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*

compiled: December 1, 2016

The adaptation of asexually reproducing organisms is determined by how they accrue beneficial mutations. In large populations, multiple beneficial mutations may arise simultaneously on different genetic backgrounds and interfere with the fixation trajectories of other competing mutations. Multiple mutations interference (MMI) theory has proven useful for investigating these interference patterns. In MMI, beneficial mutations of equal fitness effect arise on a genome with infinitely many loci. However, assuming infinite sites makes it difficult to precisely predict the fates of individual mutations, complicating the detection of MMI in sequence data. In addition, most short-term within-host adaptation of pathogens such as Human Immunodeficiency Virus (HIV) occurs at a limited number of loci under strong selection. For these reasons, we investigate how MMI shapes the genetic composition of a population with few sites under selection. Specifically, we explore the dynamics of multilocus linkage disequilibrium (MLD), a measure of multi-way associations between alleles, in a finite-sites MMI model inspired by early HIV infection. In this regime, MLD oscillates over time in a wavelet-like fashion, a consequence of the sequential acquisition of beneficial mutations. We further show that the frequency of these oscillations is proportional to the rate of adaptation. Together, these findings suggest that MLD oscillations could be used as a signature of interference among multiple equally advantageous mutations.

Introduction

Many microorganisms, viruses, and cancer types replicate asexually with large population sizes and under strong selection [1–7]. This gives rise to pronounced *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. The factors that govern the speed of this adaptation are useful targets for intervention: for example, combination drug therapy was proposed as a treatment for human immunodeficiency virus (HIV) infection to diminish the accrual rate of drug resistance mutations [17, 18].

A fundamental contribution to understanding interference is *multiple mutations interference* (MMI) [16]. MMI theory assumes that mutations of equal fitness effect s may emerge on an infinite number of sites [16]. This theory offered a new perspective on interference, complementing classical *'one-by-one' clonal interference* (CI) [4, 19], where recurring large-effect mutations temporarily diminish genetic diversity by removing mutations of lower fitness. Unlike CI, under MMI mutations on less fit backgrounds are not automatically doomed to extinction. Instead, such haplotypes can survive by acquiring additional beneficial mutations. This leads to competitive interactions beyond individual mutations, involving competition between "coalitions", or "cohorts", of mutations [4, 5, 20].

The key advantage of MMI is that it can appropriately describe more complex forms of interference, and may thus serve as a null-model for interference in general [4, 16]. In more realistic forms of interference (*complete interference* [2]) a mutation's effect size is drawn from a distribution of fitness effects (DFE). Depending on the DFE, such a system may show combined hallmarks of CI and MMI [2]. Theoretical work shows that MMI can also capture important features of complete interference if mutations' fitnesses, s, are rescaled [16, 21–23]. In fact, MMI could appropriately predict how changes in population parameter values affected the speed of adaptation of experimentally evolving yeast [4]. Crucially, MMI could explain the observations, whereas CI theory on its own could not [4].

However, it is hard to identify MMI in sequence data. Because the infinite-sites assumption focuses on occurrence rather than localization of mutations [7], it is difficult to specify tell-tale genetic signatures of MMI or CI [5]. While extensive sequencing can reveal individual mutation frequency trajectories usually associated with CI, MMI or sweeps [20, 24], these trajectories are of limited use in characterizing the system as a whole, since they are not unique to their interference sub-type [5]. This problem is exacerbated in the within-host evolution of pathogens, such as human immunodeficiency virus (HIV), where sequence data are sparse and alternative fitness measurements are infeasible. In fact, the *in vitro* replication of the highly complex *in vivo* environmental conditions shaped by the immune system is imperfect or yields different fitness-related estimates of mutations [25].

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

The infinite-sites assumption is also hard to reconcile with most short-term adaptation of pathogens to new environments occurring in a limited number of known sites [7]. Examples are drug resistance mutations or escape mutations in viruses [26], such as HIV [8, 27]. Thus, interference models that only consider a limited number of loci might be more useful to detect MMI in these cases than more coarse models with an infinite number of sites under selection.

In this study, therefore, we investigate MMI in a finite-sites context. We extend previous theory to the case where interference is rare, and the number of interfering mutations is small [19, 28–30]. In particular, we explore the potential of multi-way associations between loci or *multilocus linkage disequilibrium* (MLD) to serve as the basis for a genetic signature of finite-sites MMI.

The motivation to use MLD to characterize MMI is rooted in the tradition of the population genetics of interference: For a system of two loci, a signature of interference is negative pairwise linkage disequilibrium (LD) [12, 31–34] —the departure from random associations of alleles at two loci [34]. Negative LD is expected when two beneficial mutations transiently coexist on different backgrounds, until they are combined into the same haplotype. For this reason, strong LD accompanied by a strong phylogenetic signal has been proposed as a defining hallmark of clonal reproduction in pathogens [35].

However, pairwise LD is insufficient to identify and characterize interference in populations of microorganisms. Since multiple beneficial mutations may segregate simultaneously in microbial populations [2] and disequilibria between different pairs of alleles may not be independent of one another [31, 36], pairwise linkage disequilibria will be difficult to interpret.

MLD, a generalization of pairwise LD [37–40], has the advantage that it accounts for deviations from random association at more than two loci. MLD may thus more appropriately reflect and characterize finite-sites interference. To this end, we develop a recursive programming method to compute MLD in systems of up to seven loci. We compute MLD at multiple time points in simulated systems evolving under finite-sites MMI, in a model previously designed to capture relevant features of early HIV infection [41].

We show that the evolution of MLD over time is interpretable and largely robust to the evolutionary stochasticity arising in our simulation model. We also show that, under MMI with strong selection, MLD oscillates a finite number of times. This wavelet-like behavior ultimately results from, and is a 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 haplotypes as a result of recombination, MLD between neutral genes should decrease exponentially over time [37, 38].

