Finite-sites multiple mutations interference gives rise to wavelet-like oscillations of multilocus linkage disequilibrium

Victor Garcia,[1,*] Emily C. Glassberg,[1] Arbel Harpak,[1] and Marcus W. Feldman[1]

[1] *Department of Biology, Stanford University, 371 Serra Mall, Stanford CA-94305, USA*

Within-host adaptation of pathogens such as human immunodeficiency virus (HIV) often occurs at more than two loci. Multiple beneficial mutations may arise simultaneously on different genetic backgrounds and interfere, affecting each other's fixation trajectories. Here, we explore how these adaptive dynamics are mirrored in multilocus linkage disequilibrium (MLD), a measure of multi-way associations between alleles. In the parameter regime corresponding to HIV, we show that deterministic early infection models induce MLD to oscillate over time in a wavelet-like fashion. We find that the frequency of these oscillations is proportional to the rate of adaptation. This signature is robust to drift, but can be eroded by high variation in fitness effects of beneficial mutations. Our findings suggest that MLD oscillations could be used as a signature of interference among multiple equally advantageous mutations and may aid the interpretation of MLD in data.

INTRODUCTION

Many microorganisms, viruses, and cancer cells replicate asexually with large population sizes and under strong selection [1–7]. This gives rise to pronounced genetic *interference* [2, 6, 8, 9], where beneficial mutations can emerge on different haplotypes and compete, leading to mutual growth impairment [10–16]. Since interference determines how asexual organisms adapt, it is of particular relevance to understanding infectious disease agents.

Most recent theoretical treatments of interference rest upon the assumption that there exists an infinite supply of new, beneficial mutations arising from an infinite number of loci [2, 16–19]. Indeed, recent studies in yeast and other microorganisms suggest that beneficial mutation supply is not what limits the rate of adaptation [1–6]. However, the infinite-sites assumption might be inappropriate for understanding many asexual populations evolving over short time scales under strong selective pressure. In fact, most short-term adaptation of pathogens to new host environments occurs at a limited number of either known [7], or detectable loci [20]. Key examples are drug resistance mutations or escape mutations in viruses [21], such as HIV [8, 22]. Furthermore, the selective pressures exerted on pathogens by a finite number of immune responses during early infection typically far exceed what other stresses might shape their adaptation [20, 22–25].

Past research on finite-sites interference typically involved only a few, predominantly two loci [12, 26–31]. Traditionally, pair-wise linkage disequilibrium (LD) has been a successful summary statistic for such interference [31, 32]. For example, if two mutations are physically linked by the same background, this will result in pronounced positive LD. If they disproportionately often appear on different backgrounds, as is the case in interference scenarios, this will be reflected in negative LD [12, 15, 26, 33]. To our knowledge, no analogue to this LD behavior has been found for adaptation in multiple (> 2) sites.

Here, we aim to extend these insights on the relationship between interference and LD to a context with multiple, but not infinite sites. First, we explore ways to generalize LD to multiple loci. *Multi-locus linkage disequibrium* (MLD) [34–37], has the advantage that it accounts for deviations from random association at more than two loci. MLD may thus appropriately reflect and characterize finite-sites interference. To compute MLD, we develop a recursive programming method applicable to up to seven loci. Second, we investigate the behavior of MLD under a finite-sites model of *multiple mutations interference* (MMI) [16] –a simplified model of interference– with parameter values calibrated to match HIV early infection dynamics. We focus on MMI due to its well-established theoretical framework [16] and its ability to appropriately describe more complex forms of interference [4].

We show that the evolution of MLD over time is interpretable and largely robust to drift. In deterministic scenarios, MMI causes MLD to oscillate, with a frequency proportional to the speed of evolution. Drift causes these features to become less pronounced, but still detectable. MLD oscillations can be further eroded by variation in fitness effects. We conclude that the wavelet-like oscillatory behavior of MLD results from, and is a robust signature of, finite-sites MMI.

RESULTS

Partition based definition makes MLD computationally tractable

To analyze how MLD is affected by multilocus interference during evolutionary dynamics, we first require a method to compute MLD. MLD, as formulated by Geiringer and Bennet, generalizes the notion of linkage disequilibrium from two to multiple loci using the principle that, due to the decay of allelic associations in haplo-

* Corresponding author: victor.garcia_palencia@alumni.ethz.ch

types as a result of recombination, MLD between neutral genes should decrease exponentially over time [34, 35].

Consider L loci with alleles $i_1, i_2, \ldots, i_L$ and allele frequencies $p_{i_1}, p_{i_2}, \ldots, p_{i_L}$. Let $p_{i_1 i_2 \ldots i_L}$, denote the frequency of haplotype $\mathbf{i} = i_1 i_2 \ldots i_L$, in the population. As introduced by Bennett [35], functions of allele and haplotype (i.e. gamete) frequencies, which satisfy the aforementioned decay condition are

$$D_{i_1 i_2} = p_{i_1 i_2} - p_{i_1} p_{i_2} \tag{1}$$

$$D_{i_1 i_2 i_3} = (p_{i_1 i_2 i_3} - p_{i_1} p_{i_2 p_{i_3}}) - p_{i_1} D_{i_2 i_3} - p_{i_2} D_{i_1 i_3} - p_{i_3} D_{i_1 i_2} \tag{2}$$

$$\begin{aligned} D_{i_1 i_2 i_3 i_4} =& (p_{i_1 i_2 i_3 i_4} - p_{i_1} p_{i_2} p_{i_3} p_{i_4}) \\ & - p_{i_1} D_{i_2 i_3 i_4} - p_{i_2} D_{i_1 i_3 i_4} \\ & - p_{i_3} D_{i_1 i_2 i_4} - p_{i_4} D_{i_1 i_2 i_3} \\ & - D_{i_1 i_2} D_{i_3 i_4} - D_{i_1 i_3} D_{i_2 i_4} - D_{i_1 i_4} D_{i_2 i_3} \\ & - p_{i_1} p_{i_2} D_{i_3 i_4} - p_{i_1} p_{i_3} D_{i_2 i_4} - p_{i_1} p_{i_4} D_{i_2 i_3} \\ & - p_{i_2} p_{i_3} D_{i_1 i_4} - p_{i_2} p_{i_4} D_{i_1 i_3} - p_{i_3} p_{i_4} D_{i_1 i_2} \end{aligned} \tag{3}$$

$$D_{i_1 i_2 i_3 i_4 i_5} = (p_{i_1 i_2 i_3 i_4 i_5} - - p_{i_1} p_{i_2} p_{i_3} p_{i_4} p_{i_5}) - \ldots \tag{4}$$

In equations (1–4), the terms $(p_{i_1 \ldots i_L} - p_{i_1} \ldots p_{i_L})$ are called *Dausset's disequilibrium* [38]. MLD, defined by $D_{i_1 \ldots i_L}$, measures how much of Dausset's disequilibrium cannot be attributed to lower-order associations of alleles. What remains is the unexplained over- or under-representation of the L^{th} order haplotype $i_1 \ldots i_L$ only, or the L^{th} order MLD [34, 35]. Equations (1–4) are valid for multiple alleles at any locus j, but we will restrict our analysis to bi-allelic loci, $i_j \in \{0, 1\}$.

