71] and with experimental findings. In particular, we found strong selection for mutations that allow the virus to escape from the host immune system, in agreement with a large body of experimental work and clinical observations[25,59,65,66,76].

In this HIV-1 data set, the distribution of epistatic interactions that we inferred was peaked near zero, but with a substantial tail of strong negative epistasis. Patterns of negative epistasis have also been observed in other viruses[15]. Negative epistasis was especially common between CTL escape mutations, consistent with the finding that single mutations within an epitope typically already disrupt T cell recognition[68–70]. We also observed negative epistasis between pairs of beneficial mutations more generally. This finding is consistent with more general studies that have observed decreasing effect sizes of beneficial mutations over time[6,77–79].

Our approach to inferring epistasis from temporal data differs from some prior methods, which used statistical models to explain correlations in protein sequences collected from many individuals or species[22,23,26,80–84]. These models treat sequence data as samples from a static, equilibrium distribution, and interpret correlations between mutations as possible evidence for epistasis. Only a handful of methods allow for the possibility that linkage disequilibrium may arise from an underlying phylogenetic structure to the sequence data[85,86], or simply by chance. Nonetheless, these approaches have also been successful at tasks such as predicting the fitness effects of HIV-1 mutations in experiments[82,83] and the dynamics of immune escape within individual patients[26]. In contrast to the present work, these models offer a "global" view of epistasis averaged across many related sequences.

The HCMF approach that we have developed is general. While our study focused on HIV-1, future work could be applied to other populations, including viruses like influenza and SARS-CoV-2 (ref. [75]), experimental evolution[57], or bacteria[87,88]. As one example, recent studies have suggested that epistasis plays an important role in maintaining fitness among SARS-CoV-2 Spike mutations that escape from antibodies and control receptor binding[19,20]. More systematic studies could reveal the importance of epistasis in different aspects of SARS-CoV-2 evolution. More generally, a deeper understanding of epistasis may also improve our ability to understand and predict viral evolution.

ACKNOWLEDGEMENTS

The work of K.S.S. 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.

AUTHOR CONTRIBUTIONS

All authors contributed to research design, methods development, interpretation of results, and writing the paper. K.S.S. performed simulations as well as theoretical and computational analyses. J.P.B. supervised the project.

References

1. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**, 855–867 (2008).
2. Mackay, T. F. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics* **15**, 22–33 (2014).
3. Whitlock, M. C., Phillips, P. C., Moore, F. B.-G. & Tonsor, S. J. Multiple fitness peaks and epistasis. *Annual review of ecology and systematics* 601–629 (1995).
4. Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & De Visser, J. A. G. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment* **2013**, P01005 (2013).
5. Bank, C. Epistasis and adaptation on fitness landscapes. *Annual Review of Ecology, Evolution, and Systematics* **53**, 457–479 (2022).
6. Kryazhimskiy, S., Rice, D. P., Jerison, E. R. & Desai, M. M. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344**, 1519–1522 (2014).

7. Park, Y., Metzger, B. P. & Thornton, J. W. Epistatic drift causes gradual decay of predictability in protein evolution. *Science* **376**, 823–830 (2022).
8. Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLoS computational biology* **13**, e1005541 (2017).
9. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genetics Research* **23**, 23–35 (1974).
10. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genetics Research* **8**, 269–294 (1966).
11. Illingworth, C. J. Fitness inference from short-read data: within-host evolution of a reassortant h5n1 influenza virus. *Molecular Biology and Evolution* **32**, 3012–3026 (2015).
12. Pedruzzi, G. & Rouzine, I. M. An evolution-based high-fidelity method of epistasis measurement: Theory and application to influenza. *PLoS Pathogens* **17**, e1009669 (2021).
13. Sohail, M. S., Louie, R. H., Hong, Z., Barton, J. P. & McKay, M. R. Inferring epistasis from genetic time-series data. *Molecular Biology and Evolution* **39**, msac199 (2022).
14. Burch, C. L. & Chao, L. Epistasis and its relationship to canalization in the rna virus φ6. *Genetics* **167**, 559–567 (2004).
15. Sanjuán, R., Moya, A. & Elena, S. F. The contribution of epistasis to the architecture of fitness in an rna virus. *Proceedings of the National Academy of Sciences* **101**, 15376–15379 (2004).
16. Lyons, D. M. & Lauring, A. S. Mutation and epistasis in influenza virus evolution. *Viruses* **10**, 407 (2018).
17. Starr, T. N. *et al.* Shifting mutational constraints in the sars-cov-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424 (2022).
18. Starr, T. N. *et al.* Deep mutational scans for ace2 binding, rbd expression, and antibody escape in the sars-cov-2 omicron ba. 1 and ba. 2 receptor-binding domains. *PLoS pathogens* **18**, e1010951 (2022).
19. Moulana, A. *et al.* Compensatory epistasis maintains ace2 affinity in sars-cov-2 omicron ba. 1. *Nature Communications* **13**, 7011 (2022).
20. Moulana, A. *et al.* The landscape of antibody binding affinity in sars-cov-2 omicron ba. 1 evolution. *Elife* **12**, e83442 (2023).
21. Witte, L. *et al.* Epistasis lowers the genetic barrier to sars-cov-2 neutralizing antibody escape. *Nature Communications* **14**, 302 (2023).
22. Flynn, W. F. *et al.* Deep sequencing of protease inhibitor resistant hiv patient isolates reveals patterns of correlated mutations in gag and protease. *PLoS computational biology* **11**, e1004249 (2015).
23. Butler, T. C., Barton, J. P., Kardar, M. & Chakraborty, A. K. Identification of drug resistance mutations in hiv from constraints on natural evolution. *Physical Review E* **93**, 022412 (2016).
24. Zhang, T.-h. *et al.* Predominance of positive epistasis among drug resistance-associated mutations in hiv-1 protease. *PLoS genetics* **16**, e1009009 (2020).
25. Brockman, M. A. *et al.* Escape and compensation from early hla-b57-mediated cytotoxic t-lymphocyte pressure on human immunodeficiency virus type 1 gag alter capsid interactions with cyclophilin a. *Journal of virology* **81**, 12608–12618 (2007).
26. Barton, J. P. *et al.* Relative rate and location of intra-host hiv evolution to evade cellular immunity are predictable. *Nature communications* **7**, 11660 (2016).
27. Bollback, J. P., York, T. L. & Nielsen, R. Estimation of 2 n es from temporal allele frequency data. *Genetics* **179**, 497–502 (2008).
28. Lacerda, M. & Seoighe, C. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics* **198**, 1237–1250 (2014).
29. Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**, 599–607 (2012).
30. Mathieson, I. & McVean, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**, 973–984 (2013).
31. Feder, A. F., Kryazhimskiy, S. & Plotkin, J. B. Identifying signatures of selection in genetic time series. *Genetics* **196**, 509–522 (2014).
32. Steinrücken, M., Bhaskar, A. & Song, Y. S. A novel spectral method for inferring general diploid selection from time series genetic data. *The annals of applied statistics* **8**, 2203 (2014).
33. Mathieson, I. & Terhorst, J. Direct detection of natural selection in bronze age britain. *Genome Research* **32**, 2057–2067 (2022).
34. He, Z., Dai, X., Lyu, W., Beaumont, M. & Yu, F. Estimating temporally variable selection intensity from ancient dna data. *Molecular Biology and Evolution* **40**, msad008 (2023).
35. Kimura, M. Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232 (1964).
36. Ewens, W. J. *Mathematical population genetics: theoretical introduction*, vol. 27 (Springer, 2004).
37. Schraiber, J. G., Evans, S. N. & Slatkin, M. Bayesian inference of natural selection from allele frequency time series. *Genetics* **203**, 493–511 (2016).
38. He, Z., Dai, X., Beaumont, M. & Yu, F. Estimation of natural selection and allele age from time series allele frequency data using a novel likelihood-based approach. *Genetics* **216**, 463–480 (2020).
39. Illingworth, C. J. & Mustonen, V. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics* **189**, 989–1000 (2011).
40. Illingworth, C. J. & Mustonen, V. A method to infer positive selection from marker dynamics in an asexual population. *Bioinformatics* **28**, 831–837 (2012).
41. Illingworth, C. J., Fischer, A. & Mustonen, V. Identifying selection in the within-host evolution of influenza using viral sequence data. *PLoS computational biology* **10**, e1003755 (2014).
42. Foll, M. *et al.* Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS genetics* **10**, e1004185 (2014).
43. Terhorst, J., Schlötterer, C. & Song, Y. S. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS genetics* **11**, e1005069 (2015).
44. Tataru, P., Simonsen, M., Bataillon, T. & Hobolth, A. Statistical inference in the wright–fisher model using allele frequency data. *Systematic biology* **66**, e30–e46 (2017).
45. Tataru, P., Mollion, M., Glémin, S. & Bataillon, T. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* **207**, 1103–1119 (2017).
46. He, Z., Dai, X., Beaumont, M. & Yu, F. Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies with forward-in-time simulations. *Genetics* **216**, 521–541 (2020).
47. Sohail, M. S., Louie, R. H., McKay, M. R. & Barton, J. P. Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature biotechnology* **39**, 472–479 (2021).
48. Buffalo, V. & Coop, G. The linked selection signature of rapid adaptation in temporal genomic data. *Genetics* **213**, 1007–1045 (2019).
49. Buffalo, V. & Coop, G. Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences* **117**, 20672–20680 (2020).
50. Fisher, R. A. *The genetical theory of natural selection: a complete variorum edition* (Oxford University Press, 1999).
51. Wright, S. Evolution in mendelian populations. *Genetics* **16**, 97 (1931).
52. Crow, J. F. *An introduction to population genetics theory* (Scientific Publishers, 2017).
53. Shimagaki, K. & Barton, J. P. Bézier interpolation improves the inference of dynamical models from data. *Physical Review E* **107**, 024116 (2023).
54. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature methods* **7**, 741–746 (2010).
55. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nature protocols* **11**, 1782–1787 (2016).
56. Shimagaki, K. S., Lynch, R. M. & Barton, J. P. Parallel hiv-1 evolutionary dynamics in humans and rhesus macaques who develop broadly neutralizing antibodies. *bioRxiv* 2024–07 (2024).
57. Hong, Z., Shimagaki, K. S. & Barton, J. P. popdms infers mutation effects from deep mutational scanning data. *Bioinformatics* **40**, btae499 (2024).
58. Tomaras, G. D. *et al.* Initial b-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin m (igm) and igg antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *Journal of virology* **82**, 12449–12463 (2008).
59. Liu, M. K. *et al.* Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. *The Journal of clinical investigation* **123** (2012).
60. Zanini, F., Puller, V., Brodin, J., Albert, J. & Neher, R. A. In vivo mutation rates and the landscape of fitness costs of hiv-1. *Virus evolution* **3**, vex003 (2017).
61. Neher, R. A. & Leitner, T. Recombination rate and selection strength in hiv intra-patient evolution. *PLoS computational biology* **6**, e1000660 (2010).
62. Shimagaki, K. & Weigt, M. Selection of sequence motifs and generative hopfield-potts models for protein families. *Physical Review E* **100**, 032128 (2019).
63. Posfai, A., Zhou, J., McCandlish, D. M. & Kinney, J. B. Gauge fixing for sequence-function relationships. *bioRxiv* (2024).
64. Posfai, A., McCandlish, D. M. & Kinney, J. B. Symmetry, gauge freedoms, and the interpretability of sequence-function relationships. *bioRxiv* (2024).
65. McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N. & Haynes, B. F. The immune response during acute hiv-1 infection: clues for vaccine development. *Nature Reviews Immunology* **10**, 11–23 (2010).
66. Allen, T. M. *et al.* 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. *Journal of virology* **79**, 13239–13249 (2005).
67. Zanini, F. *et al.* Population genomics of intrapatient hiv-1 evolution. *Elife* **4**, e11282 (2015).
68. Lee, J. K. *et al.* T cell cross-reactivity and conformational changes during tcr engagement. *The Journal of experimental medicine* **200**, 1455–1466 (2004).
69. Huseby, E. S. *et al.* How the t cell repertoire becomes peptide and mhc specific. *Cell* **122**, 247–260 (2005).
70. Huseby, E. S., Crawford, F., White, J., Marrack, P. & Kappler, J. W. Interface-disrupting amino acids establish specificity between t cell receptors and complexes of major histocompatibility complex and peptide. *Nature immunology* **7**, 1191–1199 (2006).
71. Gao, Y. & Barton, J. P. A binary trait model reveals the fitness effects of hiv-1 escape from t cell responses. *Proceedings of the National Academy of Sciences* **122**, e2405379122 (2025).
72. Sanjuán, R., Cuevas, J. M., Moya, A. & Elena, S. F. Epistasis and the adaptability of an rna virus. *Genetics* **170**, 1001–1008 (2005).
73. Elena, S. F., Solé, R. V. & Sardanyés, J. Simple genomes, complex interactions: epistasis in rna virus. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20** (2010).
74. Polster, R., Petropoulos, C. J., Bonhoeffer, S. & Guillaume, F. Epistasis and pleiotropy affect the modularity of the genotype–phenotype map of cross-resistance in hiv-1. *Molecular biology and evolution* **33**, 3213–3225 (2016).
75. Lee, B. *et al.* Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. *Nature Communications* **16**, 441 (2025).
76. Liu, Y. *et al.* Selection on the human immunodeficiency virus type 1 proteome

following primary infection. *Journal of Virology* **80**, 9519–9529 (2006).
77. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332**, 1193–1196 (2011).
78. Chou, H.-H., Chiu, H.-C., Delaney, N. F., Segrè, D. & Marx, C. J. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* **332**, 1190–1192 (2011).
79. Johnson, M. S., Reddy, G. & Desai, M. M. Epistasis and evolution: recent advances and an outlook for prediction. *BMC biology* **21**, 120 (2023).
80. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67–72 (2009).
81. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293–E1301 (2011).
82. Ferguson, A. L. *et al.* Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
83. Mann, J. K. *et al.* The fitness landscape of hiv-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS computational biology* **10**, e1003776 (2014).
84. Biswas, A., Haldane, A., Arnold, E. & Levy, R. M. Epistasis and entrenchment of drug resistance in hiv-1 subtype b. *Elife* **8**, e50524 (2019).
85. Rodriguez Horta, E., Barrat-Charlaix, P. & Weigt, M. Toward inferring potts models for phylogenetically correlated sequence data. *Entropy* **21**, 1090 (2019).
86. Colavin, A., Atolia, E., Bitbol, A.-F. & Huang, K. C. Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. *Scientific Reports* **12**, 820 (2022).
87. Li, Y. & Barton, J. P. Estimating linkage disequilibrium and selection from allele frequency trajectories. *Genetics* **223**, iyac189 (2023).
88. Li, Y. & Barton, J. P. Correlated allele frequency changes reveal clonal structure and selection in temporal genetic data. *Molecular Biology and Evolution* **41**, msae060 (2024).

Supplementary Information

Methods

Model

We model the evolution of a population of N individuals subject to mutation, recombination, natural selection, and genetic drift (finite population size), following the Wright-Fisher (WF) model [1–3]. We assume that all individuals are haploids. Each genotype $\boldsymbol{g} \in \mathcal{A}^L$ is a sequence of L alleles, with the alleles considered to be categorical variables. In the main text, we assumed binary mutant/WT alleles for simplicity, but for real sequence data one could choose $\mathcal{A} = \{\text{A, T, G, C, -}\}$ for DNA or $\mathcal{A} = \{\text{A, C,}\ldots\text{, W, Y, -}\}$ for amino acids, for example. In our description below, we will keep the notation general and use alphabetical indices $a, b, c, \ldots$ for alleles. We use Greek characters $\alpha, \beta, \ldots$ to index different genotypes and indices $i, j, k, \ldots$ to index sites along the genetic sequence. As described in the main text, we write the frequency of individuals with genotype α at time t as $z_\alpha(t)$. We write the M-dimensional vector of all genotype frequencies as $\boldsymbol{z}(t)$, which satisfies $\sum_\alpha^M z_\alpha(t) = 1$. In principle, all genotype (and allele, ...) frequencies are functions of time. However, when the context is clear, we will sometimes omit the time specification for simplicity.

Fitness function

We write the fitness of a genotype α as

$$f^\alpha = f(\boldsymbol{g}^\alpha) = 1 + \sum_i s_i(g_i^\alpha) + \sum_{i<j} s_{ij}(g_i^\alpha, g_j^\alpha)\,. \tag{S1}$$

The $s_i(a)$ are additive selection coefficients that quantify the fitness effect of each allele a at each site i. Similarly, the $s_{ij}(a,b)$ are epistatic interactions between pairs of alleles a and b at sites i and j, respectively. Introducing the indicator function $g_{i,a}^\alpha$ that is equal to one if genotype α has allele a at site i and zero otherwise, this fitness function can be expressed in the same form as in Eq. (1) in the main text:

$$f^\alpha = 1 + \sum_{i,a} s_i(a) g_{i,a}^\alpha + \sum_{i<j} \sum_{a,b} s_{ij}(a,b) g_{i,a}^\alpha g_{j,b}^\alpha = 1 + \sum_e s_e g_e^\alpha + \sum_{e<f} s_{ef} g_e^\alpha g_f^\alpha\,,$$

where the generic indices e and f run over both single and pairs of site and allele indices. For example, for DNA sequences, the e and f indices would run over

$$\Big((1,A),(1,T),(1,G),(1,C),(1,-),(2,A),\ldots,(L,-),((1,A),(2,A)),((1,A),(2,T)),\ldots,((L-1,-),(L,-))\Big)\,.$$

We assume the underlying fitness parameters are constant in time, with the population's adaptation speed much faster than the rate of environmental changes.

Mutation and recombination

We write the probability per generation to mutate from genotype α to genotype β as $\mu_{\alpha\beta}$, which need not be symmetric (i.e., in general $\mu_{\alpha\beta} \neq \mu_{\beta\alpha}$).

Additionally, we use a simple recombination model, in which there is a probability r per site per generation that a recombination breakpoint will occur at that site. A recombinant sequence derived from two sequences $\boldsymbol{g}^\alpha$ and $\boldsymbol{g}^\beta$ with recombination breakpoint i then has the form $(g_1^\alpha, \ldots, g_i^\alpha, g_{i+1}^\beta, \ldots, g_L^\beta)$. We assume that the partner sequence β is always chosen randomly from the population with a probability proportional to its current frequency.

Given a population with genotype frequencies $\boldsymbol{z}$, their expected frequencies after recombination $\boldsymbol{y}$ are given by

$$y_\alpha(\boldsymbol{z}) = (1-r)^{L-1} z_\alpha + \Big(1 - (1-r)^{L-1}\Big) \sum_{\beta,\gamma} R_{\alpha,\beta\gamma} z_\beta z_\gamma\,. \tag{S2}$$

Here, $R_{\alpha,\beta\gamma}$ is the probability that the genotypes β and γ recombine to another genotype α in a recombination event that features a single breakpoint. The first term in Eq. (S2) represents the probability that the genotype remains unchanged because no recombination occurs at any possible breakpoint. The second term in Eq. (S2) represents a single recombination event that occurs with probability $1-(1-r)^{L-1}$, resulting in the genotype α as a recombinant. In principle, higher order terms could also be included that allow for the possibility of multiple breakpoints along the sequence. However, below we will assume that the recombination probability per site r is small and treat it perturbatively, suppressing terms of order r^2 and above. This approach should be modified for very distantly spaced loci or high rates of recombination.

Dynamics

Assuming a fixed population size N, the WF model dynamics are multinomial:

$$P(\boldsymbol{z}(t+1)|\boldsymbol{z}(t);N,\boldsymbol{s},\mu,r) = N!\prod_{\alpha=1}^{M}\frac{p_\alpha(\boldsymbol{z}(t))^{Nz_\alpha(t+1)}}{[Nz_\alpha(t+1)]!}\,, \tag{S3}$$

where p_α is the relative probability of "drawing" sequence α after mutation, recombination and selection, given by

$$p_\alpha(\boldsymbol{z}) = \frac{f^\alpha y_\alpha + \sum_\beta(\mu_{\beta\alpha}y_\beta - \mu_{\alpha\beta}y_\alpha)}{\sum_\beta f^\beta y_\beta}\,. \tag{S4}$$

The multinomial distribution reflects the stochastic dynamics inherent in a finite-sized population. As noted above, the y_α are genotype frequencies after recombination, which depend on the current distribution of genotype frequencies $\boldsymbol{z}$; see Eq. (S2).

Diffusion limit

To simplify the dynamics, we will assume that the mutation rate, recombination rate, and selection coefficients are of order $\mathcal{O}(1/N)$ while the population size N is large. We define the frequency change between subsequent generations t and $t+1$ as $\boldsymbol{\Delta z}(t) = \boldsymbol{z}(t+1) - \boldsymbol{z}(t)$. Under the WF model, the first and second moments of the change in genotype frequencies are (to leading order in $1/N$)

$$\begin{aligned}\langle \Delta z_\alpha(t)\rangle &= \sum_{\boldsymbol{z}(t+1)} \Delta z_\alpha(t) P(\boldsymbol{z}(t+1)\mid \boldsymbol{z}(t);N) = d_\alpha(\boldsymbol{z}(t))\,,\\ \big\langle \Delta z_\alpha(t)\Delta z_\beta(t)\big\rangle &= \sum_{\boldsymbol{z}(t+1)} \Delta z_\alpha(t)\Delta z_\beta(t) P(\boldsymbol{z}(t+1)\mid \boldsymbol{z}(t);N) = C_{\alpha\beta}(\boldsymbol{z}(t))/N\,,\end{aligned} \tag{S5}$$

with

$$d_\alpha(\boldsymbol{z}) = C_{\alpha\alpha}s_\alpha + \sum_{\beta(\neq\alpha)} C_{\alpha\beta}s_\beta + \sum_\beta(\mu_{\beta\alpha}z_\beta - \mu_{\alpha\beta}z_\alpha) - r(L-1)\left(z_\alpha - \sum_{\beta\gamma} R_{\alpha,\beta\gamma}z_\beta z_\gamma\right) + \mathcal{O}(1/N^2) \tag{S6}$$

and

$$C_{\alpha\beta}(\boldsymbol{z}) = \begin{cases} -p_\alpha p_\beta = -z_\alpha z_\beta + \mathcal{O}(1/N) & \alpha\neq\beta \\ (1-p_\alpha)p_\alpha = (1-z_\alpha)z_\alpha + \mathcal{O}(1/N) & \alpha=\beta \end{cases}\,. \tag{S7}$$

Here, as previously defined, $P(\boldsymbol{z}(t+1)|\boldsymbol{z}(t);N)$ is the multinomial distribution (other conditional arguments are omitted for simplicity), $d_\alpha(\boldsymbol{z})$ and $C_{\alpha\beta}(\boldsymbol{z})$ are the expected change in genotype frequency and the genotype covariance between α and β. As mentioned above, assuming that the rates of mutation and recombination and the fitness effects of mutations are as small as $\mathcal{O}(1/N)$, changes in genotype frequencies are not abrupt, and we can employ the diffusion approximation[4] to simplify the WF model. This results in the following Kimura's diffusion equation (Fokker-Planck equation or Kolmogorov forward equation[5]):

$$\partial_t P(\boldsymbol{z};t) = \frac{1}{2}\sum_{\alpha,\beta}\partial_\alpha\partial_\beta C_{\alpha\beta}(\boldsymbol{z})P(\boldsymbol{z};t) - \sum_\alpha \partial_\alpha d_\alpha(\boldsymbol{z})P(\boldsymbol{z};t)\,. \tag{S8}$$

The above equation represents a continuous-time Gaussian process with a drift vector $\boldsymbol{d}$ and diffusion matrix C given by Eq. (S5). When $N\gg 1$, the genotype frequency transition probability can be written as

$$P(\boldsymbol{z}(t+\Delta t)\mid \boldsymbol{z}(t)) \propto \exp\left(-\frac{N}{2\Delta t}\left(\boldsymbol{\Delta z}(t) - \Delta t\boldsymbol{d}(\boldsymbol{z}(t))\right)^\top C(\boldsymbol{z}(t))^{-1}\left(\boldsymbol{\Delta z}(t) - \Delta t\boldsymbol{d}(\boldsymbol{z}(t))\right)\right)\,, \tag{S9}$$

with $\boldsymbol{\Delta z}(t) = \boldsymbol{z}(t+\Delta t) - \boldsymbol{z}(t)$. In the main text, Eq. (3) gives the optimal selection coefficients $\hat{s}_i$ and epistatic interactions $\hat{s}_{ij}$ to maximize a posterior distribution over the above diffusion processes.

So far, we have discussed the diffusion process in the genotype distribution space. To make the expressions more transparent, we project the genotype frequency dynamics onto the allele frequency space, Eq. (2). As the detailed derivations are provided in our previous work[6], we present only the main results below. Interested readers may refer to ref.[6] for further details.

Expected frequency change due to mutation

Assuming the WF process, we can compute the expected frequency change due to mutation and recombination effects and integrate them over the generations. Since the mutation rate is small, to leading order in $1/N$, no more than one mutation occurs per generation for each individual. Therefore, only mutations between genotypes α and β that differ by a Hamming distance $D_{\alpha,\beta} = L - \sum_{i,a} g^{\alpha}_{i,a} g^{\beta}_{i,a}$ of one are possible. Suppose that mutations occur randomly at each site with a site-independent probability μ_{ab} of mutating from allele a to allele b. The expected frequency change in allele a at site i due to mutation is then

$$\begin{aligned} u_{i,a} &= \sum_{\alpha,\beta|D_{\alpha,\beta}=1} \sum_{b} g^{\alpha}_{i,a} g^{\beta}_{i,b} (\mu_{\beta\alpha} z_{\beta} - \mu_{\alpha\beta} z_{\alpha}) \\ &= \sum_{b|b\neq a} \left(\mu_{ba} x_{i,a} - \mu_{ab} x_{i,b}\right), \end{aligned} \tag{S10}$$

where the frequency of allele a at site i is $x_{i,a} = \sum_{\alpha} g^{\alpha}_{i,a} z_{\alpha}$. Similarly, for pairwise frequencies of alleles a and b at sites i and j, respectively, written $x_{ij,ab} = \sum_{\alpha} g^{\alpha}_{i,a} g^{\alpha}_{j,b} z_{\alpha}$, we obtain

$$\begin{aligned} u_{ij,ab} &= \sum_{\alpha,\beta|D_{\alpha,\beta}=1} g^{\alpha}_{i,a} g^{\alpha}_{j,b} \left(\sum_{c|c\neq a} g^{\beta}_{i,c} g^{\beta}_{j,b} + \sum_{c|c\neq b} g^{\beta}_{i,a} g^{\beta}_{j,c} \right) (\mu_{\beta\alpha} z_{\beta} - \mu_{\alpha\beta} z_{\alpha}) \\ &= \sum_{c} \Big([\mu_{bc} x_{ij,ac} + \mu_{ac} x_{ij,cb}] - [\mu_{cb} + \mu_{ca}] x_{ij,ab} \Big). \end{aligned} \tag{S11}$$

Expected frequency change due to recombination

Similarly, we project the genotype frequency changes due to recombination, as shown in the drift term Eq. (S9), into allele space. More comprehensive derivations can be found in ref.[6], Eqs. S39 – S51. By symmetry, one can show that recombination has no effect on the expected change in individual allele frequencies[7],

$$v_{i,a} = r(L-1) \sum_{\alpha} g^{\alpha}_{i,a} \left(z_{\alpha} - \sum_{\beta\gamma} R_{\alpha,\beta\gamma} z_{\beta} z_{\gamma} \right) = 0 . \tag{S12}$$

However, linkage disequilibrium is naturally diluted by recombination. For pairwise frequencies, recombination decreases correlations between mutations until they become independent. One can show that the expected change in pairwise allele frequencies due to recombination is[6]

$$\begin{aligned} v_{ij,ab} &= r(L-1) \sum_{\alpha} g^{\alpha}_{i,a} g^{\alpha}_{j,b} \left(z_{\alpha} - \sum_{\beta\gamma} R_{\alpha,\beta\gamma} z_{\beta} z_{\gamma} \right) \\ &= -r|i-j| (x_{ij,ab} - x_{i,a} x_{j,b}) . \end{aligned} \tag{S13}$$

Ref.[6] provides more detailed descriptions.

Maximum *a posteriori* solution over the path

Leveraging the analytically tractable transition probability under the diffusion limit, Eq. (S9), the maximum *a posteriori* estimate of the selection coefficients and epistatic interactions is[6]

$$\begin{aligned} \hat{\boldsymbol{s}} &= \arg\max_{s} P(\boldsymbol{s}|\boldsymbol{\gamma}) \prod_{k=0}^{K} P(\boldsymbol{z}(t_{k+1}) \mid \boldsymbol{z}(t_k); N, \boldsymbol{s}, \mu, r) \\ &= (C^{\text{int}} + \text{diag}(\boldsymbol{\gamma}))^{-1} \left(\boldsymbol{x}(t_{K+1}) - \boldsymbol{x}(t_0) - \boldsymbol{u}^{\text{int}} - \boldsymbol{v}^{\text{int}} \right) . \end{aligned} \tag{S14}$$

Here $\text{diag}(\boldsymbol{\gamma})$ is a diagonal matrix with $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L, , \gamma_{1,2}, \ldots, \gamma_{L-1,L}) = (\gamma_e)_e$, with $\gamma_e > 0$ on the diagonal and zeros elsewhere. As we defined in the main text and previous sections (Eq. (S10), Eq. (S11) and Eq. (S13)), $\boldsymbol{u}^{\text{int}}$ and $\boldsymbol{v}^{\text{int}}$ are the expected net allele frequency changes due to mutation and recombination effects over the course of evolution. $P(\boldsymbol{s} \mid \boldsymbol{\gamma})$ represents a prior distribution for the selection and epistatic coefficients, given by the normal distribution

$$P(\boldsymbol{s} \mid \boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2} \boldsymbol{s}^{\top} \text{diag}(\boldsymbol{\gamma}) \boldsymbol{s} \right) . \tag{S15}$$

Representing the integrated covariance matrix with linear interpolation by a low-rank matrix

In real data, samples are collected at discrete intervals rather than in continuous time. Interpolating both the allele frequencies and the covariance matrix C consistently improves inference accuracy in finitely sampled data. In this section, we show the expression of the integrated covariance matrix with piece-wise linear interpolation is given as the integration of covariance with a piece-wise constant interpolation and a sum of rank-one matrices. The explicit expression for the allele frequency covariance matrix is given Eq. (4). Assuming a linear interpolation of (both additive and pairwise) allele frequencies between time points, the integrated covariance matrix is

$$\begin{aligned} C^{\text{int}} &= \sum_{k=0}^{K-1} \Delta t_k \int_0^1 C^{[k,k+1]}(\tau)\, d\tau \\ C_{e,f}^{[k,k+1]}(\tau) &= x_{ef}^{[k,k+1]}(\tau) - x_e^{[k,k+1]}(\tau)\, x_f^{[k,k+1]}(\tau) \\ x_e^{[k,k+1]}(\tau) &= (1-\tau)x_e(t_k) + \tau x_e(t_{k+1})\,. \end{aligned} \tag{S16}$$

It is straightforward to verify that the following expression is identical to the diagonal of the integrated covariance matrix with the piece-wise linear interpolation given in ref.[7]:

$$\begin{aligned} &\frac{\Delta t_0}{2} C_{ii}(t_0) + \frac{\Delta t_{K-1}}{2} C_{ii}(t_K) + \sum_{k=1}^{K-1} \frac{\Delta t_k + \Delta t_{k-1}}{2} C_{ii}(t_k) + \sum_{k=1}^{K-1} \frac{\Delta t_k}{6} (\Delta x(t_k))^2 \\ &= \sum_{k=0}^{K-1} \frac{\Delta t_k}{2} \Big(x_i(t_k)(1-x_i(t_k)) + x_i(t_{k+1})(1-x_i(t_{k+1})) \Big) + \sum_{k=1}^{K-1} \frac{\Delta t_k}{6} (x_i(t_{k+1}) - x_i(t_k))^2 \\ &= \sum_{k=0}^{K-1} \Delta t_k \left(\frac{x_i(t_k) + x_i(t_{k+1})}{2} - \frac{x_i(t_{k+1})(x_i(t_k) + x_i(t_{k+1}))}{3} - \frac{x_i(t_k)^2}{3} \right) \\ &= \sum_{k=0}^{K-1} \Delta t_k \left(\frac{(3 - 2x_i(t_k))(x_i(t_k) + x_i(t_{k+1}))}{6} - \frac{x_i(t_k)^2}{3} \right) = C_{ii}^{\text{int}}\,. \end{aligned} \tag{S17}$$

For the off-diagonal case, the pairwise-frequency term $x_{ij}(t_k)$ is linear in time, and the result of the integral with the linear interpolation is the same as the integral with the piecewise-constant interpolation. Therefore, we explicitly write only the integrals that are non-linear in time:

$$\begin{aligned} &\frac{\Delta t_0}{2} x_i(t_0) x_j(t_0) + \frac{\Delta t_{K-1}}{2} x_i(t_K) x_j(t_K) + \sum_{k=1}^{K-1} \frac{\Delta t_k + \Delta t_{k-1}}{2} x_i(t_k) x_j(t_k) - \sum_{k=1}^{K-1} \frac{\Delta t_k}{6} \Delta x_i(t_k) \Delta x_j(t_k) \\ &= \sum_{k=0}^{K-1} \Delta t_k \left(\frac{x_i(t_k) x_j(t_k) + x_i(t_{k+1}) x_j(t_{k+1})}{2} - \frac{(x_i(t_{k+1}) - x_i(t_k))(x_j(t_{k+1}) - x_j(t_k))}{6} \right) \\ &= \sum_{k=0}^{K-1} \Delta t_k \left(\frac{x_i(t_k) x_j(t_k) + x_i(t_{k+1}) x_j(t_{k+1})}{3} + \frac{x_i(t_k) x_j(t_{k+1}) + x_i(t_{k+1}) x_j(t_k)}{6} \right) \\ &= -C_{ij}^{\text{int}} + (\text{integrated pairwise frequency matrix})\,. \end{aligned} \tag{S18}$$

By summarizing these equations, we represent the integrated covariance matrix (Eq. (S16)) as:

$$C^{\text{int}} = \frac{\Delta t_0}{2} C(t_0) + \frac{\Delta t_{K-1}}{2} C(t_K) + \sum_{k=1}^{K-1} \frac{t_{k+1} - t_{k-1}}{2} C(t_k) + \sum_{k=1}^{K-1} \frac{\Delta t_k}{6} \boldsymbol{\Delta x}(t_k) \boldsymbol{\Delta x}(t_k)^\top\,, \tag{S19}$$

which we can readily factorize by a matrix Ξ such that $C^{\text{int}} = \Xi\Xi^\top$. The size of the matrix Ξ is $D \times d$, where $d = \sum_{k=0}^{K} d(t_k) + K - 1$ with $d(t_k)$ denoted as a rank of $C(t_k)$. In most of the evolutionary data, the size of the higher-order covariance matrix is much larger than the effective matrix rank size; hence, typically $d \ll D$.

Inferring fitness parameters from multiple replicate trajectories

In cases where multiple ensembles of trajectories evolve under similar conditions, it is natural to extend the path likelihood to multiple ensembles. Suppose there are Q replicates, let q be the index of the q-th replicate, $(t_k)_{k=1}^{K_q}$ be a set of K_q sampling time-steps for the q-th replicate, and $\boldsymbol{x}^q(t_k)$ be the set of single and pairwise frequencies for the q-th replicate.

The maximum path likelihood solution using Q replicates can be expressed as[6]

$$\begin{aligned} \boldsymbol{s} &= \gamma^{-1}(\boldsymbol{\Delta x} - \Xi \boldsymbol{r}) \\ \boldsymbol{r} &= \left(\Xi\Xi^\top + \gamma I\right)^{-1} \Xi^\top \boldsymbol{\Delta x}\,, \end{aligned} \tag{S20}$$

where I represents the identity matrix and

$$\begin{aligned} \boldsymbol{\Delta x} &= \sum_{q=1}^{Q} \boldsymbol{\Delta x}^q \\ \Xi^\top &= \left(\Xi^{1\,\top}, \ldots, \Xi^{Q\,\top}\right) \in \mathbb{R}^{D\times B}\,. \end{aligned} \tag{S21}$$

B is the total number of samples across replicates over the evolution, formally, $B = \sum_q^Q B^q$, where B^q is the total number of samples of the q-th replicate over its evolution. Intuitively, the likelihood of multiple independent trajectories is equal to the product of each of their likelihoods individually.

Gauge transformation

The effects of natural selection are determined by differences in fitness values, such as the difference between the fitness of the wild type and a mutant. Shifting the fitness values globally by adding a constant, $F(\boldsymbol{g}) \leftarrow F(\boldsymbol{g}) + \text{const.}$, has no effect on fitness differences. In the additive fitness model, it is easy to see that shifting the selection coefficient at any locus by an arbitrary constant K_i does not alter the relative fitness landscape: $\sum_{i,a}(s_i(a) - K_i)\delta_{g_i,a} = \sum_i s_i(g_i) + \text{const}$. In other words, the effective fitness parameters can be reduced to $(q-1)L$, and the degrees of freedom that can be arbitrarily adjusted without changing the overall fitness picture are L parameters. More systematic arguments under general situations exist and are known as gauge theory in physics and mathematical physics. These concepts have been applied to many genetic sequence-based inference problems[8–10], with recent reviews for gauge theory in more complex cases[11,12]. In our study, mutation effects of the wild type or TF's allele serve as reference values; therefore, considering any effects involved with TF's alleles being zeros is a reasonable choice and makes the inference results more interpretable, as inferred parameters become sparser. To fix the gauge, we employed the following gauge transformation, which is commonly used in statistical inference for genetic sequences[10,13,14],

$$\begin{aligned} s_i(a) &\leftarrow s_i(a) - s_i(a^{\mathrm{WT}}) + \sum_{j\ (>i)} \left(s_{ij}(a, b^{\mathrm{WT}}) - s_{ij}(a^{\mathrm{WT}}, b^{\mathrm{WT}})\right) + \sum_{j\ (<i)} \left(s_{ji}(b^{\mathrm{WT}}, a) - s_{ji}(b^{\mathrm{WT}}, a^{\mathrm{WT}})\right) \\ s_{ij}(a,b) &\leftarrow s_{ij}(a,b) - s_{ij}(a^{\mathrm{WT}}, b) - s_{ij}(a, b^{\mathrm{WT}}) + s_{ij}(a^{\mathrm{WT}}, b^{\mathrm{WT}})\,, \end{aligned} \tag{S22}$$

where $a^{\mathrm{WT}}, b^{\mathrm{WT}}$ are WT (i.e., TF) alleles at locus i and j, respectively. This choice of gauge ensures $s_i(a^{\mathrm{WT}}) = s_{ij}(a^{\mathrm{WT}}, b) = s_{ij}(a, b^{\mathrm{WT}}) = 0$ for all a, b.

Further compression of Ξ

Although the size of the matrix $\Xi \in \mathbb{R}^{D\times d}$ (where D represents the size of the covariance matrix, and d denotes its matrix rank) is much smaller than the size of the full covariance matrix $C^{\mathrm{int}} \in \mathbb{R}^{D\times D}$ with $d \ll D$, still keeping Ξ can be the major bottleneck. An example is HIV-1 CH848 data, where >1200 sequences were collected sequencing more than half of the HIV-1 genome. When we naively compute Ξ storing float variables, that requires about a terabyte of memory. To further reduce memory usage, we only consider alleles with nonzero frequency change over time, satisfying $\sum_k |\Delta x_i(t_k)| > 0$.

Heterogeneous regularization

Instead of applying constant regularization across all parameters, we use a generalized heterogeneous regularization approach with the HCMF method: $\gamma \to \boldsymbol{\gamma} = (\gamma_e)_{e=1}^D$. By denoting $\Lambda_{\boldsymbol{\gamma}}$ as a diagonal matrix with $(\Lambda_{\boldsymbol{\gamma}})_{ef} = \gamma_e \delta_{ef}$, then the optimal fitness parameter becomes $\hat{s} = (C^{\mathrm{int}} + \Lambda_{\boldsymbol{\gamma}})^{-1}\boldsymbol{\Delta x}$, where C^{int} and $\boldsymbol{\Delta x}$ are integrated covariance matrix and allele frequency changes, as we defined earlier. Consequently, the expression for the efficient expression becomes:

$$\begin{aligned} \hat{\boldsymbol{s}} &= \Lambda_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\Delta x} - \Xi \boldsymbol{\Delta\eta}) \\ \boldsymbol{\Delta\eta} &= \left(\Xi^\top \Lambda_{\boldsymbol{\gamma}}^{-1} \Xi + I\right)^{-1} (\Lambda_{\boldsymbol{\gamma}}^{-1}\Xi)^\top \boldsymbol{\Delta x}\,. \end{aligned} \tag{S23}$$

Theoretical error bars

The posterior distribution is given by the Bayes' rule,

$$P(\boldsymbol{s}|(\boldsymbol{x}(t_k))_{k=0}^K) \propto \exp\left(-\frac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{s}})^\top \Sigma^{-1} (\boldsymbol{s} - \hat{\boldsymbol{s}})\right)\,, \tag{S24}$$

which is a normal distribution for the fitness parameters $\boldsymbol{s}$ with mean $\hat{\boldsymbol{s}}$ and a precision matrix $\Sigma^{-1} = (C^{\text{int}} + \gamma I)$. In a Bayesian inference framework, the uncertainty in the inferred fitness parameters is characterized by $(C^{\text{int}} + \gamma I)^{-1}$. More specifically, the diagonal entries of the covariance matrix can be thought of as the theoretical "error bars" on the inferred coefficients, as the standard deviation for s_e is given by $((C^{\text{int}} + \gamma I)^{-1})_{ee}$. Let $\tilde{\boldsymbol{\xi}}_e \in \mathbb{R}^d$ for $e \in \{1, \ldots, D\}$ be row vectors for Ξ, then by exploiting the structure of the integrated covariance matrix $C^{\text{int}} = \Xi\Xi^\top$, one can get

$$\text{Var}(s_e) = (\gamma N)^{-1} \left(1 - \tilde{\boldsymbol{\xi}}_e^\top \left(\Xi\Xi^\top + \gamma I\right)^{-1} \tilde{\boldsymbol{\xi}}_e\right). \tag{S25}$$

As the inverse in Eq. (S25) is easier to obtain, once the inverse is obtained, the variance of each s_e should be straightforwardly obtained.

While this expression can be useful to get a sense of the intrinsic uncertainty in the inferred parameters, the theoretical error term does not account for finite sampling. Thus, Eq. (S25) should be thought of as a lower bound on the true parameter uncertainty. To obtain more realistic error bars, we also performed bootstrap resampling of the data. Let n_k represent the number of sequences in a particular data set that were sampled at time t_k. To perform bootstrap resampling, at each time t_k we randomly sampled n_k sequences from the empirical distribution at the same time, with replacement. We then used these resampled sequences to estimate selection coefficients and epistatic interactions. After performing bootstrap resampling 20 times, we computed the mean and 90% confidence interval of the inferred parameters. These are the points and error bars shown in **Figs. 4-5** and **Figs. S2-3**.

HIV-1 data

We retrieved HIV-1 sequences for individuals in this study from the Los Alamos National Laboratory (LANL) HIV Sequence Database[15] (see **Table S1**). We processed the sequence data as described in ref.[7] to minimize the influence of sampling noise. Processing steps included 1) removing the sequences with large deletions, 2) removing sites with high gap frequencies (indicating rare insertions or potential alignment errors), and 3) eliminating time points with <4 sequences or ones that were obtained >200 days after the prior sampling time. In addition, we imputed ambiguous nucleotides using the most common nucleotides in the data at the same site within that individual.

Parameters for simulations and data analyses

In simulations, we used a mutation probability of $\mu = 10^{-3}$ per site per generation, such that $\mu_{\alpha\beta} = \mu^{D_{\alpha,\beta}}$, where $D_{\alpha,\beta}$ is the Hamming distance between binary genotypes α and β with mutant/wild-type alleles only; for real data, we consider mutation probabilities per site per generation μ_{ab} from allele a to allele b, which we will assume to be constant across sites. We use a simple effective recombination rate $r = 10^{-4}$ per site per generation. For HIV-1 data analysis, we used mutation rates estimated from a longitudinal virus evolution study[16], along with a constant recombination rate of $r = 10^{-5}$ per site per generation, in line with past estimates of the effective recombination rate[17–21]. This choice for representing recombination in HIV-1 is a simplification. In reality, HIV-1 recombination occurs in multiple steps: first, two different viruses must coinfect the same cell. Then, genetic material from each virus can be packaged together in the same virion. When such a virion infects a new cell, recombination can occur as the viral reverse transcriptase switches between templates. Thus, the effective HIV-1 recombination rate involves both coinfection and template switching probabilities. Recent work has also shown that the effective recombination rate can increase when viral load is higher, due to increased rates of coinfection[22]. Here we applied only the simple recombination model in which probabilities of coinfection and template switching are combined into a single effective recombination rate. Future work could relax this assumption and consider the effects of time-varying recombination rates due to fluctuations in viral load.

Data and code

Data and code used in our analysis is available in the GitHub repository located at `https://github.com/bartonlab/paper-hcmf`. This repository also contains scripts that can be run to reproduce our figures and analysis.

Supplementary References

1. Wright, S. Evolution in mendelian populations. *Genetics* **16**, 97 (1931).
2. Fisher, R. A. *The genetical theory of natural selection: a complete variorum edition* (Oxford University Press, 1999).
3. Ewens, W. J. *Mathematical population genetics: theoretical introduction*, vol. 27 (Springer, 2004).
4. Kimura, M. Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232 (1964).
5. Risken, H. & Risken, H. *Fokker-planck equation* (Springer, 1996).
6. Sohail, M. S., Louie, R. H., Hong, Z., Barton, J. P. & McKay, M. R. Inferring epistasis from genetic time-series data. *Molecular Biology and Evolution* **39**, msac199 (2022).
7. Sohail, M. S., Louie, R. H., McKay, M. R. & Barton, J. P. Mpl resolves genetic linkage in fitness inference from complex evolutionary histories. *Nature biotechnology* **39**, 472–479 (2021).
8. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67–72 (2009).

9. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293–E1301 (2011).
10. Rizzato, F. *et al.* Inference of compressed potts graphical models. *Physical Review E* **101**, 012309 (2020).
11. Posfai, A., McCandlish, D. M. & Kinney, J. B. Symmetry, gauge freedoms, and the interpretability of sequence-function relationships. *bioRxiv* (2024).
12. Posfai, A., Zhou, J., McCandlish, D. M. & Kinney, J. B. Gauge fixing for sequence-function relationships. *bioRxiv* (2024).
13. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics* **81**, 032601 (2018).
14. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **87**, 012707 (2013).
15. Los Alamos National Laboratory. Hiv sequence database (2023). URL https://www.hiv.lanl.gov. Accessed: 703010505 and 703010848 inpatient code.
16. Zanini, F., Puller, V., Brodin, J., Albert, J. & Neher, R. A. In vivo mutation rates and the landscape of fitness costs of hiv-1. *Virus evolution* **3**, vex003 (2017).
17. Sabino, E. C. *et al.* Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes b and f in two epidemiologically linked individuals from brazil. *Journal of Virology* **68**, 6340–6346 (1994).
18. Gao, F. *et al.* Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. *Journal of virology* **78**, 2426–2433 (2004).
19. Neher, R. A. & Leitner, T. Recombination rate and selection strength in hiv intra-patient evolution. *PLoS computational biology* **6**, e1000660 (2010).
20. Batorsky, R. *et al.* Estimate of effective recombination rate and average selection coefficient for hiv in chronic infection. *Proceedings of the National Academy of Sciences* **108**, 5661–5666 (2011).
21. Song, H. *et al.* Tracking hiv-1 recombination to resolve its contribution to hiv-1 evolution in natural infection. nat commun 9: 1928 (2018).
22. Romero, E. V. & Feder, A. F. Elevated hiv viral load is associated with higher recombination rate in vivo. *Molecular Biology and Evolution* **41**, msad260 (2024).

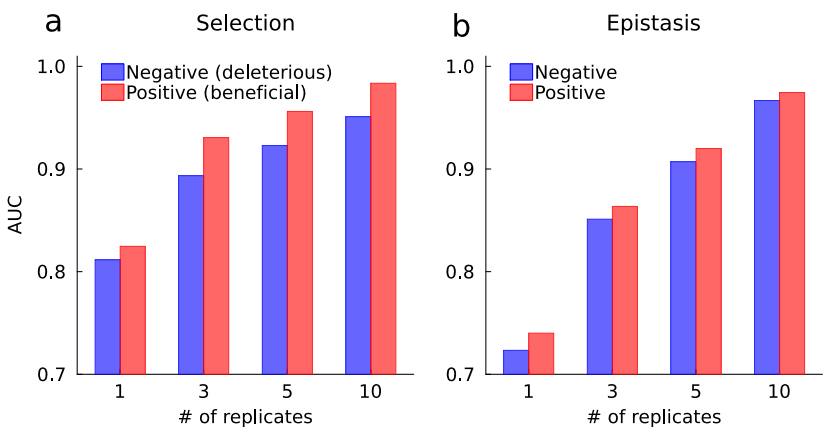

Fig. S1. Combining evolutionary replicates improves inference accuracy. **a**, AUC values for identifying selection coefficients as a function of the number of replicates. When using a single trajectory, the AUC values for beneficial and deleterious coefficients are 0.82 and 0.81, respectively. The AUC values for the inferred selection coefficients increase to 0.93 and 0.89, respectively, when combining two replicates. AUC values continue to increase as the number of replicates grows, reaching 0.98 and 0.95 for beneficial and deleterious coefficients with a set of 10 replicates. **b**, AUC values for identifying epistasic interactions from a single replicate are 0.74 for positive and 0.72 for negative epistasis. Similar to the case for selection coefficients, inference accuracy steadily improves with the addition of more replicates.

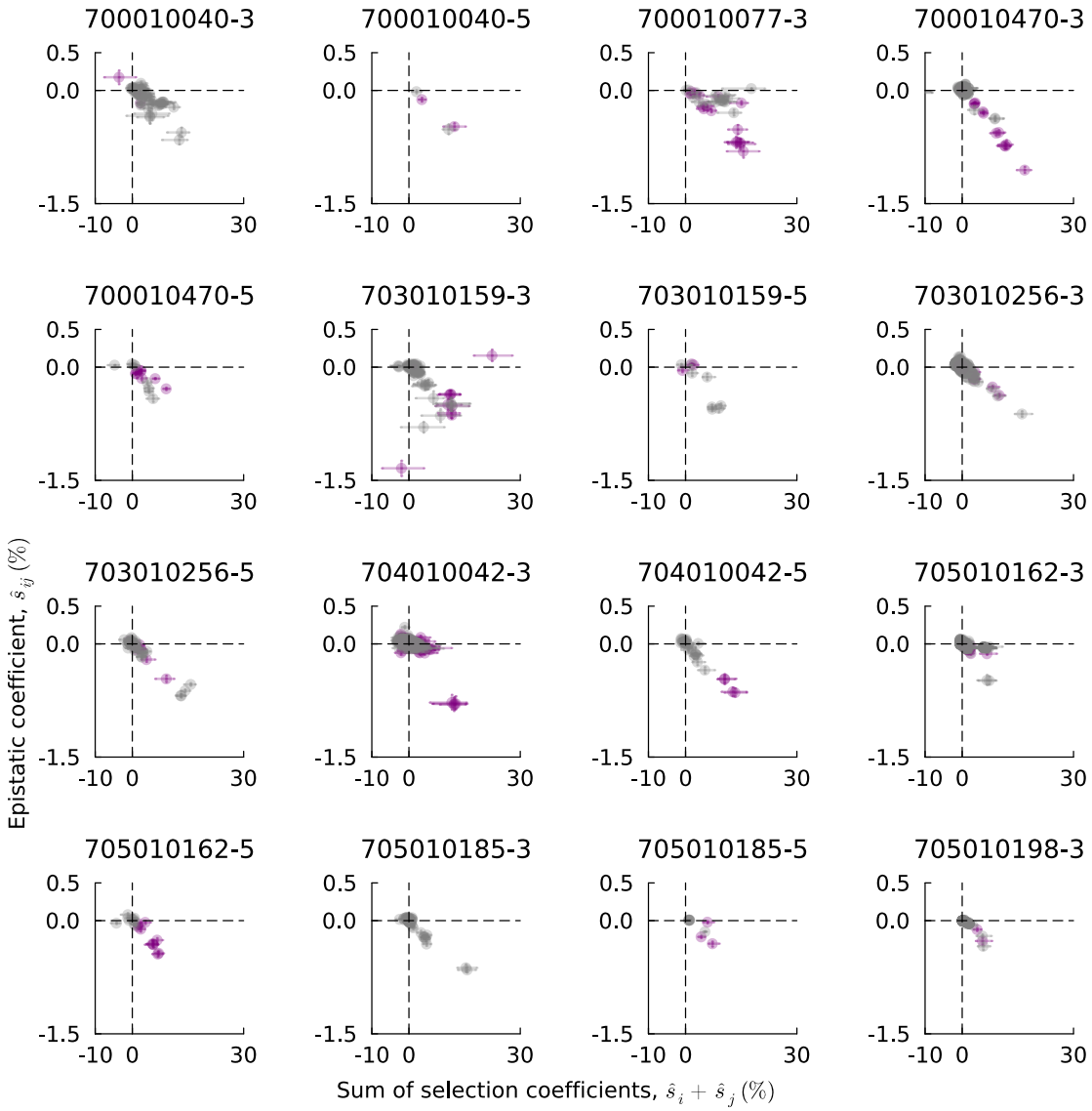

Fig. S2. Comparison of inferred epistatic coefficients and the sum of selection coefficients. These figures are analogous to **Fig. 4a** in the main text, but for all individuals and sequencing regions that we analyzed. The tendency of strong anticorrelation between epistasis and the sum of selection coefficients is widely observed across multiple individuals. Relatively strong negative epistatic coefficients are often seen among significantly beneficial mutations, many of which are involved in CTL escape. Error bars denote 90% confidence intervals across independently inferred selection and epistatic coefficients from 20 bootstrap samples.

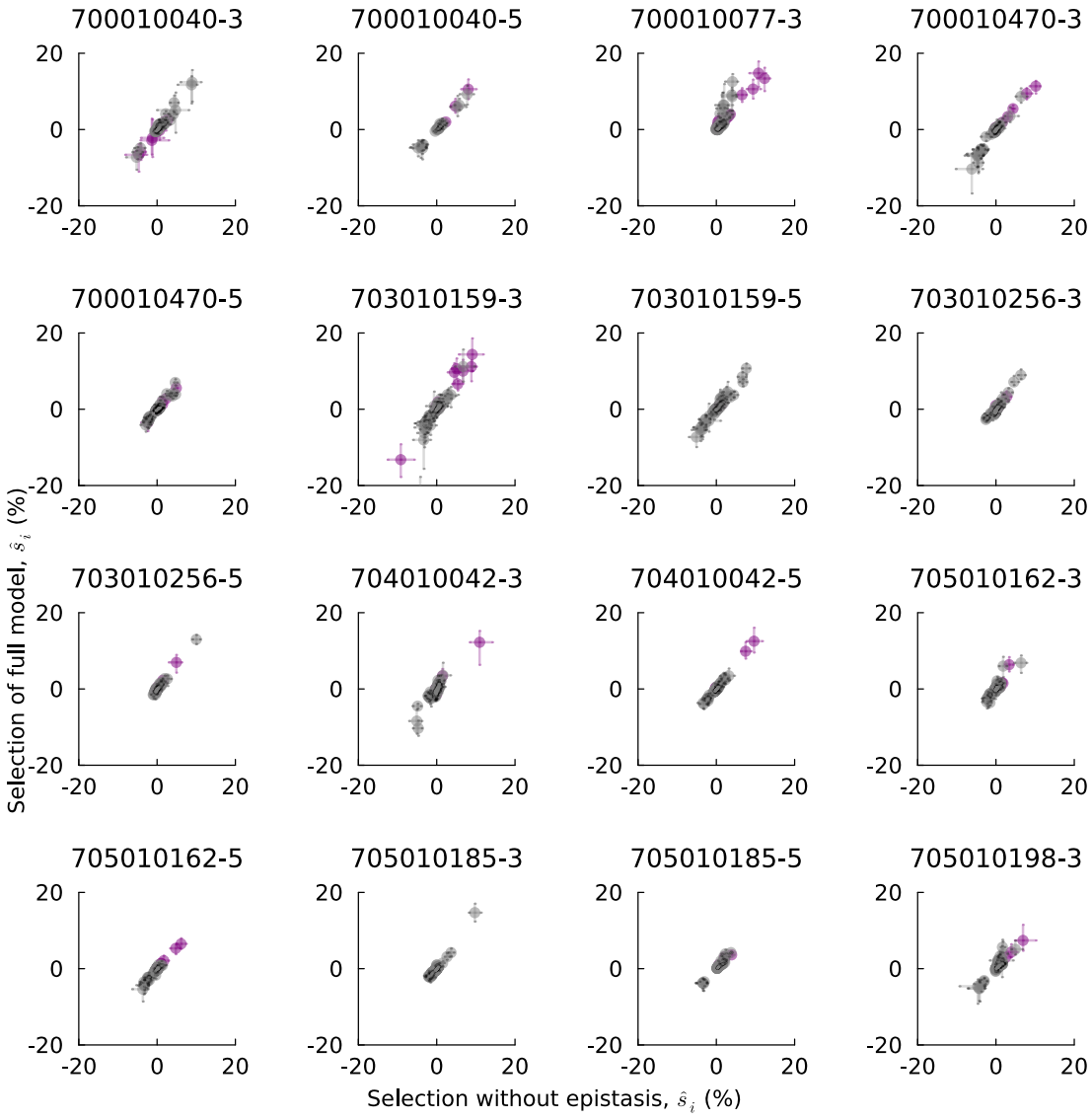

Fig. S3. Comparison of inferred selection coefficients in models with and without epistasis. This figure is analogous to **Fig. 5a** in the main text, but across other individuals and sequencing regions that we analyzed. The inferred selection coefficients learned with epistasis are globally consistent with those learned without epistasis. Relatively strong positive selection coefficients are often involved in mutations in CTL epitopes. Similar to **Fig. 5a**, error bars denote 90% confidence intervals across independently inferred selection and epistatic coefficients from 20 bootstrap samples.

ID	L	D	K	N	T_{HCMF} (sec)	M_{HCMF} (GB)	T_{Naive} (sec)	M_{Naive} (GB)
700010040-3	303	1.4×10^5	8	82	6.2	2.6	4.8×10^4	4.7×10^2
700010040-5	146	3.2×10^4	8	74	3.5	0.7	8.8×10^2	2.7×10^1
700010058-3	90	1.2×10^4	4	25	2.6	0.3	3.9×10^4	4.0×10^2
700010058-5	96	1.3×10^5	8	52	2.4	0.4	3.9×10^4	4.0×10^2
700010077-3	203	6.5×10^4	5	44	3.7	0.8	5.3×10^3	1.0×10^2
700010077-5	48	3.4×10^3	4	32	2.3	0.3	3.5×10^1	6.1×10^{-1}
700010470-3	367	2.2×10^5	6	113	10.8	4.8	1.8×10^5	1.1×10^3
700010470-5	193	5.7×10^4	7	104	5.9	1.4	3.7×10^3	8.0×10^1
700010607-3	239	8.5×10^4	4	73	5.2	1.5	1.1×10^4	1.8×10^2
700010607-5	78	9.1×10^3	4	76	2.3	0.4	1.0×10^2	2.8
703010131-3	744	8.7×10^4	9	114	50.8	19.4	1.2×10^4	1.8×10^2
703010131-5	261	9.9×10^4	9	76	5.0	1.9	1.8×10^4	2.4×10^2
703010159-3	477	3.5×10^5	8	98	15.8	7.0	7.3×10^5	2.9×10^3
703010159-5	216	7.0×10^4	8	93	5.5	1.6	6.5×10^3	1.2×10^2
703010256-3	463	3.5×10^5	6	99	14.7	6.6	7.3×10^5	2.9×10^3
703010256-5	402	2.4×10^5	6	110	14.1	5.5	2.4×10^5	1.4×10^3
704010042-3	875	1.3×10^6	6	93	51.1	21.6	3.7×10^7	3.9×10^4
704010042-5	266	1.1×10^5	6	85	5.1	2.1	2.4×10^4	2.9×10^2
705010162-3	508	4.0×10^5	5	69	12.6	5.7	1.1×10^6	3.7×10^3
705010162-5	254	9.6×10^4	5	60	4.9	1.4	1.6×10^4	2.2×10^2
705010185-3	292	1.3×10^5	5	97	7.3	2.7	3.9×10^4	4.0×10^2
705010185-5	85	1.1×10^4	3	49	2.3	0.4	1.3×10^2	3.9
705010198-3	204	6.3×10^4	3	48	4.1	0.9	4.9×10^3	9.7×10^1
705010198-5	72	7.8×10^3	3	47	2.2	0.3	8.6×10^1	2.2
706010164-3	485	3.7×10^5	6	102	19.1	7.4	8.6×10^5	3.2×10^3
706010164-5	204	6.2×10^4	6	98	5.1	1.5	4.7×10^3	9.5×10^1
cap256-3	204	1.3×10^6	6	98	5.1	1.5	3.7×10^7	3.9×10^4
703010505	1,131	2.5×10^6	23	578	319.1	215.3	2.7×10^8	1.4×10^5
703010848	2,694	1.4×10^7	31	1,205	1,736.0	395.5	4.7×10^{10}	4.5×10^6

Table S 1. Computational time and required memory for inferring epistatic and selection coefficients. The table summarizes the number of polymorphic sites (length L), the effective matrix dimension to be inverted (D), the number of time points (K), the number of obtained sequences (N), as well as the required computational time (T_{HCMF}) and memory size (M_{HCMF}) for each individual. For comparison, we estimated expected computational time (T_{Naive}) and memory usage (M_{Naive}) using the naive method (a strategy that directly obtains the higher-order covariance and its inverse.) IDs consist of the patient identifier and sequencing region (5′ or 3′ end of the genome), separated by a dash. Computations were performed using a single CPU core with a single thread. For comparison, lengths of around $L \sim 200$ are already near computational limits even using a high-performance computing system.