Consider L loci with alleles $i_1,\ i_2,\ i_3, \ldots, i_L$ and allele frequencies $p_{i_1},\ p_{i_2},\ p_{i_3}$. 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 [38], functions of allele and haplotype (i.e. gamete) frequencies, which satisfy the condition of exponential decrease over time for various numbers of loci are

$$\begin{aligned}
D_{i_1 i_2} =& p_{i_1 i_2} - p_{i_1} p_{i_2} && (1)\\
D_{i_1 i_2 i_3} =& p_{i_1 i_2 i_3} - p_{i_1} D_{i_2 i_3} && (2)\\
& - p_{i_2} D_{i_1 i_3} - p_{i_3} D_{i_1 i_2} - p_{i_1} p_{i_2} p_{i_3} \\
D_{i_1 i_2 i_3 i_4} =& p_{i_1 i_2 i_3 i_4} - p_{i_1} D_{i_2 i_3 i_4} - p_{i_2} D_{i_1 i_3 i_4} && (3)\\
& - p_{i_3} D_{i_1 i_2 i_4} - p_{i_4} D_{i_1 i_2 i_3} \\
& - D_{i_1 i_2} D_{i_3 i_4} - D_{i_1 i_3} D_{i_2 i_4} - D_{i_1 i_4} D_{i_2 i_3} \\
& - p_{i_1} p_{i_2} D_{i_3 i_4} - p_{i_1} p_{i_3} D_{i_2 i_4} - p_{i_1} p_{i_4} D_{i_2 i_3} \\
& - p_{i_2} p_{i_3} D_{i_1 i_4} - p_{i_2} p_{i_4} D_{i_1 i_3} - p_{i_3} p_{i_4} D_{i_1 i_2} \\
& - p_{i_1} p_{i_2} p_{i_3} p_{i_4} \\
D_{i_1 i_2 i_3 i_4 i_5} =& p_{i_1 i_2 i_3 i_4 i_5} - \ldots && (4)
\end{aligned}$$

In equations (1–4), the terms $(p_{i_1 \ldots i_L} - p_{i_1} \ldots p_{i_L})$ are called *Dausset's disequilibrium* [42]. 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 [37, 38]. 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 [43, 44]. We add a superscript L to indicate the LD of L^{th} order, given L loci, and write:

$$D^L_{i_1 \ldots i_L} = p_{i_1 \ldots i_L} - \sum_{c \in E} \left[\prod_{u=1}^{|c|} D^{c_u}_{I_u} \right], \qquad (5)$$

where E is the set of all compositions c of all partitions of L except $c = (L)$. A *partition* π of L is an unordered set of positive integers l_m, whose sum is L, that is $\sum_m l_m = L$ [44]. For instance, the set of all partitions of $L = 4$ is $\{\{4\}, \{3,1\}, \{2,2\}, \{2,1,1\}, \{1,1,1,1\}\}$. Each partition π can be represented by a number of ordered tuples, or *compositions*, c. For example, the compositions –or ordered triples– of the partition $\pi = \{2,1,1\}$ are given by $(2,1,1)$, $(1,2,1)$ and $(1,1,2)$ [43, 44]. The u^{th} entry of composition c is c_u, $|c|$ is the number of entries in c, and I_u is the sub-haplotype $i_{\left(1+\sum_k^{u-1} c_k\right)} \ldots i_{\left(\sum_k^{u} c_k\right)}$. Thus, MLD involves all possible sub-haplotypes of the

L^{th}-order haplotype $i_1 \dots i_L$. The disequilibrium of a single locus, $D^1_{i_j}$, is defined as the allelic frequency p_{i_j} at that locus j [43].

The partition-based definition (5) of MLD 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 [45, 46] (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 [47], defines it as the covariance between multiple alleles at multiple sites. The two approaches are not equivalent for systems involving more than 3 loci; however, both capture the deviation from random associations of alleles. The conclusions presented here apply to both definitions of MLD, although our analyses focus on the Geiringer-Bennet approach (see *SI*.2).

To study the dynamics of MLD in parallel to haplotype dynamics under MMI, we applied the recursive MLD algorithm described above to evolving haplotype data simulated under two models; a simplified ‘traveling wave’ model and the stochastic Wright-Fisher (WF) model. Due to our interest in rapid adaptation of human pathogens, simulation parameters were chosen to correspond to estimates from early HIV infection [41] (see *SI*.3, and Figure S1 for examples), because it provides an ideal model for a system where the within-host adaptation has been studied in terms of its genetic diversification.

Oscillations under the ‘traveling wave’ deterministic approximation

The first simulation framework employed treats an evolving population as a traveling fitness wave [16, 48–50]. This model reflects that, over time, selection moves the distribution of fitnesses of haplotypes steadily forward; 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].

Our ‘traveling wave’ 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*.4 for definition), 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 effects such as stochastically distributed establishment times of beneficial mutations and genetic drift, allowing us to examine the dynamics of MLD in the absence of such complications. Moreover, it ensures that all of the possible 2^L haplotypes exist at some point in the evolutionary trajectory of the simulation. These dynamics are therefore said to produce a *full escape graph*