Equations (1–4) for MLD can be expressed in a more concise fashion by means of partition theory, as shown by Gorelick and Laubichler [39, 40]. We add a superscript L to indicate the LD of L^{th} order, given L loci, and write:

$$D^L_{i_1 \ldots i_L} = p_{i_1 \ldots i_L} - \sum_{A \in E} \left[\prod_{u=1}^{|A|} D^{|a_u|}_{\mathbf{i}_{a_u}} \right], \tag{5}$$

where E is the set partition of the set $\{1, \ldots, j, \ldots, L\}$, except for the trivial cell $\{\{1, \ldots, L\}\}$. The *set partition* Ξ of a set S is a family of sets A, called *cells*, which contains all non-empty disjoint subsets of S, whose union is S. For example, the set partition Ξ of $\{1, 2, 3\}$ is $\{\{1, 2, 3\}\}$, $\{\{1, 2\}, \{3\}\}$, $\{\{1, 3\}, \{2\}\}$, $\{\{2, 3\}, \{1\}\}$ and $\{\{1\}, \{2\}, \{3\}\}$. The cell $A = \{\{1, 3\}, \{2\}\}$, has size $|A| = 2$, and elements $a_1 = \{1, 3\}$ and $a_2 = \{2\}$. $\mathbf{i}_{a_u}$ denotes a sub-haplotype: given for instance $a_u = \{1, 3\}$, then $\mathbf{i}_{a_u} = i_1 i_3$. The disequilibrium of a single locus, $D^1_{i_j}$, is defined as the allelic frequency p_{i_j} at that locus j [39].

Definition (5) allows disequilibria of higher order to be recursively defined in terms of disequilibria of lower orders. Recursive programming enabled us to computerize the algebra for higher order linkage disequilibria [41, 42] (see *Supporting Information, section 1*, (*SI*.1)). We obtained algebraic expressions for MLD, which depend only on haplotype and allele frequencies, for up to seven loci.

An alternative approach to MLD, due to Slatkin [43], defines it as the covariance between multiple alleles at multiple sites. The conclusions presented here apply to both definitions of MLD, although our analyses focus on the Geiringer-Bennet approach (see *SI*.2).

Oscillations under the deterministic finite-sites approximation

We begin by describing the behavior of MLD under a simplified model of finite-sites interference, the *deterministic finite-sites MMI* (DFMMI) model, a deterministic analogue to MMI [16] in a finite-sites context. We retain MMI's core assumption that all mutations will confer the same selective advantage, s, but remove stochasticity. Intuitively, selection moves the distribution of fitnesses of haplotypes steadily forward in the manner of a travelling wave [16, 44–46]; rapid growth of rare, fitter-than-average haplotypes expands the front of the distribution, while gradual loss of less-fit haplotypes contracts the distribution's tail [8, 10, 16]. Due to our interest in rapid adaptation of pathogens, simulation parameters were chosen to correspond to estimates from early HIV infection (see *SI*.3).

Our DFMMI simulations begin with a wildtype ancestor having a limited number L of possible beneficial mutations, which accumulate at a fixed rate; a rise in frequency of haplotypes with k mutations (k-mutants) is followed by a rise in frequency of $k+1$-mutants every time period τ_{inter} (see *SI*.3 for derivation), a constant independent of k. Within each k-mutant wave, we assume that the relative haplotype frequencies are equal. This assumption eliminates haplotype frequency imbalances stemming from genetic drift, allowing us to examine the dynamics of MLD in the absence of such complications. Moreover, it ensures that all possible 2^L haplotypes exist at some point in the evolutionary trajectory of the simulation – a *full escape graph* [47].

In each simulation, we allow a population to evolve for roughly 300 generations, calculating the L^{th} order MLD relative to the ancestral haplotype at fixed time intervals (see Figure 1). Unless otherwise noted, we subsequently refer to MLD relative to the ancestral haplotype, i.e. $D^L_{0_1 \ldots 0_L}$, where $i_j = 0_j$ denotes no mutation in the j^{th} allele.

In these DFMMI dynamics, the highest $\left(L^{th}\right)$ order MLD is initially zero. As single-mutant haplotypes appear and spread, the ancestral haplotype is outcompeted (Figs. 1A and B) and becomes under-represented relative to the expectation from random allelic associations. The MLD decreases during this process (Fig. 1C). During the remainder of the dynamics, the dominant k-mutants are replaced by successive $k+1$ mutants (see Figs. 1B and 2A). The highest order MLD correspondingly oscillates from negative to positive. We found that the number of oscillations in the highest order MLD, n_O, increases with the number of loci L simulated (Figs. 3A and C), and

