ARTICLES
https://doi.org/10.1038/s41587-020-0737-3
nature biotechnology



# MPL resolves genetic linkage in fitness inference from complex evolutionary histories

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

**Genetic linkage causes the fate of new mutations in a population to be contingent on the genetic background on which they appear. This makes it challenging to identify how individual mutations affect fitness. To overcome this challenge, we developed marginal path likelihood (MPL), a method to infer selection from evolutionary histories that resolves genetic linkage. Validation on real and simulated data sets shows that MPL is fast and accurate, outperforming existing inference approaches. We found that resolving linkage is crucial for accurately quantifying selection in complex evolving populations, which we demonstrate through a quantitative analysis of intrahost HIV-1 evolution using multiple patient data sets. Linkage effects generated by variants that sweep rapidly through the population are particularly strong, extending far across the genome. Taken together, our results argue for the importance of resolving linkage in studies of natural selection.**

Evolving populations exhibit complex dynamics. Cancers[1–6] and pathogens, such as HIV-1 (refs. [7–9]) and influenza[10,11], generate multiple beneficial mutations that increase fitness or allow them to escape immunity. Subpopulations with different beneficial mutations then compete with one another for dominance, referred to as clonal interference, resulting in the loss of some mutations that increase fitness[12]. Neutral or deleterious mutations can also hitchhike to high frequencies if they occur on advantageous genetic backgrounds[13]. Experiments have demonstrated that these features of genetic linkage are pervasive in nature[14–16].

Linkage makes distinguishing the fitness effects of individual mutations challenging because their dynamics are contingent on the genetic background on which they appear. Lineage tracking experiments can be used to identify beneficial mutations[17], but they cannot readily be applied to evolution in natural conditions, such as in cancer or in natural infection by viruses or bacteria. Most existing computational methods to infer fitness from population dynamics ignore linkage entirely[18–25]. Ignoring linkage could lead to errors when genetic hitchhiking or clonal interference are present, which frequently occur in nature. A few methods have attempted to incorporate linkage information, but these methods are exceptionally computationally intensive and may scale poorly to populations with many polymorphic variants[26–28].

Here we describe a method to infer selection from evolutionary histories, captured by genetic time series data, and demonstrate its ability to resolve linkage effects. Simulations show that our approach, which we call marginal path likelihood (MPL)[29,30], is faster and more accurate than current state-of-the-art methods for selection inference. As an example application, we use our method to reveal patterns of selection in intrahost HIV-1 evolution using 14 patient data sets. The genetic diversity exhibited in these data sets makes them exceptionally challenging to analyze using existing linkage-aware methods. With MPL, we observe strong selection for escape from CD8$^+$ T cell responses, which is partially masked by linkage due to extensive clonal interference between competing escape mutants. We further quantify the influence of linkage on inferred selection across the viral genome. Our results show that most variants have negligible effects on inferred selection at other sites, but a small minority of highly influential variants have dramatic and far-reaching effects. These highly influential variants are often ones that sweep rapidly through the population. We also find modest selection for escape from antibody responses, even in an individual who develops broadly neutralizing antibodies (bnAbs). Collectively, our results argue for the importance of accounting for genetic linkage when inferring selection, while providing a practical method for achieving this for large data sets.

## Results

**Evolutionary model incorporating linkage.** The principle idea of our inference approach is to efficiently quantify the probability of an evolutionary 'path,' defined by the set of all mutant allele frequencies at each time, using a path integral method derived from statistical physics (Methods). Path integrals for related evolutionary models have been derived under different assumptions in past work[31–33], but they have not been widely applied for inference. This method allows us to disentangle the effects of individual mutations from the sequence background, that is, genetic linkage, without making the likelihood function intractable. In fact, the path integral can be analytically inverted to find the parameters that are most likely to have generated a path.

To define the path integral, we consider Wright–Fisher (WF) population dynamics with selection, mutation and recombination, in the diffusion limit[34]. Under an additive fitness model, the fitness of any individual is a sum of selection coefficients, $s_i$, which quantify the selective advantage of mutant allele $i$ relative to wild-type (WT). The probability of an evolutionary path is then a product of

[1]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China. [2]Institute for Advanced Study, Hong Kong University of Science and Technology, Hong Kong, China. [3]The Kirby Institute, University of New South Wales, Sydney, New South Wales, Australia. [4]School of Medical Sciences, University of New South Wales, Sydney, New South Wales, Australia. [5]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong, China. [6]Department of Physics and Astronomy, University of California, Riverside, Riverside, CA, USA. [7]These authors contributed equally: Muhammad Saqib Sohail, Raymond H. Y. Louie. ✉e-mail: m.mckay@ust.hk; john.barton@ucr.edu

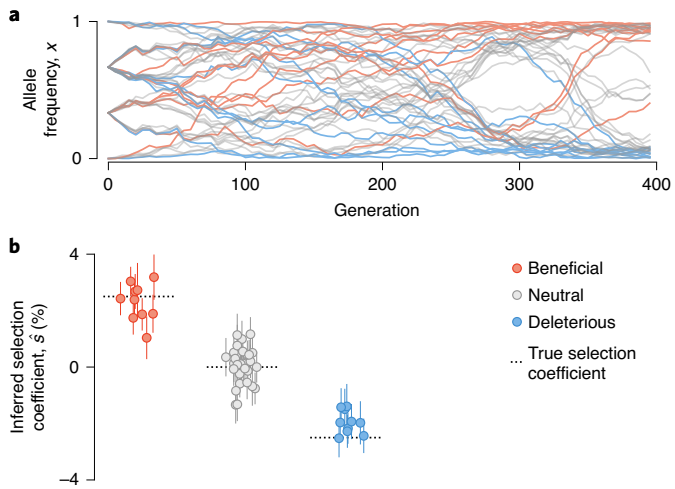



**Fig. 1 | MPL accurately recovers selection from complex dynamics. a**, Simulated allele frequency trajectories in a model with ten beneficial, 30 neutral and ten deleterious mutant alleles. The initial population is a mix of three subpopulations with random mutations. Selection is challenging to discern from individual trajectories alone. **b**, Selection coefficients inferred by MPL, presented as mean values ± 1 theoretical s.d. (Methods), are close to their true values. Simulation parameters. $L = 50$ loci with two alleles at each locus (mutant and WT): ten beneficial mutants with $s = 0.025$, 30 neutral mutants with $s = 0$ and ten deleterious mutants with $s = -0.025$. Mutation probability per locus per generation $\mu = 10^{-3}$, population size $N = 10^3$. The initial population is composed of approximately equal numbers of three random founder sequences, evolved over $T = 400$ generations.

probabilities of changes in mutant allele frequencies at each locus between successive generations, including the influence of selection at linked loci.

**Bayesian inference of selection.** Applying Bayes' theorem to the path integral likelihood leads to an analytical expression for the maximum a posteriori vector of selection coefficients $\hat{\mathbf{s}}$ corresponding to a path (Methods),

$$\hat{\mathbf{s}} = (C_{\text{int}} + \gamma I)^{-1}(\Delta \mathbf{x} - \boldsymbol{\mu}_{\text{fl}}). \quad (1)$$

The covariance matrix of mutant allele frequencies integrated over time, $C_{\text{int}}$, accounts for the speed of evolution and linkage effects. It is computed by summing the mutant allele frequency covariance matrices at each observed time point, weighted by the differences between observed time points (Methods). Here $\gamma$ quantifies the width of a Gaussian prior distribution for selection coefficients and $I$ is the identity matrix. The net change in mutant allele frequencies $\Delta \mathbf{x}$ is the difference between the frequencies at the last and first time points. The integrated mutational flux $\boldsymbol{\mu}_{\text{fl}}$ quantifies the expected cumulative increase or decrease in mutant allele frequency over time due to mutations. The difference between $\Delta \mathbf{x}$ and $\boldsymbol{\mu}_{\text{fl}}$ determines whether the dynamics of a mutant allele appear to be beneficial or deleterious when linkage is ignored. Explicit definitions of these terms are given in Methods. Because equation (1) emerges from the likelihood of allele frequency trajectories, a subset of the full genotype distribution, we refer to it as the MPL estimate of the selection coefficients.

**Inference with MPL is fast and robust.** To test the ability of MPL to uncover selection, we analyzed data from simulations of a variety of evolutionary scenarios (Supplementary Text). Even in cases with strong linkage (Fig. 1a), MPL accurately recovers true selection coefficients (Fig. 1b). We tested the robustness of these results to limited sampling by restricting both the time (measured in generations) between sampling events and the number of sequences sampled at each time point. We found that performance remains strong even when data is limited, an important practical consideration (Extended Data Fig. 1). Using as few as ten sequences per time point still allowed beneficial, neutral and deleterious mutations to be distinguished with high accuracy in the presence of strong linkage.

Next, we compared MPL to a panel of state-of-the-art methods of selection inference[21,23–26,35]. MPL was the most accurate method in terms of both classification accuracy, measured by area under the receiver operating characteristic for classifying mutant alleles as beneficial or deleterious, and in the absolute error in inferred selection coefficients (Fig. 2 and Supplementary Text) across two different simulation scenarios. The first 'simple' scenario begins with a homogeneous population. The second 'complex' scenario begins with a mixture of five random founder sequences, has stronger selection and a shorter overall time period. Notably, MPL results showed better agreement with the underlying parameters in most individual simulations as well as on average (Extended Data Fig. 2). Due to the simplicity of equation (1), MPL was fastest among the methods that we compared, with a running time roughly six orders of magnitude faster than approaches that rely on iterative Monte Carlo simulations. While these results (Figs. 1 and 2 and Extended Data Fig. 2) are for scenarios without recombination, we note that MPL performs equally well in scenarios with recombination (Extended Data Fig. 3).

**Patterns of selection in intrahost HIV-1 evolution.** We applied MPL to study the intrahost evolution of HIV-1 and to resolve complex interactions between HIV-1 and the immune system. We analyzed a variety of patient data sets, most of which sequenced half of the HIV-1 genome. Even without omitting invariant sites, the run time for MPL to analyze each data set (containing roughly $2 \times 10^4$ variants) was only around 20 min, demonstrating the scalability of our approach. Identifying selective pressures on HIV-1 gives insight into the evolutionary dynamics leading to HIV-1 escape from immune control and the development of bnAbs.

We analyzed 14 patient data sets, initially focusing on a collection of longitudinal HIV-1 half-genome sequence data sets from 13 individuals, where early-phase CD8$^+$ T cell responses were also comprehensively analyzed[36], and later on a single data set from an individual who develops bnAbs[37,38]. MPL is robust to sampling conditions similar to these 14 patient data sets (Supplementary Text and Extended Data Fig. 4). In the first set of 13 individuals, 37.8% of the top 1% most beneficial mutations reported by MPL are nonsynonymous mutations in identified[36] CD8$^+$ T cell epitopes (Fig. 3a). This is a 19-fold enrichment in mutations in T cell epitopes compared to expectations by chance (Methods). Here we observe more strongly beneficial escape mutations in subtype B viruses compared to subtype C, which is explained by the larger number of T cell epitopes targeted by individuals infected by subtype B viruses in this data set. Reversions to subtype consensus are also strongly beneficial. Nonsynonymous reversions outside T cell epitopes are 13-fold enriched in this subset. Furthermore, nonsynonymous reversions within T cell epitopes are 320-fold enriched. All enrichment values are highly significant (two-sided Fisher's exact test $P$ values of $<10^{-30}$, $<10^{-10}$ and $<10^{-19}$, respectively). These findings agree with past studies that have observed strong selection for T cell escape[8,9,39] and reversions[9].

Resolving linkage leads to substantial differences in the magnitude and distribution of selection estimates. MPL places 1.63 times as many T cell escape mutations within the top 1% most beneficial mutations as an independent model that ignores linkage between mutant alleles (Fig. 3b). Conversely, MPL ranks 0.38 times as many nonsynonymous reversions outside T cell epitopes to be strongly

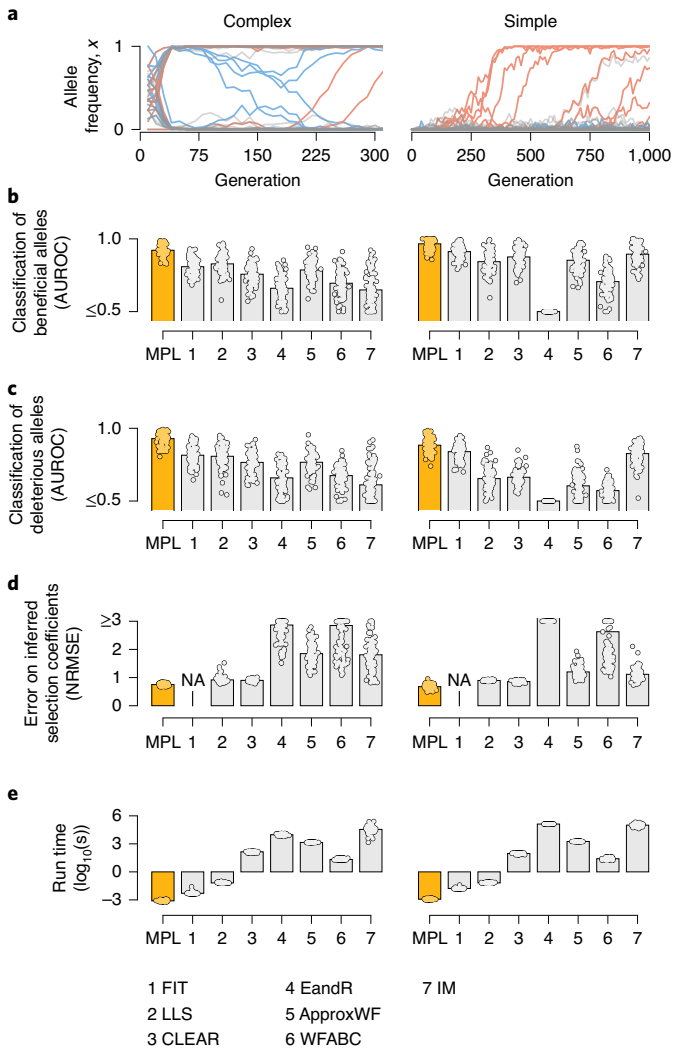



**Fig. 2 | MPL compares favorably with state-of-the-art methods. a**, We compared the ability of MPL and existing methods to infer selection from simulated test data that was rich with interference patterns and linkage, as shown in representative allele frequency trajectories. To evaluate robustness to finitely sampled data, we selected $n_s = 100$ sequences per time point for inference, with sampling time points separated by $\Delta t = 10$ generations. **b**–**e**, Performance was evaluated by comparing the successful classification of beneficial (**b**) and deleterious (**c**) mutations, error in the estimated selection coefficients (**d**) and run time (**e**), averaged over $n = 100$ replicate simulations with identical parameters. MPL achieves the highest performance in terms of classification and estimation accuracy, and in run time. Note that the frequency increment test (FIT) does not estimate selection coefficients. Simulation parameters. $L = 50$ loci with two alleles at each locus (mutant and WT): ten beneficial mutants ($s = 0.1$ for complex, $s = 0.025$ for simple), 30 neutral mutants ($s = 0$ for both scenarios) and ten deleterious mutants ($s = -0.1$ for complex, $s = -0.025$ for simple). Mutation probability $\mu = 10^{-4}$, population size $N = 10^3$. For the complex case, the initial population is composed of equal numbers of five random founder sequences, evolved over $T = 310$ generations. Recorded trajectories used for inference begin at generation 10. For the simple case, the initial population begins with all WT sequences, evolved over $T = 1{,}000$ generations. AUROC, area under the receiver operating characteristic; NA, not applicable; NRMSE, normalized root mean square error.

beneficial as the independent model does (Fig. 3c). These differences are explained by the joint resolution of genetic linkage effects, including clonal interference.

**Quantifying the contribution of linkage to inferred selection.** To dissect the contributions of linkage to estimates of selection, we computed the pairwise effects $\Delta\hat{s}_{ij}$ of each variant $i$ on the inferred selection coefficients for all other variants $j$ (Methods). We defined $\Delta\hat{s}_{ij}$ as the difference between the estimated selection coefficient $\hat{s}_j$ for variant $j$ using all of the data and the value of $\hat{s}_j$ when variant $i$ is replaced by the transmitted/founder (TF) nucleotide at the same site, thereby removing the contribution to selection from linkage with variant $i$. Positive values of $\Delta\hat{s}_{ij}$ indicate that linkage with variant $i$ increases the selection coefficient inferred for variant $j$ (for example, due to clonal interference between them). Negative values indicate that variant $i$ decreases the selection coefficient inferred for variant $j$ (for example, due to hitchhiking). Computing the $\Delta\hat{s}_{ij}$ allowed us to examine the extent to which linkage affects the inference of selection, how these effects were distributed among different genetic variants and how they depend on the distance along the genome between a pair of linked variants.

