Natural selection promotes the evolution of recombination 2: during the *process* of natural selection*

Philip J Gerrish,[1,2,3] Fernando Cordero,[4] Benjamin Galeota-Sprung,[5]
Alexandre Colato,[6] Varun Vejalla,[7] Nick Hengartner,[2] and Paul Sniegowski[5]

[1] *Department of Biology, University of New Mexico, Albuquerque, New Mexico, USA*†
[2] *Theoretical Biology & Biophysics, Los Alamos National Lab, Los Alamos, New Mexico, USA*
[3] *Instituto de Ciencias Biomédicas, Universidad Autónoma de Ciudad Juárez, México*
[4] *Biomathematics and Theoretical Bioinformatics, Technische Fakultät, Universität Bielefeld, Germany*
[5] *Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA*
[6] *Departamento de Ciências da Natureza, Matemática e Educação, Univ Fed de São Carlos, Araras SP, Brazil*
[7] *Thomas Jefferson High School for Science and Technology, Alexandria, Virginia, USA*

(Dated: 5 June 2021)

The ubiquity of sex and recombination in nature has eluded unified explanation since the time of Darwin. Conditions that promote the evolution of recombination, broadly defined as any form of genetic mixing, are fairly well understood: it is favored when genomes tend to contain more selectively mismatched combinations of alleles than can be explained by chance alone. Yet, while a variety of theoretical approaches have been put forth to explain why such conditions would have a tendency to prevail in natural populations, each has turned out to be of limited scope and applicability. In our two-part study, we show that, simply and surprisingly, the action of natural selection acting on standing heritable variation creates conditions favoring the evolution of recombination. In this paper, we focus on the mean selective advantage created by recombination between individuals from the same population. We find that the mean selective advantages of recombinants and recombination are non-negative, in expectation, independently of how genic fitnesses in the standing variation are distributed. We further find that the expected asymptotic frequency of a recombination-competent modifier is effectively equal to the probability that the fittest possible genotype is a virtual recombinant; remarkably, expected asymptotic modifier frequency is independent of the strength of selection. Taken together, our findings indicate that the evolution of recombination should be promoted in expectation wherever natural selection is operating.

I. INTRODUCTION

The oldest ideas about the evolutionary role of recombination are from Weismann (1889), who argued that sex provides increased variation for natural selection to act upon. Since then, the amount of work that has addressed the evolution of sex and recombination is spectacular. To preface our developments therefore, we will cover some essential background and will make reference to some wonderful reviews [3–9] that give a much more complete overview of the remarkable wealth of previous and current work in this area. Also, we refer the reader to our companion publications [1, 2] for additional introductory material.

Fisher [10] and Muller [11] first provided concrete mechanisms for an advantage to recombination. Muller surmised that in order for separately arising beneficial mutations to fix in the same genotype, in an asexual population they must arise in the same lineage sequentially, while in a recombining population, they may arise contemporaneously and be subsequently reshuffled into the same background. Fisher argued that a single beneficial mutation, because it arises in a single individual, has a significant probability of arising on a non-optimal genetic background. In an asexual population, the beneficial mutation is stuck with this non-optimal background, while in a recombining population, the background can be swapped out for a fitter one. If the beneficial mutation is successful, despite arising on a non-optimal background, a second beneficial mutation may eventually arise on the background of the first as it progresses toward fixation. If this happens in an asexual population, Hill and Robertson [12, 13] found that the probability of success of the second beneficial mutation will be depressed as a consequence of arising in the growing lineage founded by the first beneficial mutation. Generally speaking, genetic linkage (the absence of recombination) introduces selective interference [5] that decreases the efficiency of natural selection.

Recombination can ameliorate all of these linkage-induced hindrances to natural selection [14–17], and recombining populations should therefore adapt faster [18]. However, the magnitude of this benefit depends very much upon parameter choices [19]. More fundamentally, this process provides only a group-level benefit for sex, and group-level explanations, besides being characteristically viewed with suspicion in evolutionary biology, are unsatisfactory in that they cannot explain the origin

* This article is published in concert with two companion papers referenced as PRL [1] and PRE1 [2] and Supplemental Materials referenced by the adding the prefix "S".

† pgerrish@unm.edu

and fixation of sexual reproduction within a single population, nor explain its maintenance by evolution [20]. Therefore it is necessary to study the evolution of recombination within a single population.

To do so requires consideration of an additional "modifier" locus that determines recombination rate. An allele at this locus conferring increased recombination rate is introduced into a population at low frequency. The questions of interest are: 1) what is the selective value of this allele? and 2) what is the fate of this allele? A variety of theoretical studies have studied the evolution of such recombination modifiers. These studies have investigated mechanisms including fluctuating selection [14, 15, 21]; negative epistasis [14, 15, 22, 23]; assortative mating [24]; and finite population effects, i.e. drift [12, 17, 25–27]. Of these, the drift-based explanations have come into favor in recent years as the more promising in explaining the ubiquity of recombination [28], but the general consensus is that some fundamental piece of the puzzle is still missing [5].

To address the evolution of sex and recombination, we have taken a reductionist approach. Our aim is restricted to studying the effects of one very key process, namely natural selection, in isolation (no mutation, no drift, etc), and we distill this problem to what we believe is its most essential form: we ask, how does the action of natural selection, by itself, affect the selective value and fate of recombinants and recombination? In choosing this approach, we sought analytic tractability and transparency that might lead to robust new insights into the evolution of sex and recombination.

Our focus on *recombinants* tells us what will happen to a recombination-competent modifier (or rec^+ modifier) when linkage between the modifier and fitness loci is weak, i.e., when there is a high rate of recombination between fitness loci and the modifier locus itself. A modifier will be bumped up in frequency a notch with each advantageous recombinant it produces, before becoming dissociated from the recombinant through recombination. If recombinants are more likely to be advantageous than disadvantageous then the modifier will on average increase in frequency over time. Our later focus on *recombination* tells us what will happen to a rec^+ modifier when linkage between the modifier and fitness loci is strong, i.e., when there is a low rate of recombination between fitness loci and the modifier locus itself.

In companion papers [1] and [2], we show that natural selection acting on standing variation has an encompassing tendency to amplify selectively mismatched combinations of alleles, thereby promoting the evolution of recombination across selected genotypes. In the present study, we assess how the selective value of recombinants and recombination are affected during the process of natural selection within a population. In these combined studies, we find that recombinants are favored and recombination promoted, in expectation, as an inherent consequence of the dynamics and statistical properties of selective sorting. Our findings elucidate and unify some classical models.

II. MEASURING SELECTIVE IMBALANCE

In much of the relevant literature, the measure of selective mismatch across loci affecting the evolution of recombination is *linkage disequilibrium* (LD) [14–17, 23, 29, 30], which measures the covariance in allelic *states* across two loci [5] (i.e., it measure the bias in allelic frequencies across loci) but does not retain information about the selective value of those alleles.

Here, our measure of selective mismatch will be *covariance* between genic fitnesses. This departure from tradition is advantageous because covariance retains information about both the frequencies and selective value of alleles, and it is convenient because the mean selective advantage accrued by recombinants over the course of a single generation is equal to minus the covariance (below). Many of our results will thus be given in terms of covariance.

Discrete time

For the purpose of presentation, it is enough to consider an organism whose genome consists of just two genes, or *loci*. We let random variable X denote the fitness contribution, or *genic fitness*, of the first locus, and we let Y denote the genic fitness of the second locus. Classical population genetics was formulated in discrete time and asserted that fitness was multiplicative. The fitness of an individual organism in this case is the product XY, and the average selective advantage of recombinants is $\bar{s}_r = \mathbb{E}[X]\mathbb{E}[Y]/\mathbb{E}[XY] - 1 = -\sigma_{XY}/\bar{w}$, where σ_{XY} is covariance, and $\bar{w} = \mathbb{E}[XY]$ is mean fitness.

Continuous time

In the present study, everything is in continuous time. Redefining random variables X and Y to be continuous-time genic fitnesses at two loci, we define their cumulant-generating function (*cgf*), $\mathcal{C}_0(\varphi, \theta) = \ln \mathbb{E}[e^{\varphi X + \theta Y}]$ at time 0. In later developments and in the Supplemental Materials (SM), we show that, in continuous time, the mean selective advantage of recombinants over the course of their first generation of growth is:

$$\begin{aligned} \bar{s}_r &= \mathcal{C}_0(1,0) + \mathcal{C}_0(0,1) - \mathcal{C}_0(1,1) \\ &= \ln \frac{\mathbb{E}[e^X]\mathbb{E}[e^Y]}{\mathbb{E}[e^{X+Y}]} \approx -\sigma_{XY} \end{aligned}$$

This approximation is extremely accurate because, as we show in the SM, the two-dimensional Jensen gaps for numerator and denominator essentially cancel each other out. It is further corroborated by stochastic simulations of a Moran evolutionary process over the course of a single generation (SM).

III. FITNESS EVOLUTION

As stated above, the goal of the present study is to focus exclusively on natural selection and ask how natural selection, by itself, affects the selective value of recombinants and recombination. This goal requires a reductionist approach in which natural selection is studied in isolation. Consequently, our evolutionary models here retain only the natural selection terms; other, more complete models that incorporate selection, mutation, drift, and recombination, may be found in the SM; these are presented there for completeness and to lay the groundwork for subsequent studies.

