# Natural selection promotes the evolution of recombination 1: among selected genotypes*

Philip J Gerrish,[1,2,3] Benjamin Galeota-Sprung,[4] Paul Sniegowski,[4] Julien Chevallier,[5] and Bernard Ycart[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] *Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, USA*
[5] *Mathematique Apliquée, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France*

(Dated: 5 June 2021)

Shuffling one's genetic material with another individual seems a risky endeavor more likely to decrease than to increase offspring fitness. This intuitive argument is commonly employed to explain why the ubiquity of sex and recombination in nature is enigmatic. It is predicated on the notion that natural selection assembles selectively well-matched combinations of genes that recombination would break up resulting in low-fitness offspring – a notion so intuitive that it is often stated in the literature as a self-evident premise. We show, however, that this common premise is only self evident on the surface and that, upon closer examination, it is fundamentally flawed: we find that natural selection in fact has an encompassing tendency to assemble selectively mismatched combinations of alleles; recombination breaks up these selectively mismatched combinations (on average), assembles selectively matched combinations, and should thus be favored. The new perspective our findings offer suggests that sex and recombination are not so enigmatic but are instead natural and unavoidable byproducts of natural selection.

## INTRODUCTION

The first measurable indicator that recombination and sex could confer a fitness advantage came from agriculture: when one parent comes from one inbred lineage and the other parent from a different inbred lineage, offspring tend to have increased fitness – a phenomenon known as *hybrid vigor* or *heterosis*. Documentation of this phenomenon dates back at least as far as Darwin [3, 4]. It was later conjectured that recombination could be advantageous even within a single evolving population because it increases trait "resolution" (reduces trait granularity), making fitness peaks more accessible than they otherwise might be [5]. The two parts of the present study follow these two historical threads, addressing the effect of natural selection on across-population and within-population recombination, respectively. The details of these two parts are developed in this paper (across-population) and companion paper (within-population) [6] and an accessible presentation of a synthesis of these two parts is found in companion paper [1].

After the rediscovery of Mendel's work, two competing mechanistic explanations for heterosis emerged:

The first explanation – the *dominance hypothesis* – relied on two observations: 1) inbreeding tends to produce homozygotes, and 2) deleterious alleles tend to be recessive. If a locus is homozygous dominant (wildtype) in one population and homozygous recessive (deleterious) in the other, an across-population recombination event has probability 1/4 of producing a deleterious offspring, whereas it would have probability 3/4 in the absence of dominance.

The second explanation – the *overdominance hypothesis* relied on empirical observations of a seemingly magical phenomenon (overdominance) [7–11], where heterozygotes are fitter than either homozygote. While overdominance has been observed, and there are several famous examples, the genetic/mechanistic basis of overdominance is varied and nebulous. Perhaps the most plausible mechanism conjectured to explain overdominance generally is that the locus under consideration is really two or more loci in linkage; "allelic repulsion" [7], i.e., selectively mismatched homozygotes across populations, would then give the appearance of overdominance as fitter dominants mask less-fit recessives; this is called *pseudo-overdominance*. Allelic repulsion, however, requires an assumption of negative fitness associations across linked loci. The work we present here gives a general theoretical basis for such negative associations across linked loci and may thus provide a general explanation for pseudo-overdominance.

If across-population recombination confers a fitness advantage, it seems reasonable that across-subpopulation recombination should also confer a fitness advantage. And the same goes for across-niche or across-clone recombination. Indeed, more recent studies have shown that recombination across any sort of population structure can confer a fitness advantage. Here, we refer to all of the above as recombination across "selected genotypes". Selected genotypes can refer to any genotypes that have fixed locally within populations, subpopulations, demes, niches, or competing clones. Put differently, selected genotypes are the *products* of selection acting locally. Studying the evolution of recombination across selected genotypes constitutes part one of our two-

* This article is published in concert with [1] and [2]
† pgerrish@unm.edu

part study.

Recombinants whose parents are two distinct selected genotypes will carry an immediate selective advantage, on average, when the ensemble of selected genotypes harbors an excess of selectively antagonistic (mismatched) gene combinations and a deficit of synergistic (well-matched) combinations: by randomly shuffling alleles across loci among selected genotypes, recombination will on average increase offspring fitness. The challenge in explaining the ubiquity of sex and recombination in nature is to identify a source of this selective imbalance that is comparably ubiquitous. One feature of living things whose prevalence approximates that of sex and recombination is evolution by natural selection. In the present study, we assess the effects of natural selection by itself on selective imbalance among selected genotypes and hence on the evolution of recombination across selected genotypes. In essence, the present study assesses the selective value of recombination in structured (e.g., spatially structured) populations.

Many previous studies of the evolution of sex and recombination have, wittingly or not, adopted a causal perspective in which the emergence of recombination constitutes the proverbial "horse" (the cause) and expedited, efficient adaptation constitutes the "cart" (the effect). Here, we explore an inversion of this perspective, asking whether adaptation might instead be the "horse" and the emergence of recombination the "cart". Put differently, where many previous studies view recombination as facilitating adaptation, the present study together with its companion studies [1, 2] asks whether adaptation facilitates the evolution of recombination. This reversal in conjectured causality leads to new and surprising results that reveal the ubiquity of sex and recombination to be perhaps not so enigmatic.

## 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) [12–18], which measures the covariance in allelic *states* across two loci [19] (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 and SM). 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 $\sigma_{XY}$ is covariance, and $\bar{w} = \mathbb{E}[XY]$ is mean fitness (SM).

### *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 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.

## NATURAL SELECTION: SIMULATIONS

As an introduction to how we are modeling the selective value of recombination across selected genotypes, we begin by describing simple simulations. We encourage interested readers to perform these very simple simulations to see for themselves the counter-intuitive outcome and its remarkable robustness to the choice of distribution.