**Distribution of linkage effects on inferred selection.** Our analysis revealed that most observed variants have virtually no effect on estimates of selection at other sites, but a small minority of highly influential variants have dramatic effects (Fig. 4 and Extended Data Figs. 5 and 6). The highly influential variants are often ones that change rapidly in frequency, sweeping through the population and exerting substantial effects on linked sites (Supplementary Fig. 1). Consistent with this observation, 40% of highly influential variants are putative CD8$^+$ T cell escape mutations. This data indicates that some highly influential variants are drivers of selective sweeps.

Our results indicate that the effects of linkage on inferred selection can be highly asymmetrical. That is, a genetic variant $i$ may substantially modify the selection coefficient inferred for variant $j$, while variant $j$ has little impact on $i$. Figure 4 shows both cases where linkage effects are asymmetrical and where they are reciprocal.

**Association between linkage effects and genomic distance between variants.** While there exist some highly influential variants whose effects span across long genomic distances (Fig. 4 and Extended Data Fig. 6), in most cases, the effects of linkage on estimated selection drop off sharply with increasing distance along the genome (Extended Data Fig. 7a). The largest effects are naturally felt for variants at the same site on the genome, which are in complete competition. Linkage effects on inferred selection are most prominent up to a distance of around 10 bp between variants. Rare, strong linkage effects ($|\Delta\hat{s}_{ij}| > 1\%$) are noticeably more frequent within distances of around 30 bp, roughly the length of a CD8$^+$ T cell epitope. After this point, additional distance has little influence on the magnitude of linkage effects on inferred selection.

For this data, recombination is expected to contribute to the general decrease in strength of linkage effects on inferred selection with increasing distance along the genome. When two different viruses coinfect the same cell, distinct RNA from each of them can be packaged in new virions. Then, when these virions subsequently infect new cells, HIV-1 can undergo recombination as the reverse transcriptase switches between templates. Estimates show that the effective recombination rate for HIV-1 in vivo is high, around $10^{-5}$ per base per generation[40,41], which is comparable to the mutation rate. Recombination acts to break up linkage at long distances along the genome, leading to reduced correlations between mutant variants at more widely spaced loci. This effect is clearly evident in the HIV-1 data (Extended Data Fig. 7b). The decay of correlations with distance is smooth, although strong correlations still persist at long ranges. This further indicates the existence of long-range linkage patterns in the data, despite the action of recombination. However, the strongest effects of linkage on inferred selection are comparatively more short-ranged on average, with long-range effects being more punctuated (Extended Data Fig. 7a).

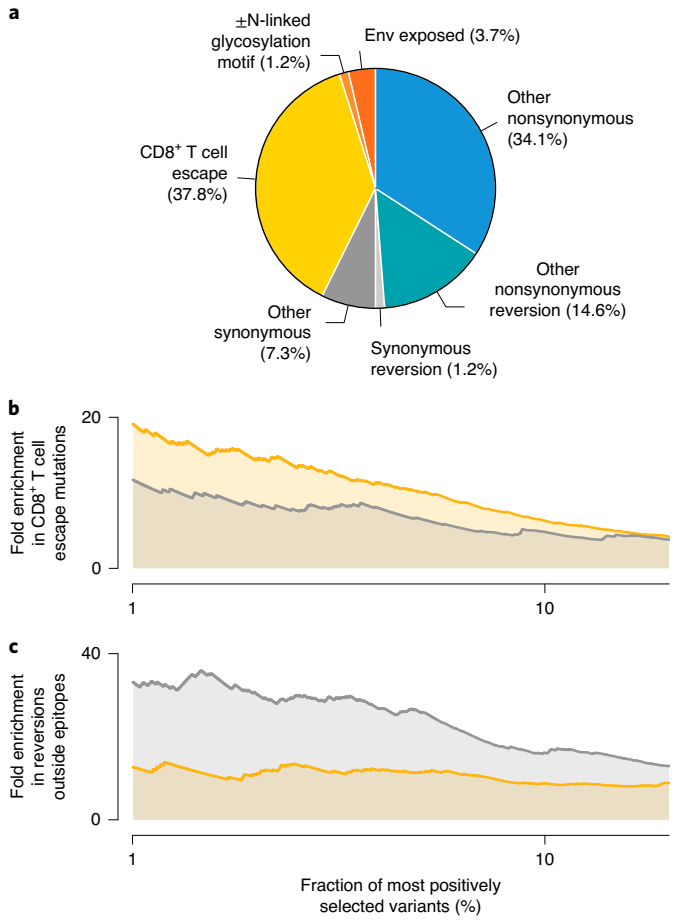



**Fig. 3 | Patterns of strong selection in intrahost HIV-1 evolution. a**, Among the top 1% most beneficial variants across individuals, mutations to escape from T cell-mediated immunity are especially common. **b**, Due to clonal interference between escape mutants, MPL identifies more escape variants to be strongly beneficial than an independent model that ignores genetic linkage. **c**, In contrast, the independent model estimates an excess in the number of strongly beneficial reversions.

**Illustration of the effects of clonal interference on inferred selection.** Viral escape from a T cell response targeting the Nef KF9 epitope in individual CH77 provides a clear example of clonal interference (Fig. 5a). MPL infers strong positive selection for all escape variants. In contrast, when linkage is ignored, escape variants that are lost are inferred to be neutral, and the magnitude of selection for 9040C decreases substantially (Fig. 5b). Experimental tests have shown that most nonsynonymous mutations within CD8+ T cell epitopes are escape mutations, which limit the ability of T cells to kill the mutant form of the virus[7]. Such mutations are likely to be beneficial to viral replication in vivo. Ignoring linkage thus leads to selection estimates that are qualitatively and quantitatively suspect. We observe similar instances of clonal interference in other epitopes (see Extended Data Figs. 8 and 9 for examples).

In the case of KF9, competition between the different escape variants increases the estimated selection coefficient for each of them (Fig. 5c). The interaction between variants 9040C and 9044G, which compete in the viral population at later times, is particularly strong. Inferred selection is also influenced by linkage with other mutations outside the KF9 epitope (Fig. 5c,d). For example, 9040C is inferred to be more beneficial due to its competition with the DI9 escape mutation 6021C. The selection coefficient for 9044G, in turn, is reduced due to positive linkage with 8719G, which is the dominant escape mutation in the nearby Env DR9 epitope.

**Modest selection for HIV-1 escape from antibody responses.** The HIV-1 surface protein Env is targeted by antibodies that can block or disrupt infection. Some strongly selected mutations lie in regions of Env that are exposed to antibodies, or in N-linked glycosylation motifs that affect the area of Env that is accessible to antibodies (Fig. 3a). However, these mutations are infrequent compared to others in T cell epitopes. As an example, for the case of CH77, one observes little positive selection in Env outside T cell epitopes (Fig. 5d). Overall, selection for escape from antibody responses appears to be weaker or less frequent than CD8+ T cell-mediated selection.

We asked, therefore, whether strong antibody-mediated selection would be observed in individuals who generate bnAbs. To explore this question, we studied HIV-1 evolution in individual CAP256 who developed the VRC26 lineage of bnAbs[37,38]. This case is particularly challenging for inference because of a superinfection event 15 weeks after initial infection (Fig. 6a). Superinfection led to intense and complex patterns of linkage as the superinfecting strain recombined and competed with the primary infecting strain (Fig. 6b and Supplementary Fig. 2). For this reason, estimates of selection that ignore linkage are exceptionally poor. Most (six out of 11) of the top 1% most beneficial mutations inferred by the independent model are from the background of the superinfecting strain and are synonymous. In contrast, none of the most beneficial mutations inferred by MPL are synonymous.

We found that selection for known VRC26 resistance mutations[37,38] is modest (Fig. 6c). The most strongly selected mutation in the VRC26 epitope region is 6709C ($\hat{s}=0.041$) in codon 162 in Env, a variant present in the superinfecting strain that completes an N-linked glycosylation motif that is absent from the primary infecting virus. However, this modification makes the virus more sensitive to VRC26 (refs. [37,38]). We observed selection against 6717T ($\hat{s}=-0.012$), corresponding to the Env 165L variant in the superinfecting strain. Reversion of this residue to V, the variant in the primary infecting strain, improves resistance to early VRC26 antibodies[38]. We also observed modest positive selection for nonsynonymous variation at codon 169 in Env (maximum $\hat{s}=0.010$), where mutations lead to complete resistance to VRC26 lineage antibodies[38]. Thus, even the most strongly selected resistance mutations fall outside the top 5% most strongly selected mutations in the larger sample of 13 individuals.

Weak selection on the virus for antibody escape may, in fact, facilitate the development of bnAbs. Multiple escape variants, as well as variants that are sensitive to the antibody, can readily coexist for long times when escape is weakly selected. This coexistence increases the diversity of the viral population. Pressure on antibodies to bind to multiple variants can then select for breadth[42]. Indeed, viral diversification has been observed to precede bnAb development[38,43]. Stronger pressure on the virus for escape could instead reduce viral diversity due to rapid fixation of beneficial escape variants and the elimination of sensitive ones.

## Discussion

We developed an efficient approach to infer the fitness effects of mutations from time series sequence data that accounts for the confounding effects of genetic linkage. MPL successfully infers selection from simulation data, is robust to sampling constraints and it performs favorably compared to state-of-the-art approaches to this problem. Notably, MPL is also fast, easily extending to systems with tens of thousands of genetic variants. Our method is general and should be widely applicable to investigate selection in evolving populations.

Our application of MPL to intrahost HIV-1 evolution demonstrated the importance of resolving linkage due to clonal interference

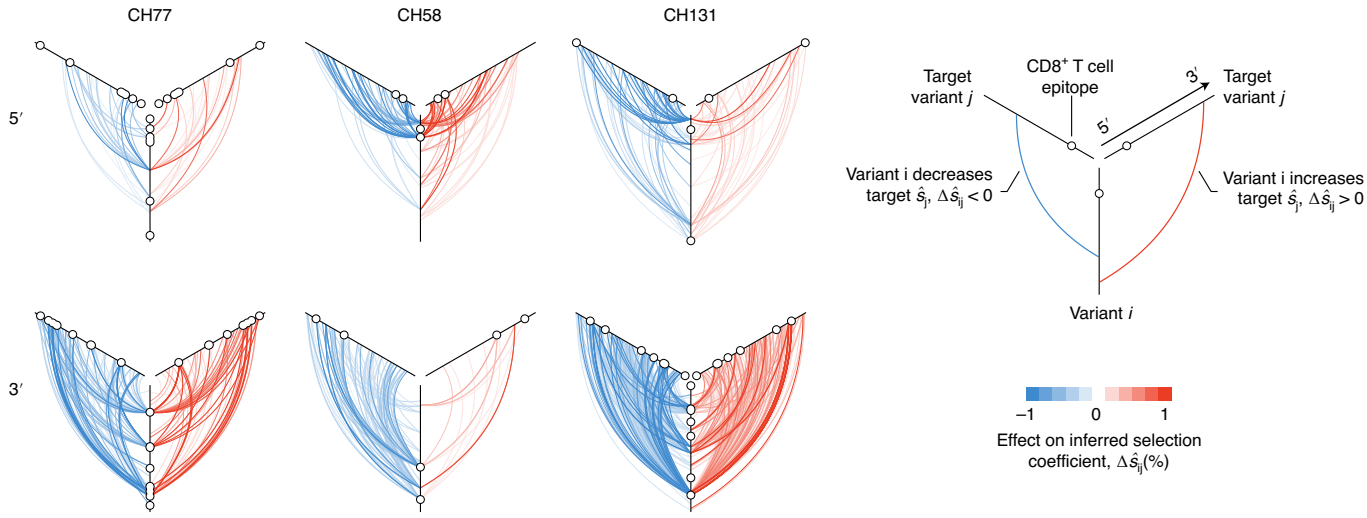



**Fig. 4 | Maps of strong contributions of linkage to inferred selection.** Plot of all large ($|\Delta\hat{s}_{ij}| > 0.004$) linkage effects on inferred selection coefficients, $\Delta\hat{s}_{ij}$, for three individuals. One plot is shown for each sequencing region, for each individual. Strong effects of linkage on inferred selection coefficients can span the viral half-genome. Maps of inferred selection for these three individuals are presented in Fig. 5 (CH77, 3′ region), Extended Data Fig. 8 (CH58, 5′ region) and Extended Data Fig. 9 (CH131, 3′ region). Maps of strong contributions of linkage to inferred selection for all individuals are shown in Extended Data Fig. 6.

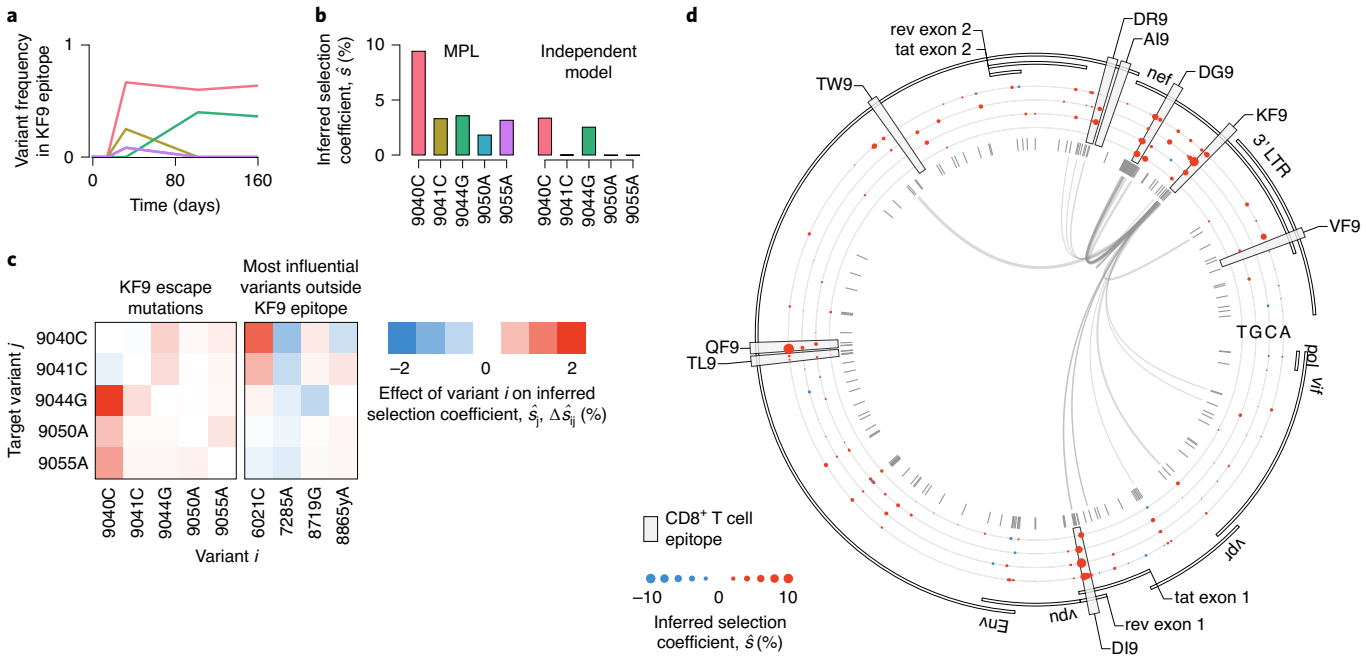



**Fig. 5 | Estimates of selection coefficients for viral escape mutations must account for clonal interference. a**, Across the viral population, multiple escape mutations appear in the T cell epitope KF9, targeted by individual CH77 and exhibit clonal interference. **b**, Using the full half-genome-length sequence data as input, MPL infers that all KF9 escape variants are positively selected. In contrast, estimates based solely on the trajectories of individual variants only uncover substantial positive selection for the 9040C and 9044G variants that coexist at the final time point. Furthermore, the independent model infers attenuated estimates of selection because it does not account for competition with other beneficial mutations, including other escape mutations within the same epitope. **c**, Linkage effects on inferred selection coefficients for KF9 escape mutations. Effects shown here are due to variants within the KF9 epitope and the top four most influential variants outside the KF9 epitope, defined as the variants $i$ for which $\sum_j |\Delta\hat{s}_{ij}|$ is the largest. All of these influential variants lie within other T cell epitopes (6021C lies in DI9, 7285A in QF9, 8719G in DR9 and 8865yA in DG9). **d**, Inferred selection in the HIV-1 half-genome sequence for CH77. Inferred selection coefficients are plotted in tracks. Coefficients of TF nucleotides are normalized to zero. Tick marks denote polymorphic sites. Inner links, shown for sites connected to the KF9 epitope, have widths proportional to matrix elements of the inverse of the integrated covariance (equation (1)). LTR, long terminal repeat.