One locus

This model is a continuous-time formulation of evolution by natural selection; the model and its analyses are not new and have close parallels in [31–36]. We let $u_t(x)$ denote probability density in fitness x at time t (i.e., $\int_x u_t(x) = 1$ for all t) for an evolving population. Dropping the subscript t, we have that, under selection and mutation, u evolves as:

$$\partial_t u(x) = (x - \bar{x})u(x)$$

where $\bar{x}$ is mean fitness ($\bar{x} = \int_x x u_t(x)$). Let $M(\varphi)$ denote the moment-generating function (*mgf*) for $u(x)$, i.e., $M(\varphi) = \mathbb{E}_u[e^{\varphi X}]$. The transformed equation is:

$$\partial_t M(\varphi) = \partial_\varphi M(\varphi) - \partial_\varphi M(0) M(\varphi).$$

We define cumulant-generating function (*cgf*) $\mathcal{C}(\varphi) = \ln M(\varphi)$; noting that $\partial_\varphi \mathcal{C}(\varphi) = (\partial_\varphi M(\varphi))/M(\varphi)$, and $\partial_t \mathcal{C}(\varphi) = (\partial_t M(\varphi))/M(\varphi)$ we find that the *cgf* evolves as:

$$\partial_t \mathcal{C}(\varphi) = \partial_\varphi \mathcal{C}(\varphi).$$

This equation is a variant of the transport equation; it is immediately apparent that the solution will be of the form $\mathcal{C}_t(\varphi) = F(\varphi + t)$, where F is an arbitrary function. When boundary condition $\mathcal{C}_t(0) = 0 \; \forall \; t$ is applied, it has solution:

$$\mathcal{C}_t(\varphi) = \mathcal{C}_0(\varphi + t) - \mathcal{C}_0(t) \tag{1}$$

where the subscripts are now necessary again: $\mathcal{C}_t(\varphi)$ is the *cgf* of the fitness distribution $u_t(x)$ at time t. We note that the fitness evolution of a population can thus be projected into the future based only on the present fitness distribution (i.e., at $t = 0$).

Two loci

While many parts of our analyses are true for general fitness functions, where total fitness is some arbitrary function $z = \phi(x, y)$, we here and in other parts of our analyses restrict ourselves to additive fitness $z = x + y$. Remarkably, however, this restriction to additive fitness will cease to be necessary in some expressions and key findings downstream from here. Also, generalizations of these evolution equations to non-additive functions are found in the SM.

We now suppose that there are two "genes" that determine fitness. Letting fitness contributions of the two genes be denoted by x and y, respectively, the total fitness is $z = x + y$. The extension of the previous one-dimensional *pde* is immediate:

Let $u_t(x, y)$ denote probability density in fitness contributions x and y at time t for an evolving population. Dropping the subscripts again, under selection and mutation, u evolves as:

$$\partial_t u(x, y) \;=\; (x + y - \bar{x} - \bar{y})u(x, y)$$

The *cgf* is now two-dimensional: $\mathcal{C}(\varphi, \theta) = \ln \mathbb{E}[e^{\varphi X + \theta Y}]$, and evolves as:

$$\partial_t \mathcal{C}(\varphi, \theta) = \partial_\varphi \mathcal{C}(\varphi, \theta) + \partial_\theta \mathcal{C}(\varphi, \theta) - \partial_\varphi \mathcal{C}(0, 0) - \partial_\theta \mathcal{C}(0, 0). \tag{2}$$

This equation is a two-dimensional variant of the transport equation and has more possible solution forms than the one-dimensional case, namely, solutions can be of the form: $F(t + \varphi, \theta - \varphi)$, $F(t + \theta, \varphi - \theta)$, or $F(t + \varphi, t + \theta)$. The consistent solution form is the last of these. When boundary condition $\mathcal{C}_t(0, 0) = 0 \; \forall \; t$ is applied, it has solution:

$$\mathcal{C}_t(\varphi, \theta) = \mathcal{C}_0(\varphi + t, \theta + t) - \mathcal{C}_0(t, t) \tag{3}$$

where the subscripts have again become necessary. We again note that the evolution of a population can thus be projected into the future based only on the present fitness distribution (i.e., at $t = 0$).

Covariance dynamics

Infinitely-many genotypes

Covariance dynamics are immediate from Eq (3):

$$\sigma_{XY}(t) = \mathcal{C}_0^{(1,1)}(t, t) \tag{4}$$

where $\mathcal{C}_t^{(i,j)}(\varphi, \theta)$ denotes the $(i, j)^{\text{th}}$ derivative of the *cgf* for genic fitnesses at time t. We note that if the initial fitness distribution is normal, covariance remains constant over time: $\sigma_{XY}(t) \;=\; \sigma_{XY}(0) \quad \forall \; t$.

In real populations, the number of genotypes is finite. Given the marked qualitative difference between the finite- and infinite-genotypes cases, we will focus primarily on the finite-genotypes case.

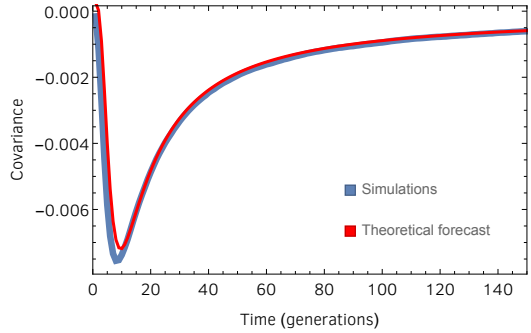

FIG. 1. Covariance forecast. Blue curve is the average of 1000 stochastic, individual-based simulations. Red curve is the theoretical prediction given by $\mathbb{E}[\tilde{\mathcal{C}}_0^{(1,1)}(t,t)]$ (Eq (7)). We call it a forecast because it is based solely on the distribution of fitnesses at $t = 0$. Fitnesses were drawn from a bivariate normal distribuiton with means $(0, 0)$ and standard deviations $(0.2, 0.2)$ and zero correlation. The initial population consisted of ten distinct genotypes in a population of size $N = 2000$.

Finitely-many genotypes

While Eq (4) is expressed in terms of the analytical *cgf*, it is equally applicable when the empirical *cgf* is employed. The empirical *mgf* and *cgf* are defined as:

$$\tilde{M}_t(\varphi,\theta) = \sum_{j=1}^{n} f_j(t) e^{\varphi x_j + \theta y_j} \tag{5}$$

$$\tilde{\mathcal{C}}_t(\varphi,\theta) = \ln \tilde{M}_t(\varphi,\theta) \tag{6}$$

where x_j and y_j are genic fitnesses of genotype j, n is number of genotypes, and $f_j(t)$ is the frequency of genotype j at time t. In what follows, f_j written as such with no argument denotes initial frequency, i.e., $f_j := f_j(0)$. In many of the developments that follow, we will assume that genotypes are initially found at equal frequency so that $f_j = 1/n$ for all j.

Replacing x_j and y_j in Eq (5) with random variables X_j and Y_j, the future dynamics of expected covariance may be computed as:

$$\mathbb{E}[\tilde{\sigma}_{XY}(t)] = \mathbb{E}[\tilde{\mathcal{C}}_0^{(1,1)}(t,t)] \quad . \tag{7}$$

This covariance-forecasting equation is shown to be very accurate in Fig 1. From this, we directly have the dynamics of recombinant advantage, because $\bar{s}_r(t) = -\sigma_{XY}(t)$.

IV. NATURAL SELECTION MAKES RECOMBINANTS ADVANTAGEOUS

Without epistasis

We recall that recombinant advantage is $-\sigma_{XY}$. Here, we examine the simplest scenario of two loci and two genotypes. We study how the selection-driven changes in the frequencies of types (x_1, y_1) and (x_2, y_2) *within a single unstructured population* change covariance $\sigma_{XY} = \sigma_{XY}(t)$ over time. We are interested in the net effect of these changes, given by $\int_0^\infty \sigma_{XY}(t)dt$; in particular, we are interested in knowing whether this quantity is positive (net recombinant disadvantage) or negative (net recombinant advantage) in expectation.

PROPOSITION 1. *Within-population covariance integrated over time is:*

$$\int_0^\infty \sigma_{XY}(t)dt \;=\; q\frac{(x_2 - x_1)(y_2 - y_1)}{|z_2 - z_1|} \tag{8}$$

where q is the initial frequency of the inferior genotype. And $z_i = x_i + y_i$.