In order to isolate the effects of natural selection, we must assume the population size to be infinite so that dynamics are deterministic (as stated in companion studies [1, 2]). We will assume the organism in question has two loci. The simulations begin by generating a set of $n$ distinct genotypes; this is achieved simply by drawing $n$ genic fitness pairs $(x_i, y_i)$, $i = 1, 2, ..., n$ from some bivariate distribution. The bivariate distribution can be any distribution with any covariance.

Next, the simulation simply records the $(x_i, y_i)$ pair whose sum $x_i + y_i$ is the largest and puts this pair into a new array that we will denote by $(\hat{x}_j, \hat{y}_j)$. This mimics natural selection acting in an infinite population; in an

infinite population there is no role for chance and natural selection thus deterministically fixes the fittest genotype.

The procedure is then repeated a few thousand times, so that there are a few thousand entries in the $(\hat{x}_j, \hat{y}_j)$ array of "winners", or "selected genotypes". The covariance of the $(\hat{x}_j, \hat{y}_j)$ array is then computed. This covariance will always be less than the covariance of the initial bivariate distribution used to generate the $(x_i, y_i)$. In particular, if the covariance of the initial bivariate distribution is zero (i.e., if $X$ and $Y$ are independent), the covariance of the "selected genotypes" will always be negative (i.e., the mean value of recombinants across selected genotypes will always be positive). The interested reader may want to explore this case first, because: 1) you will see that any bivariate distribution from uniform to Cauchy gives this result, and 2) this is the case that is the primary focus of the following mathematical developments. An example set of such simulations where $X$ and $Y$ are skew-normal is plotted in Fig 1.

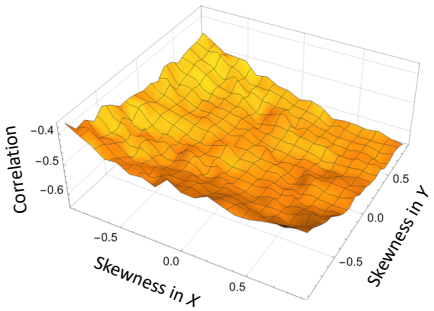


Figure 1. Correlation between genic fitness $\hat{x}_i$ and $\hat{y}_i$ among selected genotypes in simple simulations. A set of 20 $x$-values was drawn from a skew-normal distribution with mean $-0.1$, standard deviation 0.1 and skewness indicated by the $x$-axis. A set of 20 $y$-values was drawn from a skew-normal distribution with mean $-0.1$, standard deviation 0.1 and skewness indicated by the $y$-axis. These $x$ and $y$ values were paired up to form an array of 20 $(x, y)$ pairs. The pair whose sum $x + y$ was the largest was selected and its values appended to a new array $(\hat{x}, \hat{y})$ of selected genotypes. This was repeated 5000 times. The correlation between $\hat{x}$ and $\hat{y}$ was computed and plotted for each pair of skewness values.

## NATURAL SELECTION: ANALYSIS

### General setting: $m$ loci, $n$ alleles

Let $n$ and $m$ be two positive integers. Let $(X_{i,j})_{1 \leqslant i \leqslant n; 1 \leqslant j \leqslant m}$ be a rectangular array of independent random variables. For our purposes, each $X$ quantifies a fitness-related phenotype encoded at one locus. Each row represents an individual's haploid genome and each column represents a locus on that genome. See Fig. 2. We shall denote by $X_{(1)} = (X_{i,j})_{1 \leqslant j \leqslant m}$ the $i$-th row of the array (the $i$-th individual in a population). Let $\phi$ be

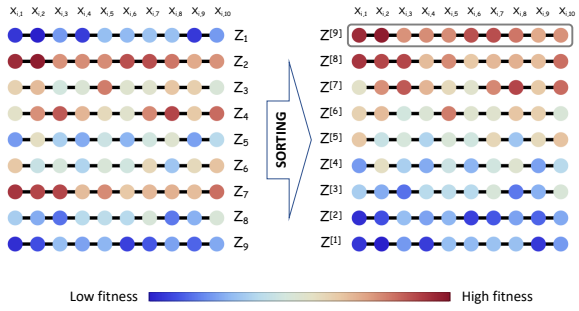


Figure 2. **General setting.** The population here consists of $n = 9$ individuals (9 genomes) represented by the 9 rows, each of which carries a genome with $m = 10$ loci represented by the 10 columns. Each dot represents a locus on an individual genome and its color indicates its genic fitness. The total fitness of the $i^{th}$ individual is $Z_i = \phi(X_{i,1}, X_{i,2}, ..., X_{i,m})$, where $X_{i,j}$ is the genic fitness of $j^{th}$ locus in the $i^{th}$ individual. Strictly speaking, $\phi$ can be any function, but our developments eventually require that it be some increasing function of the genic fitnesses, $X_{i,j}$. To give a simple and useful example, $\phi$ may be defined simply as the sum of its arguments. We employ this definition of $\phi$ extensively in the main text and in our analyses, both because of its simplicity and because of its connection to classical population genetics and notions of additive fitness. On the left-hand side, the genomes are not sorted in any order; on the right-hand side, the same genomes are sorted (ranked) by their total fitness, $Z$, such that $Z^{[1]}$ is the genome of lowest fitness and $Z^{[n]}$ is the genome of highest fitness. If selection were deterministic, the fittest genome ($Z^{[n]}$, highlighted by a frame) would eventually displace all other genomes. The statistical properties of the genic fitnesses of this fittest genome are thus of special interest from an evolutionary perspective. In particular, we are here interested in any statistical associations among these genic fitnesses: if that association tends to be negative, then recombination will be favored.

a measurable function from $\mathbb{R}^m$ into $\mathbb{R}$. For $i = 1, \ldots, n$, denote by $Z^{[1]}$ the image by $\phi$ of the $i$-th row of the array.