between strongly selected mutations. For some variants, the effects of linkage can extend far across the genome despite frequent recombination. The ability to quantify how linkage affects inferred selection also aids in the interpretability of our results. Our analysis emphasized the central role of T cell escape mutations in HIV-1 evolution, while revealing a modest selection for escape from

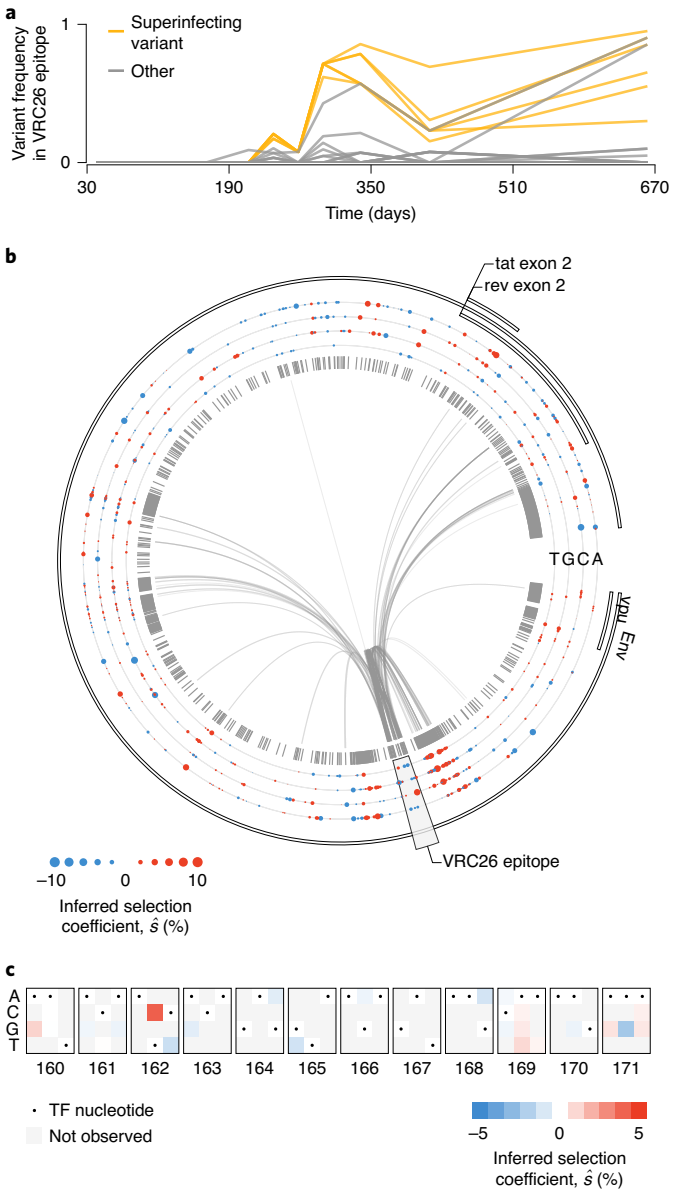



**Fig. 6 | Complex patterns of selection in HIV-1 Env following superinfection in an individual who develops broadly neutralizing antibodies. a**, Multiple variants, including several from the superinfecting strain of the virus, rise and fall in frequency within the epitope targeted by the VRC26 lineage of antibodies. **b**, Inferred selection in CAP256 HIV-1 Env sequences. Inferred selection coefficients are plotted in tracks. Coefficients of TF nucleotides are normalized to zero. Tick marks denote polymorphic sites. Inner links, shown for sites connected to the VRC26 epitope, have widths proportional to matrix elements of the inverse of the integrated covariance. Linkage is extensive due to the struggle for dominance in the viral population between the TF, superinfecting and recombinant strains. **c**, Map of inferred selection within the VRC26 epitope, consisting of codons 160–171 in Env.

antibody responses, even in an individual who develops bnAbs. The polyclonality of the antibody response may contribute to weaker overall selection due to conflicting pressures for escape from different antibodies.

The role of CD8$^+$ T cell escape in HIV-1 evolution has also been analyzed in previous studies using techniques and metrics that are distinct from ours. For example, past work estimated selection for T cell escape variants using a simulation-based procedure and a single-locus evolutionary model that does not account for genetic linkage[39]. T cell escape rates, which are related to (but distinct from) selection coefficients, have also been investigated for HIV-1 in multiple studies. These studies often use specialized differential equation-based models of HIV evolution[44–46], and generally do not account for genetic linkage[44] or account for it only approximately[46–48]. Related studies provide evidence for selection for T cell escape by observing increased nonsynonymous variation within T cell epitopes during within-host HIV-1 evolution[8,9].

A key distinction of our study, relative to previous studies, is that we provide an unbiased quantification of selection, and how it is affected by linkage, across large stretches of the HIV-1 genome. Our approach is unbiased in the sense that we consider all observed HIV-1 genetic variation, rather than focusing specifically on, for example, T cell escape mutations. While we find that many escape mutations are strongly selected, they still represent a minority of the most beneficial mutations that we observe. Our analysis accounts for and quantifies the interactions between variants observed during within-host HIV-1 evolution, including competition and synergistic interactions within and between CD8$^+$ T cell escape mutations. The ability to quantify selection at the allele level while accounting for linkage effects across large genomic regions is a unique feature of our study.

Our analysis emphasizes the importance of accounting for genetic linkage when inferring selection during HIV-1 evolution. Linkage effects can strongly bias selection inference, despite being dominated by a small subset of variants (Extended Data Fig. 5). Our aggregate analysis shows, for example, that selection for T cell escape is much stronger than would be expected if one were to ignore linkage, while the opposite is true for reversions toward subtype consensus (Fig. 3). The consequences of ignoring linkage are particularly evident for the analysis of CAP256, where, when linkage is disregarded, most variants estimated to have the strongest selection are synonymous. In contrast, all of the most beneficial variants inferred by MPL are nonsynonymous.

Constraints on the type and quality of data necessary for reliable inference place some limitations on the application of our method. While MPL could easily be applied to single-locus data, knowledge of pairwise variant frequencies is needed to disentangle the confounding effects of genetic linkage. Algorithms for estimating genotype distributions, such as those used for haplotype reconstruction in virus populations[49] and clone frequency inference in cancer[50], could be used to estimate mutant pair correlations in situations where complete information is unavailable. New computational methods that explicitly incorporate temporal information[51] would be ideal for reconstructing maps of genetic linkage across time. The continuing development of long-read sequencing technologies will also make pairwise variant frequency data more accessible. As with any inference method, MPL is also limited by the quantity and extent of data available. Genetic variation that lies outside the sequencing region, or undetected genetic alterations (for example, copy number variation) could potentially affect inference results. However, here we found that in HIV-1 evolution data only a small minority of genetic variants strongly affect selection coefficients inferred at other sites (Extended Data Fig. 5). In cases where an important genetic driver is missed, we anticipate that its selective effect will be distributed among linked variants. Limitations in the temporal resolution of sequence data also affect the strength of selection that can reliably be inferred. In particular, selection coefficients for variants that arise and completely fix in between two sampling events are likely to be underestimated.

The evolutionary model that we have used could be extended. While our model is general, it does not yet account for features such as epistasis, time-varying selection or migration. Future work will consider these important questions. The development of practical,

efficient algorithms to reveal epistasis in large-scale data remains a particular challenge because the number of parameters to infer grows quadratically with the genome length. Efficient statistical methods, possibly incorporating sparse model selection, will likely be required. In the case of time-varying selection, the selection coefficients that we infer are likely to be similar to the average strength of selection during the time over which that variant was observed. In the case of HIV-1, this may occur, for example, when the magnitude of the immune response against the virus shifts over time. Viral load also undergoes substantial shifts during the course of HIV-1 infection. Although the relationship between viral load and effective population size is complicated[52], changes in the number of infected cells could lead to different relative strengths of genetic drift during different stages of infection. Future work will extend MPL to population models with time-varying parameters. While MPL works well with limited sequence data (Extended Data Fig. 1), Bayesian methods to integrate over uncertainty in variant frequency trajectories due to finite sampling could further improve the robustness of our approach.

Our analysis reveals an intriguing link between population genetics and coevolutionary methods[53] that have enjoyed great success in predicting protein structure[54–56] and fitness[57–65] based on sequence information. Coevolutionary methods use a statistical model to capture the low-order moments of the distribution of mutations in a set of sequences, whose parameters can then be related to structure and fitness. So far, there has been no convincing theoretical explanation for the success of coevolutionary methods, or why only the low-order moments are necessary. Here we discovered that, while the evolutionary dynamics of the WF model are defined naturally at the level of genotypes, MPL estimates of fitness only depend on trajectories of the low-order moments of the sequence distribution, at least for the additive fitness landscape that we consider (Methods). Higher-order moments contain no further information about fitness. Energy parameters from Gaussian or standard mean field coevolutionary models[53] also have a similar dependence on the inverse of the variant frequency covariance matrix as the selection coefficients inferred by MPL. A mathematical connection between these two frameworks may point to an underlying evolutionary reason for why the low-order statistics used by coevolutionary models are sufficient to capture rich biological information.

Substantial effort in the biomedical sciences is dedicated to identifying the underlying genetic drivers of disease. Notable examples include mutations that promote cancer progression and immune evasion, or mutations that confer drug resistance to bacteria. In the right environment, these mutations confer survival benefits to the pathogens that carry them. However, it can be challenging to separate adaptive mutations from random genetic variation in a complex evolving population. MPL provides a method to infer the fitness effects of individual mutations at large scales even in the face of pervasive genetic linkage. Given the potential pitfalls of ignoring linkage that we have demonstrated, our results call for a greater focus on resolving linkage effects in studies of selection.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0737-3.

Received: 16 September 2019; Accepted: 14 October 2020;
Published online: 30 November 2020

## References

1. Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
2. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
3. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
4. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
5. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
6. Łuksza, M. et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
7. 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. *Nat. Rev. Immunol.* **10**, 11–23 (2010).
8. 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. *J. Virol.* **79**, 13239–13249 (2005).
9. Zanini, F. et al. Population genomics of intrapatient HIV-1 evolution. *eLife* **4**, e11282 (2015).
10. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics* **192**, 671–682 (2012).
11. Łuksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).
12. Muller, H. J. The relation of recombination to mutational advance. *Mut. Res.* **1**, 2–9 (1964).
13. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
14. Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**, 1615–1617 (2006).
15. Lang, G. I. et al. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**, 571–574 (2013).
16. Tenaillon, O. et al. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature* **536**, 165–170 (2016).
17. Levy, S. F. et al. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
18. Bollback, J. P., York, T. L. & Nielsen, R. Estimation of $2N_es$ from temporal allele frequency data. *Genetics* **179**, 497–502 (2008).
19. 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).
20. Mathieson, I. & McVean, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**, 973–984 (2013).
21. Feder, A. F., Kryazhimskiy, S. & Plotkin, J. B. Identifying signatures of selection in genetic time series. *Genetics* **196**, 509–522 (2014).
22. Lacerda, M. & Seoighe, C. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics* **198**, 1237–1250 (2014).
23. Foll, M., Shim, H. & Jensen, J. D. WFABC: a Wright–Fisher ABC–based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Mol. Ecol. Resour.* **15**, 87–98 (2015).
24. Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D. & Wegmann, D. An approximate Markov model for the Wright–Fisher diffusion and its application to time series data. *Genetics* **203**, 831–846 (2016).
25. Taus, T., Futschik, A. & Schlötterer, C. Quantifying selection with Pool-Seq time series data. *Mol. Biol. Evol.* **34**, 3023–3034 (2017).
26. Illingworth, C. J. R. & Mustonen, V. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics* **189**, 989–1000 (2011).
27. Illingworth, C. J. R., Fischer, A. & Mustonen, V. Identifying selection in the within-host evolution of influenza using viral sequence data. *PLoS Comput. Biol.* **10**, e1003755 (2014).
28. Terhorst, J., Schlötterer, C. & Song, Y. S. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genet.* **11**, e1005069 (2015).
29. Sohail, M. S., Louie, R. H. Y., McKay, M. R. & Barton, J. P., MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Github* https://github.com/bartonlab/paper-MPL-inference (2020).
30. Sohail, M. S., Louie, R. H. Y., McKay, M. R. & Barton, J. P., MPL resolves genetic linkage in fitness inference from complex evolutionary histories. *Code Ocean* https://doi.org/10.24433/CO.1795728.v1 (2020).
31. Mustonen, V. & Lässig, M. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl Acad. Sci. USA* **107**, 4248–4253 (2010).
32. Illingworth, C. J. R., Parts, L., Schiffels, S., Liti, G. & Mustonen, V. Quantifying selection acting on a complex trait using allele frequency time series data. *Mol. Biol. Evol.* **29**, 1187–1197 (2011).
33. Schraiber, J. G. A path integral formulation of the Wright–Fisher process with genic selection. *Theor. Popul. Biol.* **92**, 30–35 (2014).

34. Ewens, W. J. *Mathematical Population Genetics 1: Theoretical Introduction* (Springer Science & Business Media, 2012).
35. Iranmehr, A., Akbari, A., Schlötterer, C. & Bafna, V. CLEAR: Composition of likelihoods for evolve and resequence experiments. *Genetics* **206**, 1011–1023 (2017).
36. Liu, M. K. P. et al. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J. Clin. Invest.* **123**, 380–393 (2013).
37. Moore, P. L. et al. Multiple pathways of escape from HIV broadly cross-neutralizing V2-dependent antibodies. *J. Virol.* **87**, 4882–4894 (2013).
38. Doria-Rose, N. A. et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509**, 55–62 (2014).
39. Liu, Y. et al. Selection on the human immunodeficiency virus type 1 proteome following primary infection. *J. Virol.* **80**, 9519–9529 (2006).
40. Neher, R. A. & Leitner, T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput. Biol.* **6**, e1000660 (2010).
41. Batorsky, R. et al. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc. Natl Acad. Sci. USA* **108**, 5661–5666 (2011).
42. Wang, S. et al. Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell* **160**, 785–797 (2015).
43. Liao, H.-X. et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476 (2013).
44. Ganusov, V. V. et al. Fitness costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL escape during acute and chronic phases of HIV Infection. *J. Virol.* **85**, 10518–10528 (2011).
45. Ganusov, V. V., Neher, R. A. & Perelson, A. S. Mathematical modeling of escape of HIV from cytotoxic T lymphocyte responses. *J. Stat. Mech.: Theory Exp.* **2013**, P01010 (2013).
46. Kessinger, T., Perelson, A. & Neher, R. Inferring HIV escape rates from multi-locus genotype data. *Front. Immunol.* **4**, 252 (2013).
47. Pandit, A. & de Boer, R. J. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. *Retrovirology* **11**, 11–56 (2014).
48. Leviyang, S. & Ganusov, V. V. Broad CTL response in early HIV infection drives multiple concurrent CTL escapes. *PLoS Comput. Biol.* **11**, e1004492 (2015).
49. Beerenwinkel, N., Günthard, H. F., Roth, V. & Metzner, K. J. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* **3**, 329 (2012).
50. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
51. Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E. & Desai, M. M. The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).
52. Kouyos, R. D., Althaus, C. L. & Bonhoeffer, S. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends Microbiol.* **14**, 507–511 (2006).
53. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.* **81**, 032601 (2018).
54. Socolich, M. et al. Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).
55. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
56. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
57. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B. & Ranganathan, R. Natural-like function in artificial WW domains. *Nature* **437**, 579–583 (2005).
58. Ferguson, A. L. et al. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
59. 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 Comput. Biol.* **10**, e1003776 (2014).
60. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* **33**, 268–280 (2015).
61. Barton, J. P. et al. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. *Nat. Commun.* **7**, 11660 (2016).
62. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
63. Louie, R. H. Y., Kaczorowski, K. J., Barton, J. P., Chakraborty, A. K. & McKay, M. R. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl Acad. Sci. USA* **115**, E564–E573 (2018).
64. Quadeer, A. A., Louie, R. H. Y. & Mckay, M. R. Identifying immunologically-vulnerable regions of the HCV E2 glycoprotein and broadly neutralizing antibodies that target them. *Nat. Commun.* **10**, 2073 (2019).
65. Quadeer, A. A., Barton, J. P., Chakraborty, A. K. & McKay, M. R. Deconvolving mutational patterns of poliovirus outbreaks reveals its intrinsic fitness landscape. *Nat. Commun.* **11**, 377 (2020).