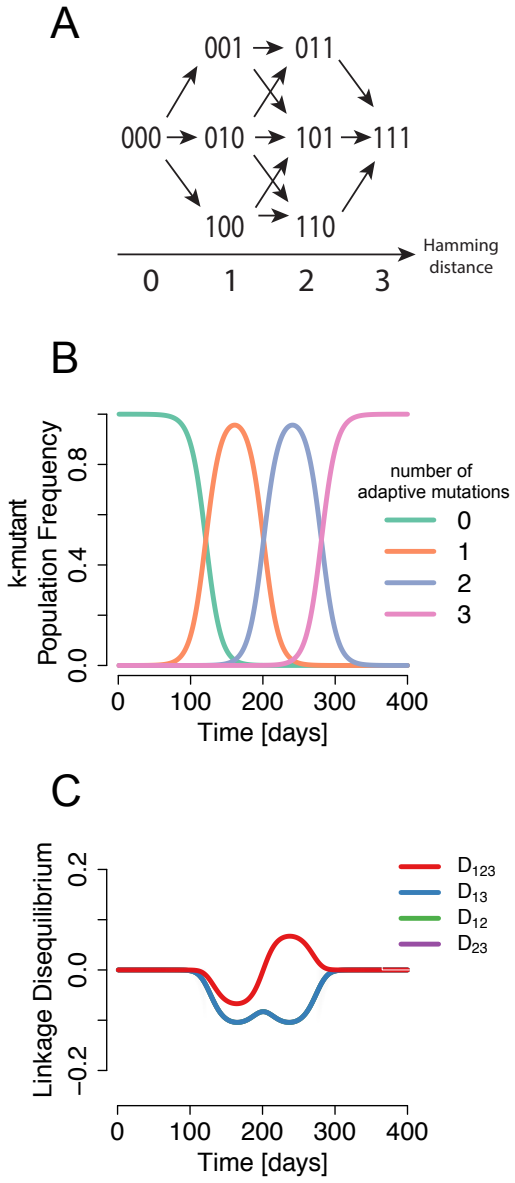

FIG. 1. **Origin of oscillations in multilocus linkage disequilibrium (MLD).** A) The space of all possible haplotypes, starting from the wildtype (no mutations: all zeros). B) As evolution pushes the fitness distribution to higher Hamming distances, it generates a signature of over-representation of haplotypes with equal Hamming distance. This is reflected by the sequential rise and fall of k-mutant waves. C) Pairwise and three-locus Geiringer-Bennett linkage disequilibria, measured with the wildtype 000 as reference, over the course of the simulation (all the pairwise disequilibria overlap). When taken as a reference haplotype, all haplotypes with the same number of adaptive mutations produce an MLD of equal sign.

follows the simple relationship:

$$n_O = \frac{L-1}{2}. \tag{6}$$

MLD oscillations reflect DFMMI dynamics

The observed oscillations in the highest order MLD can be explained by the temporary dominance of k-mutant haplotypes in the population. In fact, the oscillations in highest order MLD reflect the acquisition of beneficial mutations. In the following, we refer to the highest order MLD simply as MLD.

As shown in Figure 1B, at any point during the dynamics, the population will consist mainly of haplotypes containing k mutants; i.e. k-mutant haplotypes will be over-represented. Therefore, the MLD relative to all k-mutant haplotypes will be positive. As mutation and selection push the population to higher fitness levels, $k+1$-mutants spread. Then, the MLD relative to $k+1$-mutant haplotypes will increase until it becomes positive.

A useful property of MLD in bi-allelic systems allows us to relate the MLD relative to a k mutant haplotype to the MLD relative to the ancestral haplotype: $\forall j$: $j \in \{1, \ldots, L\}$;

$$D^L_{i_1 i_2 \ldots 0_j \ldots i_L} = -D^L_{i_1 i_2 \ldots 1_j \ldots i_L}. \tag{7}$$

Equation (7) (see *SI*.2 for proof) can be interpreted as follows: if a reference haplotype is over-represented relative to our expectation, each haplotype with the opposite allele to the reference at a given locus must be equally under-represented.

Therefore, at any point during the dynamics, MLD relative to haplotypes containing a single beneficial allele (that is, single mutants) will be of equal magnitude, but opposite sign to MLD relative to the ancestral haplotype. Further, MLD relative to double-mutant haplotypes will be of equal magnitude, but opposite sign to MLD relative to single-mutant haplotypes; this also implies that MLD relative to double mutant haplotypes is equal to MLD relative to the ancestral haplotype. We conclude that when single or odd-k mutant haplotypes are over-represented (i.e. positive MLD), the MLD relative to the ancestral haplotype will be negative. In the same way, when double or even-k mutant haplotypes are over-represented, the MLD relative to the ancestral haplotype will be positive (see also *SI*.4, and Fig. S1).

Therefore, as the 'traveling wave' accrues subsequent beneficial mutations and the set of haplotypes that are over-represented (i.e., those haplotypes with positive MLD) shifts, the sign of the MLD relative to the ancestral haplotype also shifts. This explains both the observed oscillation in MLD and the relationship between the number of possible beneficial alleles and the number of observed oscillations (Eq (6)); there are $L-1$ soft sweeps as additional beneficial mutations appear, and each sweep is reflected in a MLD half-oscillation.

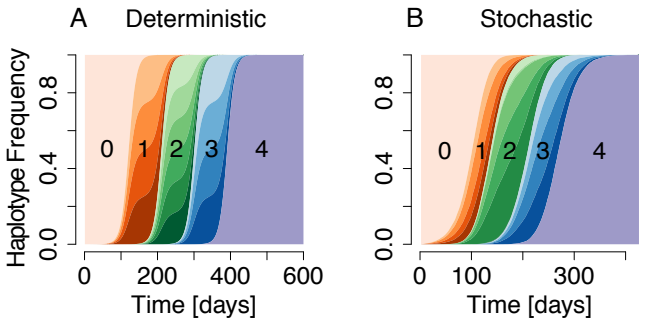

FIG. 2. **Haplotype dynamics of DFMMI and FSMMI simulations.** A) Haplotype frequencies over the course of a DFMMI simulation with $L = 4$ loci. Beneficial mutations arise every $\tau_{inter} = 100$ days (see *SI*.3) and begin to sweep at a rate $\epsilon = 0.095$ (see *SI*.3, eqn. (S11)). Colors indicate haplotypes with an equal number of mutations k. B) Haplotype frequencies over the course of a simulation of the FSMMI model with $L = 4$ selected loci, selection coefficients per mutation $s = 0.1$, population size $N = 10^5$ and beneficial mutation rate $\mu_b = 10^{-4}$ per locus per generation.

The DFMMI model generates oscillations in MLD in another, rapidly evolving regime (when τ_{inter} is very small, see *SI*.5)). However, this particular MLD pattern is expected to be rare, and can be neglected when applying appropriate checks in data.

Speed of evolution and MLD dynamics

As half-oscillations in MLD reflect partial sweeps of sequential layers of k-mutant haplotypes, we expect the frequencies of the MLD wavelets to correlate with the rate of evolution of the system. Let us assume that beneficial mutations accumulate at a stable rate, that is, that the population's fitness wave proceeds at a well-defined constant speed v through fitness space. Then, the time for the fitness wave to accumulate one beneficial mutation corresponds to the time it takes for half an oscillation of the highest order MLD, $T/2$, where T is the MLD oscillation period. Thus, the speed of evolution of the fitness wave v must be related to the oscillation frequency of the MLD as follows:

$$v = \frac{s}{(T/2)} = 2sf, \tag{8}$$

where f is the frequency of the oscillations.

Retention of MLD oscillation properties under drift

Next, we tested whether MLD oscillations appear and can be detected and analyzed in the presence of drift, using a Wright-Fisher model with selection. To this end, we adapted the MMI model [16]: like MMI, our WF-model only considers drift-prone, beneficial mutations, each with the same effect, but our model considers beneficial mutations at finitely many loci. This model, previously employed in other studies [48, 49], is termed *finite-sites MMI* (FSMMI) model (see *SI*.6, and Fig. S2 for an example simulation).