$$Z^{[1]} = \phi(X_{(1)}) \ .$$

$Z^{[1]}$ represents the total fitness of individual $i$. Denote by $\sigma \in \mathcal{S}_n$ the random permutation such that

$$\min_{i=1}^{n} Z^{[1]} = S_{\sigma(1)} \leqslant \cdots \leqslant S_{\sigma(n)} = \max_{i=1}^{n} Z^{[1]} \ .$$

The permutation $\sigma$ is uniquely defined up to the usual convention of increasing order for indices corresponding to ties. Deterministically, natural selection will cause the genome of highest fitness ($S_{\sigma(n)} = \max_{i=1}^{n} Z^{[1]}$) to fix. We are interested in the statistical properties of the $X_{\sigma(n),j}$ ; in particular, we are interested in any associations that might arise across loci (across different values of $j$) in this winning genotype. If these associations are

negative, recombination – which alleviates negative associations across loci – should be favored.

For $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m$, define:

$$A_{i,j} = X_{\sigma(i),j} \, .$$

For $1 \leqslant i \leqslant n$, $A_i = (A_{i,j})_{1 \leqslant j \leqslant m}$ is that row in the array $(X_{i,j})$ which ranks $i$-th in the order of images by $\phi$.

*Density*

PROPOSITION 0. *Assume that for $j = 1, \dots, m$, $X_{i,j}$ has pdf $f_j$, for all $i = 1, \dots, n$. Denote by $H$ the common cdf of $Z^{[1]}$'s and assume that $H$ is continuous over its support. The joint pdf of $A_i$ is:*

$$n \, f_1(x_1) \cdots f_m(x_m) \binom{n-1}{i-1} H^{i-1}(\phi(x_1, \dots, x_m))(1 - H(\phi(x_1, \dots, x_m)))^{n-i} \, .$$

*Proof*: For any continous bounded function $\Psi$ of $m$ variables:

$$\begin{aligned} \mathbb{E}(\Psi(A_i)) &= \sum_{\ell=1}^{n} \frac{1}{n} \mathbb{E}(\Psi(X_\ell) \,|\, \sigma(i) = \ell) \\ &= \mathbb{E}(\Psi(X_1) \,|\, \sigma(i) = 1) \, . \end{aligned}$$

Thus the distribution of $A_i$ and the conditional distribution of $X_1$ given that $\Phi(X_1)$ ranks $i$-th, are the same. The pdf of $X_1$ is $f_1(x_1) \cdots f_m(x_m)$. The probability of the event $\sigma(i) = 1$ is $1/n$. Conditioning on $X_1 = (x_1, \dots, x_m)$, the probability that $X_1$ ranks $i$-th is the probability that among $S_2, \dots, S_n$, $i-1$ are below $\phi(x_1, \dots, x_m)$ and $n - i$ are above. The probability for $S_\ell$ to be below $\phi(x_1, \dots, x_m)$ is $H(\phi(x_1, \dots, x_m))$. Hence the result. □

Observe that the average of the densities of $A_i$'s is the common density of all $X_{(1)}$'s, *i.e.* $f_1(x_1), \dots, f_m(x_m)$. This was to be expected, since choosing at random one of the $A_i$'s is equivalent to choosing at random one of the $X_{(1)}$'s. The question is whether the $A_i$'s are negatively associated in the sense of Joag-Dev and Proschan [20]; this seems a reasonable conjecture in light of Theorems 2.8 and also examples (b) and (c) of section 3.2 in that reference. However we have not been able to prove it. We now focus on the simplest possible scenario:

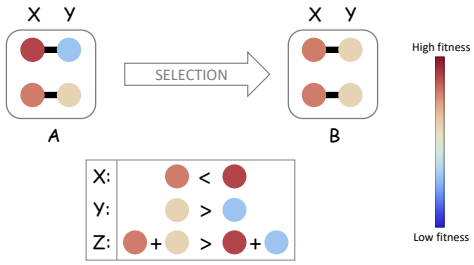


Figure 3. **Two loci, two alleles.** Here, a large (infinite) population consists of individuals whose genome has only two loci $x$ and $y$, each of which carries one of two alleles: genotype 1 carries allele $X_1$ at the $x$ locus and $Y_1$ at the $y$ locus, and genotype 2 carries allele $X_2$ at the $x$ locus and $Y_2$ at the $y$ locus. An individual's fitness is simply the sum of its genic fitnesses, $Z = X + Y$, so that the fitnesses of genotypes 1 and 2 are $Z_1 = X_1 + Y_1$ and $Z_2 = X_2 + Y_2$, respectively. The fitter of these two genotypes has total fitness denoted $Z^{[2]}$ (i.e., $Z^{[2]} = \text{Max}\{Z_1, Z_2\}$) and genic fitnesses $X_{(2)}$ and $Y_{(2)}$ (i.e., $Z^{[2]} = X_{(2)} + Y_{(2)}$). Similarly, the less-fit of these two genotypes has total fitness $Z^{[1]} = X_{(1)} + Y_{(1)}$. We note: $Z^{[2]} > Z^{[1]}$ by definition, but this does *not* guarantee that $X_{(2)} > X_{(1)}$ or that $Y_{(2)} > Y_{(1)}$, as illustrated in the lower box. The population labeled $A$ consists of two distinct genotypes but selection acts to remove the inferior genotype leaving a homogeneous population in which individuals are all genetically identical (with fitness $Z^{[2]}$) as illustrated in the population labeled $B$.