## Methods

Our method makes use of the diffusion approximation, widely used in population genetics[34,66–69], and is a path integral-based framework for statistical inference for a generalized multi-locus model. While familiar in physics[70], the path integral approach is less widely used in population genetics, although exceptions exist. Past work has derived path integrals for more specific models and for purposes other than inference[31,33], or ignored genetic drift and relied on numerical methods for solution[32]. The multi-locus model that we use accounts for the effects of selection and mutation, with the key novelty that it also accounts for the effects of linkage, recombination and incomplete temporal sampling. Notably, our approach gives a closed-form solution for the selection coefficients that are most likely to underlie a given evolutionary history.

**Evolutionary model.** Our inference approach is based on the standard WF model of population genetics, which describes the stochastic dynamics of an evolving population of $N$ individuals. Each individual is represented by a genetic sequence of length $L$. The population evolves in discrete, nonoverlapping generations subject to the forces of selection, mutation and recombination. For simplicity, we begin by describing the model with two alleles per locus, WT and mutant. Thus, there are $M=2^L$ unique genotypes. Later, we show that our approach readily generalizes to consider multiple alleles per locus.

The state of the population at a generation $t$ is given by the genotype frequency vector $\mathbf{z}(t)=(z_1(t), \ldots, z_M(t))$, where $z_a(t)$ denotes the frequency of individuals with genotype $a$. Conditioned on $\mathbf{z}(t)$, the probability that the genotype frequency vector in the next generation is $\mathbf{z}(t+1)$ is multinomial:[34]

$$P(\mathbf{z}(t+1)|\mathbf{z}(t)) = N! \prod_{a=1}^{M} \frac{(p_a(\mathbf{z}(t)))^{Nz_a(t+1)}}{(Nz_a(t+1))!}, \tag{2}$$

with

$$p_a(\mathbf{z}(t)) = \frac{y_a(t)f_a + \sum_{b\neq a}(\mu_{ba}y_b(t)f_b - \mu_{ab}y_a(t)f_a)}{\sum_{b=1}^{M} y_b(t)f_b}. \tag{3}$$

Here $f_a$ denotes the fitness of genotype $a$, and $\mu_{ab}$ is the probability of genotype $a$ mutating to genotype $b$. For simplicity we will assume at first that the mutation probability $\mu$ is the same at all loci, and that the probability of mutating from WT to mutant is the same as that from mutant to WT. In equation (3),

$$y_a(t) = (1-r)^{L-1}z_a(t) + \left(1-(1-r)^{L-1}\right)\psi_a(t) \tag{4}$$

is the frequency of genotype $a$ after recombination. Here $r$ is the probability of recombination per locus per generation, and $\psi_a(t)$ is the probability that randomly recombining any two individuals in the population results in an individual of genotype $a$ (Supplementary Text). Although equations (3) and (4) appear complex, they have an intuitive interpretation. The first term in equation (3) reflects the fact that fitter individuals reproduce more efficiently and are therefore more likely to be observed in future generations. Mutations, captured through the second term, lead to conversions from genotype $a$ to other genotypes and vice versa. The denominator in equation (3) provides an overall normalization and indicates that relative fitness is important: for a particular genotype to reliably grow in frequency, its fitness should be higher than the average fitness of all individuals in the population. The first term of equation (4) gives the proportion of individuals of genotype $a$ not undergoing recombination, while the second term accounts for the net inward flow due to recombination from all other genotypes to genotype $a$.

We assume that data consists of sets of genetic sequences obtained from a population at multiple time points $t_k$, $k\in\{0, 1, \ldots, K\}$. For such a population evolving under the WF model, the probability that the genotype frequency vector follows a particular evolutionary path $(\mathbf{z}(t_1), \mathbf{z}(t_2), \ldots, \mathbf{z}(t_K))$, conditioned on the initial state $\mathbf{z}(t_0)$, is

$$P\left((\mathbf{z}(t_k))_{k=1}^{K}|\mathbf{z}(t_0)\right) = \prod_{k=0}^{K-1} P(\mathbf{z}(t_{k+1})|\mathbf{z}(t_k)). \tag{5}$$

This expression is difficult to work with for parameter inference. This is due in part to the high dimensionality of the vector $\mathbf{z}$, which scales exponentially with the length of the genetic sequence. Thus, in most real data sets, the sequence space is dramatically under-sampled. The functional form of equation (2) is also complex.

Our approach circumvents these issues by using two approximations. First, we obtain a simplified version of equation (5) by using a path integral. Path integral expressions for evolutionary models have also been derived under different assumptions in past work[31–33], but they have not been widely applied for inference. We also assume that fitness is additive, such that the total fitness of each genotype $a$ is just given by the sum of the selection coefficients $s_i$ for mutant alleles at each locus $i$,

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

Here $g_i^a$ is 1 if genotype $a$ has a mutant allele at locus $i$ and 0 otherwise. These assumptions will substantially simplify the expression for equation (5).

**Path integral for mutant allele frequencies.** In this section we will develop a simplified version of equation (5) defined at the level of allele frequencies rather than genotype frequencies. Later we will demonstrate that, if the fitness effects of mutations are additive as assumed above, this approach will lead to no loss of information for estimating the selection coefficients from data. We begin by using the WF dynamics above, which are defined for genotype frequencies, to compute the expected changes in frequency of mutant alleles. The mutant allele frequency $x_i$ at locus $i$ is

$$x_i(t) = \sum_{a=1}^{M} g_i^a z_a(t).$$

Following the assumptions above, and in the WF diffusion limit[34], one can show that the probability density for mutant allele frequencies $\mathbf{x}(t)=(x_1(t), x_2(t), \ldots, x_L(t))$ follows a Fokker–Planck equation with a drift vector having entries

$$d_i(\mathbf{x}(t)) = \mu(1-2x_i(t)) + x_i(t)(1-x_i(t))s_i + \sum_{j\neq i}\left(x_{ij}(t) - x_i(t)x_j(t)\right)s_j \tag{6}$$

and diffusion matrix with entries $C_{ij}/N$, where

$$C_{ij}(\mathbf{x}(t)) = \begin{cases} x_i(t)(1-x_i(t)) & i=j \\ x_{ij}(t) - x_i(t)x_j(t) & i\neq j. \end{cases} \tag{7}$$

Here $x_{ij}$ is the frequency of individuals in the population with mutant alleles at both loci $i$ and $j$. The drift vector describes the expected change in mutant allele frequencies in time. Note that the last term in equation (6) quantifies linked selection, that is, how the dynamics of mutant allele frequencies are affected by the average genetic background on which they appear. The drift vector should not be confused with genetic drift, the fluctuation in allele frequencies due to the inherent stochasticity of replication, which is instead described by the diffusion matrix. The diffusion matrix is simply the covariance matrix of mutant allele frequencies divided by the population size $N$. It therefore depends on the double mutant frequencies $x_{ij}$, but we will use the shortened notation $C_{ij}(\mathbf{x}(t))$ for brevity.

Applying standard methods from statistical physics[70], the Fokker–Planck equation can be converted into a path integral that quantifies the probability density for 'paths' of mutant allele frequencies $(\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_K))$. This expression will allow us to efficiently estimate the parameters that are most likely to have generated a specific path (Supplementary Text). The probability for a path is

$$\begin{aligned} P\left((\mathbf{x}(t_k))_{k=1}^{K}|\mathbf{x}(t_0), N, \mu, \mathbf{s}\right) &\approx \left(\prod_{k=0}^{K-1}\frac{1}{\sqrt{\det C(\mathbf{x}(t_k))}}\left(\frac{N}{2\pi\Delta t_k}\right)^{L/2}\prod_{i=1}^{L}dx_i(t_{k+1})\right) \\ &\times\left(\prod_{k=0}^{K-1}\exp\left(-\tfrac{N}{2}S\left((\mathbf{x}(t_k))_{k=0}^{K}\right)\right)\right) \\ S\left((\mathbf{x}(t_k))_{k=0}^{K}\right) &= \sum_{k=0}^{K-1}\sum_{i=1}^{L}\sum_{j=1}^{L}\tfrac{1}{\Delta t_k}\left[x_i(t_{k+1}) - x_i(t_k) - \Delta t_k d_i(\mathbf{x}(t_k))\right] \\ &\times\left(C^{-1}(\mathbf{x}(t_k))\right)_{ij}\left[x_j(t_{k+1}) - x_j(t_k) - \Delta t_k d_j(\mathbf{x}(t_k))\right], \end{aligned} \tag{8}$$

where $\Delta t_k = t_{k+1} - t_k$. In the language of physics, $S\left((\mathbf{x}(t_k))_{k=0}^{K}\right)$ is referred to as the action. The population size $N$ is analogous to the inverse temperature in statistical physics. The action penalizes deviation of the change in mutant frequencies between generations from the expectation given by the drift vector at the previous generation. This is normalized by the diffusion matrix, which quantifies the magnitude of typical changes in mutant frequencies due to random replication alone (that is, genetic drift). The path integral equation in (8) follows the Itô convention.

**MPL estimate of the selection coefficients.** Given an observed path of mutant allele frequencies, we can use Bayesian inference to determine the maximum a posteriori selection coefficients $\hat{\mathbf{s}}$ corresponding to the data, assuming that the population size $N$ and mutation probability $\mu$ are known. In practice, our data consists of sets of genetic sequences obtained from a population at multiple time points $t_k$, $k\in\{0, 1, \ldots, K\}$. These sequences can be used to compute the path of mutant allele frequencies $(\mathbf{x}(t_0), \mathbf{x}(t_1), \ldots, \mathbf{x}(t_K))$ as well as the double mutant frequencies $x_{ij}(t_k)$, which also appear in equation (8). We will assume that the observed mutant allele frequencies are equal to the true ones, which simplifies the inference procedure. Our tests indicate that our results are robust to errors in the frequencies due to finite sampling (Extended Data Fig. 1). Future work will relax this assumption.

In total, the posterior probability of the selection coefficients $\mathbf{s}=(s_1, s_2, \ldots, s_L)$ given the observed path $(\mathbf{x}(t_0), \mathbf{x}(t_1), \ldots, \mathbf{x}(t_K))$ is

$$P\left(\mathbf{s}|(\mathbf{x}(t_k))_{k=0}^{K}\right) = P\left((\mathbf{x}(t_k))_{k=1}^{K}|\mathbf{x}(t_0)\right)\times P_{\text{prior}}(\mathbf{s}), \tag{9}$$

where $P\left((\mathbf{x}(t_k))_{k=1}^{K}|\mathbf{x}(t_0)\right)$ is the probability of the path (given by equation (8), but extended to arbitrary sampling times as shown in Supplementary Text) and $P_{\text{prior}}(\mathbf{s})$ is the prior probability for the selection coefficients. Equation (9) is a complicated expression of the mutant allele frequencies. However, solving for the selection

coefficients that maximize the posterior probability is straightforward because equation (8) is a Gaussian function of the selection coefficients. Taking $P_{\text{prior}}(\mathbf{s})$ to be Gaussian with mean zero and covariance matrix $\sigma^2 I$, where $I$ is the identity matrix, the selection coefficients that maximize equation (9) are given by equation (1), with

$$\begin{aligned} C_{\text{int}} &= \sum_{k=0}^{K-1} \Delta t_k C(\mathbf{x}(t_k)), \\ \Delta \mathbf{x} &= \mathbf{x}(t_K) - \mathbf{x}(t_0), \\ \boldsymbol{\mu}_{\text{fl}} &= \mu \sum_{k=0}^{K-1} \Delta t_k (1 - 2\mathbf{x}(t_k)). \end{aligned} \tag{10}$$

Collectively, this gives

$$\begin{aligned} \hat{s}_i = \; & \sum_{j=1}^{L} \left[ \sum_{k=0}^{K-1} \Delta t_k C(\mathbf{x}(t_k)) + \gamma I \right]^{-1}_{ij} \\ & \times \left[ x_j(t_K) - x_j(t_0) - \mu \sum_{k=0}^{K-1} \Delta t_k (1 - 2x_j(t_k)) \right], \end{aligned} \tag{11}$$

where $\gamma = 1/N\sigma^2$. We refer to this as the MPL estimate of selection coefficients. Because of the Gaussian form of equation (9), the maximum a posteriori estimates of the selection coefficients are the same as their posterior means. The theoretical covariance in the inferred selection coefficients can also be computed from equation (9), which is given by $C_{\text{int}}^{-1}$.

Equation (11) can be readily interpreted. Let us start by considering the vector term in the 'numerator' of equation (11) that multiplies the matrix inverse. Here the first terms quantify how the frequency of each mutant allele has changed between the initial and final generations. Naturally, alleles that increase in frequency over time are more likely to be beneficial. The remaining terms quantify the integrated mutational flux, that is, population flow from mutant to WT (or vice versa) due to mutation. Net mutational flux from mutant to WT is also associated with higher fitness for the mutant allele. This is because this indicates that the mutant state maintained higher frequency than the WT over the trajectory, despite the force of mutation that drives the frequencies toward the same value. Together, these terms in the numerator of equation (11) determine whether a mutant allele is inferred to be beneficial or deleterious, at least when the off-diagonal elements of the matrix that it multiplies are zero.

While the numerator of equation (11) roughly determines the sign of selection, the denominator determines the strength of the inferred selection coefficient. Let us refer to $\sum_{k=0}^{K-1} \Delta t_k C(\mathbf{x}(t_k))$ as the integrated covariance matrix, $C_{\text{int}}$. From equation (7) we see that the entries of $C_{ij}(\mathbf{x}(t))$ are small when the mutant frequency is near the boundaries (0 or 1). Thus, the dominant contribution to the integrated covariance matrix comes from points on the path where the mutant frequency is far from the boundaries. If selection is strong, so that the mutant allele is much fitter than the WT (or vice versa), then we expect that a large portion of the path will be spent with the mutant allele frequency close to the boundary. In such cases the diagonal part of the integrated covariance will be small, and we correctly infer strong selection. The prior distribution for the selection coefficients simply adds a constant to the diagonal of the integrated covariance, which both shrinks selection estimates toward zero and ensures that the matrix is invertible. The off-diagonal terms of the integrated covariance matrix capture linkage effects, that is, how much of the change in the mutant frequency at a locus can be attributed to the average sequence background on which the mutant appears.

The effect of recombination is notably absent from equation (11). While the evolutionary model (equations (3) and (4)) incorporates recombination, the recombination term cancels out during the genotype to allele transformation, and thus the MPL estimate is independent of the recombination probability $r$ under the additive fitness model assumed here (Supplementary Text). While recombination certainly affects the types of evolutionary history that are likely to be observed (by reducing linkage disequilibrium, see Extended Data Fig. 3), it does not affect the selection coefficients that we estimate conditioned on a particular evolutionary history.

**Equivalence of genotype- and allele-level analyses.** In the preceding section we derived an estimate for the selection coefficients most likely to have generated an observed evolutionary path. To do this we used an expression for the likelihood of a path of mutant allele frequencies that depended on the mutant frequencies $x_i(t)$ and their pairwise correlations $x_{ij}(t)$, but not on higher-order correlations of the full genotype distribution. However, the WF dynamics is defined at the level of genotypes.