As in the DFMMI model, our FSMMI framework and parameters are chosen to capture some features of early HIV within-host evolution, when HIV undergoes very rapid adaptation to the host environment [48, 49]. Specifically, we focus on regimes in which the population size is around $N = 10^5$, the beneficial mutation rate per locus per generation is $\mu_b = 10^{-4}$ [33, 48–51], and each beneficial mutation carries the same selective advantage s between $0.01 - 0.3$ [22, 51, 52]. The simulations are run with a population size N and selection acts on all loci from the start.

Unlike the DFMMI model described above, in FSMMI simulations beneficial mutations establish stochastically, breaking the symmetry in k-mutant haplotype frequencies. Thus, a full escape graph is not guaranteed. Despite this added stochasticity, beneficial mutations are still typically accrued in a sequential fashion (see Fig. 2), with subsequent k-mutants rising and falling in frequency. This is a prerequisite for MLD oscillations.

In fact, both wavelet-based statistical tests and Fisher tests for hidden periodicities indicate that oscillations in the highest order MLD persist under FSMMI (see *SI*.7, Figs. S3 and S4). However, as expected, the oscillations tend to be less precise than in the DFMMI case (Fig. 3A vs 3D) and n_O full oscillations are not always realized. This dampening of signal is likely due to portions of the haplotype space remaining unexplored in stochastic simulations.

We further investigated whether the frequencies of these MLD signals may be estimated. To this end, we computed the wavelet power spectrum [53–55], of the simulated dynamics (Fig. 3B,3E) and used it to infer the frequency at which MLD oscillates (see *Materials and Methods*). As shown in Fig. 3C for a DFMMI benchmark, even in stochastic simulations (Fig. 3F), wavelet analysis can confidently reconstruct the frequency of MLD oscillations.

We proceeded to examine whether the MLD oscillations under FSMMI also retain other features displayed under DFMMI, such as equation (8). To this end, we compared the rate of evolution estimated using the MLD oscillation frequency, $\hat{v}$, to the rate of evolution expected under infinite-sites MMI theory, v_{MMI} [16] (see Fig. 4, *SI*.8). v_{MMI} serves as benchmark because there does not exist a clear ground truth for the speed of evolution under finite-sites.

Figure 4A shows that the rate of evolution inferred by MLD dynamics from our simulated FSMMI model, $\hat{v}$ from (8) with $N = 10^5$ is close to v_{MMI}, the predicted rate of evolution in infinite-sites MMI theory, [16].

When N varies from 10^3 to 10^6, the mismatch between

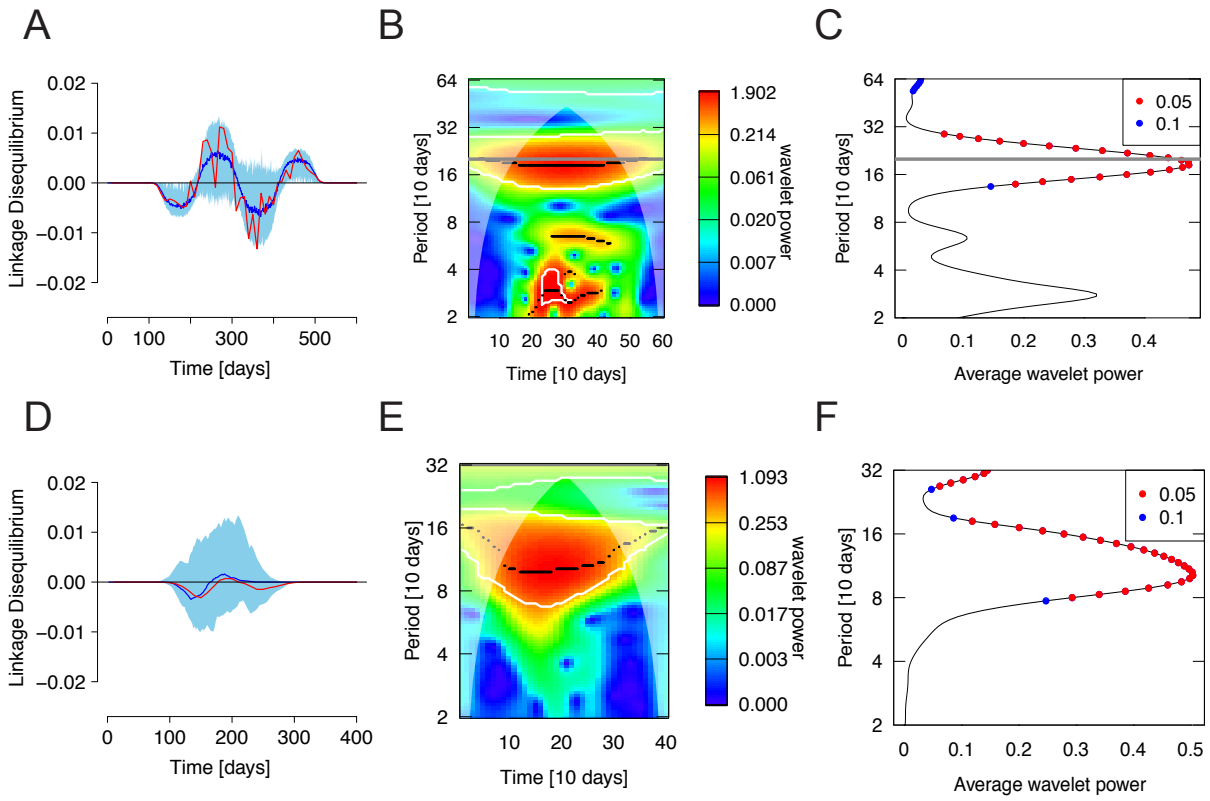

FIG. 3. **MMI-induced MLD oscillations are still detectable under drift.** A) Oscillation of the fifth order MLD in a symmetric full escape graph. The dark blue line is the median of a set of 200 runs, and the upper and lower bounds of the light blue area represent the 2.5 and 97.5 percentiles of all measured LDs. The MLD was calculated every 10 days using a sample size of 20 haplotypes. The red trajectory represents the measured LD from one particular repeat. B) The wavelet power spectrum in the time-period domain of the fifth order MLD values obtained with the sampling points of the red line in A) [53, 54]. The horizontal grey line is the true oscillation period of the red time series in A). The white contour lines indicate regions where the power spectrum values are significantly ($< 5\%$) non-random. The black lines indicate local power spectrum maxima. The half-transparent region demarcates a low-confidence wavelet power region. C) The time-averaged wavelet power spectrum. The red and blue dots indicate whether the null-hypothesis that the the time-averaged wavelet power may have been generated by white noise is rejected at below 0.05 or 0.1 significance levels, respectively. The maximum spectral density is attained close to the simulated period of $T = 200$ days (horizontal thick grey line) of the oscillations. D) The analogous situation to A) for 100 simulation runs of the FSMMI model with selection, run with parameters $L = 4$, $N = 10^5$, $\mu_b = 10^{-4}$ and $s = 0.1$. Samples are taken every 5 generations or 10 days. E) Wavelet power spectrum of one randomly chosen MLD trajectory (red line in D)). F) Analogous to C), but without the horizontal line indicating expected value.