### Two loci, two alleles

No hypothesis on the ranking function $\phi$ is made at this point, apart from being measurable. Notations will be simplified as follows: $(X_1, Y_1, X_2, Y_2)$ are i.i.d.; $(X_{(1)}, Y_{(1)})$ (the *infimum*) denotes that couple $(X_1, Y_1)$ or $(X_2, Y_2)$ whose value by $\phi$ is minimal; $(X_{(2)}, Y_{(2)})$ (the *supremum*) denotes that couple $(X_1, Y_1)$ or $(X_2, Y_2)$ whose value by $\phi$ is maximal.

PROPOSITION 1. *Let $\psi$ be any measurable function from $\mathbb{R}^2$ into $\mathbb{R}$. Then: $\frac{1}{2}\mathbb{E}(\psi(X_{(1)}, Y_{(1)})) + \frac{1}{2}\mathbb{E}(\psi(X_{(2)}, Y_{(2)})) = \mathbb{E}(\psi(X_1, Y_1))$ . In particular, the arithmetic mean of $\mathbb{E}(X_{(1)})$ and $\mathbb{E}(X_{(2)})$ is $\mathbb{E}(X_1)$.*

*Proof*: Consider a random index $I$, equal to "(1)" or "(2)" each with probability 1/2, independent from $(X_1, Y_1, X_2, Y_2)$. By an argument used in the previous section, the couple $(X_I, Y_I)$ is distributed as $(X_1, Y_1)$. Hence, $\mathbb{E}(\psi(X_I, Y_I)) = \mathbb{E}(\psi(X_1, Y_1))$ , however,

$$\mathbb{E}(\psi(X_I,Y_I)) = \mathbb{E}(\mathbb{E}(\psi(X_I,Y_I)\,|\,I)) = \frac{1}{2}\mathbb{E}(\psi(X_{(1)},Y_{(1)})) + \frac{1}{2}\mathbb{E}(\psi(X_{(2)},Y_{(2)}))\ .$$

□

PROPOSITION 2. *We have:* $\mathrm{Cov}(X_{(1)},Y_{(1)}) + \mathrm{Cov}(X_{(2)},Y_{(2)}) = -(\mathrm{Cov}(X_{(1)},Y_{(2)}) + \mathrm{Cov}(X_{(2)},Y_{(1)})) = -\frac{1}{2}\mathbb{E}(X_{(2)}-X_{(1)})\mathbb{E}(Y_{(2)}-Y_{(1)})$ .

*Proof*: Consider again the same random index $I$, equal to "(1)" or "(2)" each with probability 1/2, independent from $(X_1,Y_1,X_2,Y_2)$. The couples $(X_I,Y_I)$ and $(X_I,Y_{3-I})$ are both distributed as $(X_1,Y_1)$. Therefore their covariances are null. These covariances can also be computed by conditioning on $I$ (see *e.g.* formula (1.1) in [20]). For $(X_I,Y_I)$: $\mathrm{Cov}(X_I,Y_I) = \mathbb{E}(\mathrm{Cov}(X_I,Y_I|I)) + \mathrm{Cov}(\mathbb{E}(X_I|I),\mathbb{E}(Y_I|I))$ . On the right-hand side, the first term is: $\mathbb{E}(\mathrm{Cov}(X_I,Y_I|I)) = \frac{1}{2}\mathrm{Cov}(X_{(1)},Y_{(1)}) + \frac{1}{2}\mathrm{Cov}(X_{(2)},Y_{(2)})$ . The second term is: $\mathrm{Cov}(\mathbb{E}(X_I|I),\mathbb{E}(Y_I|I)) = \frac{1}{4}\mathbb{E}(X_{(2)}-X_{(1)})\mathbb{E}(Y_{(2)}-Y_{(1)})$ . Similarly, we have: $\mathrm{Cov}(X_I,Y_{3-I}) = \mathbb{E}(\mathrm{Cov}(X_I,Y_{3-I}|I)) + \mathrm{Cov}(\mathbb{E}(X_I|I),\mathbb{E}(Y_{3-I}|I))$ . The first term in the right-hand side is: $\mathbb{E}(\mathrm{Cov}(X_I,Y_{3-I}|I)) = \frac{1}{2}\mathrm{Cov}(X_{(1)},Y_{(2)}) + \frac{1}{2}\mathrm{Cov}(X_{(2)},Y_{(1)})$ . The second term in the right-hand side is: $\mathrm{Cov}(\mathbb{E}(X_I|I),\mathbb{E}(Y_{3-I}|I)) = -\frac{1}{4}\mathbb{E}(X_{(2)}-X_{(1)})\mathbb{E}(Y_{(2)}-Y_{(1)})$ . Hence the result. □

PROPOSITION 3. *Assume that the ranking function $\phi$ is symmetric: $\phi(x,y) = \phi(y,x)$. Then the couple $(X_{(1)},Y_{(2)})$ has the same distribution as the couple $(Y_{(1)},X_{(2)})$.*

As a consequence, $X_{(1)}$ and $Y_{(1)}$ have the same distribution, so do $X_{(2)}$ and $Y_{(2)}$. Thus: $\mathbb{E}(X_{(2)}-X_{(1)}) = \mathbb{E}(Y_{(2)}-Y_{(1)}) = \frac{1}{2}\mathbb{E}(Z^{[2]}-Z^{[1]})$ . Another consequence is that: $\mathrm{Cov}(X_{(1)},Y_{(2)}) = \mathrm{Cov}(X_{(2)},Y_{(1)})$ . Thus by Proposition 2: $\mathrm{Cov}(X_{(1)},Y_{(2)}) = \mathrm{Cov}(X_{(2)},Y_{(1)}) = \frac{1}{16}\mathbb{E}^2(Z^{[2]}-Z^{[1]})$ .