It can be shown that the use of equation (8) does not result in any loss in information beyond the approximations inherent in the WF diffusion limit. In the WF diffusion limit, the same steps as those applied to derive equation (8) can be performed for the genotype frequencies (Supplementary Text). This results in a path integral expression that quantifies the probability density of genotype frequency paths. As in the allele-level analysis, the estimated selection coefficients are those that maximize

$$P\left(\mathbf{s} | (\mathbf{z}(t_k))_{k=0}^{K}\right) = P\left((\mathbf{z}(t_k))_{k=1}^{K} | \mathbf{z}(t_0)\right) \times P_{\text{prior}}(\mathbf{s}),$$

where $P\left((\mathbf{z}(t_k))_{k=1}^{K} | \mathbf{z}(t_0)\right)$ is the probability density of the genotype frequency path. The full expression is more complicated, and less transparent, than the allele-level equivalent. Nonetheless, one can show that the expression for the selection coefficients that maximize the posterior probability above is exactly the same as equation (11). Full details of this derivation are given in the Supplementary Text. This result is important because it shows that, following the assumptions of the WF diffusion limit and assuming that the fitness effects of mutations are additive, higher-order mutational correlations contain no further information about the fitness effects of mutations.

**Extension to multiple alleles per locus and asymmetric mutation probabilities.** The MPL framework extends readily to models with $\ell$ alleles per locus, as well as asymmetric mutation probabilities. Let $x_{i,\alpha}(t)$ denote the frequency of allele $\alpha$ at locus $i$ at generation $t$, and denote $\mu_{\alpha\beta}$ as the mutation probability per locus from allele $\alpha$ to allele $\beta$. Now, the trajectory of allele frequency vectors is ($\mathbf{x}(t_0)$, $\mathbf{x}(t_1)$, …, $\mathbf{x}(t_K)$), where $\mathbf{x}(t_k) = \left(x_{1,1}(t_k), x_{1,2}(t_k), \ldots, x_{1,\ell}(t_k), x_{2,1}(t_k), \ldots, x_{L,\ell}(t_k)\right)$.

Following parallel arguments to before (Supplementary Text), the MPL estimate of the selection coefficient $\hat{s}_{i,\alpha}$ for allele $\alpha$ at locus $i$ is

$$\begin{aligned} \hat{s}_{i,\alpha} = \; & \sum_{j=1}^{L} \sum_{\beta=1}^{\ell} \left[ \sum_{k=0}^{K-1} \Delta t_k C(\mathbf{x}(t_k)) + \gamma I \right]^{-1}_{ij,\alpha\beta} \\ & \times \left[ x_{j,\beta}(t_K) - x_{j,\beta}(t_0) - \sum_{k=0}^{K-1} \Delta t_k \sum_{\delta=1}^{\ell} \left( \mu_{\delta\alpha} x_{j,\delta}(t_k) - \mu_{\alpha\delta} x_{j,\alpha}(t_k) \right) \right], \end{aligned} \tag{12}$$

where $\gamma = 1/N\sigma^2$ as before. Off-diagonal entries of the covariance matrix $C(\mathbf{x}(t_k))$ are given by

$$C_{ij,\alpha\beta}(\mathbf{x}(t_k)) = x_{ij,\alpha\beta}(t_k) - x_{i,\alpha}(t_k) x_{j,\beta}(t_k),$$

where $x_{ij,\alpha\beta}(t_k)$ is the frequency of sequences with alleles $\alpha$ and $\beta$ at loci $i$ and $j$, respectively, at time $t_k$.

**Simulation data.** We implemented the WF model with discrete generations in Python. Briefly, we evolved populations of sequences according to equation (2) over $T = t_K$ generations, recording the entire evolutionary history. To mimic finite sampling in real data, we randomly selected $n_s$ sequences from the population to be used for analysis every $\Delta t$ generations. For example, if $n_s = 100$ and $\Delta t = 10$, we would select 100 sequences at random from the population every ten generations for the purposes of estimating selection coefficients. In cases where we show data from multiple trials, this data is obtained from independent simulations with the same underlying parameters. The initial population and simulation parameters are described in the figure captions. For MPL, we computed the single $x_i$ and double $x_{ij}$ mutant frequencies from the sampled sequences and used them to infer the selection coefficients with equation (11). We used this program to record 100 evolutionary histories each for three different choices of the underlying parameters. Parameter values are detailed in Extended Data Figs. 1 and 2. For all simulations we assumed only two alleles per site. The simulation code, code for analysis and original simulation data are contained in the GitHub repository.

**Other time series inference methods.** The independent model that we compared MPL against in the main text is a single-locus variant of MPL in which the off-diagonal elements of the integrated covariance matrix are set to zero.

The seven additional time series-based inference methods that we compared MPL against are frequency increment test (FIT)[21], linear least squares (LLS)[25], composition of likelihoods for evolve and resequence experiments (CLEAR)[35], evolve and resequence (EandR)[28], approximate Wright–Fisher (ApproxWF)[24], Wright–Fisher approximate Bayesian computation (WFABC)[23] and Illingworth and Mustonen's method (IM)[26]. Where available, we used the scripts provided by the authors. Some of these methods required preprocessing of the time series data to obtain valid estimates of selection coefficients. See Supplementary Text for details on implementation and data processing.

**Patient cohort.** We studied HIV-1 sequence data obtained from 14 individuals recruited under the CHAVI 001 and CAPRISA 002 studies in the United States, Malawi and South Africa. The locations of CD8$^+$ T cell epitopes were experimentally[36] or computationally[61] determined in 13 of the 14 individuals. In the remaining individual, CAP256, experimental studies identified the VRC26 family of antibodies and mapped the epitope location on Env[38].

**HIV-1 sequence data.** Multiple sequence alignments of HIV-1 nucleotide sequences for all individuals were obtained from the Los Alamos National Laboratory (LANL) HIV Sequence Database (www.hiv.lanl.gov; accessed 19 October 2018). Sequences labeled as problematic were not downloaded. For the 13 individuals with identified T cell epitopes, sets of 3′ and 5′ half-genome sequences were obtained, which were approximately 4,500 bp in length. Only Env sequences were available for CAP256 (approximately 2,500 bp in length). All sequences were aligned with the HXB2 reference sequence (GenBank accession number K03455) for numbering, and with subtype consensus sequences to determine reversions,

using the LANL HIVAlign tool[71]. A summary of the data is given in Supplementary Table 1.

For each patient data set, except for CAP256, the TF virus was deduced from single genome amplification and sequencing of plasma virus taken in acute infection close to peak viremia[36]. For CAP256, the transmitted founder sequence was not determined, and analysis started with sequences obtained 42 d after infection[38]. In our analysis, we defined the TF nucleotide at each site based on TF sequences in the LANL HIV Sequence Database. In cases where the TF sequence was not available, we used the most frequently observed nucleotide at each site at the earliest sequencing time point.

Accurately inferring selection requires balancing the benefits of additional data versus the potential drawbacks of statistical noise due to, for example, small numbers of available sequences or large time gaps. For this reason, we applied various selection criteria to limit the influence of noise and other artifacts in the data.

*Maximum number of gaps.* Sequences with large numbers of gaps may have large deletions or sequencing regions that only partly overlap with the region of interest. We therefore removed sequences with >200 gaps in excess of the subtype consensus sequence from the analysis.

*Maximum gap frequency.* Rare insertions, which would appear in the data as sites with high gap frequencies, may represent misalignments. We conservatively removed sites with >95% gaps from our analysis.

*Minimum number of sequences.* Additional time points are helpful for identifying selection, but variant frequencies at time points with small numbers of sequences are poorly constrained by the data. This finite sampling noise may make it difficult to reliably infer selection. We therefore dropped time points where fewer than four sequences were observed from our analysis.

*Maximum time gap.* Large time gaps can degrade performance if they cause us to miss evolutionary dynamics that are relevant for inferring selection. Here we dropped time points that were separated by >300 d from the last included time point.

*Imputation of ambiguous nucleotides.* To include sequences that contain some ambiguous nucleotides in our analysis, we imputed ambiguous nucleotides by replacing them with the most frequently observed nucleotide at the same site from that patient. Imputations were also constrained by the identity of the ambiguous nucleotide. For example, an R would be replaced by either A or G, depending on which nucleotide was more frequently observed at that site in the same patient.

For all of these exclusion criteria, different thresholds could reasonably have been chosen. Extended Data Fig. 10 shows that the selection coefficients that we infer are robust to the specific data processing choices that we have made.

In the course of data processing we also determined the number of open reading frames in which each substitution was nonsynonymous, whether it occurred within an identified CD8$^+$ T cell epitope that was actively targeted during the time for which sequence samples were available, whether it occurred within the exposed surface of Env (using surface residues as identified in ref. [63]), and whether it may have plausibly affected Env glycosylation by completing or disrupting an N-linked glycosylation motif. These analyses were performed using custom Python scripts available in the GitHub repository.

Variant indices were labeled relative to the standard HXB2 reference sequence of HIV-1. Insertions relative to HXB2 are labeled with lowercase alphabetical indices per standard conventions[72]. For example, if three nucleotides were inserted relative to HXB2 after site 1, these would be labeled 1a, 1b and 1c, respectively.

**Enrichment analyses.** We used fold enrichment values to determine the relative excess or lack of particular types of mutation among the HIV-1 variants that were inferred to be the most beneficial. For a given set of $N_{\mathrm{sel}}$ selected mutations (for example, corresponding to the top 1% most beneficial), we computed the number $n_{\mathrm{sel}}$ of these mutations that have a particular property. This may represent, for example, the number of nonsynonymous mutations within identified CD8$^+$ T cell epitopes, or the number of nonsynonymous reversions. The ratio $n_{\mathrm{sel}}/N_{\mathrm{sel}}$ then represents the fraction of the selected mutations having the specified property. This number was compared with $n_{\mathrm{null}}/N_{\mathrm{null}}$, where $N_{\mathrm{null}}$ is the total number of non-TF variants across all individuals and sequencing regions of the HIV-1 genome and $n_{\mathrm{null}}$ is the number of these variants that share the specified property.

The fold enrichment of the selected set for a specified property is then naturally defined as $(n_{\mathrm{sel}}/N_{\mathrm{sel}})/(n_{\mathrm{null}}/N_{\mathrm{null}})$. A fold enrichment value greater than one indicates a larger proportion of mutants in the selected set that have the given property than expected by chance, while a value less than one indicates a smaller proportion than expected by chance.

**Selection inference with MPL.** We implemented the MPL method as described above in C++ and applied it to infer selection coefficients from the HIV-1 sequence data and from simulations. The original code used for inference is included in the GitHub repository. For the HIV-1 data, we assumed a regularization strength of $\gamma = 5$. We also used mutation probabilities estimated in ref. [73]. as input. Mutation probabilities to and from gap states, representing deletions and insertions, respectively, were assumed to be very small ($\mu = 10^{-9}$). For the simulated data, we used a smaller regularization strength of $\gamma = 1$ due to the greater sampling depth.

For time intervals $\Delta t \gg 1$, naïve evaluation of $C_{\mathrm{int}}$ and $\boldsymbol{\mu}_{\mathrm{fl}}$ may give results that are inconsistent with more realistic, smoothly varying allele frequencies. For example, if a variant rises from frequency zero to a nonzero frequency in the final time step, the diagonal part of the integrated covariance $C_{\mathrm{int}}$ from equation (10) would formally be zero for this variant. To increase robustness and avoid unnatural covariance and flux terms, we assumed that the true underlying allele frequency trajectories were piecewise linear and replaced the sums over time in equation (12) with integrals. Following the assumption of piecewise linearity, these integrals can be computed analytically. Specifically, the contribution of the mutational term to the numerator is then

$$-\sum_{k=0}^{K-1}\Delta t_k\sum_{\delta=1}^{\ell}\left(\mu_{\delta\alpha}\frac{x_{i,\delta}(t_k)+x_{j,\delta}(t_{k+1})}{2}-\mu_{\alpha\delta}\frac{x_{j,\alpha}(t_k)+x_{j,\alpha}(t_{k+1})}{2}\right),$$

the diagonal terms of the integrated covariance matrix are

$$\sum_{k=0}^{K-1}\Delta t_k\left(\frac{\left(3-2x_{i,\alpha}(t_{k+1})\right)\left(x_{i,\alpha}(t_k)+x_{i,\alpha}(t_{k+1})\right)}{6}-\frac{x_{i,\alpha}(t_k)x_{i,\alpha}(t_k)}{3}\right),$$

and the off-diagonal terms of the integrated covariance matrix are

$$\sum_{k=0}^{K-1}\Delta t_k\frac{x_{ij,\alpha\beta}(t_k)+x_{ij,\alpha\beta}(t_{k+1})}{2}$$
$$-\sum_{k=0}^{K-1}\Delta t_k\left(\frac{x_{i,\alpha}(t_k)x_{j,\beta}(t_k)}{3}+\frac{x_{i,\alpha}(t_{k+1})x_{j,\beta}(t_{k+1})}{3}+\frac{x_{i,\alpha}(t_k)x_{j,\beta}(t_{k+1})}{6}+\frac{x_{i,\alpha}(t_{k+1})x_{j,\beta}(t_k)}{6}\right).$$

After selection coefficients were inferred, we normalized them such that the TF (HIV-1) or WT (simulation) allele had a selection coefficient of zero.

**Calculation of effects of linkage on inferred selection.** Due to the inverse of the integrated covariance matrix in equation (11), the selection coefficients estimated by MPL are affected by linkage. To quantify the effects of linkage on inferred selection during HIV-1 evolution, we computed the pairwise effects $\Delta\hat{s}_{ij}$ of each variant $i$ on selection for other variants $j$, as described in the main text. Here, for ease of notation, each effective index $i$ or $j$ represents a single non-TF nucleotide at a particular site on the genome. That is, the indices incorporate both the label for the locus and for the allele.

To compute $\Delta\hat{s}_{ij}$, we iteratively select each nucleotide at each site, which together are represented by the index $i$, and generate a modified version of the sequence data in which variant $i$ is replaced by the TF nucleotide at the same site. In this way, linkage between the masked variant $i$ and all other variants $j$ is eliminated. We then infer the selection coefficients again for all variants $j$ using the data where variant $i$ has been replaced by TF, denoted as $\hat{s}_j^{\setminus i}$. Then we define

$$\Delta\hat{s}_{ij} = \hat{s}_j - \hat{s}_j^{\setminus i}.$$

Positive values of $\Delta\hat{s}_{ij}$ thus indicate that linkage with variant $i$ increases the selection coefficient inferred for variant $j$. This may be due, for example, to clonal interference between variants $i$ and $j$. Negative values indicate that variant $i$ decreases the selection coefficient inferred for variant $j$. This may occur, for example, if variant $j$ hitchhikes on a beneficial genetic background that includes variant $i$.

**Statistics and reproducibility.** Details of enrichment analysis are given in Methods. The $P$ values were calculated using the two-sided Fisher's exact test. Simulation results in Fig. 2 and Extended Data Figs. 1–4 were computed on 100 evolutionary histories each obtained from an independent Monte Carlo run. The theoretical covariance in the inferred selection coefficients can be computed from equation (9), which is given by $C_{\mathrm{int}}^{-1}$. In Fig. 1 and Extended Data Fig. 1b, we show theoretical standard deviations on the inferred selection coefficients, computed by the square root of the diagonal entries of $C_{\mathrm{int}}^{-1}$.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Raw data used in our analysis is available in the GitHub repository located at https://github.com/bartonlab/paper-MPL-inference. Source data are provided with this paper.

## Code availability

Code used in our analysis is available in the GitHub repository located at https://github.com/bartonlab/paper-MPL-inference. The repository also contains

Jupyter notebooks that can be run to reproduce the results presented here. The source code is shared under GPL-3.0 license https://github.com/bartonlab/paper-MPL-inference/blob/master/LICENSE-GPL. An executable version is also provided on Code Ocean at https://codeocean.com/capsule/3400567/tree (ref. [30]), distributed under the GPL-3.0 license https://opensource.org/licenses/gpl-license/.

## References


66. Kimura, M. Diffusion models in population genetics. *J. Appl. Probab.* **1**, 177–232 (1964).
67. Tataru, P., Bataillon, T. & Hobolth, A. Inference under a Wright-Fisher model using an accurate beta approximation. *Genetics* **201**, 1133–1141 (2015).
68. He, Z., Beaumont, M. & Yu, F. Effects of the ordering of natural selection and population regulation mechanisms on Wright-Fisher models. *G3: Genes, Genomes, Genetics* **7**, 2095–2106 (2017).
69. Tataru, P., Simonsen, M., Bataillon, T. & Hobolth, A. Statistical inference in the Wright-Fisher model using allele frequency data. *Syst. Biol.* **66**, e30–e46 (2017).
70. Risken, H. *The Fokker–Planck Equation: Methods of Solution and Applications* 2nd edn (Springer, 1989).
71. Gaschen, B., Kuiken, C., Korber, B. & Foley, B. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics* **17**, 415–418 (2001).
72. Korber, B. et al. in *Human Retroviruses and AIDS* (eds Korber, B. et al.) 102–111 (Los Alamos National Laboratory, 1998)..
73. 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 Evol.* **3**, vex003 (2017).