Proof: We employ Eq (7) to give us covariance dynamics as a function of the initial two genotypes. We let p denote initial frequency of the superior of the two genotypes, and we let $q = 1 - p$ denote initial frequency of the inferior genotype. Time-integrated covariance is:

$$\int_0^\infty \sigma_{XY}(t)dt = \tag{9}$$

$$(x_2 - x_1)(y_2 - y_1)\int_0^\infty \frac{pq\mathrm{e}^{(z_1+z_2)t}}{(p\mathrm{e}^{z_1 t} + q\mathrm{e}^{z_2 t})^2}dt$$

We note the convenient fact that the integral depends only on the z_i, allowing us to relax the assumption of additive fitness and revert to the general definition $z_i = \phi(x_i, y_i)$ – a fact we will leverage below. Integration by parts yields:

$$\int_0^\infty \sigma_{XY}(t)dt = q\frac{(x_2 - x_1)(y_2 - y_1)}{|z_2 - z_1|}$$

where q in Prop 1 is written as $1 - p$. □

Now to find the expectation of time-integrated covariance, we must replace x_i and y_i with random variables X_i and Y_i. A *spacing* is a difference between two draws from the same distribution. For example $\Delta X = X_1 - X_2$ is a spacing. The expectation of time-integrated covariance is:

$$\mathbb{E}[\int_0^\infty \sigma_{XY}(t)dt] = q\mathbb{E}[\frac{\Delta X \Delta Y}{|\Delta Z|}]$$

where $\Delta X = X_2 - X_1$, $\Delta Y = Y_2 - Y_1$ and $\Delta Z = Z_2 - Z_1$ are all spacings. Spacings are themselves random variables and their distribution is always symmetric about zero. This fact is what allows us to make the following remarkable statement:

PROPOSITION 2. *We define spacings $\Delta X = X_2 - X_1$, $\Delta Y = Y_2 - Y_1$, and $\Delta Z = Z_2 - Z_1 = \Delta X + \Delta Y$. If the pairs (X_i, Y_i) are independently drawn from any distribution, then ΔX and ΔY are symmetric about zero, and time-integrated covariance is unconditionally non-positive in expectation:*

$$\mathbb{E}[\int_0^\infty \sigma_{X,Y}(t)dt] \;=\; q\mathbb{E}[\frac{\Delta X \Delta Y}{|\Delta Z|}] \;\leq\; 0$$

Proof: The bivariate distribution governing X and Y is irrelevant to this proof: no assumption about this distribution is required. There is also no need to assume that $(\Delta X, \Delta Y)$ has a density. ΔX and ΔY are two real-valued random variables such that: $(-\Delta X, \Delta Y)$ has the same distribution as $(\Delta X, \Delta Y)$. This is guaranteed by the fact that ΔX and ΔY are spacings. We have:

$$\begin{aligned}
&\mathbb{E}[\Delta X \Delta Y / |\Delta X + \Delta Y|] = \\
&\mathbb{E}[\mathbb{1}_{\Delta X \Delta Y > 0} \Delta X \Delta Y / |\Delta X + \Delta Y|] \\
&\qquad + \mathbb{E}[\mathbb{1}_{\Delta X \Delta Y < 0} \Delta X \Delta Y / |\Delta X + \Delta Y|] \\
&= \mathbb{E}[\mathbb{1}_{\Delta X \Delta Y > 0} \Delta X \Delta Y / |\Delta X + \Delta Y|] \\
&\qquad + \mathbb{E}[\mathbb{1}_{-\Delta X \Delta Y < 0} (-\Delta X) \Delta Y / |\Delta Y - \Delta X|] \\
&= \mathbb{E}[\mathbb{1}_{\Delta X \Delta Y > 0} \Delta X \Delta Y / |\Delta X + \Delta Y|] \\
&\qquad - \mathbb{E}[\mathbb{1}_{\Delta X \Delta Y > 0} \Delta X \Delta Y / |\Delta Y - \Delta X|] \\
&= \mathbb{E}[\mathbb{1}_{\Delta X \Delta Y > 0} \Delta X \Delta Y (1/|\Delta X + \Delta Y| - 1/|\Delta Y - \Delta X|)] \\
&\leq 0
\end{aligned}$$

When ΔX and ΔY have the same sign as imposed by the indicator function in the last expectation, we have $|\Delta X + \Delta Y| > |\Delta Y - \Delta X|$, from which the inequality derives. □

COROLLARY 1. *Proposition 2 holds for divergent expectations.*

Proof: Set $U = |\Delta X|$ and $V = |\Delta Y|$; $\Lambda = \mathrm{Max}(U, V)$, $\lambda = \mathrm{Min}(U, V)$. Then you can rewrite the expectation as:

$$\begin{aligned}
&\mathbb{E}[UV\{1/(U+V) - 1/(|U-V|)\}] \\
&\qquad = \mathbb{E}[\lambda\Lambda\{-2\lambda/(\Lambda^2 - \lambda^2)\}] \\
&\qquad = -2\mathbb{E}[\Lambda\lambda^2/(\Lambda^2 - \lambda^2)] \leq 0
\end{aligned}$$

Indeed, if the expectation is divergent, then it is always $-\infty$. This approach removes the need to make the argument that $U + V > |U - V|$ and avoids the need to take a difference of expectations. An alternative approach is given in an expanded statement and proof of Proposition 2 in the SM. □

With epistasis

We now leverage the aforementioned fact that fitness additivity is not required here and we employ the general definition of z, namely $z = \phi(x, y)$ which we insert into Eq (9), leading to the following corollary to Proposition 2:

COROLLARY 2. *For any real number ξ, let us consider a fitness function of the form $\phi_\xi(x, y) = aX + bY + \xi g(X, Y)$, where $a, b > 0$ and g is a function independent of ξ. Let $\Delta Z(\xi) = \phi_\xi(X_2, Y_2) - \phi_\xi(X_1, Y_1)$. Assume that for some $\varepsilon > 0$,*

$$\mathbb{E}\left[\sup_{|\xi|<\varepsilon} \frac{|\Delta X \Delta Y|}{|\Delta Z(\xi)|}\right] < \infty, \tag{10}$$

and that $\mathbb{P}(\Delta X \Delta Y = 0) < 1$. Then, there is $\varepsilon_0 \in (0, \varepsilon)$, such that for all $\xi \in (-\varepsilon_0, \varepsilon_0)$, we have

$$\mathbb{E}\left[\frac{\Delta X \Delta Y}{|\Delta Z(\xi)|}\right] < 0. \tag{11}$$

Proof. Condition (10) implies that the function $h : (-\varepsilon, \varepsilon) \to \mathbb{R}$ defined via

$$h(\xi) = \mathbb{E}\left[\frac{\Delta X \Delta Y}{|\Delta Z(\xi)|}\right]$$

is continuous. Moreover, since $\mathbb{P}(\Delta X \Delta Y = 0) < 1$, proceeding as in the proof of Proposition 2, we obtain that $h(0) < 0$. Hence, by continuity of h, we infer that there is $\varepsilon_0 \in (0, \varepsilon)$ such that h is negative in $(-\varepsilon_0, \varepsilon_0)$, which concludes the proof. □

Figure 2 plots the left-hand side of the inequality in Eq (11) with generalized fitness function:

$$Z(\xi) = \phi_\xi(X, Y) = aX + bY + \xi XY \tag{12}$$

with $a, b > 0$ and $\xi \in \mathbb{R}$ is an epistasis parameter. This figure reveals where the interval $(-\varepsilon_0, \varepsilon_0)$ lies for different correlation coefficients. The predicted symmetry of this interval about zero is corroborated with both Montecarlo expectations of the left-hand side of Eq (11) as well as fully-stochastic evolutionary simulations.

We now turn our attention to the analysis of time-integrated covariance with epistasis for the special case where total fitness Z is given by Eq (12). As before, we let $\Delta Z(\xi) = \phi_\xi(X_2, Y_2) - \phi_\xi(X_1, Y_1) = (a + \xi Y_1)\Delta X + (b + \xi X_2)\Delta Y$. The case where the random variables $(|\Delta X \Delta Y|/|\Delta Z(\xi)|)_{\xi \in (-\varepsilon, \varepsilon)}$ are uniformly integrable (i.e. condition (10) is satisfied) is covered already by Corollary 2. If it is not uniformly integrable, we have the following:

COROLLARY 3. *Assume that the distribution of (X_i, Y_i) has finite support, i.e. there is $K > 0$ such that $\mathbb{P}(X_i \in [-K, K], Y_i \in [-K, K]) = 1$ and that $|\xi| < (a \wedge b)/K$, where $a \wedge b$ denotes the minimum between a and b. The we have:*

$$\mathbb{E}\left[\frac{|\Delta X \Delta Y|}{|\Delta Z(\xi)|}\right] = \infty \quad \Rightarrow \quad \mathbb{E}\left[\frac{\Delta X \Delta Y}{|\Delta Z(\xi)|}\right] = -\infty. \tag{13}$$

The proof for this corollary is in the SM.

V. NATURAL SELECTION MAKES RECOMBINATION ADVANTAGEOUS

Evolution with recombination

Let $u(x, \cdot)$ denote the marginal density of individuals carrying genic fitness x at the first locus, and let $u(\cdot, y)$ denote the marginal density of individuals carrying genic fitness y and the second locus. Independent evolution at the two loci means that $u(x, y) = u(x, \cdot)u(\cdot, y)$. Let

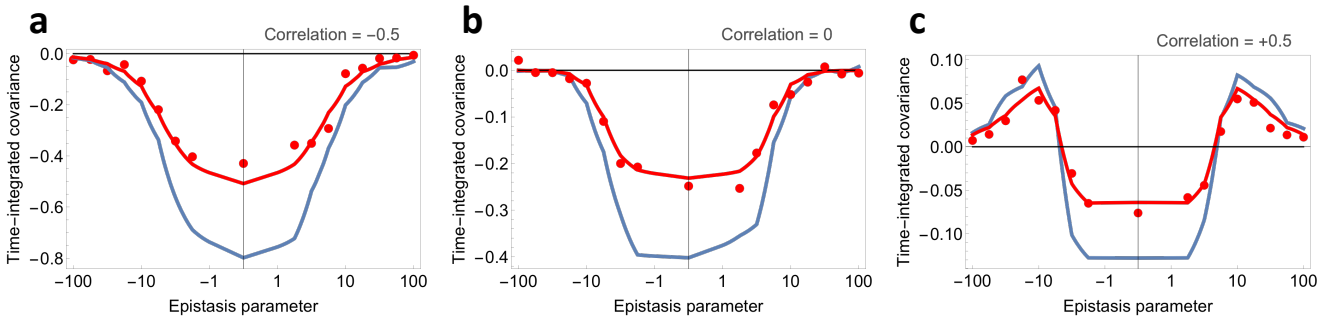

FIG. 2. Effect of epistasis ξ (horizontal axes) on time-integrated covariance (vertical axes) when correlation between genic fitnesses in the initial population is: **a)** -0.5, **b)** 0, and **c)** $+0.5$. Theoretical predictions (solid curves) plot Montecarlo expectations $\mathbb{E}\left[\Delta X \Delta Y / |\Delta Z(\xi)|\right]$, the left-hand side of Eq (11), where $N = 2000$ (red) and $N = 10^5$ (blue). Red points plot the means of 500 stochastic simulations with $N = 2000$. For these plots, we employ the general fitness function given by Eq (12). These plots support the validity of our conjecture that Z in Eq (9) can be generalized by allowing $Z = \phi(X, Y)$. They also corroborate our analyses showing that time-integrated covariance is unconditionally negative in the absence of epistasis. They further corroborate our finding that time-integrated covariance is negative in an epistasis interval that is symmetric about zero. Finally, they indicate that epistasis must be fairly strong (either positive or negative) to make time-integrated covariance non-negative.

$\bar{x}$ denote the mean of x ($\bar{x} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xu(x,y)dxdy$). Under recombination, we have the two loci evolving independently as follows:

$$\partial_t u(x,\cdot) = (x - \bar{x})u(x,\cdot) \quad (14)$$
$$\partial_t u(\cdot,y) = (y - \bar{y})u(\cdot,y) \quad (15)$$

In the foregoing sections, we have shown that a *mgf*-transformed version of the above evolution equations has a closed-form analytical solution. Letting $\mathcal{C}_t(\varphi,\theta) = \ln \mathbb{E}_t[e^{\varphi X + \theta Y}]$, the cumulant-generating functions (*cgf*) for density $u(x,y)$ at time t. The solution to the *mgf* transform of equations (14) and (15) gives:

$$\mathcal{C}_t(\varphi,0) = \mathcal{C}_0(\varphi + t, 0) - \mathcal{C}_0(t,0)$$
$$\mathcal{C}_t(0,\theta) = \mathcal{C}_0(0, \theta + t) - \mathcal{C}_0(0,t)$$

under free recombination. Mean fitness of the rec^+ subpopulation at time t is therefore:

$$\bar{z}_r(t) = \mathcal{C}_t^{(1,0)}(0,0) + \mathcal{C}_t^{(0,1)}(0,0) = \mathcal{C}_0^{(1,0)}(t,0) + \mathcal{C}_0^{(0,1)}(0,t)\,. \quad (16)$$

Evolution without recombination

For the non-recombining wildtype subpopulation, the evolution equation has already been presented and analyzed above. The transformed equation has solution:

$$\mathcal{C}_t(\varphi,\theta) = \mathcal{C}_0(\varphi + t, \theta + t) - \mathcal{C}_0(t,t)\,.$$

Mean fitness at time t is now:

$$\bar{z}(t) = \mathcal{C}_t^{(1,0)}(0,0) + \mathcal{C}_t^{(0,1)}(0,0) = \mathcal{C}_0^{(1,0)}(t,t) + \mathcal{C}_0^{(0,1)}(t,t) \quad (17)$$

Evolution of fitness differential between rec^+ modifier and wildtype

The fitness differential between the rec^+ modifier and wildtype at time t is:

$$\begin{aligned} s_r(t) &= \bar{z}_r(t) - \bar{z}(t) \\ &= \mathcal{C}_0^{(1,0)}(t,0) + \mathcal{C}_0^{(0,1)}(0,t) - \mathcal{C}_0^{(1,0)}(t,t) - \mathcal{C}_0^{(0,1)}(t,t) \end{aligned} \quad (18)$$

where $\bar{z}_r(t)$ and $\bar{z}(t)$ are given by Eqs (16) and (17). This may be equivalently written as:

$$s_r(t) = \frac{d}{dt}\left(\mathcal{C}_0(t,0) + \mathcal{C}_0(0,t) - \mathcal{C}_0(t,t)\right) \quad (19)$$

or as

$$\begin{aligned} s_r(t) = \frac{d}{dt}(&\mathcal{C}_0(t,0,...,0) + \mathcal{C}_0(0,t,...,0) + \mathcal{C}_0(0,0,...,t) \\ &- \mathcal{C}_0(t,t,...,t)) \end{aligned}$$

for more than two loci.

We note that future dynamics of the fitness differential is predicted based simply on the distribution of fitnesses in the initial variation upon which natural selection acts, i.e., $s_r(t)$ depends only on $\mathcal{C}_0(\varphi,\theta)$, the *cgf* of the initial fitness distribution.

Asymptotic fitness differential between rec^+ modifier and wildtype

PROPOSITION 3. *A population initially consists of n distinct genotypes at equal frequency (this assumption is relaxed later) characterized by their genic fitness vector $(x_{i1}, x_{i2}, ..., x_{im})$, $i = 1, 2, ..., n$. These values may be*

drawn from any multivariate distribution, continuous or not. The action of natural selection by itself will cause the fitness of a rec$^+$ *modifier to increase relative to its non-recombining counterpart by an amount given by:*

$$s_r(t) \xrightarrow{t\to\infty} \sum_{j=1}^{m} \max_i(x_{ij}) - \max_i(\sum_{j=1}^{m} x_{ij}) \;\geq\; 0 \qquad (20)$$

Proof: We employ the fitness differential as defined by Eq (18) and insert empirical *cgf*'s as defined by Eq (6), giving:

$$s_r(t) \;=\; \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} x_{ij} e^{x_{ij}t}}{\sum_{i=1}^{n} e^{x_{ij}t}} - \frac{\sum_{i=1}^{n}(\sum_{j=1}^{m} x_{ij}) e^{\sum_{j=1}^{m} x_{ij}t}}{\sum_{i=1}^{n} e^{\sum_{j=1}^{m} x_{ij}t}}$$

which gives the result by inspection. Inspection is facilitated by examining the case of two loci whose genic fitnesses we deonte x_i and y_i:

$$s_r(t) \;=\; \frac{\sum_{i=1}^{n} x_i e^{x_i t}}{\sum_{i=1}^{n} e^{x_i t}} + \frac{\sum_{i=1}^{n} y_i e^{y_i t}}{\sum_{i=1}^{n} e^{y_i t}} - \frac{\sum_{i=1}^{n}(x_i+y_i) e^{(x_i+y_i)t}}{\sum_{i=1}^{n} e^{(x_i+y_i)t}} \xrightarrow{t\to\infty} \max_i(x_i) + \max_i(y_i) - \max_i(x_i+y_i) \;\geq\; 0$$

□

Corollary 4. *For the case of two genotypes and two loci ($n = 2$ and $m = 2$) the asymptotic fitness differential given by Proposition 3 can be rewritten as:*

$$2s_r(t) \xrightarrow{t\to\infty} |\Delta x| + |\Delta y| - |\Delta x + \Delta y| \;\geq\; 0$$

where $\Delta x = x_2 - x_1$ and $\Delta y = y_2 - y_1$.

Remark 1: The couples (x_1, y_1) and (x_2, y_2) are two independent draws from some unspecified bivariate distribution. This fact guarantees that Δx and Δy are symmetric, from which it is apparent that the asymptotic fitness differential will be zero half of the time.

Corollary 5. *We now define random variable S_r to denote the asymptotic fitness differential as defined above. Here we generalize Proposition 3 to m loci. Here, fitness at each of the m loci is given by random variable X_j $j = 1, 2, ..., m$. Expected asymptotic fitness differential is:*

$$\mathbb{E}[S_r] \;=\; \sum_{j=1}^{m} \mathbb{E}[X_j^{[n]}] - \mathbb{E}[Z^{[n]}] \;\geq\; 0$$

where $X_j^{[n]}$ denotes the nth order statistic (i.e., maxima) of X_j, j denotes the jth locus, and $Z^{[n]}$ denotes the nth order statistic (i.e., maxima) of $Z = \sum_{j=1}^{m} X_j$.

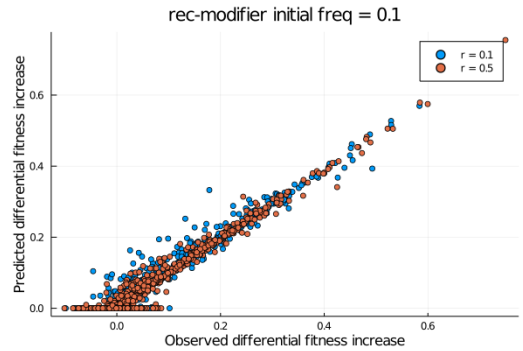

FIG. 3. Fitness differential between *rec*$^+$ modifier and wild-type after one bout of selection: observed (horizontal axis) and predicted (vertical axis). Initial modifier frequency was 0.1, and plotted are values for recombination rates of $r = 0.1$ (blue) and $r = 0.5$ (red). Observed values come from fully-stochastic simulations. Predicted values are computed using Eq (20)

The foregoing expressions very accurately predict the fitness differential after one bout of selection in simulations (Fig 3), and is surprisingly robust to recombination rate and initial *rec*$^+$ frequency.

Remark 2: We have shown that, after one "bout" of selection has completed, the *rec*$^+$ modifier fitness advantage is non-negative. This is indeed suggestive of natural selection's favorable effect on recombination, but it only gives information about the modifier's fitness advantage at the *end* of the bout of selection. It does not guarantee, for example, that the modifier's fitness advantage did not become negative over the course of the bout of selection and consequently suppress the modifier's frequency in the process. This concern is especially relevant in light of our observation in Remark 1 above that, for the case of two loci and two alleles, the modifier's fitness advantage after the bout of selection has completed is zero half of the time.

VI. NATURAL SELECTION PROMOTES THE EVOLUTION OF RECOMBINATION

To understand how natural selection affects the *evolution* of recombination, the more directly relevant question to ask is how natural selection affects the *frequency* of a *rec*$^+$ modifier.

Dynamics of a *rec*$^+$ modifier

If a lineage has time-dependent selective advantage $s(t)$, its frequency $\rho(t)$ evolves as the solution to the logistic equation:

$$\rho'(t) \;=\; s(t)\rho(t)(1-\rho(t))$$

which, with initial condition $\rho(t) = \rho_0$, has solution:

$$\rho(t) \;=\; \left(1 + (\rho_0^{-1} - 1)e^{-\int_0^t s(u)du}\right)^{-1} \tag{21}$$

Our time-dependent selection coefficient is of course $s(t) = s_r(t)$ which is worked out above. It remains to find:

$$\int_0^t s_r(u)du$$

Conveniently, we immediately have the anti-derivative of $s_r(t)$ from Eq (19) from which we have the definite integral:

$$\int_0^t s_r(u)du \;=\; \mathcal{C}_0(t,0) + \mathcal{C}_0(0,t) - \mathcal{C}_0(t,t) \tag{22}$$

Recalling that $\mathcal{C}_t(\theta,\phi) \;=\; \ln M_t(\theta,\phi)$, and substituting Eq (22) into Eq (21), the expression for modifier dynamics becomes:

$$\rho(t) \;=\; \left(1 + (\rho_0^{-1} - 1)\frac{M_0(t,t)}{M_0(0,t)M_0(t,0)}\right)^{-1}$$

This expression gives the generalized case for two loci and holds for any number n of alleles per locus.

The further generalization to m loci and n alleles per locus is immediate:

$$\rho(t) \;=\; \left(1 + (\rho_0^{-1} - 1)\frac{M_0(t,t,...,t)}{M_0(t,0,...,0)M_0(0,t,...,0)...M_0(0,0,...,t)}\right)^{-1} \tag{23}$$

Infinitely-many alleles

The purpose of this subsection is primarily to reveal what appears to be a fundamental qualitative difference between the case of finitely-many vs infinitely-many alleles. Assuming there are infinitely-many alleles, and supposing that the distribution of X and Y is bivariate normal, we have:

$$\rho(t) \;=\; \left(1 + (\rho_0^{-1} - 1)e^{\sigma_{XY}t^2}\right)^{-1} \tag{24}$$

so modifier dynamics depend critically on the sign of the covariance of the initial fitnesses: when $\sigma_{XY} < 0$, then $\rho(t) \xrightarrow{t\to\infty} 1$; when $\sigma_{XY} > 0$, then $\rho(t) \xrightarrow{t\to\infty} 0$; and when $\sigma_{XY} = 0$, then $\rho(t) = f(0)\ \forall\, t$. This finding stands in stark contrast to the case of finitely-many alleles in which, as we shall see, modifier frequency for all practical purposes always ends up at higher frequency than where it started, quite independently of the bivariate distribution governing X and Y (even when the initial population has strongly positive correlation between X and Y).

Remark: In general, Eq (24) may be written for any distribution:

$$\rho(t) \;=\; \left(1 + (\rho_0^{-1} - 1)e^{h(t)}\right)^{-1} \tag{25}$$

where:

$$h(t) \;=\; \sum_{i=2}^{\infty}\frac{t^i}{i!}\sum_{j=1}^{i-1}\binom{i}{j}\kappa_{j,i-j}$$

and $\kappa_{i,j}$ is the $(i,j)^{th}$ cumulant of the initial bivariate distribution of genic fitnesses X and Y. This expression is obtained by Taylor expansion of Eq (22). The normal case can be gleaned from this general expression by recalling that for the normal distribution, $\kappa_{i,j} = 0$ when $i + j > 2$, so that $h(t) = \kappa_{1,1}t^2 = \sigma_{XY}t^2$.

Finitely-many alleles

If the number of alleles is finite, we employ the empirical *cgf*, $\tilde{\mathcal{C}}_t(\theta,\phi)$, as defined by Eq (6), and the empirical *mgf*, $\tilde{M}_t(\theta,\phi) = e^{\tilde{\mathcal{C}}_t(\theta,\phi)}$. Assuming n genotypes are present in the population in question, and replacing $\mathcal{C}$ with $\tilde{\mathcal{C}}$ in Eq (22), we have:

$$\int_0^t s_r(u)du \;=\; \sum_{j=1}^{m}\ln\left[\sum_{i=1}^{n} f_i e^{x_{ij}t}\right] - \ln\left[\sum_{i=1}^{n} f_i e^{\sum_{j=1}^{m} x_{ij}t}\right]\;. \tag{26}$$

Assuming for now that $f_i = 1/n$, the expected frequency of the modifier is:

$$\mathbb{E}[\rho(t)] \;=\; \mathbb{E}\left[\left(1 + n^{m-1}(\rho_0^{-1}-1)\frac{\sum_{i=1}^{n} e^{\sum_{j=1}^{m} X_{ij}t}}{\prod_{j=1}^{m}(\sum_{i=1}^{n} e^{X_{ij}t})}\right)^{-1}\right] \quad (27)$$

The notation in this expression is already messy and would only get messier were we to proceed with the general m-locus case. We therefore restrict ourselves to the case of $m = 2$ loci for the sake of presentation. The general m-locus case is a rather trivial (albeit messy) extension of these developments. For the two-locus case, Eq (27) becomes:

$$\mathbb{E}[\rho(t)] \;=\; \mathbb{E}\left[\left(1 + n(\rho_0^{-1}-1)\frac{\sum_{i=1}^{n} e^{(X_i+Y_i)t}}{(\sum_{i=1}^{n} e^{X_i t})(\sum_{i=1}^{n} e^{Y_i t})}\right)^{-1}\right] \quad (28)$$

which can be rewritten as:

$$\mathbb{E}[\rho(t)] \;=\; \mathbb{E}\left[\frac{\sum_{i,j} e^{(X_i+Y_j)t}}{\sum_{i\neq j} e^{(X_i+Y_j)t} + (n(\rho_0^{-1}-1)+1)\sum_i e^{(X_i+Y_i)t}}\right] \quad (29)$$

Equation (28) accurately predicts rec^+ modifier dynamics as show in Fig 4.

We note that, for the case $n = 2$, Eq (28) has the curious alternative form:

$$\mathbb{E}[\rho(t)] \;=\; \mathbb{E}\left[\left(\rho_0^{-1} + (\rho_0^{-1}-1)\tanh(\Delta X t/2)\tanh(\Delta Y t/2)\right)^{-1}\right]$$

where $\Delta X = X_2 - X_1$ and $\Delta Y = Y_2 - Y_1$.

Asymptotic modifier frequency

Proposition 4. *A non-recombining population initially consists of n distinct genotypes. The ith genotype is characterized by the random vector of genic fitness $(X_{i1}, X_{i2}, ..., X_{im})$, where m is the number of loci under selection. This random vector may have any multivariate distribution, continuous or not. A* rec$^+$ *modifier is introduced into the population at frequency $\rho(0)$. The action of natural selection by itself will cause the frequency of the modifier to converge in expectation to:*

$$\mathbb{E}[\rho(t)] \xrightarrow{t\to\infty} \mathbb{E}[\mathbb{1}_{NR}/(n^{(m-1)}(\rho_0^{-1}-1)+1) + \mathbb{1}_R] \quad (30)$$

where conditions NR and R are met when the maximum-fitness genotype is a non-recombinant and recombinant, respectively. Specifically, NR is met when the maximum-fitness genotype has the following property: for any pair of genic fitnesses (X_{ij}, X_{kl}), it is the case that $i = k$. Condition R is met when NR is not true.

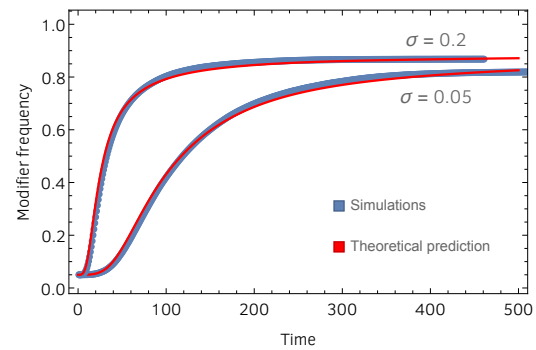

FIG. 4. Recombination modifier dynamics over the course of one bout of selection. Red curves plot theoretical predictions given by Eq (28). Blue curves plot mean trajectory observed in 500 replicate simulations. Simulations were fully stochastic, individual-based, with a population size of $N = 20,000$. The recombination-competent modifier conferred a recombination rate of $r = 0.2$. The distribution of genic fitnesses X and Y in the initial population had a bivariate normal distribution with zero means, standard deviations $\sigma_X = \sigma_Y = \sigma = 0.2$ for the upper curves and $\sigma_X = \sigma_Y = \sigma = 0.05$ for the lower curves, and zero correlation. The initial population consisted of $n = 10$ distinct genotypes. We note that upper and lower curves eventually converge to the same asymptotic frequency, as our theory predicts, despite very different strengths of selection (very different σ's).

Proof:
The proof is by inspection of Eq (29) and its full m-locus extrapolation. (Inspection is facilitated by considering the case $m = 2$.)

□

From here, it's easy to see that *in theory* the modifier can decrease in frequency to below its initial frequency. This happens under the worst-case scenario for the modifier, which is when the correlation coefficient becomes extremely close to $+1$. When the correlation is exactly equal to $+1$, we have:

$$\mathbb{E}[\rho(t)] \xrightarrow{t\to\infty} 1/(n^{(m-1)}(\rho_0^{-1}-1)+1) \;<\; \rho(0) \quad (31)$$

where $n \geq 2$ is the number of genotypes in the initial population and m the number of loci constituting a genotype. Numerical solution of Eq (30), however, reveals that the correlation coefficient has to be unrealistically close to one for the modifier to decrease in frequency (Fig 5).

Corollary 6. *We generalize Proposition 4 by allowing each genotype i to have its own starting frequency f_i. Until now, we have assumed that $f_i = 1/n$. Expected asymptotic modifier frequency in this generalized case is:*

$$\mathbb{E}[\rho(t)] \xrightarrow{t\to\infty} \mathbb{E}[\mathbb{1}_{NR}\left(\frac{\mathcal{F}^{m-1}}{\rho_0^{-1} - 1 + \mathcal{F}^{m-1}}\right) + \mathbb{1}_R] \quad (32)$$

where conditions NR and R are as defined above, and random variable $\mathcal{F}$ is starting frequency (the f_i are instances of $\mathcal{F}$).

If $\mathcal{F}$ is exponential with mean $1/n$ and $m = 2$, Eqs (30) and (32) are, for all practical purposes, equivalent. For $m > 2$, we have found numerically that:

$$\mathbb{E}[\frac{\mathcal{F}^{m-1}}{\rho_0^{-1} - 1 + \mathcal{F}^{m-1}}] \ > \ 1/(n^{(m-1)}(\rho_0^{-1} - 1) + 1)$$

but the left-hand side is still very small, validating the following Corollary for equal or random starting frequencies.

COROLLARY 7. *If a* rec$^+$ *modifier is initially at low frequency in a population, it's final (asymptotic) frequency is well approximated by the probability that the maximum-fitness genotype is a virtual recombinant.*

More specifically, given that a population initially consists of n genotypes carrying random vectors of genic fitnesses $(X_{i1}, X_{i2}, ..., X_{im})$, $i = 1, 2, ..., n$, the final expected modifier frequency is effectively equal to the probability of condition R defined in Proposition 4 above.

Proof:

This corollary comes about by noting that the first term on the right-hand side of Eq (30) is typically much smaller than the second term:

$$\mathbb{E}[\mathbb{1}_{NR}/(n^{(m-1)}(\rho_0^{-1} - 1) + 1)] \ \ll \ \mathbb{E}[\mathbb{1}_R]$$

for the case of equal starting frequencies, or:

$$\mathbb{E}[\mathbb{1}_{NR}\left(\frac{\mathcal{F}^{m-1}}{\rho_0^{-1} - 1 + \mathcal{F}^{m-1}}\right)] \ \ll \ \mathbb{E}[\mathbb{1}_R]$$

for the case of random starting frequencies. This fact is corroborated by Montecarlo expectations plotted in Fig 5 where this approximation appears indistinguishable from the exact solution.

□

Corollary 7 allows us to say some things about non-extreme cases as well. When the correlation coefficient is zero, Corollary 7 together with simple combinatorics tell us the asymptotic frequency is:

$$\sigma_{XY} = 0 \quad \Rightarrow \quad \mathbb{E}[\rho(\infty)] \ = \ 1 - n^{-(m-1)} \qquad (33)$$

If the heritable variation upon which natural selection acts is itself a product of previous selection, our companion papers [1, 2] show that the genic fitness correlation between loci will be negative. For this case, asymptotic modifier frequency is even higher:

$$\sigma_{XY} < 0 \quad \Rightarrow \quad \mathbb{E}[\rho(\infty)] \ > \ 1 - n^{-(m-1)} \qquad (34)$$

from which it is apparent that asymptotic modifier frequency quickly gets very close to one as number of genotypes and number of loci increase. For example, two genotypes and ten loci will have an asymptotic modifier frequency greater than 0.998 when the correlation is negative. Significantly, Eqs (33) and (34) appear to be fairly robust to epistasis (SM).

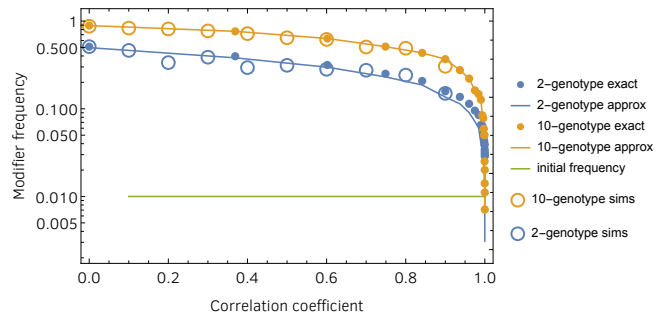

FIG. 5. Asymptotic modifier frequency as a function of positive correlation between genic fitnesses X and Y among the genotypes initially present in the population, as computed by Eq (30). Only when the correlation gets unrealistically close to one does the asymptotic modifier frequency dive to values that can, in theory, dip below the initial frequency. Asymptotic modifier frequency for negative correlations is not plotted; it becomes increasingly close to one as correlation becomes increasingly negative. Open circles plot final modifier frequency in simulated populations of size 2000 with recombination rate of the modifier of $r = 0.1$.

A surprising feature of the asymptotic modifier frequency is the absence of any requirement for information about the magnitude of selective differences among alleles at the different loci. This independence of strength of selection is illustrated in Fig. 4, where it is corroborated with simulations. An implication of this finding is that modifier frequency will converge to the same value even under very weak selection. This fact may speak to concerns, expressed in previous work [7, 37], about the strength of selection required for recombination (and sex) to evolve.

And perhaps even more surprising is the fact, shown in Fig 5, that increase in modifier frequency is substantial even when the correlation between X and Y is strongly positive in the initial population. It is only when this correlation gets unrealistically close to $+1$ that increase in modifier frequency is substantially reduced. This is very surprising because a strongly positive correlation between genic fitness is precisely the condition that one would expect to suppress, not favor, recombination (discussed in [1, 2]). The reason that modifier frequency increases despite positive correlation has to do with the dynamics of selective sorting, and the fact that these dynamics cause recombinants to be favored, on average (covariance to be negative on average), independently of the correlation between genic fitnesses X and Y in the initial population or any other feature of those fitnesses for that matter, as proved in Proposition 2 above.

Recombination and evolutionary dynamics

Let $u_t(x, y)$ denote probability density in fitness contributions x and y at time t for an evolving population.

Dropping the subscripts, under selection, u evolves as:

$$\partial_t u(x,y) = (x+y-\bar{x}-\bar{y})u(x,y) + R(u(x,\cdot)u(\cdot,y) - u(x,y))$$

where R is recombination rate. The transformed equation is:

$$\partial_t \mathcal{C}(\varphi,\theta) = \partial_\varphi \mathcal{C}(\varphi,\theta) + \partial_\theta \mathcal{C}(\varphi,\theta) - \partial_\varphi \mathcal{C}(0,0) - \partial_\theta \mathcal{C}(0,0) + R(e^{\mathcal{C}(\varphi,0)+\mathcal{C}(0,\theta)-\mathcal{C}(\varphi,\theta)} - 1). \quad (35)$$

whose solution is given by the $\mathcal{C}_t(\varphi,\theta)$ which satisfies:

$$\mathcal{C}_t(\varphi,\theta) = \mathcal{C}_0(\varphi+t,\theta+t) - \mathcal{C}_0(t,t) + R\int_0^t (e^{\mathcal{C}_s(\varphi,0)+\mathcal{C}_s(0,\theta)-\mathcal{C}_s(\varphi,\theta)} - 1)ds, \quad (36)$$

and boundary condition, $\mathcal{C}_t(0,0) = 0 \ \forall \ t$. This equation can be solved iteratively for $\mathcal{C}_t(\varphi,\theta)$.

These developments lead to a variant of Proposition 1:

PROPOSITION 5. *A first iteration of Eq* (36) *yields a modification of the time-integrated within-population covariance given in Proposition 1:*

$$\int_0^\infty \sigma_{XY}(t)dt \ \approx \ q(1-R)\frac{(x_2-x_1)(y_2-y_1)}{|z_2-z_1|} \quad (37)$$

where q is the initial frequency of the inferior genotype, and continuous parameter $R \in [0,1]$ is recombination rate.

Proof: For the first iteration, we simply replace the $\mathcal{C}_s(\varphi,\theta)$ in the exponent by $\mathcal{C}_0(\varphi+s,\theta+s) - \mathcal{C}_0(s,s)$, giving:

$$\mathcal{C}_t(\varphi,\theta) = \mathcal{C}_0(\varphi+t,\theta+t) - \mathcal{C}_0(t,t) + R\int_0^t (e^{\mathcal{C}_0(\varphi+s,s)+\mathcal{C}_0(s,\theta+s)-\mathcal{C}_0(\varphi+s,\theta+s)-\mathcal{C}_0(s,s)} - 1)ds. \quad (38)$$

which satisfies the boundary condition, $\mathcal{C}_t(0,0) = 0 \ \forall \ t$. Then,

$$\int_0^\infty \sigma_{XY}(t)dt \ = \ \int_0^\infty \tilde{\mathcal{C}}_0^{(1,1)}(t,t)dt \ = \ q(1-R)\frac{(x_2-x_1)(y_2-y_1)}{|z_2-z_1|}$$

□

If this approximation is accurate, it follows that Proposition 2 also holds for already-recombining populations, but the magnitude of the negative time-integrated covariance (i.e., the magnitude of the average recombinant advantage) decreases as recombination rate increases. This is evidenced by the new factor $(1-R)$ introduced here. However, we do not state this formally, as we have not conducted a thorough analysis and/or exploration of parameter space to determine the accuracy of this approximation.

From Eq (35), the role of recombination in the evolution of total mean fitness, $\bar{z} = \bar{x} + \bar{y}$, is elucidated by the derivative expressions:

$$\begin{aligned} \partial_t \bar{x} &= \sigma_X^2 + \sigma_{XY} \\ \partial_t \bar{y} &= \sigma_Y^2 + \sigma_{XY} \\ \partial_t \sigma_{XY} &= \kappa_{12} + \kappa_{21} - R\sigma_{XY} \end{aligned}$$

In the absence of selection, the κ's will be zero, giving the prediction that $\sigma_{XY}(t) = \sigma_{XY}(0)e^{-Rt}$ under neutral evolution. This equation accurately predicts covariance dynamics in simulations of neutral evolution (Fig 6). With selection, the κ's will be non-zero and covariance dynamics are given by:

$$\sigma_{XY}(t) = \sigma_{XY}(0)e^{-Rt} \ + \ \int_0^t e^{R(\gamma-t)}(\kappa_{12}(\gamma)+\kappa_{21}(\gamma))d\gamma \quad (39)$$

which also shows good agreement with simulations (Fig 7).

Effect of increasing recombination rate in an already recombining population

Until now, our focus has been on rec^+ modifiers in an otherwise rec^- population. We now examine the case in which the resident population is rec^+ and the modifier introduced has a higher recombination rate than the resident population. The evolutionary dynamics model developed above allows us to address this question. Specifically, we ask how a further increase in recombination rate increases the rate of increase in mean fitness. Mean fitness increases as $\partial_t \bar{z} = \partial_t \bar{x} + \partial_t \bar{y} = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$, which increases in recombination rate as $\partial_R \partial_t \bar{z} \ = \ 2\partial_R \partial_t \sigma_{XY}$, where σ_{XY} is given by Eq (39). This quantity is positive, as seen in Fig 8.

VII. STATISTICAL MECHANICS OF SEX

We now note a curious connection between the developments presented here and statistical mechanics. This section may serve as a springboard for further work. We believe there is no existing analogue to multilocus evolution in statistical mechanics, in which case these directions may bring something new to both fields.

Statistical mechanics of single-locus evolution

Based on our developments and our use of the empirical *cgf* (Eqs (5) and (6)), when there is only one locus under selection, the expected number of individuals with genotype i at time t may be written as:

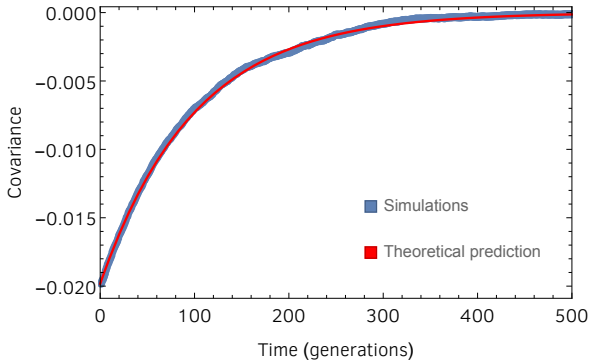

FIG. 6. Covariance dynamics under neutral evolution with recombination. Blue dots plot the average of 200 fully stochastic simulations. The red curve plots our theoretical prediction, $\sigma_{XY}(t) = \sigma_{XY}(0)e^{-Rt}$, which derives from Eq (35). Parameters are: $N = 10,000$, $R = 0.01$.

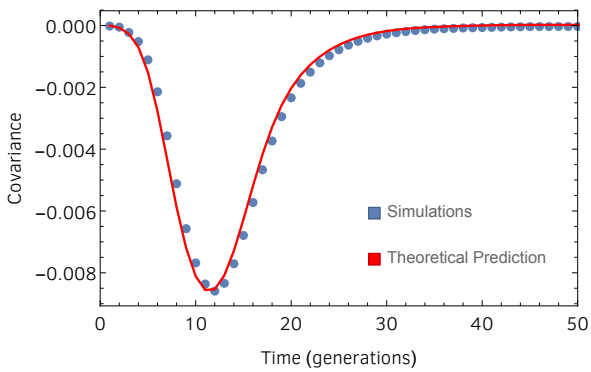

FIG. 7. Covariance dynamics under adaptive evolution with recombination. Blue dots plot the average of 200 fully stochastic simulations. The red curve plots the theoretical prediction given by Eq (39). Parameters are: $N = 10,000$, $R = 0.01$.

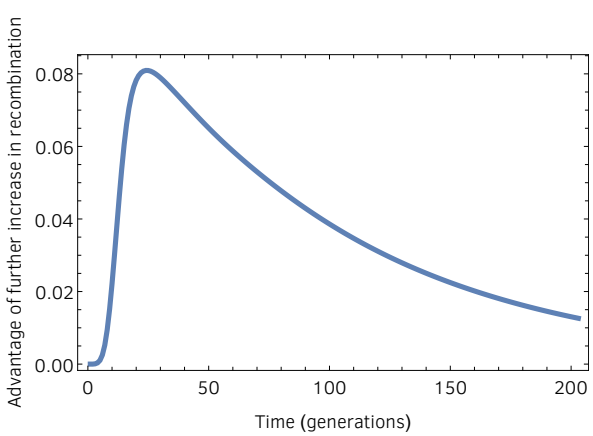

FIG. 8. Advantage of further increase in recombination rate in an already-recombining population. This is not selective advantage as it is commonly defined. Instead, it is how fast the rate of fitness increase grows with increase in recombination rate, calculated at the resident recombination rate. Specifically, it is the quantity $\partial_R \partial_t \bar{z}|_{R=R_0}$, where R_0 is the resident recombination rate. Parameters are: $N = 10,000$, $R_0 = 0.01$.

$$\langle N_i \rangle \;=\; \frac{N}{\mathcal{Z}} f_i e^{t x_i}$$

where N = total population size, f_i is the initial frequency of genotype i, x_i is the fitness of genotype i, and:

$$\mathcal{Z} = \sum_{i=1}^{n} f_i e^{t x_i} \;.$$

This expression, derived directly from our initial evolution equation, makes the connection between Darwinian evolution and statistical mechanics explicit. We compare these expressions with the Maxwell-Boltzmann equation for number of particles in energy state ε_i:

$$\langle N_i \rangle \;=\; \frac{N}{\mathcal{Z}} g_i e^{-\varepsilon_i/kT}$$

where

$$\mathcal{Z} = \sum_{i=1}^{n} g_i e^{-\varepsilon_i/kT} \;,$$

the partition function.

The analogous quantities are shown in Table I. Fitness is analogous to minus the energy. This makes sense, because fitness will tend to increase while energy will tend to decrease. Time is analogous to the inverse temperature. So the evolutionary asymptote as time goes to infinity is analogous to decreasing the temperature to absolute zero.

TABLE I. A side-by-side comparison of analogous quantities in evolutionary dynamics and statistical mechanics.

Evolutionary Dynamics	Statistical Mechanics
fitness, x_i	energy, $-\varepsilon_i$
time, t	inverse temperature, $1/kT$
$t \to \infty$	$T \to 0$

Statistical mechanics of multilocus evolution

Now we extrapolate to multilocus evolution, that is, evolution in which individuals have two or more loci under selection. Here, each genome has m loci, each of which contributes to overall fitness ϕ. Specifically, genotype i has fitness $\phi(\mathbf{x}_i)$, where vector $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{im})$ quantifies fitness contributions from each of the m loci. The expected number of individuals with fitness $\phi(\mathbf{x}_i)$ at time t is now:

$$\langle N_i \rangle \;=\; \frac{N}{\mathcal{Z}} f_i e^{t\phi(\mathbf{x}_i)}$$

where the partition function is now:

$$\mathcal{Z} = \sum_{i=1}^{n} f_i e^{t\phi(\mathbf{x}_i)}$$

We have found that, as time increases, a population will tend to evolve negative associations among the fitness contributions at the different loci. These negative associations build up across populations [2] and are generated within a single population, in expectation, as time passes. The analogy to statistical mechanics would be that each energy level has some number m of contributing factors that determine the energy of that level. As temperature is reduced, negative associations among the contributing factors will be generated. For the system to achieve the lowest possible total energy, the contributing factors would somehow need to be shuffled (analogous to recombination).

VIII. DISCUSSION

A note about epistasis

Epistasis is non-additivity in genic contributions to fitness, and negative epistasis is concave non-additivity, such that total fitness is always less than the sum of genic fitnesses. A population at equilibrium will harbor negative LD if its constituents exibit negative epistasis across loci. It was once though that the source of selective imbalance with the strongest causal link to the evolution of sex and recombination was negative, or synergistic, epistasis. However, under negative epistasis, recombinants are initially selectively *suppressed* on average, because recombination breaks up well-matched alleles across loci (it increases diffusion on a concave surface). This initial suppression of recombinants is eventually reversed, and recombinants are eventually amplified by selection, on average, owing to the fact that their fitnesses come from a distribution with larger variance [38]. Whether or not recombinants are ultimately successful therefore depends on the dynamics of recombinant fitness and whether or not the selective reversal is quick enough to rescue and amplify recombinants. Barton [15] identified a small interval of weakly negative epistasis, below zero and not containing it, for which the initial suppression of recombinants was small enough and the selective reversal quick enough to make the recombinants successful on average [3]. As there is no compelling evidence to suspect that epistasis in nature tends to fall within this "goldilocks zone" (or tends to show any general bias away from from zero) [3, 39, 40], epistasis-based explanations for the evolution of sex fell out of favor.

We examine epistasis in the context of our work. We find that, under natural selection, there is an interval for epistasis outside of which the evolution of recombination would not be favored, but: 1) this interval can be much bigger than that identified under the equilibrium / negative epistasis explanatory framework and, more importantly, 2) it always contains zero (Fig 2). Point 2 is especially relevant because it means that epistasis is not necessary for the evolution of recombination where natural selection is acting. Point 1 suggests that even if a persistent bias in epistasis is demonstrated across the tree of life, it would not invalidate our theory, as long as the bias is not too large.

Future directions

The previous section making a connection to statistical physics are starting points for two areas we feel merit fuller treatment. Additionally we would like to explore the use of approximations afforded by *Haldane linearization* [41, 42] as well as exact probabilistic treatments of recombination in multilocus systems, primarily derived by Baake & Baake [43–46], to address different variants of the questions posed here.

Acknowledgements Much of this work was performed during a CNRS-funded visit (P.G.) to the Laboratoire Jean Kuntzmann, University of Grenoble Alpes, France, and during a visit to Bielefeld University (P.G.) funded by Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) via Priority Programme SPP 1590 Probabilistic Structures in Evolution, grants BA 2469/5-2 and WA 967/4-2. P.G. and A.C. received financial support from the USA/Brazil Fulbright scholar program. P.G. and P.S. received financial support from National Aeronautics and Space Administration grant NNA15BB04A. The authors thank S. Otto and N. Barton for their thoughts on early stages of this work. Special thanks go to E. Baake for her thoughts on later stages of this work and help with key mathematical aspects. The authors thank D. Chencha, J. Streelman, R. Rosenzweig and the Biology Department at Georgia Institute of Technology for critical infrastructure and computational support.

[1] P. J. Gerrish, B. Galeota-Sprung, F. Cordero, P. Sniegowski, A. Colato, N. Hengartner, V. Vejalla, J. Chevallier, and B. Ycart, Natural selection and the advantage of recombination, Phys. Rev. Lett. **In Review** (2021).

[2] P. J. Gerrish, B. Galeota-Sprung, P. Sniegowski, J. Chevallier, and B. Ycart, Natural selection promotes the evolution of recombination 1: among selected genotypes, Physical Review E **In Review** (2021).

[3] S. P. Otto and T. Lenormand, enResolving the paradox

of sex and recombination, Nat. Rev. Genet. **3**, 252 (2002).
[4] N. H. Barton and B. Charlesworth, enWhy sex and recombination?, Science **281**, 1986 (1998).
[5] S. P. Otto, enSelective interference and the evolution of sex, J. Hered. **112**, 9 (2021).
[6] S. p. Otto, The evolutionary enigma of sex, Am. Nat. **174**, S1 (2009).
[7] J. A. G. M. de Visser and S. F. Elena, enThe evolution of sex: empirical insights into the roles of epistasis and drift, Nat. Rev. Genet. **8**, 139 (2007).
[8] M. Hartfield and P. D. Keightley, enCurrent hypotheses for the evolution of sex and recombination, Integr. Zool. **7**, 192 (2012).
[9] S. C. Lee, M. Ni, W. Li, C. Shertz, and J. Heitman, enThe evolution of sex: a perspective from the fungal kingdom, Microbiol. Mol. Biol. Rev. **74**, 298 (2010).
[10] R. A. Fisher, *The genetical theory of natural selection* (Oxford Clarendon Press, 1930) p. 302.
[11] H. J. Muller, Some genetic aspects of sex, Am. Nat. **66**, 118 (1932).
[12] D. Roze and N. H. Barton, enThe Hill-Robertson effect and the evolution of recombination, Genetics **173**, 1793 (2006).
[13] W. G. Hill and A. Robertson, enThe effect of linkage on limits to artificial selection, Genet. Res. **8**, 269 (1966).
[14] N. H. Barton, enLinkage and the limits to natural selection, Genetics **140**, 821 (1995).
[15] N. H. Barton, enA general model for the evolution of recombination, Genet. Res. **65**, 123 (1995).
[16] N. H. Barton, enGenetic linkage and natural selection, Philos. Trans. R. Soc. Lond. B Biol. Sci. **365**, 2559 (2010).
[17] S. P. Otto and N. H. Barton, enThe evolution of recombination: removing the limits to natural selection, Genetics **147**, 879 (1997).
[18] J. F. Crow and M. Kimura, Evolution in sexual and asexual populations, Am. Nat. **99**, 439 (1965).
[19] Y. Kim and H. A. Orr, enAdaptation in sexuals vs. asexuals: clonal interference and the Fisher-Muller model, Genetics **171**, 1377 (2005).
[20] J. M. Smith and J. Maynard-Smith, *The evolution of sex*, Vol. 4 (Cambridge University Press Cambridge, 1978).
[21] J. Maynard Smith, The evolution of sex, .
[22] B. Charlesworth, enMutation-selection balance and the evolutionary advantage of sex and recombination, Genet. Res. **89**, 451 (1990).
[23] S. P. Otto and M. W. Feldman, enDeleterious mutations, variable epistatic interactions, and the evolution of recombination, Theor. Popul. Biol. **51**, 134 (1997).
[24] A. Blachford and A. F. Agrawal, enAssortative mating for fitness and the evolution of recombination, Evolution **60**, 1337 (2006).
[25] N. H. Barton and S. P. Otto, enEvolution of recombination due to random drift, Genetics **169**, 2353 (2005).
[26] S. P. Otto and N. H. Barton, enSelection for recombination in small populations, Evolution **55**, 1921 (2001).
[27] P. D. Keightley and S. P. Otto, enInterference among deleterious mutations favours sex and recombination in finite populations, Nature **443**, 89 (2006).
[28] S. P. Otto, enThe evolutionary enigma of sex, Am. Nat. **174 Suppl 1**, S1 (2009).
[29] J. Felsenstein, enThe evolutionary advantage of recombination, Genetics **78**, 737 (1974).
[30] M. Slatkin, Linkage disequilibrium—understanding the evolutionary past and mapping the medical future, Nat. Rev. Genet. **9**, 477 (2008).
[31] G. Martin and L. Roques, enThe nonstationary dynamics of fitness distributions: Asexual model with epistasis and standing variation, Genetics **204**, 1541 (2016).
[32] M.-E. Gil, F. Hamel, G. Martin, and L. Roques, Mathematical properties of a class of integro-differential models from population genetics, SIAM J. Appl. Math. **77**, 1536 (2017).
[33] Y. Anciaux, A. Lambert, O. Ronce, L. Roques, and G. Martin, enPopulation persistence under high mutation rate: From evolutionary rescue to lethal mutagenesis, Evolution **73**, 1517 (2019).
[34] P. J. Gerrish and P. D. Sniegowski, enReal time forecasting of near-future evolution, J. R. Soc. Interface **9**, 2268 (2012).
[35] R. Bürger, enMoments, cumulants, and polygenic dynamics, J. Math. Biol. **30**, 199 (1991).
[36] M. Smerlak and A. Youssef, enLimiting fitness distributions in evolutionary dynamics, J. Theor. Biol. **416**, 68 (2017).
[37] S. P. Otto and T. Lenormand, enResolving the paradox of sex and recombination, Nat. Rev. Genet. **3**, 252 (2002).
[38] B. Charlesworth and N. H. Barton, Recombination load associated with selection for increased recombination (1996).
[39] S. P. Otto, enUnravelling gene interactions, Nature **390**, 343 (1997).
[40] W. R. Rice, enExperimental tests of the adaptive significance of sexual recombination, Nat. Rev. Genet. **3**, 241 (2002).
[41] E. Baake and M. Baake, Haldane linearisation done right: Solving the nonlinear recombination equation the easy way, Discrete & Continuous Dynamical Systems - A **36**, 6645 (2016).
[42] D. Mchale and G. A. Ringwood, enHaldane linearisation of baric algebras, J. Lond. Math. Soc. **s2-28**, 17 (1983).
[43] M. Baake and E. Baake, An exactly solved model for mutation, recombination and selection, Canad. J. Math. **55**, 3 (2003).
[44] E. Baake, enMutation and recombination with tight linkage, J. Math. Biol. **42**, 455 (2001).
[45] M. Esser, S. Probst, and E. Baake, enPartitioning, duality, and linkage disequilibria in the moran model with recombination, J. Math. Biol. **73**, 161 (2016).
[46] E. Baake, Deterministic and stochastic aspects of single-crossover recombination, in *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)* (Published by Hindustan Book Agency (HBA), India. WSPC Distribute for All Markets Except in India, 2011) pp. 3037–3053.

Supplementary Materials

Methods
Supplementary Text
Figs S1 to S12
Tables S1 to S3
References (31 to 40)

Author contributions P.G. conceived the theory conceptually; P.G., P.S., B.S. and A.C. developed the theory verbally and with simulation; P.G, B.Y. and J.C. developed the theory mathematically; B.Y. and J.C.

provided mathematical proofs for the across-population part; P.G., V.V., F.C. and N.H. provided mathematical proofs for the within-population part. P.G. wrote the paper with critical help and guidance from B.S., P.S. and B.Y.

Competing interests The authors declare no competing interests.

Correspondence and requests for materials should be addressed to P.G.