*Proof*: Since $\phi$ is symmetric, the change of variable $(X_1,Y_1,X_2,Y_2) \mapsto (Y_1,X_1,Y_2,X_2)$ leaves unchanged the couple $(S_1,S_2)$. □

PROPOSITION 4. *Assume that the ranking function $\phi$ is the sum: $\phi(x,y) = x+y$. Then: $\mathbb{E}(X_{(1)}) = \mathbb{E}(Y_{(1)})$ , $\mathbb{E}(X_{(2)}) = \mathbb{E}(Y_{(2)})$ , and $\mathbb{E}(X_{(1)}) < \mathbb{E}(X_{(2)})$ .*

*Proof*: The first two equalities come from Proposition 3. By definition, $\mathbb{E}(X_{(1)}+Y_{(1)}) < \mathbb{E}(X_{(2)}+Y_{(2)})$. Hence the inequality. □

PROPOSITION 5. *Assume that the ranking function $\phi$ is the sum, and that the common distribution of $X_1,Y_1,X_2,Y_2$ is symmetric: there exists a such that $f(x-a) = f(a-x)$. Then $(a-X_{(1)},a-Y_{(1)})$ has the same distribution as $(X_{(2)}-a,Y_{(2)}-a)$.*

As a consequence, $\mathrm{Cov}(X_{(1)},Y_{(1)}) = \mathrm{Cov}(\mathrm{X}_{(2)},\mathrm{Y}_{(2)})$.

*Proof*: The change of variable $(X_1,Y_1,X_2,Y_2) \mapsto (2a-X_1,2a-Y_1,2a-X_2,2a-Y_2)$ leaves the distribution unchanged. It only swaps the indices $i$ and $s$ of minimal and maximal sum. □

If we summarize Propositions 1, 2, 3, 4, 5 for the case where the ranking function is the sum, and the distribution is symmetric, one gets:

$$\mathrm{Cov}(X_{(1)},Y_{(1)}) = \mathrm{Cov}(X_{(2)},Y_{(2)}) < 0$$
$$\mathrm{Cov}(X_{(1)},Y_{(2)}) = \mathrm{Cov}(X_{(2)},Y_{(1)}) > 0$$
$$|\mathrm{Cov}(X_{(1)},Y_{(1)})| = \mathrm{Cov}(X_{(1)},Y_{(2)}) = \frac{1}{16}\mathbb{E}^2(Z^{[2]}-Z^{[1]})\ .$$

**Two loci, $n$ alleles**

Let $n$ be an integer larger than 1. For $i=1,\dots,n$, let $(X_i,Y_i)$ be i.i.d. couples of random variables. For $i=1,\dots,n$, let $Z_i = X_i+Y_i$, and:

$$P_i = \frac{\mathrm{e}^{\beta Z_i}}{\sum_{j=1}^{n}\mathrm{e}^{\beta Z_j}}\ .$$

Let $U$ be a random variable, independent from $(X_i,Y_i), i=1,\dots,n$, uniformly distributed over $(0,1)$. Define the random index $I$ in $\{1,\dots,n\}$ as:

$$I = \begin{cases} 1 & \text{if } U \leqslant P_1\ , \\ \vdots & \\ i & \text{if } P_1+\cdots+P_{i-1} < U \leqslant P_1+\cdots+P_i\ , \\ \vdots & \\ n & \text{if } P_1+\cdots+P_{n-1} < U\ . \end{cases}$$

Finally, let $(X,Y) = (X_I,Y_I)$. The goal is to say something about the first and second order moments of $X$ and $Y$.

For this, conditioning over two embedded $\sigma$-algebras, denoted by $\mathcal{F}_{2n}$ and $\mathcal{F}_n$, will be used.

$$\mathcal{F}_{2n} \text{is generated by} (X_i,Y_i)\ ,\ i=1,\dots,n\ ,$$
$$\mathcal{F}_n \text{is generated by} Z_i\ ,\ i=1,\dots,n\ .$$

If $A$ is any random variable:

$$\mathbb{E}\,(A) = \mathbb{E}\,(\mathbb{E}\,(A\,|\,\mathcal{F}_n)) = \mathbb{E}\,(\mathbb{E}\,(\mathbb{E}\,(A\,|\,\mathcal{F}_{2n})\,|\,\mathcal{F}_n))\ . \quad (1)$$

Conditioning functions of $(X,Y)$ over $\mathcal{F}_{2n}$ and $\mathcal{F}_n$ works as follows.

LEMMA 1. *Let $\phi$ be any real valued function of two variables. Provided the following expectations exist, one has:*

$$\mathbb{E}(\phi(X,Y)\,|\,\mathcal{F}_{2n}) = \sum_{i=1}^{n} P_i \phi(X_i,Y_i)\ ,$$
$$\mathbb{E}(\phi(X,Y)\,|\,\mathcal{F}_{n}) = \sum_{i=1}^{n} P_i \mathbb{E}(\phi(X_i,Y_i)\,|\,Z_i)\ .$$

For second order moments, the following well known lemma on conditional covariances will be used.

LEMMA 2. *Let $(A,B)$ be a pair of real-valued random variables on $(\Omega,\mathcal{F},\mathbb{P})$, and let $\mathcal{F}_1 \subseteq \mathcal{F}_2$ be two $\sigma$-fields on $\Omega$. Then:*

$$\mathrm{cov}\,(A,B\,|\,\mathcal{F}_1) =$$
$$\mathbb{E}\left(\mathrm{cov}\,(A,B\,|\,\mathcal{F}_2)\ |\,\mathcal{F}_1\right) + \mathrm{cov}\left(\mathbb{E}(A\,|\,\mathcal{F}_2)\,,\,\mathbb{E}(B\,|\,\mathcal{F}_2)\,|\,\mathcal{F}_1\right)\ .$$

*In particular, when $\mathcal{F}_1 = \{\emptyset,\Omega\}$:*

$$\mathrm{cov}\,(A,B) = \mathbb{E}\left(\mathrm{cov}\,(A,B\,|\,\mathcal{F}_2)\right) + \mathrm{cov}\left(\mathbb{E}(A\,|\,\mathcal{F}_2)\,,\,\mathbb{E}(B\,|\,\mathcal{F}_2)\right)\ .$$

Lemma 3 relates the moments of $X+Y$ to the $Z_i$'s and $P_i$'s.