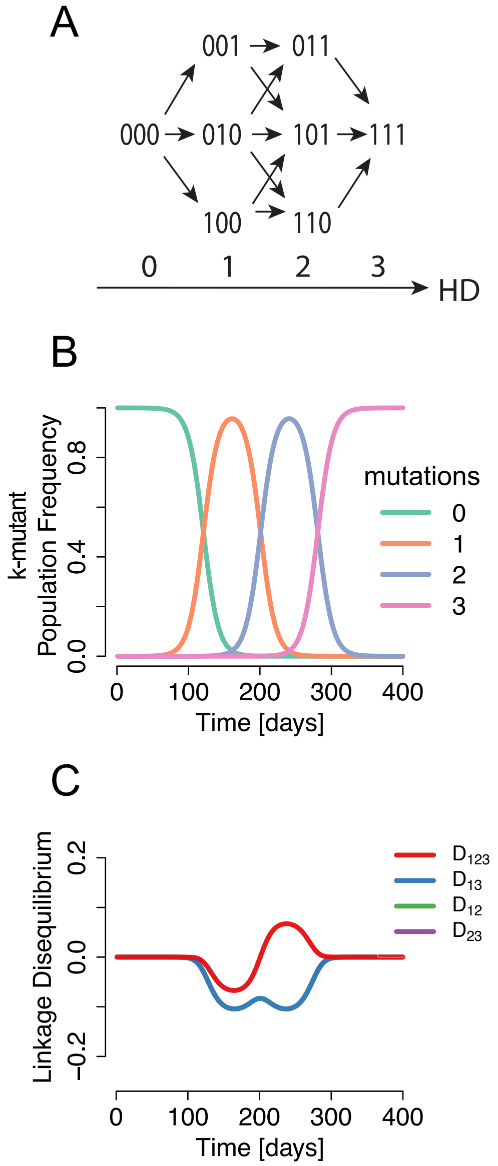

Fig. 1. **Origin of oscillations in multilocus linkage disequilibrium (MLD).** A) The space of possible haplotypes comprises disjunct layers of equal Hamming distance to the wildtype with no mutations (all zeros). As evolution pushes the fitness distribution to higher Hamming distances, it generates a signature of over-representation of the haplotypes in the corresponding layer of equal Hamming distance. This excess is captured by the MLD: When taken as a reference haplotype, each haplotype within the same layer produces an MLD of equal sign. B) Sequential rise and fall of k-mutant waves, comprising all haplotypes with k mutations. C) Pairwise and three-locus Geiringer-Bennett linkage disequilibria, measured with the wildtype 000 as reference, over the course of the simulation (the pairwise disequilibria overlap).

[51].

In each simulation, we allow a 'traveling wave' 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 deterministic, full escape dynamics, the highest $\left(L^{th}\right)$ order MLD is originally zero. As single-mutant haplotypes appear and grow, 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. Empirically, it appears that in general, the number of oscillations in the highest order MLD, n_O, increases with the number of loci L simulated (Figs. 3A and C), and follows the simple relationship:

$$n_O = \frac{L-1}{2}. \qquad (6)$$

These oscillations in MLD reflect the acquisition of subsequent layers of beneficial mutations and eventual convergence to the fittest haplotype.

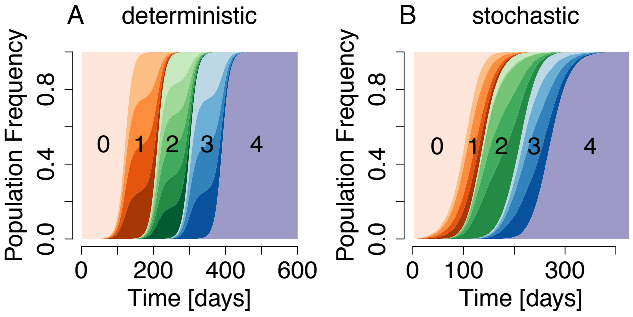

Fig. 2. **Haplotype dynamics of symmetric 'traveling wave' and Wright-Fisher (WF) simulations sequentially accrue beneficial mutations** A) Haplotype frequencies over the course of a symmetric 'traveling wave' simulation with selection, assuming $L = 4$ loci. Beneficial mutations arise every $\tau_{inter} = 100$ (see *SI*.4) days and begin to sweep at a rate $\epsilon = 0.095$ (see *SI*.4, eqn. (S9)). Colors indicate haplotypes with an equal number of mutations $k = 0, 1, 2, 3, 4$. B) Haplotype frequencies over the course of a simulation of the WF model with selection, assuming a system of $L = 4$ loci, selection coefficient per mutation $s = 0.1$, population size $N = 10^5$ and beneficial mutation rate $\mu_b = 10^{-4}$ per locus per generation. Colors of k-mutants are as in A).

MLD oscillations reflect MMI dynamics

The oscillations in the highest order MLD can be explained by the temporal dominance of k-mutant haplotypes in the population.

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:

$\forall j$: $j \in \{1, \ldots, L\}$;

$$\sum_{i_j \in \{0,1\}} D^L_{i_1 i_2 \ldots i_j \ldots i_L} = 0. \qquad (7)$$

This equation also holds for Slatkin's linkage disequilibria [47] (see *SI*.2), and multiway-associations in loglinear analysis [52, 53] – the main approaches used to identify inter-allelic associations in the literature, and the following arguments apply for both of these. Equation (7) 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*.5).

Given the relationships described above between MLDs relative to different haplotypes, 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 the observed oscillation in MLD. It also explains the relationship between the number of possible beneficial alleles among the L loci and the number of observed oscillations in MLD; there exist $L-1$ soft sweeps as additional beneficial mutations appear, and each sweep constitutes a half-oscillation in MLD.

To test whether there exist situations in which similar MLD patterns are generated, we varied τ_{inter}. With decreasing τ_{inter}, the MLD oscillations decrease in amplitude, until eventually, MLD is almost zero over the whole time course. However, if τ_{inter} is decreased even further, the system enters another regime. When τ_{inter} is very small, a single L-mutant haplotype will sweep through a population that consists largely of wildtype haplotypes. If, additionally, L is odd, such an L-mutant haplotype will generate an MLD signal that can be reminiscent of the MMI-based one. However, if L is even, the MLD pattern generated by an L-mutant differs from the MMI-based one: the initial deviation of the MLD is always positive, and the order of the sign changes is reversed. These L-mutant sweep patterns show strongly asymmetrical half-oscillations. Also the mechanism for these patterns is unrelated to the advance of the fitness wave, but stems from the combined effect of strongly positive pairwise disequilibria (see *SI*.6). This particular MLD pattern is expected to be rare, and can be neglected when applying appropriate checks in data (see *Discussion*).

Robustness of MLD oscillations to increased stochasticity

Having characterized the behavior of highest-order MLD under deterministic dynamics, we tested whether such oscillations can be detected in the presence of drift, using the WF model with selection.

As in the 'traveling wave' model, our WF framework and parameters (see *Materials and Methods*) are chosen to capture some features of early HIV within-host evolution, when HIV is hypothesized to undergo very rapid adaptation to the host environment [41]. Specifically, we focus on regimes in which the population size is $N = 10^5$, the beneficial mutation rate per locus per generation is $\mu_b = 10^{-4}$ [41, 54–56], and each beneficial mutation carries the same selective advantage s between $0.01 - 0.3$ [27, 55, 57]. The simulations begin at a population size N and selection acts on all loci from the start.

Unlike the 'traveling wave' model described above, in the WF simulations the beneficial mutations establish stochastically, breaking the symmetry in k-mutant haplotype frequencies (compare Fig.2 and Figs. 3A and 3C). Further, in the simulations, a full escape graph [51] is not guaranteed. Despite the stochasticity, beneficial mutations are still typically accrued in a sequential fashion, with subsequent k-mutants rising and falling in frequency. This can be observed by comparing an example set of simulated haplotype dynamics under the 'traveling wave' model to one simulated under this WF model, as shown in Figure 2A and 2B.

Thus, despite the increased complexity of the WF simulations, oscillations in the highest order MLD persist (Fig. 3C). However, as expected, the oscillations tend to be less precise. Specifically, as the dynamics progress and some portions of genotypic space remain unexplored, the signal is dampened. Hence, the preservation of the oscillatory MLD pattern may be interpreted as a measure of the symmetry of escape graphs.

Speed of evolution and MLD dynamics

To measure oscillation frequencies in MLD, we computed the periodograms of the simulated dynamics, which mark the proportion of the input signal explained by each frequency (see *Materials and Methods*, Fig. 3B for deterministic, and 3D WF-dynamics). These periodograms were smoothed to filter out noise introduced by the stochasticity of the WF process as well as by sampling a limited number of haplotypes at different time intervals, which we implemented to mimic empirical studies. Periodograms were subsequently used to predict the frequency at which MLD oscillates throughout the simulated dynamics. In the following, we restrict our analysis to periodograms from MLD time series of WF simulations only.

The signal processing first allows us to assess the robustness of the oscillatory signal in MLD to evolutionary stochasticity in WF simulations. Second, it suggests that MLD patterns are connected to important evolutionary parameters. Here, we demonstrate both using the rate of evolution as an example.

The MLD oscillatory pattern is related to the speed of evolution as follows: 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. Specifically, if beneficial mutations accumulate at a stable rate, the population's fitness wave proceeds at a well-defined constant speed v through fitness space. 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, $T/2$. Thus, the speed of evolution of the fitness wave must be related to the oscillation frequency of the MLD as follows:

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

where f is the frequency of the oscillations.

To further test the robustness of the MLD oscillations to evolutionary stochasticity, we used equation (8) to compare the MLD oscillation frequency in each WF-simulation estimated by a periodogram to the rate of evolution predicted by population genetic theory (see *SI*.7). This comparison tells us how well observed MLD oscillations capture the underlying haplotype dynamics, and retain information about it.

Figure 4 shows that the rate of evolution inferred by MLD dynamics from our simulated WF model and v_e in (8) is very close to the predicted rate of evolution in the Crow-Kimura-Felsenstein (CKF) theory [58]. Some mismatch between theory and inferred values is expected, since the predictions of the theory are derived under assumptions that are not satisfied in our WF model. Most importantly, the supply of beneficial mutations in our WF model is limited by the finite sites. Second, the population in our simulations is probably undergoing a short phase of acceleration in its evolutionary rates, biasing speeds of evolution [16]. Furthermore, because the

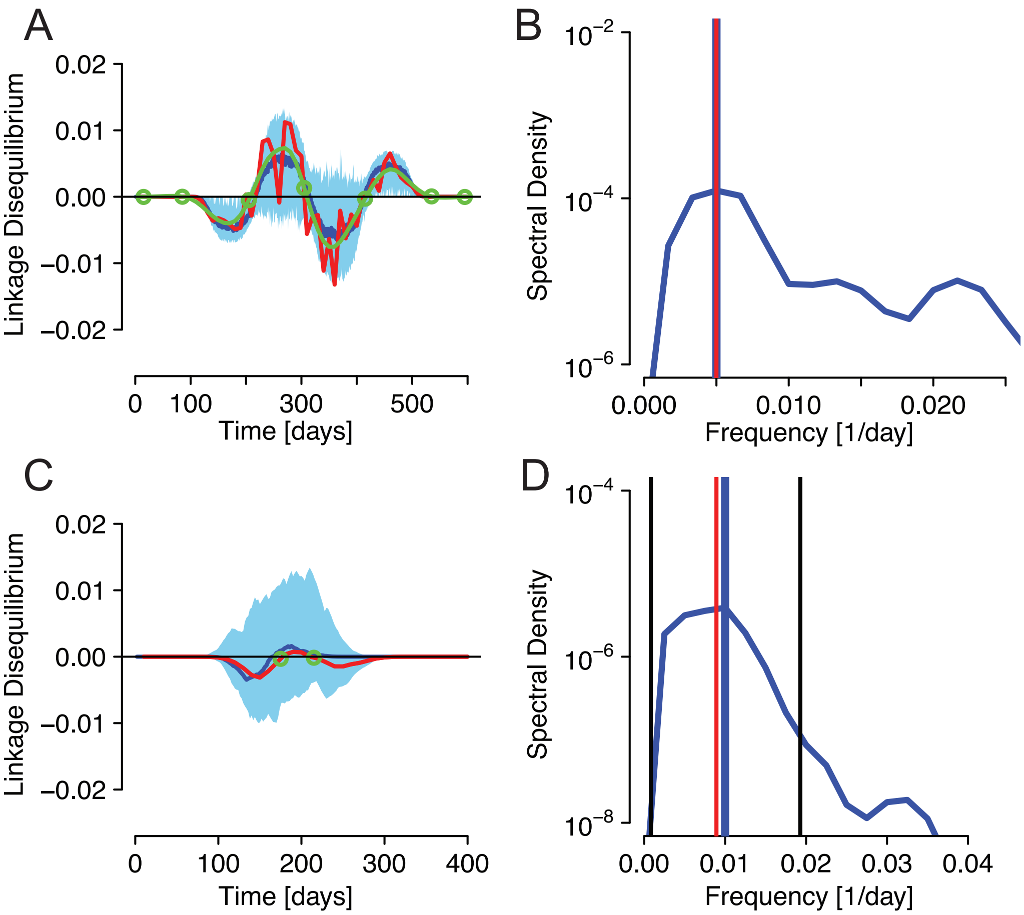

Fig. 3. **MLD oscillation frequency estimated by the power spectrum.** A) Oscillation of the fifth order LD in a symmetric full escape graph. The blue line is the median of a set of 200 runs, and the upper and lower bounds of the light blue area represent the 2.5 and 97.5 percentiles of all measured 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. The green line is a smoothing spline laid through the measured LD data, with green circles at the zeros. B) The periodogram of the fifth order MLD values (blue, non-vertical line) obtained with the sampling points of the red line in A). The maximum spectral density (vertical thick blue line) is attained exactly at the inverse of the simulated period of $T = 200$ days of the oscillations (thin vertical red line at frequency 0.005 per day). The thin red line is on top of the thicker blue. C) The analogous situation to A) for 100 simulation runs of the Wright-Fisher 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. D) Periodogram (non-vertical blue line) of one randomly chosen MLD trajectory in C) (red line in C)). The vertical thin red line is the theoretical value from CKF theory [58], whereas the thick vertical blue line goes through the maximum spectral density attained. The vertical black lines are ω_{low} and ω_{upp} (see *Materials and Methods*).

stochasticity in the WF dynamics leads to a dampening of the wavelet-like MLD patterns, the signal processing is very likely to under-estimate the true frequency. Nevertheless, the theoretical values are close to the median, and always stay within boundaries that include 95% of an analyzed signal (see *Materials and Methods*).

Implications for inference

To explore whether an MLD-based method for interference identification could be applied in real systems, it is important to understand the robustness of oscillations in MLD not only to evolutionary stochasticity (as described above), but also to randomness introduced by sampling procedures that could apply in real systems. Such sources of randomness include haplotype sample size, the sampling frequency, and, if the evolutionary parameters of a system are consistent across populations, the number of replicates examined.

We investigated the effects of sampling frequency on MLD frequency inference. As a test for the detectability of the MLD pattern, we analyzed how estimates of the speed of evolution based on equation (8) compare to CKF-theory predictions. The speeds of evolution were estimated during simulation time periods with largely non-zero genetic diversity to reduce computational overhead (see *SI*.7 and *SI*.8).

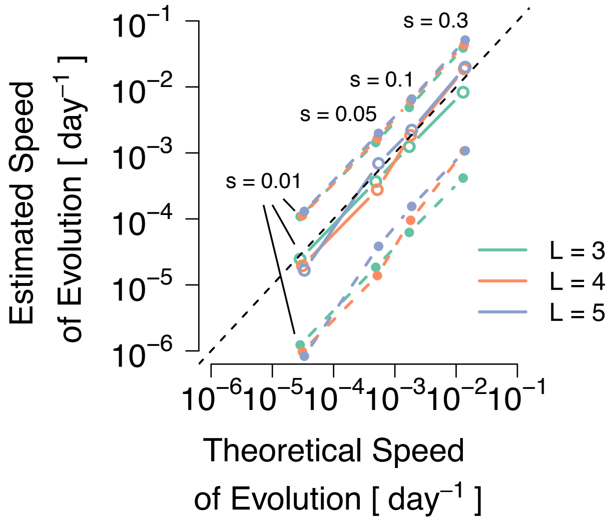

Fig. 4. **Estimates of speed of evolution based on MLD-oscillations versus theoretical predictions in the MMI regime [58].** Estimates of the speed of evolution v_e obtained by spectral analysis for Wright-Fisher model simulations run for $L = 3$, $L = 4$ and $L = 5$ loci, and selective coefficients $s \in \{0.01, 0.05, 0.1, 0.3\}$. We use these s values to estimate v_e with equation (8), where for f we use the estimated oscillation frequency derived from the spectral analysis of MLD time series. The inter-sampling period was $\Delta t = 2$. For each parameter combination a set of one hundred simulation repeats was run. The colored open circles are the median speed of evolution estimate per set. The medians of ω_{low} as well as ω_{upp} of each set –see *Materials and Methods*– are shown as filled colored circles. For each L, the medians of ω_{low} and ω_{upp} values of each set with different s are connected by dashed colored lines, and the median estimates are connected by colored lines. The dashed black line is the line of equality of x- and y-axis values. Perfect haplotype frequency information is assumed.

As long as sampling is more frequent than the expected length of an MLD half-oscillation, the accuracy of MLD-derived estimates of the rate of evolution appear largely robust to sampling frequency (see Fig. S3). This is because for simulations run with very high selection coefficients, e.g. $s = 0.3$, the period T of an oscillation is still of the order of 10^2 days. In the particular case of HIV, sampling bi-weekly should still preserve much of the signal. This should also hold for other pathogens with smaller evolutionary rates. Such inter-sampling periods lie within the means of modern empirical approaches [59–61].

To further examine the fidelity of the signal processing approach, we performed a Fisher test for hidden periodicities on all MLD time series [62, 63], where the null hypothesis is that the time series is generated by Gaussian White Noise (see *SI*.8). Since the strict assumptions for performing this test are invalid in our simulations, the results can only serve as a conservative indicator for the presence of wavelet-like oscillations. At the lowest inter-sampling duration of 2 days, the null was rejected at a 5% significance level in more than 95% of cases (except $L = 5$). At larger between-sampling periods of 10 days, the p-value medians of simulations with $s \in \{0.01, 0.05, 0.1\}$ increase, but are below or at the significance level (see *SI*.8). As a further cautionary measure, we therefore suggest applying signal processing to genetic data only in those MLD time series where Fisher's test detects oscillations.

We also investigated the effects of haplotype sample sizes on MLD frequency inference. Unlike for between-sampling duration, the accuracy of the estimates rapidly deteriorates with smaller sample sizes (see *SI*.9), and requires rather accurate haplotype information in order to obtain satisfactory median estimates and confidence bounds. Samples of about 5-20 haplotypes, as obtained in some early HIV infection studies [59], do not suffice to reveal MLD oscillations.

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 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 can 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 (7). Finally, we demonstrate that these oscillations are robust to some evolutionary stochasticity and their frequency is proportional to the rate of evolution of a population in the MMI regime. In essence, this indicates that MLD is an appropriately tuned statistic to assess interference effects in short-term adaptation; it is robust enough to small haplotype frequency changes to be unaffected by evolutionary stochasticity, but sensitive to more structural modifications in the genetic composition of the population. We conclude that the wavelets in MLD over time are a hallmark of the MMI –as opposed to one-by-one clonal interference– regime. Thus, MLD dynamics may contain information relevant to the study of the short-term evolution of microorganisms, including human pathogens like HIV, in which a finite number of loci experience strong selection. Crucially, an MLD-wavelet based MMI criterion does not require costly and time consuming competition assays, but can establish the presence of MMI in a system using haplotype information only.

There remain several possible caveats when using oscillations in MLD to detect multiple mutations interference in natural populations.

First, there is a non-interference edge case that can generate MLD-wavelets similar to those studied here. In particular, if odd-numbered L-mutant haplotypes sweep through a large, mostly wildtype population, MLD may briefly oscillate. This scenario may become relevant when $N >> 1/(L\mu_b)^L$; that is, the chance of pre-existing L-mutants at the onset of selection is high (see *SI*.6). Simple checks in data, such as the detection of L-mutant sweeps or very low haplotype diversity, can differentiate these MLD wavelets from evidence for interference.

Second, the detection of MLD oscillations depends on accurate haplotype frequency estimates and to a lesser degree on how frequently data are sampled. The continuous improvement of sequencing technologies is likely to allow for deep and dense sampling in the future, producing appropriate datasets. For example, in HIV, single genome amplification (SGA) techniques are likely to become more cost-effective over time. Moreover, methodology for the attainment of higher haplotype frequency resolution is under development [64, 65].

Third, epistatic effects were ignored. The reason for this choice is that in escape mutations of HIV, which inspired this work, we are unaware of evidence for epistatic interactions. However, other intragenic mutations are likely to give rise to epistasis [66–69]. More generally, if selection dominates over epistasis (magnitude epistasis) [70], the population will preferentially evolve along certain mutational pathways to the full escape genotype. This will break the symmetry in k-mutant representations even more than expected by drift, further dampening the oscillatory MLD signal. When epistasis dominates (sign epistasis), the evolutionary dynamics are likely to halt at a local or global fitness peak (i.e. not the full escape haplotype) [70]. In such a scenario, at mutation-selection balance, an MLD signal should be maintained that is constant and not oscillatory. This may serve to differentiate epistasis-dominated from weak or no-epistasis scenarios. Our work may thus also help to emphasize the consequences of linkage on epistatic interactions, which are commonly overlooked [71].

Fourth, we assume that all beneficial alleles confer the same selective advantage s. In natural populations beneficial mutations confer a range of selective advantages. Under *one-by-one clonal interference* [16, 19, 72] a strongly beneficial mutations may appear while weakly beneficial mutations are sweeping. The haplotype containing the strongly selected mutation will come to dominate the population. In the process, a large portion of the haplotype diversity that would give rise to MLD oscillations is destroyed. Thus, the broad underexponential distribution of fitness effects associated with CI is likely to weaken or eliminate the oscillatory MLD signal [2, 16, 21, 73]. This property of CI makes full MLD oscillations an MMI-specific signature. MLD oscillations will therefore provide a conservative categorization of the system as subject to MMI, rather than CI.

Despite its limitations, our approach to the study of interference has clear benefits. One advantage 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 [34]. This may be due to different definitions of linkage disequilibrium at multiple loci [37, 38, 47, 74, 75]. The crucial advantage of the Geiringer-Bennet MLD is that its maximum likelihood estimate always exists [76], a very useful property for estimation. We did not investigate statistical issues of MLD inference [77] or visualization (see [78] for techniques of an MLD definition similar to Slatkin's).

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 here described, suggests that the population under study is evolving under a 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. These parameters, especially those regarding the distribution of fitness effects of beneficial alleles, are relevant for the efficiency with which the immune system or sets of drugs [8] are able to target evolving haplotypes and clear infections.

A broad range of microorganisms may experience simultaneous strong selection at multiple loci [79] and, more specifically, MMI. In this study, we have shown that finite-sites MMI shows distinct MLD dynamics. The tools and concepts used in this approach may help elucidate the mode of adaptation of a wide array of microorganisms, particularly those that evolve in complex environments, such as within human hosts.

Materials and Methods

Wright-Fisher Simulation Model

To study the behavior of multilocus LD in the MMI regime we used a Wright-Fisher (WF) model with selection and mutation in a finite population, described in detail in [41]. Our model was designed to simulate specific aspects of the within-host replication and genetic diversification of HIV during early infection, in particular the changes in its genetic composition, and is rooted in a well-established tradition of population genetic HIV models [8, 50, 55, 80, 81]. The purpose of the model is not to give a maximally accurate description of the population dynamics of HIV, as it does not incorporate viral expansion and contraction phases. Nor is it intended to perfectly replicate putative differences in selection benefits incurred from acquiring mutations, which are likely. Instead, our goal is to focus on the changes in genetic structure of a population under MMI in a widely used model (WF) to identify aspects of HIV's within host evolution that might be translated to other, structurally similar scenarios.

The model tracks the frequency dynamics of cells infected by viral RNA haplotypes $\mathbf{i}$, which are represented as binary sequences of length L: $\mathbf{i} = i_1 i_2 \dots i_L$, with $i_j \in \{0, 1\}$. Simulations begin with an ancestral wildtype, represented by a sequence of zeros (the likely situation in most HIV infections after a population bottleneck due to transmission [80–82]). Zeros can mutate into ones at a rate $\mu_b = 10^{-4}$ per replication per cell, the beneficial mutation rate, which corresponds to estimates for the mutation rate of *epitopes* under selection in HIV [55]. Epitopes are stretches of RNA of about 8-10 codons in length coding for immunogenic virion peptide pieces, and are hypothesized to confer strong selective advantages when altered in physical shape due to their escape from recognition by the immune system. No back mutations are considered. One replication is assumed to take two days [83–86]. Simulations take a populations size N, and proceed by resampling from the previous generation while applying selection.

Selection alters genotype frequencies by affecting their resampling probabilities, and acts from the beginning of the simulation. The inter-generational growth factor of a haplotype $w_{\mathbf{i}}$, is tied to its relative growth rate; $s_{\mathbf{i}}$, $w_{\mathbf{i}} = e^{s_{\mathbf{i}}}$. All mutations are assumed to confer equal additive selective advantages s, which is multiplicative in w, and thus a log-fitness. For example, a haplotype $\mathbf{i}$ with k mutations would have a log-fitness ks, and correspondingly, $w_{\mathbf{i}} = e^{ks}$. At each generation, after mutations have been incorporated, haplotype $\mathbf{i}$ is resampled from the last generation with probability $p_{\mathbf{i}}^{\text{sel}} = \frac{e^{s_{\mathbf{i}}}}{\langle e^s \rangle} p_{\mathbf{i}}$ [87], where $\langle e^s \rangle = \sum_{\mathbf{i}} e^{s_{\mathbf{i}}} p_{\mathbf{i}}$ is the average fitness of the population, and $p_{\mathbf{i}}$ is the population frequency of $\mathbf{i}$. Selection starts in the first generation.

By default and if not otherwise stated, simulations were run for 2000 generations. The program was written in the C# programming language.

Oscillation estimation by means of signal processing techniques

To identify oscillations of MLD in the simulation data, we developed a detection scheme based on spectral 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 $n \in \{0, \dots, M_s - 1\}$. Sample data x_n are assumed to have been obtained at constant inter-sampling periods, and can be expressed as a vector $\mathbf{x}$ with entries x_n. We analyze the spectral density of the signal $\mathbf{x}$. An oscillating LD measure of L loci will maximally generate $L-1$ half-oscillations, starting with a negative half-oscillation. Even if dampened, such wavelet-like oscillations should leave traces in the frequency spectrum that are close to the frequency of a full period, T.

The Fourier transform coefficients of the data x_n are defined by:

$$X_k \doteq \sum_{n=0}^{M_s-1} x_n \cdot e^{-i2\pi kn/M_s}, \; k \in \mathbb{N}, \tag{9}$$

where $\mathbb{N}$ is the set of all natural numbers and the X_k are complex valued numbers.

In signal processing, the energy of a signal refers to the quantity $\sum_n |x_n|^2$, which is assumed to be finite in time-limited data. Parseval's theorem relates the energy of the signal to the Fourier coefficients:

$$\frac{1}{M_s} \sum_k |X_k|^2 = \sum_n |x_n|^2, \tag{10}$$

implying that the energy of the signal is preserved by the Fourier transform, except for a normalization factor $\frac{1}{M_s}$. Thus, a suitable measure of how the signal is distributed over Fourier space is the *power spectral density* or *power spectrum* $S(\omega)$, a concept that defines how a signal is decomposed into its frequency components [88]. The value $S(\omega)$ may be interpreted as the fraction of the signal that is explained per unit frequency at ω. The spectrum may also be treated analogously to a probability density function. In theory, the power spectrum is defined by an infinite number of signal points. In practice, the power spectrum is estimated by the *periodogram* [63, 88]:

$$I(\omega_k) = \frac{\Delta t}{M_s} | \sum_{n=0}^{M_s-1} x_n e^{i\omega n} |^2, \tag{11}$$

where $\omega_k = 2\pi k/M_s$ is the frequency of the wave component $e^{i\omega n}$, Δt is the inter-sampling period, and i is the imaginary unit. $I(\omega_k)$ can be interpolated between values of ω_k, and we will therefore drop k in the following. We thus assume $S(\omega) \approx I(\omega)$.

For each simulation, we obtained the frequency at which the periodogram was maximal: $\omega_{\max} = \max_{\omega} S(\omega)$, providing the largest contribution to the signal represented as a sum of waves. This served as our frequency estimate.

We then defined bounds for $\omega_{\max}$, by calculating the two values ω_{low} and ω_{upp} that enclose 95% of the density distribution of the periodogram. To do this, we also calculated the analogue of the cumulative distribution function of the periodogram, $W(f) = \sum_{\omega<f} S(\omega)$ for every simulation, where $W(f)$ gives the area under the function $S(\omega)$ from zero to f. We then used $W(f)$ to obtain its inverse, $W^{-1}(z) = f$, where z gives the fraction of the total area enclosed by $S(\omega)$ at smaller values than f. With this, we define $p_w \doteq W(\omega_{\max})$, which allows us to find bounds: $\omega_{\text{low}} = W^{-1}(p_w - 2p_w\Delta)$ and $\omega_{\text{upp}} = W^{-1}(p_w + 2(1 + p_w)\Delta)$, where Δ is the deviation from p_w. Defined in this way, these bounds account for the fact that $\omega_{\max}$ is not the median, and deviates from p_w in fractions of p_w. Note that if $\Delta = 0.5$,

$p_w - 2p_w\Delta = 0$, placing ω_{low} at the lowest possible frequency, and conversely, ω_{upp} at the highest possible one. By default, $\Delta = 0.475$.

For values between the discrete frequencies given by the periodogram, we used the interpolation function *approxfun* from the *stats* package in R to approximate ω_{low} at low values of z. As a backup, if this procedure failed, we ran a regression line through the 5 lowest z values in the inverse cumulative frequency distribution. The value of ω_{low} could then be calculated by evaluating the regression at $w - 2w\Delta$. If the extrapolation led to $f_{\text{extrapolation}} = W^{-1}_{\text{extrapolation}}(w - 2w\Delta) < 0$, then $f_{\text{extrapolation}}$ was automatically set to zero.

Because the the periodogram was formed on the adjusted data $x_n - \langle x \rangle$, where $\langle x \rangle$ is the average of the data, the periodogram value at frequency zero must necessarily be zero, which is not given by default in the R function *spectrum* [45]. To perform our analyses, we thus added it to improve the functioning of the interpolation function.

The periodogram is not a consistent estimator [63, 88]. To address this problem, a smoothing function is typically applied on the signal. We used the default modified Daniell kernel in *R* [45], with $m = 1$. Further, we applied the default split cosine bell taper to a 10% proportion of the data at the beginning and end of the signal series.

Acknowledgments

The authors thank Fabio Zanini, Roland Regoes, Frederic Bertels 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). MWF was supported in part by The Stanford Center for Computational, Evolutionary and Human Genomics, and by the Morrison Institute for Population and Resource Studies.

References

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

[2] Neher RA (2013) Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics* 44(1):195–215.

[3] Tenaillon O et al. (2012) The molecular diversity of adaptive convergence. *Science* 335(6067):457–461.

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

[5] Lang GI, Botstein D, Desai MM (2011) Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* 188(3):647–661.

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

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

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

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

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

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

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

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

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

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

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

[17] Bonhoeffer S, May RM, Shaw GM, Nowak MA (1997) Virus dynamics and drug therapy. *Proc Natl Acad Sci USA* 94(13):6971–6.

[18] Bonhoeffer S, Lipsitch M, Levin BR (1997) Evaluating treatment protocols to prevent antibiotic resistance. *Proc Natl Acad Sci USA* 94(22):12106–11.

[19] Gerrish P, Lenski R (1998) The fate of competing beneficial mutations in an asexual population. *Genetica* 102:127–144.

[20] Lang GI et al. (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* 500(7464):571–4.

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

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

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

[24] Miller CR, Joyce P, Wichman HA (2011) Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* 187(1):185–202.

[25] Regoes R, Yates A, Antia R (2007) Mathematical models of cytotoxic T-lymphocyte killing. *Immunology and Cell Biology* 85(4):274.

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

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

[28] Gillespie JH (2000) Genetic drift in an infinite population. the pseudohitchhiking model. *Genetics* 155(2):909–19.

[29] Kim Y, Stephan W (2003) Selective sweeps in the presence of interference among partially linked loci. *Genetics*

164(1):389–98.

[30] Park SC, Krug J (2007) Clonal interference in large populations. *Proc Natl Acad Sci USA* 104(46):18135–18140.

[31] Barton N (2000) Estimating multilocus linkage disequilibria. *Heredity* 84(3):373–389.

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

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

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

[35] Tibayrenc M, Ayala FJ (2012) Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proc Natl Acad Sci USA* 109(48):E3305–3313.

[36] Thomson G, Bodmer W (1979) Hla haplotype associations with disease. *Tissue antigens* 13(2):91–102.

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

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

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

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

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

[42] Dausset J et al. (1978) A haplotype study of HLA complex with special reference to the HLA-DR series and to Bf. C2 and glyoxalase I polymorphisms. *Tissue Antigens* 12(4):297–307.

[43] Gorelick R, Laubichler M (2004) Decomposing multilocus linkage disequilibrium. *Genetics* 166(3):1581–1583.

[44] Andrews GE (1976) The Theory of Partitions, volume 2 of Encyclopedia of Mathematics and its Applications.

[45] R Development Core Team (2012) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria). ISBN 3-900051-07-0.

[46] Hankin RKS (2006) Additive integer partitions in R. *Journal of Statistical Software, Code Snippets* 16.

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

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

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

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

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

[52] Agresti A (1992) A survey of exact inference for contingency tables. *Statistical Science* pp. 131–153.

[53] Bishop YM, Fienberg SE, Holland PW (2007) *Discrete multivariate analysis: theory and practice.* (Springer).

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

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

[56] Garcia V, Regoes RR (2015) The effect of interference on the CD8+ T cell escape rates in HIV. *Frontiers in Immunology* 5(661).

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

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

[59] Goonetilleke N et al. (2009) The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J Exp Med* 206(6):1253–1272.

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

[61] Salazar-Gonzalez JF et al. (2009) Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* 206(6):1273–1289.

[62] Ahdesmaki M, Fokianos K, Strimmer. K (2012) *GeneCycle: Identification of Periodically Expressed Genes.* R package version 1.1.2.

[63] Brockwell PJ, Davis RA (1991) *Time series: Theory and Methods.* (Springer).

[64] Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2010) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 17(3):417–428.

[65] Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12:119.

[66] Hinkley T et al. (2011) A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase. *Nat Genet* 43(5):487–9.

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

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

[69] Wang K, Mittler JE, Samudrala R (2006) Comment on ”evidence for positive epistasis in hiv-1”. *Science* 312(5775):848; author reply 848.

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

[71] Phillips PC (2008) Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9(11):855–67.

[72] Gerrish P (2001) The rhythm of microbial adaptation. *Nature* 413(6853):299–302.

[73] Fogle C, Nagle J, Desai M (2008) Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics* 180(4):2163–2173.

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

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

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

[77] Weir BS, Wilson SR (1986) Log-linear models for linked loci. *Biometrics* 42(3):665–670.

[78] Mourad R, Sinoquet C, Dina C, Leray P (2011) Visualization of pairwise and multilocus linkage disequilibrium structure using latent forests. *PLoS One* 6(12):e27320.

[79] Messer PW, Ellner SP, Hairston NG (2016) Can Population Genetics Adapt to Rapid Evolution? *Trends Genet.*

[80] Keele B et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA* 105(21):7552.

[81] Lee HY et al. (2009) Modeling sequence evolution in acute HIV-1 infection. *J Theor Biol* 261(2):341–360.

[82] Abrahams MR et al. (2009) Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J Virol* 83(8):3556–3567.

[83] Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271(5255):1582–1586.

[84] Rodrigo AG et al. (1999) Coalescent estimates of HIV-1 generation time in vivo. *Proc Natl Acad Sci USA* 96(5):2187–2191.

[85] Markowitz M et al. (2003) A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *J Virol* 77(8):5037–5038.

[86] Murray JM, Kelleher AD, Cooper DA (2011) Timing of the components of the HIV life cycle in productively infected CD4+ T cells in a population of HIV-infected individuals. *J Virol* 85(20):10798–10805.

[87] Neher RA, Shraiman BI (2011) Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics* 83(4):1283.

[88] Shumway RH, Stoffer DS (2010) *Time series analysis and its applications: with R examples.* (Springer Science & Business Media).