## Acknowledgements

We thank A.K. Chakraborty, C.J.R. Illingworth, B. Lee and J.G. Schraiber for helpful discussions and comments on the manuscript. The work of M.S.S., R.H.Y.L. and M.R.M. was supported by the Hong Kong Research Grants Council under grant number 16234716. M.S.S. and M.R.M. were also supported by the Hong Kong Research Grants Council under grant number 16201620, while R.H.Y.L. was also supported by Australia's National Health and Medical Research Council under grant number APP1121643. The work of 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 R35GM138233.

## Author contributions

All authors designed research, developed methods, analyzed data, interpreted results and wrote the paper.

## Competing interests

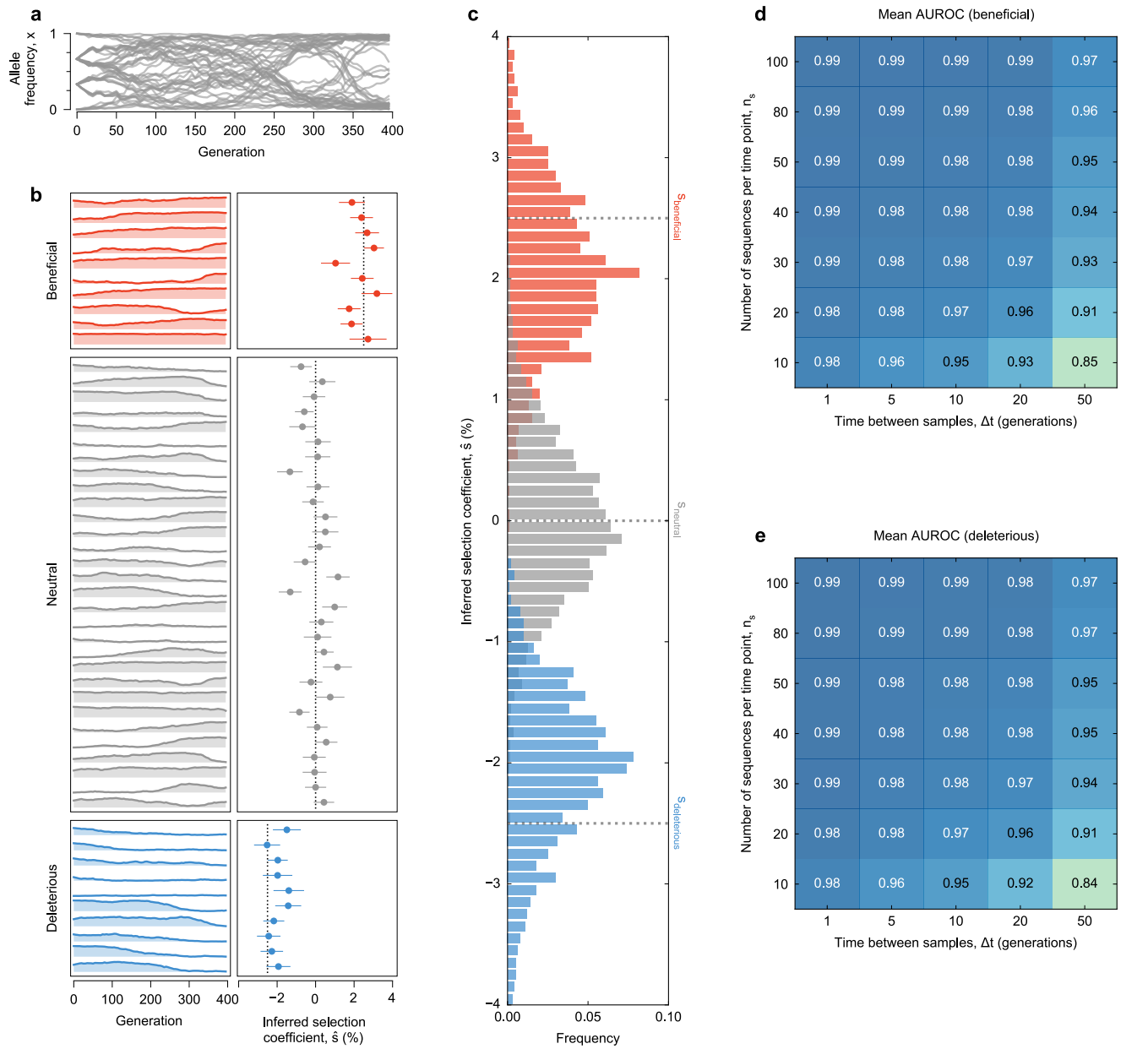
The authors declare no competing interests.

## Additional information

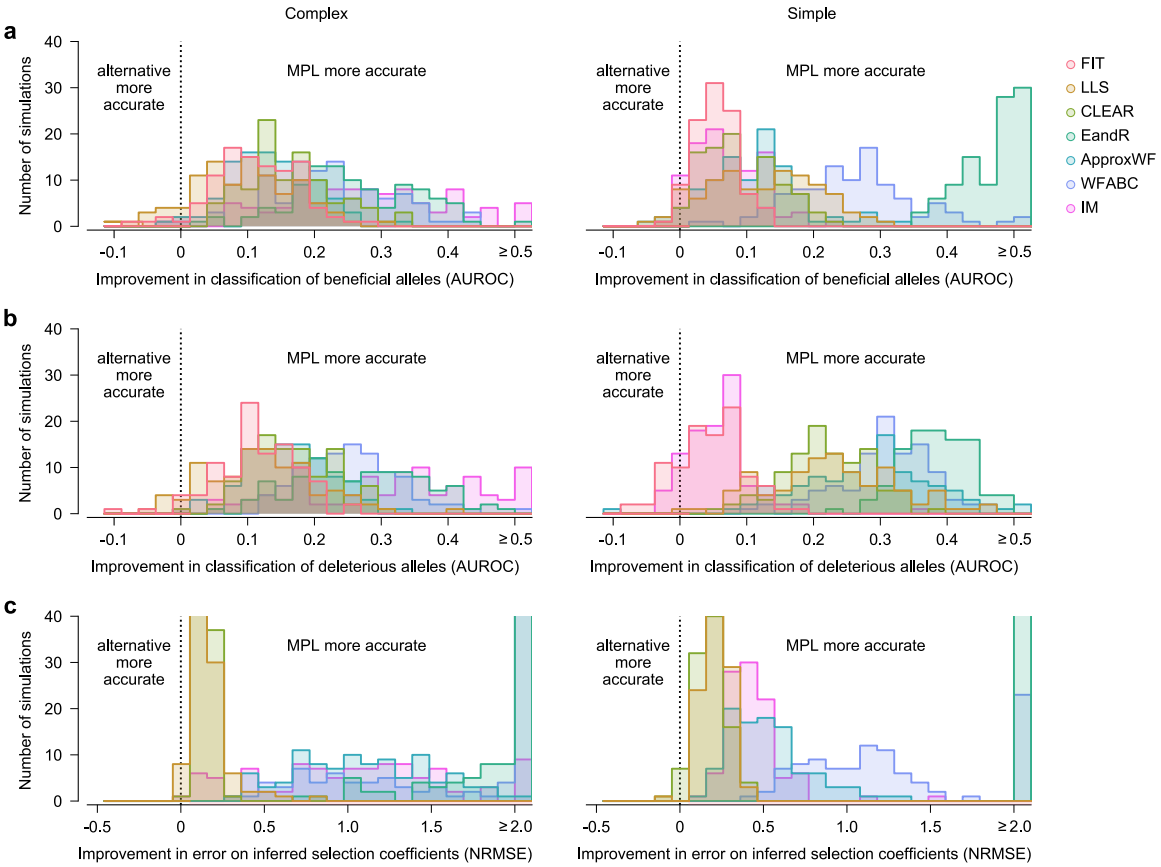
**Extended data** is available for this paper at https://doi.org/10.1038/s41587-020-0737-3.

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41587-020-0737-3.

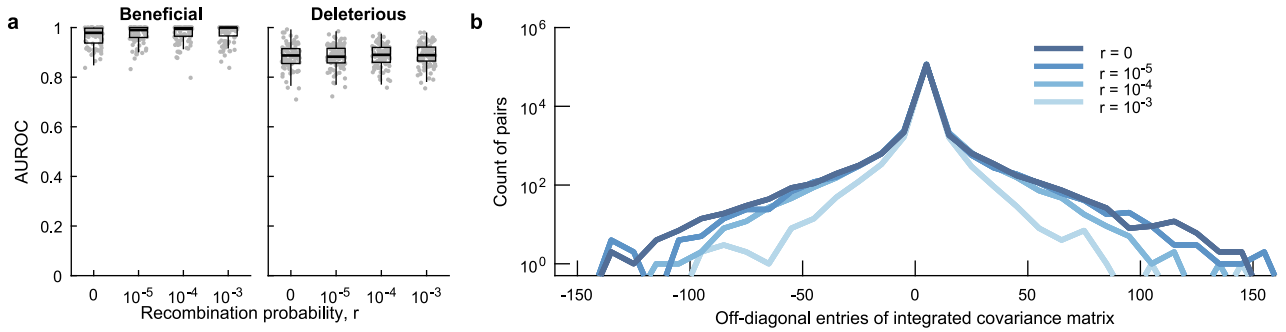
**Correspondence and requests for materials** should be addressed to M.R.M. or J.P.B.



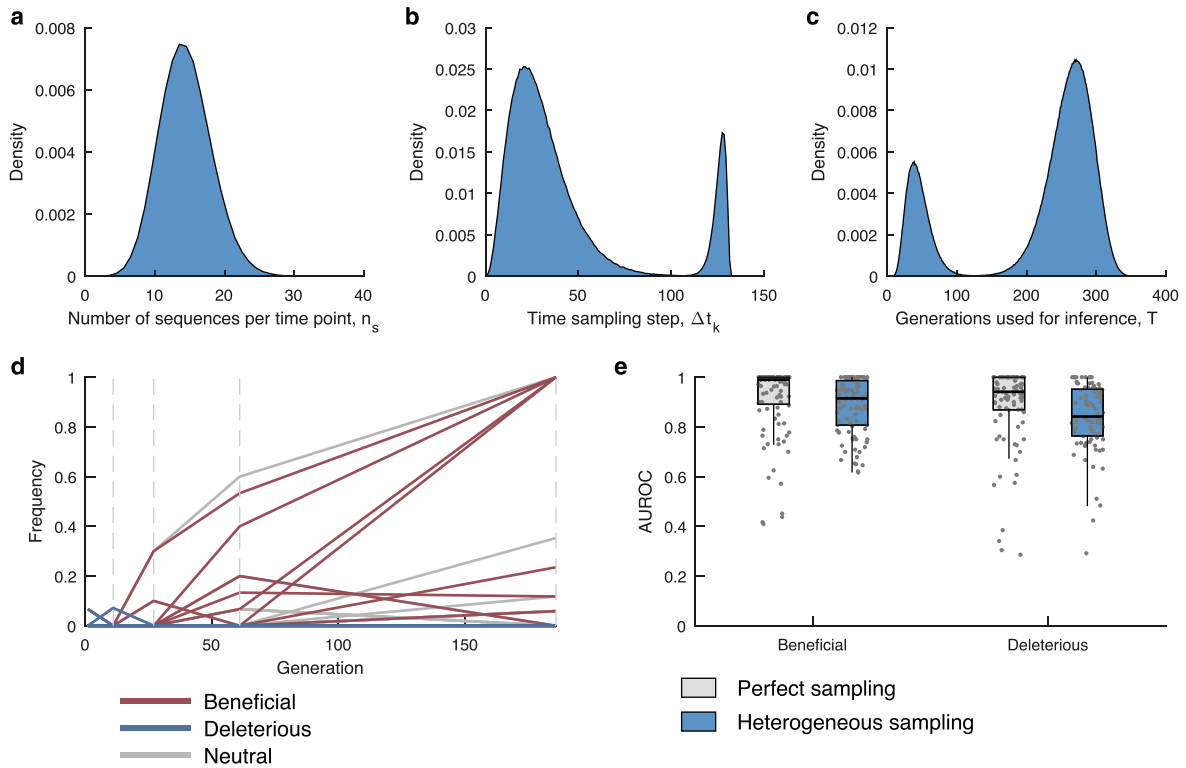

**Extended Data Fig. 1 | MPL accurately recovers selection coefficients from complex simulated evolutionary trajectories. a**, Trajectories of mutant allele frequencies over time exhibit complex dynamics in a WF simulation with a simple fitness landscape. **b**, Separate views of individual trajectories for beneficial, neutral, and deleterious mutants (*left panel*) and inferred selection coefficients (*right panel*) for a single simulation run. Note that many neutral mutations exhibit temporal variation similar to beneficial or deleterious mutations. MPL estimates the underlying selection coefficients used to generate these trajectories, presented as mean values ± one theoretical standard deviation, and distinguishes between beneficial, neutral, and deleterious mutations, using Eq. (11). Dashed lines mark the true selection coefficients. **c**, Distributions of selection coefficient estimates across $n = 100$ replicate simulations with identical parameters in the special case of perfect sampling. MPL is also robust to finite sampling constraints, accurately classifying beneficial (**d**) and deleterious (**e**) mutants even when the number of sequences sampled per time point $n_s$ is low, and the spacing between time samples $\Delta t$ is large. *Simulation parameters*. $L = 50$ loci with two alleles at each locus (mutant and WT): ten beneficial mutants with $s = 0.025$, 30 neutral mutants with $s = 0$, and ten deleterious mutants with $s = -0.025$. Mutation probability $\mu = 10^{-3}$, population size $N = 10^3$. Initial population composed of approximately equal numbers of three random founder sequences, evolved over $T = 400$ generations.