LEMMA 3. *Denote by $\overline{Z}$ and $V$ the mean and variance of $Z$ with respect to $P$:*

$$\overline{Z} = \sum_{i=1}^{n} P_i Z_i \quad \textit{and} \quad V = \left(\sum_{i=1}^{n} P_i Z_i^2\right) - \overline{Z}^2\ .$$

*Then:*

$$\mathbb{E}(X+Y) = \mathbb{E}\left(\overline{Z}\right)\ , \tag{2}$$
$$\mathrm{var}(X+Y) = \mathrm{var}\left(\overline{Z}\right) + \mathbb{E}(V)\ . \tag{3}$$

*Proof*: It turns out that $\overline{Z}$ is the conditional expectation of $X+Y$ with respect to $\mathcal{F}_n$, because:

$$\begin{aligned}\mathbb{E}(X+Y\,|\,\mathcal{F}_n) &= \mathbb{E}(\mathbb{E}(X+Y\,|\,\mathcal{F}_{2n})\,|\,F_n)\\ &= \mathbb{E}\left(\sum_{i=1}^{n} P_i Z_i\,|\,\mathcal{F}_n\right)\\ &= \sum_{i=1}^{n} P_i Z_i = \overline{Z}\end{aligned}$$

Hence: $\mathbb{E}(X+Y) = \mathbb{E}\left(\overline{Z}\right)$. Similarly, $V$ is the conditional variance of $X+Y$, given $\mathcal{F}_n$. By Lemma 2:

$$\mathrm{var}(X+Y) = \mathrm{var}\left(\overline{Z}\right) + \mathbb{E}(V)\ .$$

□

From now on, it will be assumed that the common distribution of $(X_i,Y_i)$, for $i=1,\ldots,n$, is bivariate normal.

LEMMA 4. *Let $(X_1,Y_1)$ be a couple of random variables, having bivariate normal distribution $\mathcal{N}_2(\mu,K)$, with expectation $\mu=(\mu_x,\mu_y)$, covariance matrix:*

$$K = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\ ,$$

*where $\sigma_x>0$, $\sigma_y>0$, $|\rho|<1$.*
*Denote:*

$$\eta_x = \frac{\sigma_x^2+\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\ ;\quad \eta_y = \frac{\sigma_y^2+\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\ ,$$

*and also:*

$$\delta = \mu_x\eta_y - \mu_y\eta_x\ ,\quad \gamma = \frac{\sigma_x^2\sigma_y^2(1-\rho^2)}{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\ .$$

*Let $Z_1 = X_1+Y_1$. The conditional distribution of $(X_1,Y_1)$ given $Z_1=z$ is bivariate normal, with expectation:*

$$(\delta+\eta_x z\ ,\ -\delta+\eta_y z)\ ,$$

*covariance matrix:*

$$\gamma\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\ .$$

*Proof*: The vector $(X_1,Y_1,Z_1)$ has normal distribution with expectation $(\mu_x,\mu_y,\mu_x+\mu_y)$, and covariance matrix:

$$\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y & \sigma_x^2+\rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 & \sigma_y^2+\rho\sigma_x\sigma_y \\ \sigma_x^2+\rho\sigma_x\sigma_y & \sigma_y^2+\rho\sigma_x\sigma_y & \sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y \end{pmatrix}\ .$$

The conditional distribution of $(X_1,Y_1)$ given $Z_1=z$ is again normal. The conditional expectation of $X_1$ is:

$$\begin{aligned}\mathbb{E}(X_1\,|\,Z_1=z) &= \mu_x + \frac{z-(\mu_x+\mu_y)}{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y}\left(\sigma_x^2+\rho\sigma_x\sigma_y\right)\\ &= \delta+\eta_x z\ .\end{aligned}$$

The conditional expectation of $Y_1$ is symmetric:

$$\mathbb{E}(Y_1\,|\,Z_1=z) = -\delta+\eta_y z\ .$$

The covariance matrix does not depend on $z$:

$$\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} - \frac{1}{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y} \begin{pmatrix} (\sigma_x^2+\rho\sigma_x\sigma_y)^2 & (\sigma_x^2+\rho\sigma_x\sigma_y)(\sigma_y^2+\rho\sigma_x\sigma_y) \\ (\sigma_x^2+\rho\sigma_x\sigma_y)(\sigma_y^2+\rho\sigma_x\sigma_y) & (\sigma_y^2+\rho\sigma_x\sigma_y)^2 \end{pmatrix} .$$

After simplification one gets:

$$\frac{\sigma_x^2\sigma_y^2(1-\rho^2)}{\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} .$$

□

Theorem 1 below gives the first and second order moments of the random couple $(X, Y)$, when the common distribution of the $(X_i, Y_i)$ is that of Lemma 4.

THEOREM 1. *Assume that for $i = 1, \dots, n$, the distribution of $(X_i, Y_i)$ is bivariate normal $\mathcal{N}_2(\mu, K)$. With the notations of Lemma 4:*

$$\mathbb{E}(X) = \delta + \eta_x \mathbb{E}(X+Y) , \tag{4}$$
$$\mathbb{E}(Y) = -\delta + \eta_y \mathbb{E}(X+Y) , \tag{5}$$

$$\mathrm{var}(X) = \gamma + \eta_x^2 \mathrm{var}(X+Y) , \tag{6}$$
$$\mathrm{var}(Y) = \gamma + \eta_y^2 \mathrm{var}(X+Y) , \tag{7}$$
$$\mathrm{cov}(X, Y) = -\gamma + \eta_x\eta_y \mathrm{var}(X+Y) . \tag{8}$$

Observe that, since $\eta_x + \eta_y = 1$, the first two equations add to identity, and so do the last three, the last one being doubled.