$\hat{v}$ and v_{MMI} first decreases, and then begins to increase again (see *SI* Fig. S5, left column). As expected, when population size values fall to 10^3 or lower, the interference effects fade [16], and the $\hat{v}$-to-v_{MMI} differences increase markedly. For populations sizes around 10^6, $\hat{v}$ becomes smaller than v_{MMI}, but remains within confidence intervals for almost all studied cases.

We performed an analogous test using Crow-Kimura-Felsentein (CKF) theory, v_{CKF} [56] (see *SI* Fig. S6, left column). Apart from a systematic positive bias of $\hat{v}$ relative to v_{CKF}, the above patterns are largely retained.

Retention of MLD oscillation properties under drift and dissimilar fitness effects

Despite its usefulness for mathematical analysis, the assumption that all loci have identical fitness effects is unlikely to be perfectly satisfied in biological systems. To address this issue, we devised two further models, *narrow FSMMI* and *broad FSMMI* (see *Materials and Methods*). In the narrow FSMMI model, the selective coefficients s_j of a mutated locus j are drawn from a Gamma distribution with average $\bar{s}$. The broad FSMMI model adopts a distribution of fitness effects (DFE) which falls off in an over-exponential manner [57]. Again, the

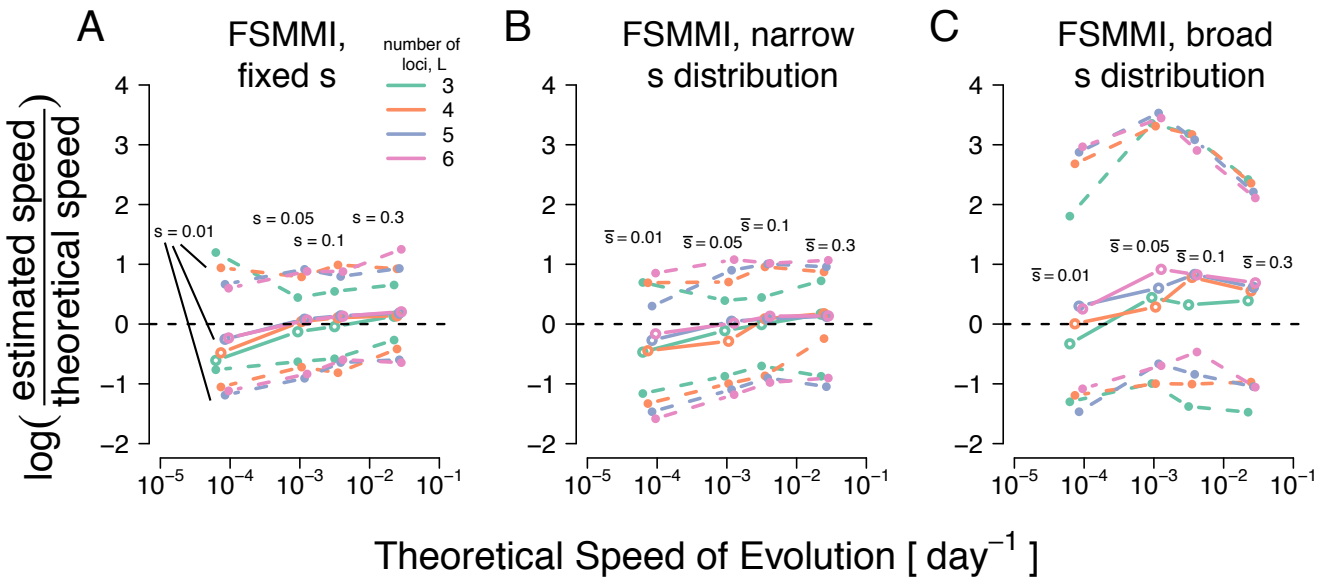

FIG. 4. **Estimates of speed of evolution based on MLD-oscillations versus MMI theory [16].** A) Estimates of the speed of evolution $\hat{v}$ obtained by wavelet analysis for FSMMI model simulations run for $L = 3, 4, 5$ and 6 loci, and selection coefficients $s \in \{0.01, 0.05, 0.1, 0.3\}$ with population size $N = 10^5$. v is estimated with equation (8) (s is known), where for f we use the MLD-based oscillation frequency estimate (Fig.3). The inter-sampling period was $\Delta t = 2$ days. Colored open circles and filled circles correspond to medians and nonparametric confidence intervals (95%), respectively, from those simulations among 100 runs that displayed significant (< 0.05 level for wavelet-based test) oscillations. B) Analogous figure for narrow FSMMI model. Here, to compute the speed of evolution v from the inferred oscillation frequency f we use the average $\bar{s}$ of the Gamma distribution from which the selection coefficients were sampled ($\bar{s}$ values equal to s values in FSMMI). C) Analgous to B), but for the broad FSMMI model.

average s_j is predefined.

We observed that narrow FSMMI simulations produced detectable oscillations nearly as often as FSMMI (see *SI*7, Figs. S7). Significant signals of MLD oscillation were also often detected under broad FSMMI (see *SI*7, Figs. S8). However, due to exacerbated symmetry breaking of the escape graph, these were substantially less frequent than in the simulations under narrow FSMMI. Further analysis is concentrated on simulations with non-random MLD time series behavior.

Figure 4B shows that the narrow FSMMI model leads to MLD-based estimates similar to the those from the FSMMI model. whereas broad-FSMMI estimates (4C) deviate more strongly from theory. Confidence intervals around estimates broaden as selective coefficients tend to become more dissimilar. Varying N affects narrow and broad FSMMI estimates similarly as it does FSMMI estimates (see *SI* Fig. S5). A similar picture emerges when using CKF-theory as benchmark (see *SI* Fig. S6).

To further corroborate these results, we also compared $\hat{v}$ values with estimates from alternative methods of inference of the speed of evolution (see *SI*9). The performance of MLD-based estimates relative to other, reasonable estimates of the speed of evolution ($\hat{v}_m$) is generally very similar to their performance relative to predictions from MMI theory (see *SI*, Fig. S9, and S10 for a systematically biased estimator).