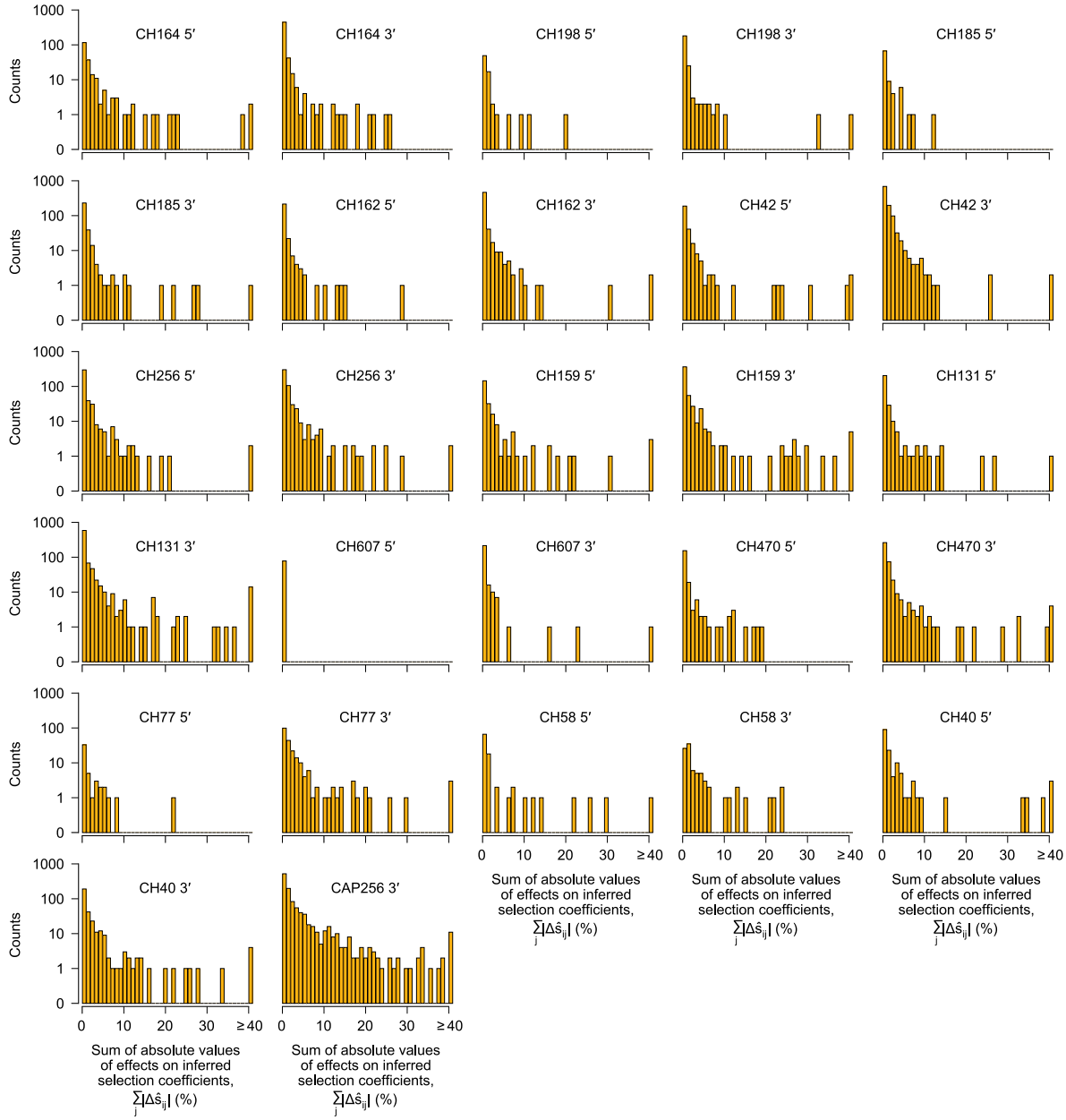

**Extended Data Fig. 2 | MPL improves selection inference for simulated data sets.** In Fig. 2, we showed the performance of MPL and existing methods on simulated test data, averaged over $n = 100$ replicate simulations with identical parameters. Here we show the improvement of MPL over existing methods for the classification of beneficial (**a**) and deleterious (**b**) mutations, and for the error in the estimated selection coefficients (**c**), for each individual simulation. Selection is more difficult to infer in some simulated data sets, but results from MPL show better agreement with the true parameters in the vast majority of simulations. *Simulation parameters*. $L = 50$ loci with two alleles at each locus (mutant and WT): ten beneficial mutants (with $s = 0.1$ for complex, $s = 0.025$ for simple), 30 neutral mutants ($s = 0$ for both scenarios), and ten deleterious mutants ($s = -0.1$ for complex, $s = -0.025$ for simple). Mutation probability $\mu = 10^{-4}$, population size $N = 10^3$. For the complex case, the initial population is composed of equal numbers of five random founder sequences, evolved over $T = 310$ generations. Recorded trajectory used for inference begins at generation 10. For the simple case, the initial population begins with all WT sequences, evolved over $T = 1000$ generations.



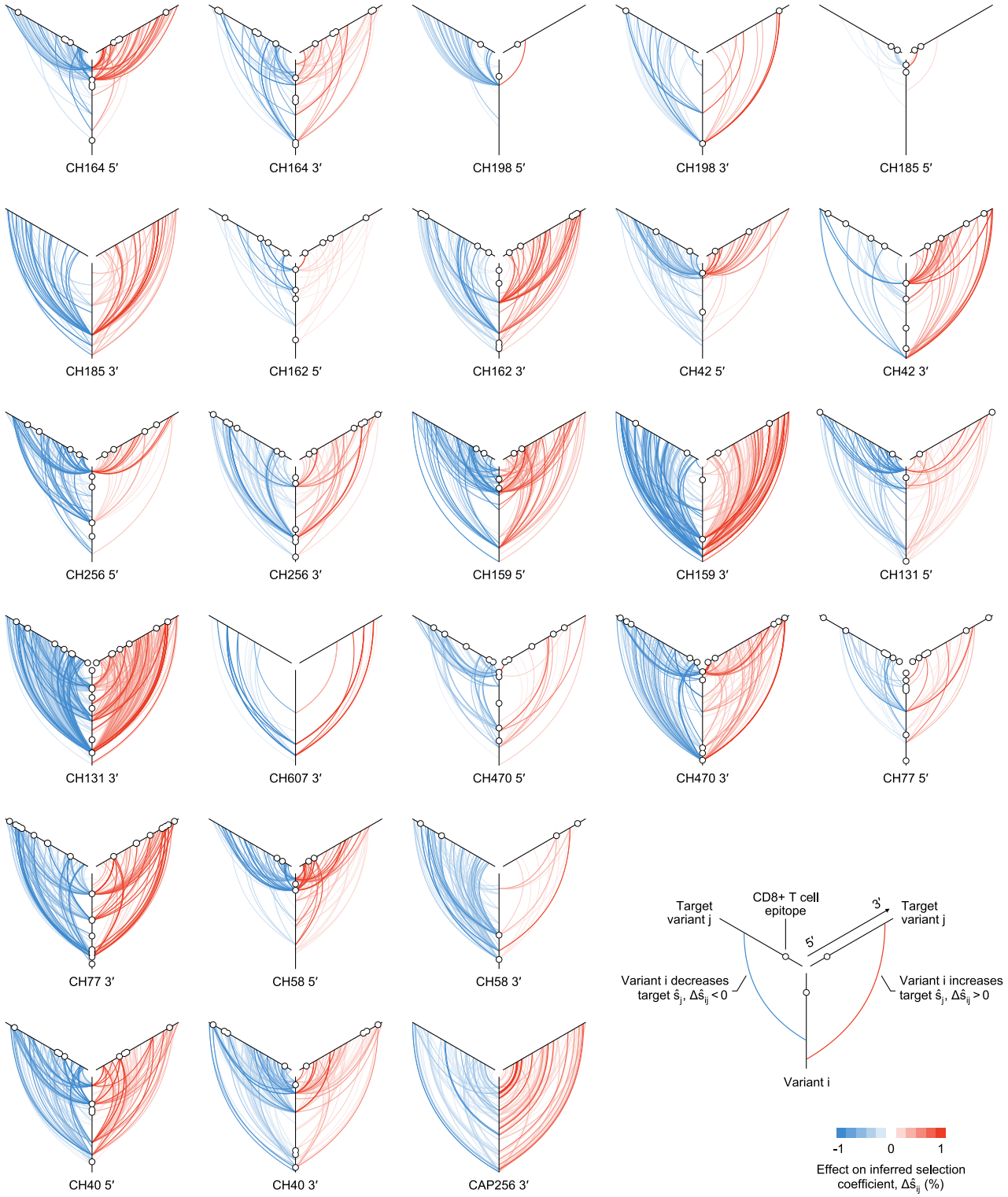

**Extended Data Fig. 3 | MPL performs well in the presence of recombination. a**, Classification performance of MPL is robust to variation in per locus recombination probability, r. Results are shown for $n = 100$ independent Monte-Carlo runs. The lower and upper edge of the boxplot correspond to the 25th to 75th percentiles, the bar corresponds to the median while the top and bottom whiskers show the maximum and minimum value within 1.5× the interquartile range from the boxplot. Linkage effects in the data decrease as the recombination probability increases. As a measure of the linkage disequilibrium in the data, we plot the histograms (**b**) of the covariance ($x_{ij} - x_i x_j$) of mutant allele frequencies integrated over time (300 generations) for a range of recombination probabilities. The number of mutant pairs with strong pairwise covariance values decrease with increasing values of $r$, indicating lower linkage disequilibrium. *Simulation parameters.* Same as those of simple scenario used in Fig. 2, that is, $L = 50$ loci with two alleles at each locus (mutant and WT): ten beneficial mutants ($s = 0.025$), 30 neutral mutants ($s = 0$), and ten deleterious mutants ($s = -0.025$). Mutation probability $\mu = 10^{-4}$, population size $N = 10^3$, $r = \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$. The initial population begins with all WT sequences, evolved over $T = 300$ generations.



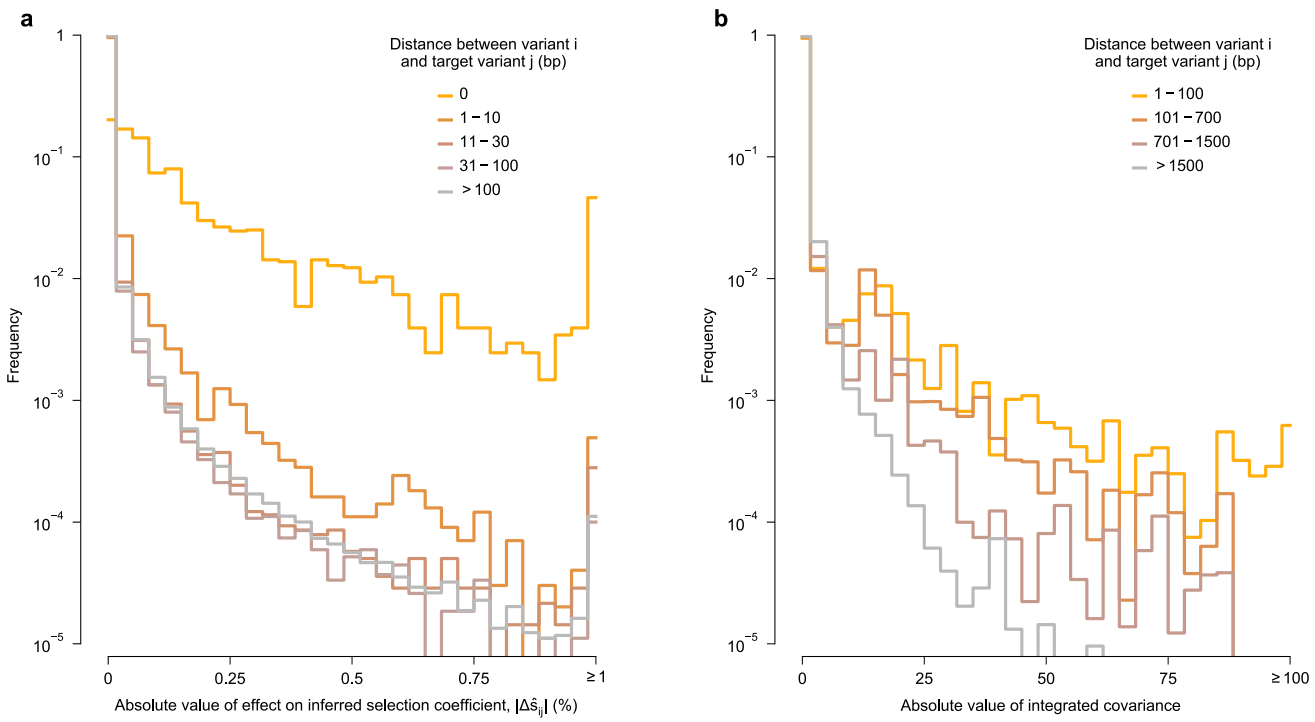

**Extended Data Fig. 4 | Performance of MPL on data with HIV-1-like sampling profiles. a**, The number of sequences per time point $n_s$ are drawn from a binomial distribution with $n = 1000$ and $p = 0.0139$, with the same mean as that of the HIV data. **b**, The time between samples is drawn from a mixture of two gamma distributions $f(x;k,\theta)$, where $k$ and $\theta$ are the shape and scale parameters. The mixture distribution has the form $w_1 \times (f(x;k_1,\theta_1) + m_1) + w_2 \times (f((k_2\theta_2 + m_2 - x);k_2,\theta_2) + m_2)$ where $m_1 = 0$, $m_2 = 120$, are constants added to shift the mean, $k_1 = 3.5$, $k_2 = 3$, $\theta_1 = 8.4$, $\theta_2 = 2$, while $w_1 = 0.87$, and $w_2 = 0.13$ are the mixing weights. The parameters were chosen to mimic the distribution of the time between samples of the HIV data analyzed in the manuscript (Supplementary Table 1). **c**, The number of generations used for inference is also drawn from a mixture of two gamma distributions, having the form given above and with parameters $k_1 = 5.5$, $k_2 = 15$, $\theta_1 = 7.2$, $\theta_2 = 8$, $m_1 = 5$, $m_2 = 143$, $w_1 = 0.21$, and $w_2 = 0.79$. The parameters were chosen to mimic the distribution of the trajectory lengths of the HIV data analyzed in the manuscript (Supplementary Table 1). **d**, A typical sampled trajectory of allele frequencies: beneficial (red), deleterious (blue) and neutral (gray). Dashed lines indicate the sampling time-points. **e**, The AUROC performance of identifying beneficial and deleterious selection coefficients under perfect and heterogeneous sampling scenarios. Results are evaluated for those sites that are polymorphic in the heterogeneous sampling case. Results are shown for $n = 100$ independent Monte-Carlo runs. The lower and upper edge of the boxplot correspond to the 25th to 75th percentiles, the bar corresponds to the median while the top and bottom whiskers show the maximum and minimum value within 1.5× the interquartile range from the boxplot. *Simulation parameters*: population size $N = 1000$, $L = 50$ loci with two alleles at each locus (mutant and WT), ten beneficial mutants with selection coefficients $s$ uniformly distributed over the range [0.075, 0.125], 30 neutral mutants with $s = 0$, and ten deleterious mutants with selection coefficients uniformly distributed over the range [-0.125, -0.075], mutation probability per site per generation $\mu = 10^{-4}$, and recombination probability per site per generation $r = 10^{-4}$.



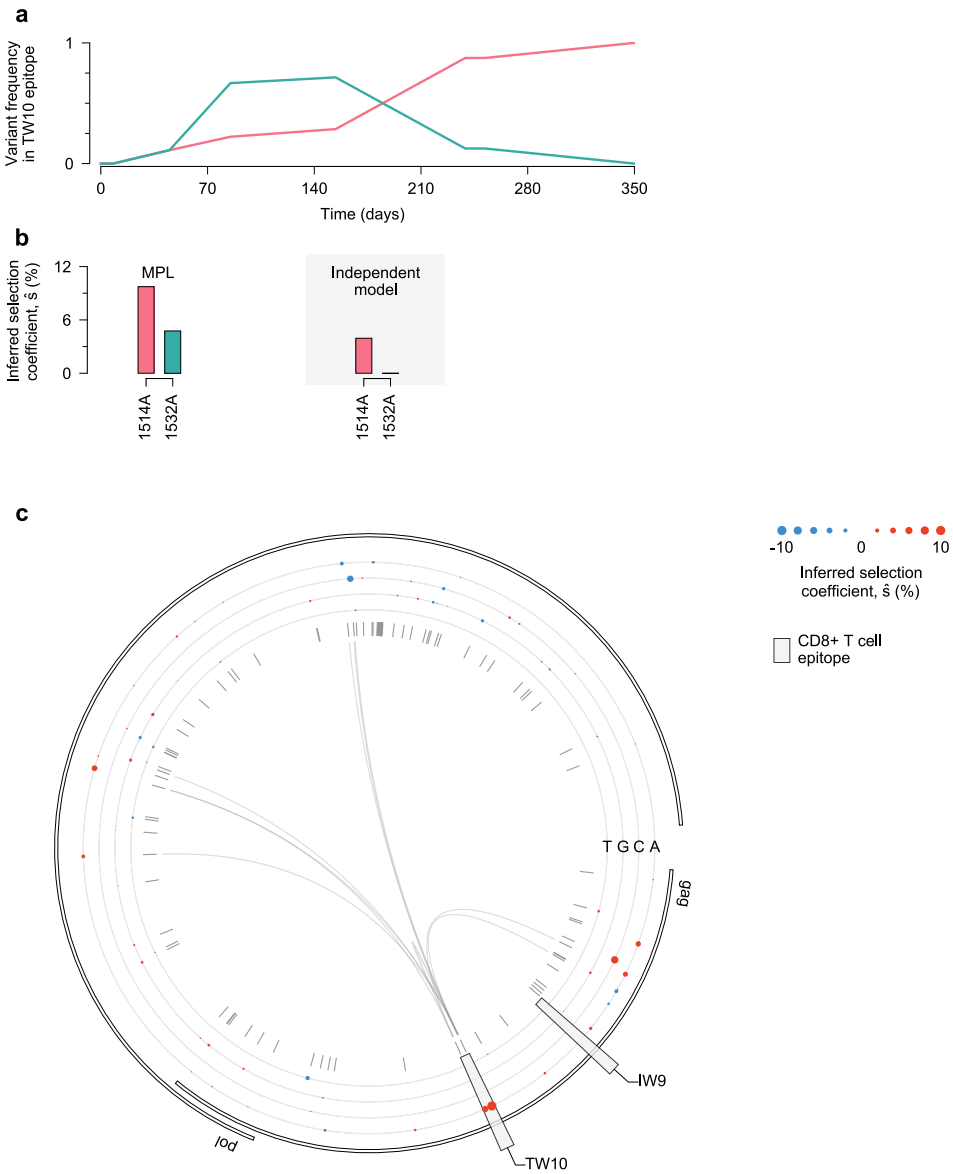

**Extended Data Fig. 5 | Most genetic variants have little effect on inferred selection at other sites, but a small minority have strong effects.** After computing the pairwise effects $\Delta\hat{s}_{ij}$ of each variant *i* on the inferred selection coefficient for each other variant *j*, referred to as the target, we summed the absolute value of the $\Delta\hat{s}_{ij}$ values over all target variants *j* to quantify the influence of each variant *i* on selection at other sites. One histogram is shown for each sequencing region, for each individual. For the vast majority of variants, the total effect on selection at other sites is near zero. However, a small minority have strong effects. We defined a variant to be 'highly influential' if the sum of the absolute values of the $\Delta\hat{s}_{ij}$ over all targets *j* was larger than 0.4 (=40%).