*Proof*: By Lemma 1,

$$\mathbb{E}(X \mid \mathcal{F}_n) = \sum_{i=1}^n P_i \mathbb{E}(X_i \mid Z_i) .$$

By Lemma 4,

$$\mathbb{E}(X_i \mid Z_i) = \delta + \eta_x Z_i .$$

Hence:

$$\mathbb{E}(X \mid \mathcal{F}_n) = \delta + \eta_x \overline{Z} .$$

Similarly:

$$\mathbb{E}(Y \mid \mathcal{F}_n) = -\delta + \eta_y \overline{Z} .$$

Let us now compute $\mathrm{var}(X)$. By Lemma 2:

$$\mathrm{var}(X) = \mathbb{E}(\mathrm{var}(X \mid \mathcal{F}_n)) + \mathrm{var}(\mathbb{E}(X \mid \mathcal{F}_n)) .$$

By Lemma 1,

$$\mathrm{var}(X \mid \mathcal{F}_n) = \sum_{i=1}^n P_i \mathrm{var}(X_i \mid Z_i) .$$

But by Lemma 4, $\mathrm{var}(X_i \mid Z_i)$ is the constant $\gamma$, independently on $Z_i$. Thus:

$$\mathrm{var}(X \mid \mathcal{F}_n) = \sum_{i=1}^n P_i \gamma = \gamma .$$

Now by Lemma 1:

$$\mathrm{var}(\mathbb{E}(X \mid \mathcal{F}_n)) = \sum_{i=1}^n P_i \mathrm{var}(\mathbb{E}(X_i \mid Z_i)) .$$

By Lemma 4, $\mathbb{E}(X_i \mid Z_i) = \delta + \eta_x Z_i$, hence:

$$\mathrm{var}(\mathbb{E}(X \mid \mathcal{F}_n)) = \sum_{i=1}^n P_i \eta_x^2 \mathrm{var}(Z_i) = \eta_x^2 \mathrm{var}(X+Y \mid \mathcal{F}_n) .$$

Joining both results through Lemma 2:

$$\mathrm{var}(X) = \gamma + \eta_x^2 \mathrm{var}(X+Y) .$$

Similarly:

$$\mathrm{var}(Y) = \gamma + \eta_y^2 \mathrm{var}(X+Y) .$$

Let us now turn to $\mathrm{cov}(X, Y)$: By Lemma 2:

$$\mathrm{cov}(X, Y) = \mathbb{E}(\mathrm{cov}(X, Y \mid \mathcal{F}_n) + \mathrm{cov}(\mathbb{E}(X \mid \mathcal{F}_n), \mathbb{E}(X \mid \mathcal{F}_n)) .$$

By Lemma 1,

$$\mathrm{cov}(X, Y \mid \mathcal{F}_n) = \sum_{i=1}^n P_i \mathrm{cov}(X_i, Y_i \mid Z_i) .$$

But by Lemma 4, $\mathrm{cov}(X_i, Y_i \mid Z_i)$ is the constant $-\gamma$, independently on $Z_i$. Thus:

$$\mathrm{cov}(X, Y \mid \mathcal{F}_n) = \sum_{i=1}^n P_i(-\gamma) = -\gamma .$$

Now by Lemma 1:

$$\mathrm{cov}(\mathbb{E}(X \mid \mathcal{F}_n), \mathbb{E}(Y \mid \mathcal{F}_n)) = \sum_{i=1}^n P_i \mathrm{cov}(\mathbb{E}(X_i \mid Z_i), \mathbb{E}(Y_i \mid Z_i)) .$$

By Lemma 4, $\mathbb{E}(X_i \mid Z_i) = \delta + \eta_x Z_i$, and $\mathbb{E}(Y_i \mid Z_i) = -\delta + \eta_y Z_i$. Hence:

$$\mathrm{cov}(\mathbb{E}(X \mid \mathcal{F}_n), \mathbb{E}(Y \mid \mathcal{F}_n)) = \eta_x\eta_y \mathrm{var}(X+Y \mid \mathcal{F}_n) .$$

Joining both results through Lemma 2:

$$\mathrm{cov}(X, Y) = -\gamma + \eta_x\eta_y \mathrm{var}(X+Y) .$$

□

## DISCUSSION

### What if one bout of selection is not enough?

If the initial correlation between $X$ and $Y$ is strongly positive, it may be the case that covariance is not suppressed to negative values (recombinant fitness is not elevated to positive values) in the first bout of selection. To explore the effects of multiple bouts of selection, we performed simulations in which the set of selected genotypes is used to seed a new bout of selection. To this end, $n$ genotypes were chosen at random from the large number of selected genotypes generated from the first bout of selection.

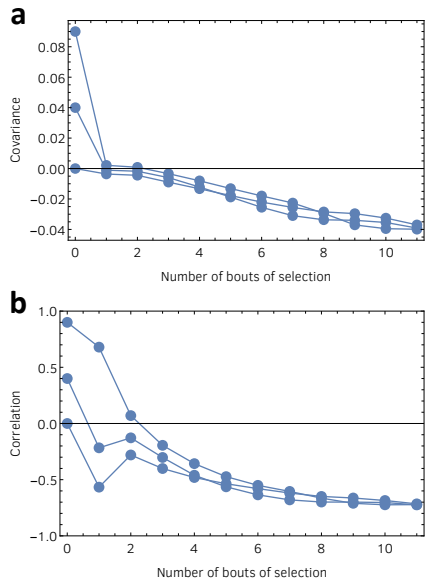


Figure 4. Covariance and correlation as functions of the number of bouts of selection. Employing the notation of the *Simulations* section above, genic fitness pairs $(x_i, y_i)$, $i = 1, 2, ..., n$ are chosen by drawing at random from the $(\hat{x}_j, \hat{y}_j)$ array, and a new selection process produces a new array of "winners" or "selected genotypes", denoted $(\hat{x}'_j, \hat{y}'_j)$, the covariance of this new array will be lower than the covariance of the previous $(\hat{x}_j, \hat{y}_j)$ array. This models a second bout of natural selection. The process is repeated to model more bouts of selection.

Specifically, employing the notation of the *Simulations* section above, genic fitness pairs $(x_i, y_i)$, $i = 1, 2, ..., n$ are chosen by drawing at random from the $(\hat{x}_j, \hat{y}_j)$ array, and a new selection process produces a new array of "winners" or "selected genotypes", denoted $(\hat{x}'_j, \hat{y}'_j)$, the covariance of this new array will be lower than the covariance of the previous $(\hat{x}_j, \hat{y}_j)$ array. This models a second bout of natural selection. If the number of initial replicate simulations is not very large, it may be necessary to add noise (mutation) to the $(\hat{x}'_j, \hat{y}'_j)$ array to avoid covariance quickly becoming zero. If the covariance of the initial bivariate distribution is strongly positive, it may take two or three bouts of selection, but the covariance always becomes negative eventually.

### Concluding remarks

To summarize what exactly has been modeled in this paper, we revisit our definition of "selected genotypes". These are genotypes that are locally prevalent, due to natural selection. Selected genotypes can include locally-prevalent genotypes in populations, subpopulations, demes, niches, or competing clones. A spatially-structured population, for example, can have many spatially separated subpopulations. After selection has been operating in these subpopulations for some time, if an individual from one subpopulation recombines with an individual from another subpopulation, our findings show that the offspring will be fitter, on average, than both parents.

Technically speaking, our findings can be seen as a kind of heterosis, because the term heterosis was originally coined to refer to the simple observation that out-crossed hybrids tend to be superior to their parents. Our findings are not generally consistent, however, with modern notions or explanations of heterosis which are intimately related to, and rely on, theories of dominance and/or overdominance. Where our findings can offer new insights is in the theory of pseudo-overdominance underlying heterosis, which relies on negative fitness associations in linked regions of the genome. The existence of such associations has been seen as somewhat mysterious. Our findings provide a theoretical basis for the kinds of negative associations required by the pseudo-overdominance theory and just might, therefore, provide a novel theoretical basis heterosis generally.

Finally, our findings correct a straw-man argument commonly used to demonstrate why sex and recombination are enigmatic. The premise of this argument is that natural selection will tend to amplify genotypes that carry "good" (selectively well-matched) combinations of genes; this premise seems so intuitive that it should require no proof. Recombination, in this case, would break up the good gene combinations amplified by selection and should thus be selectively suppressed. Independently of biological specifics, our findings show that in fact the opposite is true quite generally. We show that natural selection has an encompassing tendency to amplify genotypes carrying "bad" (selectively mis-matched) combinations of genes. Recombination on average breaks up bad combinations and assembles good combinations, and is thus selectively favored.

## 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, F. Cordero, B. Galeota-Sprung, A. Colato, V. Vejalla, and P. Sniegowski, Natural selection promotes the evolution of recombination 2: during the selective process, Physical Review E **In Review** (2021).
[3] C. Darwin, *The effects of cross and self fertilization in the vegetable kingdom* (John Murray, London, 1876).
[4] G. H. Shull, The genotypes of maize, Am. Nat. **45**, 234 (1911).
[5] A. Weismann, Significance of sexual reproduction, in *Essays upon heredity and kindred biological problems* (Clarendon Press, Oxford, 1889).
[6] 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).
[7] J. F. Crow, The rise and fall of overdominance, Plant Breed. Rev. (2000).
[8] J. A. Birchler, H. Yao, and S. Chudalayandi, Unraveling the genetic basis of hybrid vigor, Proc. Natl. Acad. Sci. U. S. A. **103**, 12957 (2006).
[9] J. F. Crow, Alternative hypotheses of hybrid vigor, Genetics **33**, 477 (1948).
[10] M. R. Labroo, A. J. Studer, and J. E. Rutkoski, Heterosis and hybrid crop breeding: A multidisciplinary review, Front. Genet. **12**, 234 (2021).
[11] Z. B. Lippman and D. Zamir, Heterosis: revisiting the magic, Trends Genet. **23**, 60 (2007).
[12] N. H. Barton, A general model for the evolution of recombination, Genet. Res. **65**, 123 (1995).
[13] N. H. Barton, Linkage and the limits to natural selection, Genetics **140**, 821 (1995).
[14] N. H. Barton, Genetic linkage and natural selection, Philos. Trans. R. Soc. Lond. B Biol. Sci. **365**, 2559 (2010).
[15] J. Felsenstein, The evolutionary advantage of recombination, Genetics **78**, 737 (1974).
[16] S. P. Otto and M. W. Feldman, Deleterious mutations, variable epistatic interactions, and the evolution of recombination, Theor. Popul. Biol. **51**, 134 (1997).
[17] S. P. Otto and N. H. Barton, The evolution of recombination: removing the limits to natural selection, Genetics **147**, 879 (1997).
[18] M. Slatkin, Linkage disequilibrium—understanding the evolutionary past and mapping the medical future, Nat. Rev. Genet. **9**, 477 (2008).
[19] S. P. Otto, Selective interference and the evolution of sex, J. Hered. **112**, 9 (2021).
[20] K. Joag-Dev and F. Proschan, Negative association of random variables with applications, Ann. Statist. **11**, 286 (1983).
[21] W. J. Ewens, *Mathematical Population Genetics: I. Theoretical Introduction* (Springer, New York, NY, 2004).