Lastly, as the assumptions of infinite-site MMI theory are not met in our FSMMI simulations, we assessed v_{MMI}'s appropriateness as a benchmark. To do this, we compared v_{MMI} against v_m. The match between MMI theory and empirical estimates is generally good (see *SI*, Fig. S11), suggesting that the infinite-sites assumption in MMI theory does not adversely affect its predictive utility for finite-sites systems within the parameter ranges here studied. This justifies the use of MMI theory as a benchmark for wavelet-based estimates.

DISCUSSION

We have developed new computational tools to calculate multilocus linkage disequilibrium (MLD), a statistic that quantifies the nonrandomness of allelic associations across loci, accounting for contributions to haplotype structure stemming from subgroups of loci. We show that, in simulated deterministic haplotype dynamics with (i) rapid accrual of a finite number of strongly beneficial mutations with similar fitness effects and (ii) tight linkage between loci (i.e. a MMI regime), MLD dynamics display a wavelet-like temporal pattern. We find that these oscillations can be explained by successive sweeps

by haplotypes containing increasing numbers of beneficial mutations in combination with specific mathematical properties of MLD expressed in Eq. 7. We demonstrate that the frequency of these oscillations is proportional to the rate of evolution. Finally, we show these oscillations are robust to evolutionary stochasticity and some degree of variation in the fitness effects of mutations. However, these properties are gradually lost as the fitness effects become more dissimilar. Thus, MLD dynamics may contain information relevant to the study of the short-term evolution of microorganisms under very strong selection, including human pathogens such as HIV, in which a finite number of loci experience strong selection.

Moreover, the detection of MLD oscillations depends on accurate haplotype frequency estimates, not obtained in most within-host evolution studies, and in HIV in particular. However, the continuous improvement of sequencing technologies is likely to allow for deep and dense sampling in the future, producing appropriate datasets.

While the MLD behavior we describe exhibits important evolutionary phenomena, the wavelet-based approach we present for inferring the speed of evolution will likely be inefficient in natural populations. Rather than providing a new estimation method, this study aims to elucidate how evolutionary dynamics are simultaneously manifest in both haplotype frequency dynamics and multilocus linkage disequilibria. This is a necessary first step before the MLD perspective on evolutionary dynamics can offer broader applicability.

Further, our method currently ignores the role of epistasis. In escape mutations of HIV, the pathogen which inspired this work, we are unaware of evidence for epistatic interactions. However, other intragenic mutations are likely to give rise to epistasis [58–61]. If epistasis dominates over selection (sign epistasis), the evolutionary dynamics are likely to halt at a local or global fitness peak (i.e. not the full escape haplotype) [62]. Then, at mutation-selection balance, an MLD signal should be maintained that is constant and not oscillatory. Extensions of this work may thus help to differentiate epistasis-dominated from weak- or no-epistasis scenarios.

We also assume that loci under selection are readily detectable. This is true in the case of epitopes targeted by cytotoxic T lymphocytes in HIV-infected patients (see [25]), but may not be true elsewhere. We did not investigate the scenario where only L' are tracked, but where $L > L'$ are under selection. Unknown loci may exacerbate escape graphs' asymmetries, thereby further dampening any MLD signal. Further work is needed to fully ascertain the impact of these effects.

Another benefit of our approach to interference is that it draws from an underexplored perspective on evolution that considers the role of linkage disequilibria, and its important statistical inference machinery. In fact, very little use has been made of MLD in the context of population genetics, in particular the study of interference [32]. This may be due to different definitions of linkage disequilibrium at multiple loci [34, 35, 43, 63, 64]. The crucial advantage of the Geiringer-Bennet MLD is that its maximum likelihood estimate always exists [65], a very useful property for estimation.

The other central benefit is MLD's capacity to characterize a population as evolving under MMI. Most simply, the presence of MLD oscillations of the type described here suggests that the population under study is evolving under an MMI regime. MMI occurs in populations with specific characteristics; namely (i) a large supply of beneficial mutations [16] (ii) beneficial mutations that confer similar, strong selective advantages [16], and (iii) low enough recombination rates that beneficial mutations are likely to compete rather than recombine onto a single haplotype. Therefore, observed MLD oscillations provide valuable information with respect to these critical population genetic parameters.

MATERIALS AND METHODS

Oscillation estimation by means of signal processing techniques

To identify oscillations of MLD in the simulation data, we developed a detection scheme based on wavelet analysis. For each run, we calculated the highest order linkage disequilibrium at each of M_s sample points from the sampled data, that is, M_s MLD-values $\{x_n\}$, where $n \in \{0, \ldots, M_s - 1\}$. Sample data x_n are assumed to have been obtained at constant inter-sampling periods dt, and can be expressed as a vector $\mathbf{x}$ with entries x_n.

We analyzed the wavelet power spectrum of $\mathbf{x}$ (R-package *WaveletComp*, [66]). An oscillating LD measure of L loci will maximally generate $L - 1$ half-oscillations, starting with a negative half-oscillation. Even if damped, such wavelet-like oscillations should leave traces in the frequency spectrum that are close to the frequency of a full period, T.

The *wavelet transform* of the data $\mathbf{x}$ is given by:

$$W(\tau, a) \doteq \sum_{n=0}^{M_s-1} x_n \cdot \frac{1}{\sqrt{a}} \psi^* \left(\frac{n - \tau}{a} \right), \tag{9}$$

where

$$\psi(t) \doteq \pi^{-\frac{1}{4}} e^{i\omega_0 t} e^{-\frac{t^2}{2}}, \tag{10}$$

is the Morlet wavelet. τ is called the *translation time*, whereas a is termed *scaling factor*. The superscript $*$ denotes the complex conjugate. The nondimensional frequency ω_0 is set to 6, such that the scale a becomes almost identical to the Fourier period.

To compute all values of $W(\tau, a)$ and its derivates, we used the package *WaveletComp* [66]. The wavelet transform is computed over a standard set of values of τ and a. τ is varied from 0 to $M_s - 1$; that is, by multiples of the time increment dt. The scaling factors a determine the coverage in the period domain. They are set to vary

as $a_{\min} \cdot 2^{j \cdot dj}$, with $j = 0, \ldots, J$ and where $a_{\min} = 2$ is the minimum scaling factor used, $dj = 1/20$ is the number of steps per analyzed octave and the maximum period analyzed, $a_{\min} \cdot 2^{J \cdot dj}$, is $2^{\log_2(M_s)}$ (which determines J).

The *wavelet power spectrum* is defined by:

$$P(\tau, a) \doteq \frac{1}{a} \mid W(\tau, a) \mid^2 . \tag{11}$$

The value of $P(\tau, a)$ at a coordinate pair (τ, a) serves as a measure of confidence that the time series $\mathbf{x}$ is oscillating at a frequency corresponding to a at translation time τ.

To measure the frequency f of an MLD time series, we first identify the value pair $(\tau_{\max}, a_{w,\max})$ for which $P(\tau, a)$ was maximal. A p-value for the null-hypothesis that there is no periodicity in $\mathbf{x}$ is provided by *WaveletComp*, using a statistical test based on the work of Cazelles et al. [53–55].

To reduce computational burden, the time series $\{x_n\}$ was trimmed for analysis. Simulation runs were all performed over the simulation time of 2000 generations. However, since L-mutants frequently fix well before 2000 generations have elapsed, most late x_n values are zero. To speed up computations, we identified the last non-zero x_n value for each series, x_z, and replaced the series $\{x_n\}$ by $\{x_0, \ldots, x_z, 0, \ldots, 0\}$, concatenating 20 zero values to the end of the series. This modified series was then used for further analysis.

When stochastic simulations were run, MLD could be zero for the entire time course at low population sizes and very different selection coefficients. In these cases, the oscillation frequencies were set to correspond to zero. For Fisher's hidden periodicities test the p-values for the null that no periodicities exist in the all-zero signal were set to unity.

Distribution of fitness effects employed in finite-sites MMI models

In this study, three different distributions of fitness effects were used to run Wright-Fisher simulations with a finite number of loci to examine the robustness of (8). The first set of simulations were run with the simplest possible assumption: all loci confer the same selective advantages.

The second set of simulations were run with selection coefficients drawn from a Gamma distribution:

$$\rho_g(s) = \frac{1}{\Gamma(k)\theta^k} s^{k-1} e^{-\frac{s}{\theta}}, \tag{12}$$

where Γ is the Gamma function, k is the shape parameter and θ is the scale parameter. For our simulations, we used the values $k = 400$, and $\theta = \{7.5 \cdot 10^{-4}, 2.5 \cdot 10^{-4}, 1.25 \cdot 10^{-4}, 2.5 \cdot 10^{-5}\}$. The average s under that distribution is $\bar{s} = k\theta$. The variance is always set to 10% of $\bar{s}$.

The third set of simulations were run by drawing the selection coefficients from an exponential-like distribution of fitness effects with parameters that favour MMI conditions [49, 57]. Specifically, the distribution of fitness effects used was:

$$\rho_e(s) = \frac{1}{\sigma} \frac{e^{-\left(\frac{s}{\sigma}\right)^{\beta}}}{\Gamma\left(1 + \beta^{-1}\right)}, \tag{13}$$

where β is a steepness parameter that indicates whether the distribution follows an over or under-exponential decline as s increases, and σ roughly corresponds to the inverse of the rate parameter in an exponential distribution. The average selection coefficient sampled, $\bar{s}$, is given by $\sigma\Gamma(2/\beta)/\Gamma(1/\beta)$. When β is one, $\rho_e(s)$ is exponentially distributed. In this study, we used parameters for average values of s that are $\beta = 1.4$ and $\sigma = \{0.432, 0.144, 0.072, 0.0144\}$.

ACKNOWLEDGMENTS

The authors thank Fabio Zanini, Roland Regoes, Frederic Bertels, Massimo Maiolo and the Feldman lab members for stimulating discussions. This work was supported by by the Swiss National Science Foundation (grant number P2EZP3_162257 to VG). AH was supported in part by The Stanford Center for Computational, Evolutionary and Human Genomics (CEHG) doctoral fellowship at Stanford. MWF was supported in part by CEHG, and by the Morrison Institute for Population and Resource Studies.

[1] Miralles R, Gerrish PJ, Moya A, Elena SF. Clonal interference and the evolution of RNA viruses. Science. 1999;285(5434):1745.

[2] Neher RA. Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation. Annual Review of Ecology, Evolution, and Systematics. 2013;44(1):195–215. Available from: `http://dx.doi.org/10.1146/annurev-ecolsys-110512-135920`.

[3] Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, et al. The molecular diversity of adaptive convergence. Science. 2012 Jan;335(6067):457–461.

[4] Desai MM, Fisher DS, Murray AW. The speed of evolution and maintenance of variation in asexual populations. Curr Biol. 2007;17(5):385–394.

[5] Lang GI, Botstein D, Desai MM. Genetic variation and the fate of beneficial mutations in asexual populations.

Genetics. 2011;188(3):647–661.

[6] Kao KC, Sherlock G. Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. Nat Gen. 2008;40(12):1499–1504.

[7] Wichman HA, Badgett MR, Scott LA, Boulianne CM, Bull JJ. Different trajectories of parallel evolution during viral adaptation. Science. 1999 Jul;285(5426):422–4.

[8] Rouzine IM, Weinberger LS. The quantitative theory of within-host viral evolution. Journal of Statistical Mechanics: Theory and Experiment. 2013;2013(01):P01009.

[9] de Visser JAG, Rozen DE. Clonal interference and the periodic selection of new beneficial mutations in Escherichia coli. Genetics. 2006;172(4):2093–2100.

[10] Fisher R. The genetical theory of natural selection. Oxford: Clarendon; 1930.

[11] Muller HJ. Some genetic aspects of sex. Am Nat. 1932;66(703):118–138.

[12] Hill W, Robertson A, et al. The effect of linkage on limits to artificial selection. Genet Res. 1966;8(3):269–294.

[13] Otto SP, Barton NH. The evolution of recombination: removing the limits to natural selection. Genetics. 1997;147(2):879–906.

[14] Maynard Smith J. What use is sex? J Theor Biol. 1971;30(2):319–335.

[15] Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737–756.

[16] Desai MM, Fisher DS. Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics. 2007 Jul;176(3):1759–1798.

[17] Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci USA. 2012;109(13):4950–4955.

[18] Hegreness M, Shoresh N, Hartl D, Kishony R. An equivalence principle for the incorporation of favorable mutations in asexual populations. Science. 2006;311(5767):1615–1617.

[19] Kosheleva K, Desai MM. The dynamics of genetic draft in rapidly adapting populations. Genetics. 2013 Nov;195(3):1007–25.

[20] Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J Exp Med. 2009;206(6):1273–1289.

[21] Cobey S, Koelle K. Capturing escape in infectious disease dynamics. Trends Ecol Evol. 2008 Oct;23(10):572–7.

[22] Asquith B, Edwards CT, Lipsitch M, McLean AR. Inefficient cytotoxic T lymphocyte–mediated killing of HIV-1–infected cells in vivo. PLoS Biol. 2006;4(4):e90.

[23] Turnbull EL, Wong ML, Wang S, Wei X, Jones NA, Conrod KE, et al. Kinetics of expansion of epitope-specific T cell responses during primary HIV-1 infection. The Journal of Immunology. 2009;182(11):7131–7145.

[24] Goonetilleke N, Liu MKP, Salazar-Gonzalez JF, Ferrari G, Giorgi E, Ganusov VV, et al. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. J Exp Med. 2009;206(6):1253–1272.

[25] Henn Mea. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathogens. 2012;8(3):e1002529.

[26] Felsenstein J. The effect of linkage on directional selection. Genetics. 1965 Aug;52(2):349–63.

[27] Hill W, Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 1968;38(6):226–231.

[28] Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998;102:127–144.

[29] Barton NH. Linkage and the limits to natural selection. Genetics. 1995;140(2):821–841.

[30] Kim Y, Stephan W. Selective sweeps in the presence of interference among partially linked loci. Genetics. 2003 May;164(1):389–98.

[31] Barton N. Genetic linkage and natural selection. Philos Trans R Soc Lond B Biol Sci. 2010;365(1552):2559–2569.

[32] Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008;9(6):477–485.

[33] Garcia V, Regoes RR. The effect of interference on the CD8+ T cell escape rates in HIV. Frontiers in Immunology. 2015;5(661). Available from: `http://www.frontiersin.org/hiv_and_aids/10.3389/fimmu.2014.00661/abstract`.

[34] Geiringer H. On the probability theory of linkage in Mendelian heredity. The Annals of Mathematical Statistics. 1944;15(1):25–57.

[35] Bennett J. On the theory of random mating. Annals of Eugenics. 1952;17(1):311–317.

[36] Hill WG. Disequilibrium among several linked neutral genes in finite population I. Mean changes in disequilibrium. Theor Pop Biol. 1974;5(3):366–392.

[37] Hill WG. Disequilibrium among several linked neutral genes in finite population: II. Variances and covariances of disequilibria. Theor Pop Biol. 1974;6(2):184–198.

[38] Dausset J, Legrand L, Lepage V, Contu L, Marcelli-Barge A, Wildloecher I, et al. A haplotype study of HLA complex with special reference to the HLA-DR series and to Bf. C2 and glyoxalase I polymorphisms. Tissue Antigens. 1978 Oct;12(4):297–307.

[39] Gorelick R, Laubichler MD. Decomposing multilocus linkage disequilibrium. Genetics. 2004;166(3):1581–1583.

[40] Andrews GE. The Theory of Partitions, volume 2 of Encyclopedia of Mathematics and its Applications. Addison-Wesley Boston,(MA); 1976.

[41] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2012. ISBN 3-900051-07-0. Available from: `http://www.R-project.org/`.

[42] Hankin RKS. Additive integer partitions in R. Journal of Statistical Software, Code Snippets. 2006 May;16.

[43] Slatkin M. On treating the chromosome as the unit of selection. Genetics. 1972;72(1):157–168.

[44] Tsimring LS, Levine H, Kessler DA. RNA virus evolution via a fitness-space model. Phys Rev Lett. 1996;76(23):4440–4443.

[45] Rouzine I, Coffin J. Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc Natl Acad Sci USA. 1999;96(19):10758–10763.

[46] Rouzine IM, Wakeley J, Coffin JM. The solitary wave of asexual evolution. Proc Natl Acad Sci USA. 2003;100(2):587–592.

[47] Leviyang S. The Coalescence of Intrahost HIV Lineages Under Symmetric CTL Attack. Bull Math Biol. 2012;74(8):1818–1856.

[48] Garcia V, Feldman MW, Regoes RR. Investigating the Consequences of Interference between Multiple CD8+ T Cell Escape Mutations in Early HIV Infection. PLoS Comput Biol. 2016 Feb;12(2):e1004721.

[49] Garcia V, Feldman MW. Within-Epitope Interactions Can Bias CTL Escape Estimation in Early HIV Infection. Front Immunol. 2017;8:423.

[50] Ganusov VV, Neher RA, Perelson AS. Mathematical modeling of escape of HIV from cytotoxic T lymphocyte responses. Journal of Statistical Mechanics: Theory and Experiment. 2013;2013(01):P01010.

[51] Kessinger TA, Perelson AS, Neher RA. Inferring HIV escape rates from multi-locus genotype data. Frontiers in Immunology. 2013;1:0.

[52] Asquith B, McLean AR. In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. Proc Natl Acad Sci USA. 2007 Apr;104(15):6365–6370.

[53] Cazelles B, Chavez M, Magny GCd, Guégan JF, Hales S. Time-dependent spectral analysis of epidemiological time-series with wavelets. J R Soc Interface. 2007 Aug;4(15):625–36.

[54] Cazelles B, Chavez M, Berteaux D, Ménard F, Vik JO, Jenouvrier S, et al. Wavelet analysis of ecological time series. Oecologia. 2008 May;156(2):287–304.

[55] Cazelles B, Cazelles K, Chavez M. Wavelet analysis in ecology and epidemiology: impact of statistical tests. J R Soc Interface. 2014 Feb;11(91):20130585.

[56] Park SC, Simon D, Krug J. The speed of evolution in large asexual populations. J Stat Phys. 2010;138(1-3):381–410.

[57] Fogle CA, Nagle JL, Desai MM. Clonal interference, multiple mutations and adaptation in large asexual populations. Genetics. 2008;180(4):2163–2173.

[58] Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM, et al. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nat Genet. 2011 May;43(5):487–9.

[59] Otwinowski J, Plotkin JB. Inferring fitness landscapes by regression produces biased estimates of epistasis. Proc Natl Acad Sci USA. 2014 Jun;111(22):E2301–9.

[60] Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ. Evidence for positive epistasis in HIV-1. Science. 2004 Nov;306(5701):1547–50.

[61] Wang K, Mittler JE, Samudrala R. Comment on "Evidence for positive epistasis in HIV-1". Science. 2006 May;312(5775):848; author reply 848.

[62] de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 2014 Jul;15(7):480–90.

[63] Mueller JC. Linkage disequilibrium for different scales and applications. Brief Bioinform. 2004;5(4):355–364.

[64] Nothnagel M, Fürst R, Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered. 2003;54(4):186–198.

[65] Weir BS, Ott J. Genetic data analysis II. Trends Genet. 1997;13(9):379.

[66] Roesch A, Schmidbauer H. WaveletComp: Computational Wavelet Analysis; 2014. R package version 1.0. Available from: `http://CRAN.R-project.org/package=WaveletComp`.