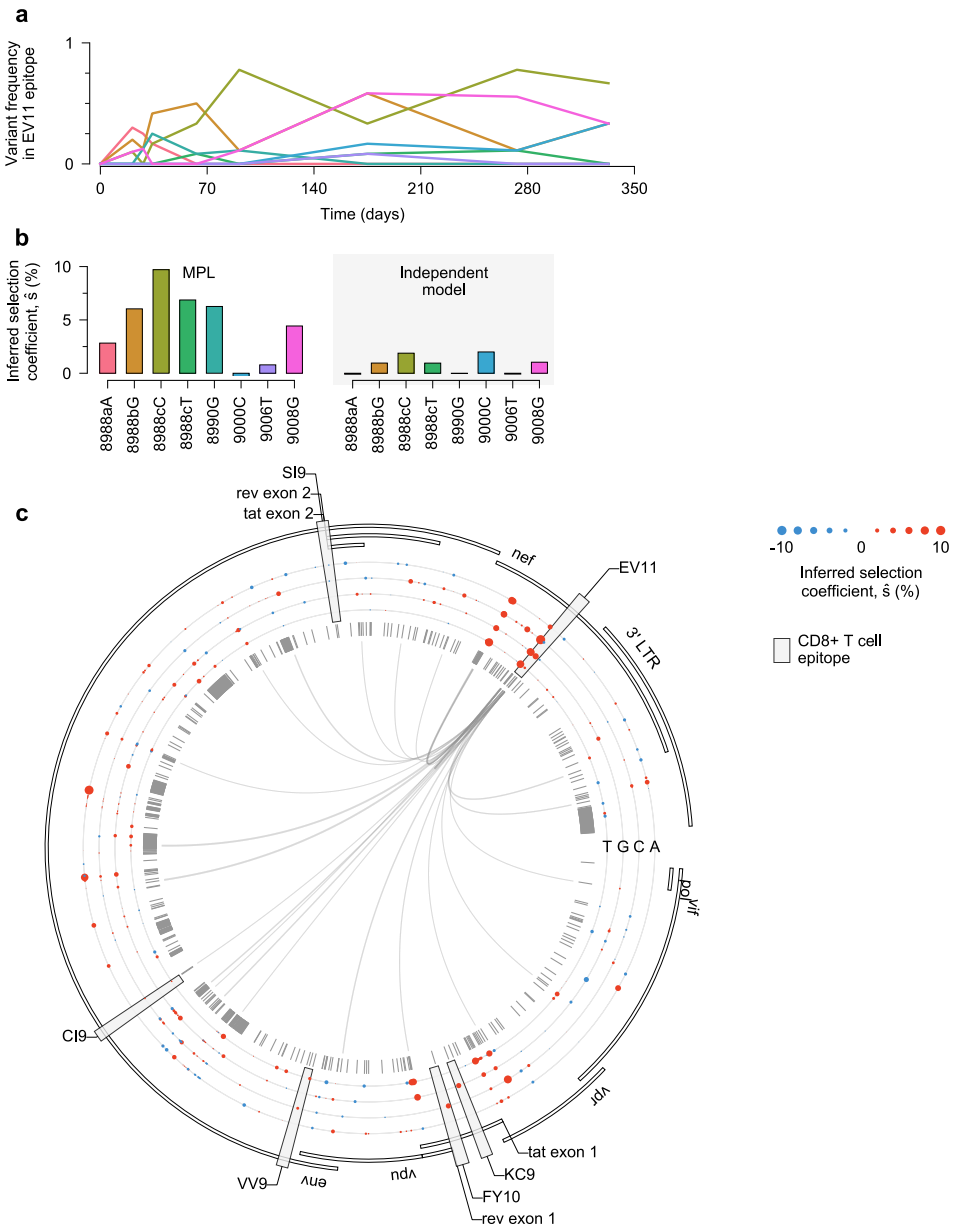

**Extended Data Fig. 6 | Variants that strongly influence inferred selection at other sites often act across large genomic distances.** Plot of all linkage effects on inferred selection coefficients $\Delta\hat{s}_{ij}$ for which $|\Delta\hat{s}_{ij}| > 0.004$. One plot is shown for each sequencing region, for each individual. These strong effects of linkage on inferred selection coefficients can act at long range across the genome. Approximately 40% of highly influential variants, characterized by strong effects on inferred selection at other sites, lie within identified CD8$^+$ T cell epitopes. The 5′ region for individual CH607 is not shown because no $\Delta\hat{s}_{ij}$ values are larger than the cutoff.



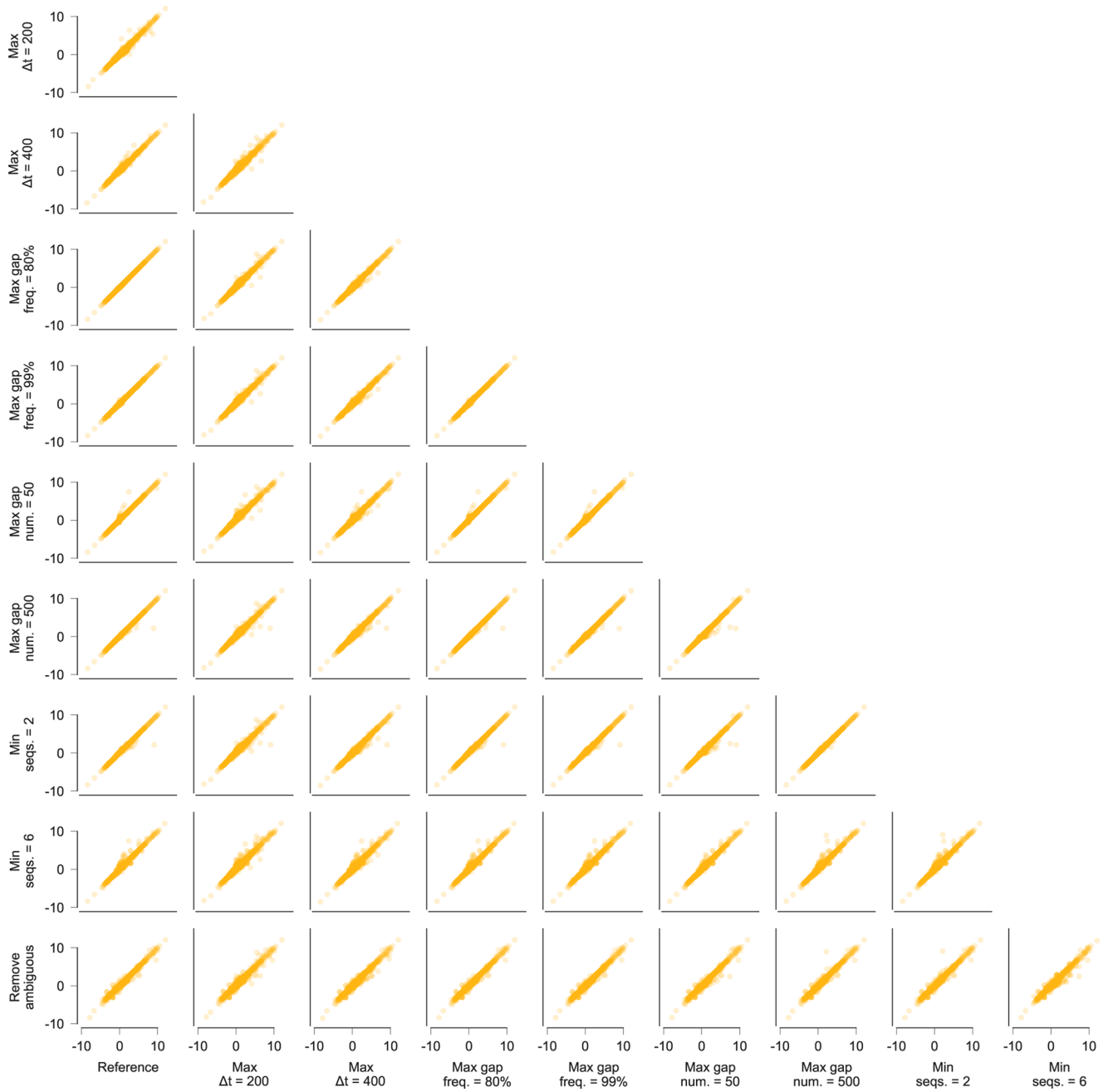

**Extended Data Fig. 7 | For most variants, effects on inferred selection coefficients for other variants, and linkage disequilibrium, are stronger at smaller genomic distances. a**, Histogram of the absolute value of linkage effects on inferred selection coefficients for other variants $|\Delta\hat{s}_{ij}|$, divided into subgroups based on the distance along the genome between variant *i* and target variant *j*. Consistent with intuition, the large effects on inferred selection coefficients occur most frequently for different variants that occur at the same site on the genome (that is, distance equal to zero). 'Interactions' between such variants are necessarily perfectly competitive because only a single nucleotide is allowed at each position in the genetic sequence. For most variants, stronger linkage effects on inferred selection coefficients are more frequently observed for other variants within a distance of ten base pairs (bp). Large linkage effects for pairs of variants within a distance of 30 bp, the approximate length of a linear T cell epitope, occur appreciably more frequently than for pairs of variants at greater genomic distances. However, there is little difference in the distribution of linkage effect sizes for pairs of variants that are between 31 bp and 100 bp apart compared to pairs of variants that are more than 100 bp apart. Nonetheless, some strong linkage effects on inferred selection are observed at long genomic distances (see Fig. 4 and Supplementary Fig. 5). **b**, Linkage disequilibrium, measured by the absolute value of the off-diagonal entries of the integrated allele frequency covariance matrix, $C_{\text{int}}$. Like the $|\Delta\hat{s}_{ij}|$, linkage decays along with the distance between variants along the genome. However, we note that linkage disequilibrium values in general appear to be more long-ranged.




**Extended Data Fig. 8 | Estimates of selection coefficients in a simple example of clonal interference. a**, Two escape mutations arise in the TW10 epitope targeted by individual CH58 and compete for dominance. **b**, MPL infers that both TW10 escape variants are positively selected. Estimates based on trajectories of individual variants only infer substantial positive selection for the 1514A variant that fixes. The magnitude of selection inferred with the independent model is also smaller than that inferred by MPL. **c**, Inferred selection in the HIV-1 5′ half-genome sequence for CH58. Inferred selection coefficients are plotted in tracks. Coefficients of transmitted/founder nucleotides are normalized to zero. Tick marks denote polymorphic sites. Inner links, shown for sites connected to the TW10 epitope, have widths proportional to matrix elements of the inverse of the integrated covariance. Linked sites affect selection estimates within the epitope.




**Extended Data Fig. 9 | Estimates of selection coefficients in a complex example of clonal interference. a**, Multiple escape variants for the Nef epitope EV11, targeted by individual CH131, interfere with one another over the course of nearly one year. Here we have omitted the trajectories for transient variants with a deletion at sites 8988a-8988c, which are insertions with respect to the HXB2 reference sequence. **b**, MPL infers that all nonsynonymous EV11 escape variants are positively selected. Variants 9000C and 9006T are both synonymous, and are inferred to be nearly neutral by MPL. As in previous examples, inferences using only the trajectories of individual variants only infer substantial positive selection for variants that are polymorphic at the final time point, or where the transmitted/founder (TF) allele at the same site appears strongly selected against. In the latter case, positive selection is inferred because all selection coefficients are normalized such that the selection coefficient for the TF variant is zero. This is why the independent model infers 8988T to be beneficial despite its low frequency at the final time point. Note that the independent model also infers the synonymous mutation 9000C to be beneficial. **c**, Inferred selection in the HIV-1 3′ half-genome sequence for CH131. Inferred selection coefficients are plotted in tracks. Coefficients of TF nucleotides are normalized to zero. Tick marks denote polymorphic sites. Inner links, shown for sites connected to the EV11 epitope, have widths proportional to matrix elements of the inverse of the integrated covariance. Linked sites affect selection estimates within the epitope.




**Extended Data Fig. 10 | Inferred selection coefficients across patients using different conventions for data processing.** Inferred selection coefficients are highly similar following different choices for processing the sequence data. Pearson $R^2$ values between inferred selection coefficients range from 0.97 to 1.00, with an average of 0.99. Data processing conventions. *Reference*: current data processing conventions. Max $\Delta t$ = 200/400: remove time points that are more than 200/400 days beyond the last included time point (reference: 300 days). *Max gap freq. = 80%/99%:* remove sites where >80%/99% of observed variants are gaps (reference: 95%). *Max gap num. = 50/500*: remove sequences with >50/500 gaps in excess of subtype consensus (reference: 200). *Min seqs. = 2/6*: remove time points with <2/6 available sequences (reference: 4). *Remove ambiguous*: remove sequences that contain ambiguous nucleotides if any other nucleotide variation is observed at the same site. LTR, long terminal repeat.

nature research

Corresponding author(s): Matthew R McKay and John R Barton

Last updated by author(s): Oct 12, 2020

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Sequence data was downloaded from Los Alamos National Laboratory HIV Sequence Database (www.hiv.lanl.gov; accessed October 19, 2018). |
| Data analysis | We implemented the Wright-Fisher model with discrete generations in Python v3.7.<br>We implemented the MPL method as described above in C++. All codes and dependencies are provided in GitHub repository located at https://github.com/bartonlab/paper-MPL-inference<br>Comparison with other methods in literature:<br>Third party codes for WFABC, ApproxWF, LLS, CLEAR, EandR were obtained from web sources cited in the respective papers and installed following the instructions therein. The exact procedure followed is given in the Jupyter notebooks uploaded at https://github.com/bartonlab/paper-MPL-inference.<br>The code that executes the methods IM, WFABC, FIT, ApproxWF was written in MATLAB 2017. The script that runs the IM method requires the "global optimization toolbox" of Matlab. All details of running these codes are provided in the GitHub repository located at https://github.com/bartonlab/paper-MPL-inference. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw data and code used in our analysis is available in the GitHub repository located at https://github.com/bartonlab/paper-MPL-inference. This repository also contains Jupyter notebooks that can be run to reproduce the results presented here.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | We describe a new framework that infers selection from evolutionary histories by accounting for genetic linkage. |
| Research sample | We studied HIV-1 sequence data obtained from 14 individuals recruited under the CHAVI 001 and CAPRISA 002 studies in the United States, Malawi, and South Africa. |
| Sampling strategy | We analyzed a previously reported data set and thus sampling was not in our control. |
| Data collection | We downloaded time-series genomic data from Los Alamos National Laboratory HIV Sequence Database (www.hiv.lanl.gov). |
| Timing and spatial scale | The data typically spanned from initial infection of the patient up to several years. |
| Data exclusions | In order to assure sequence quality, we 1) removed sequences with ≥ 200 gaps with respect to clade consensus, 2) removed sites with > 95% gaps in the alignment, 3) imputed ambiguous nucleotides, and 4) removed time points where < 4 sequences were sampled, or where the time gap between successive samples exceeded 300 days. |
| Reproducibility | Data and scripts required to reproduce the results are available at https://github.com/bartonlab/paper-MPL-inference |
| Randomization | Not applicable |
| Blinding | Not applicable |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

### Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |