Natural selection promotes the evolution of recombination 1: across 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

It seems recombination should be disadvantageous most of the time. High-fitness genotypes that are amplified by natural selection, or *selected genotypes*, it seems, should carry "good" (selectively well-matched) combinations of genes. And recombination, which shuffles genes across individuals, should only break up these good combinations and thus should be evolutionarily suppressed. In this light, the overwhelming prevalence of recombination across the tree of life is a mystery.

The foregoing paragraph outlines the standard line of reasoning commonly employed to demonstrate why the ubiquity of sex and recombination is enigmatic. The premise of this line of reasoning – that natural selection will tend to amplify genotypes carrying "good" (selectively well-matched) combinations of genes – is so intuitive that it is considered self-evident in much of the literature and has largely gone unquestioned.

We define *selected genotype* to mean any genotype that has become locally prevalent at any scale – e.g., population, subpopulation, deme, niche, competing clone, etc. – through the local action of natural selection. Put differently, selected genotypes are the products of selection acting locally.

Recombination can only have an effect on offspring fitness if the genetic makeup of the parents differ. In a structured population, local evolution can lead to divergence in genetic makeup. If two parents come from two different locally-evolved subpopulations, therefore, they will likely differ in their genetic makeup. The question then becomes, will they differ in such a way that enhances or diminishes the fitness of their offspring? In other words, will the offspring of different *selected genotypes* (as defined above) tend to be superior or inferior to their parents? The standard argument outlined in the first paragraph would seem to imply that offspring fitness should be inferior to parent fitness, because recombination would break up good gene combinations that each parent had acquired in their local environments.

An answer to this question came early on from agriculture. That the out-crossing of inbred lineages tends to confer *vigorous* offspring (later dubbed "hybrid vigor" or "heterosis" [3–6]) is an observation that has likely been part of farmer folklore for centuries. The earliest known systematic study of this phenomenon was conducted by Darwin himself [7, 8]; his study was perhaps motivated, at least in part, by his search for a theory of inheritance that was consistent with his theory of natural selection (Darwin's big headache [8]). Observations of hybrid vigor gave his "blending" theory of inheritance a plausible foothold: either chance differences in the founding individuals of two locally-evolving subpopulations or divergent selection pressures in these subpopulations would give rise to persistent genetic differences across subpopulations despite local blending within each subpopulation. Sex between individuals from two different subpopulations thus had at least the possibility of offspring whose fitness differed from its parents.

Translated to the language used in the present study, what Darwin was documenting in these early studies was that recombination between two *selected genotypes* tends to produce high-fitness offspring. Since these early studies of heterosis, several more recent studies have shown that different kinds of population structure can

* This article is published in concert with [1] and [2]

† pgerrish@unm.edu

create conditions that make recombination favorable [9–17]. These studies find that population structure can help to maintain the variation without which recombination would be neither advantageous nor disadvantageous but neutral, and they identify conditions under which recombination is advantageous. Spatially heterogeneous selection can, for example, create negative fitness associations if selection is stronger for one locus in one spatial "patch" and stronger for another locus in a neighboring patch [12, 13, 17]. The negative fitness associations that arise in such a scenario would be broken up by recombination thus giving recombination-competent (rec^+) lineages a selective advantage. The prevalence of such scenarios in nature, however, is unknown and questionable [13].

Generally speaking, 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. We find that natural selection by itself has an encompassing tendency to amplify selectively mismatched combinations of alleles, thereby promoting the evolution of recombination across selected genotypes.

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, 18–23], which measures the covariance in allelic *states* across two loci [24] (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 (see [2] and SM). Our results will thus be given in terms of covariance.

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 assume the population size to be infinite so that dynamics are deterministic (as stated in companion studies [1, 2]). As we will show later, however, our findings are quite robust to relaxation of this assumption. 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 $(x_i, y_i),\ i = 1, 2, ..., n$ at random 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. Remarkably, this covariance is always less than the covariance of the initial bivariate distribution used to generate the (x_i, y_i). 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) she/he 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.

NATURAL SELECTION: ANALYSIS

We now turn to mathematical analyses of the procedure described above for simulations. We begin with a generalization of what we describe above: here, instead of two loci with genic fitnesses x_i and y_i for the ith genotype, we have m loci and a vector of genic fitnesses $(x_{i1}, x_{i2}, ..., x_{im})$. Next, we zero in on analyses of the simplest scenario of two loci and two genotypes. Extrapolation of our qualitative results from this simplest-case scenario to more loci and more genotypes is corroborated by simulation (SM). As will become apparent, the mathematical analyses eventually require some restrictions on the bivariate distribution governing genic fitnesses in the

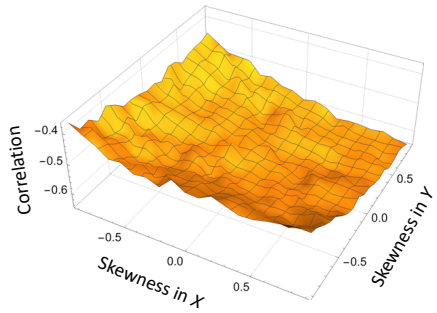

Figure 1. Correlation between genic fitness $\hat{x}_i$ and $\hat{y}_i$ among selected genotypes in simple simulations. A set of 20 x-values was drawn from a skew-normal distribution with mean -0.1, standard deviation 0.1 and skewness indicated by the x-axis. A set of 20 y-values was drawn from a skew-normal 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.

initial population. Simulations, however, show that our findings hold qualitatively for essentially any distribution chosen.

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 $X_{(i)} = (X_{i,j})_{1\leqslant j\leqslant m}$ the i-th row of the array (the i-th individual in a population). Let ϕ be a measurable function from $\mathbb{R}^m$ into $\mathbb{R}$. For $i = 1, \ldots, n$, denote by Z_i the image by ϕ of the i-th row of the array.

$$Z_i = \phi(X_i) \ .$$

Z_i represents the total fitness of genotype i. Denote by $\sigma \in \mathcal{S}_n$ the random permutation such that

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

The permutation σ 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 ($S_{\sigma(n)} = \max_{i=1}^{n} Z_i$) 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.

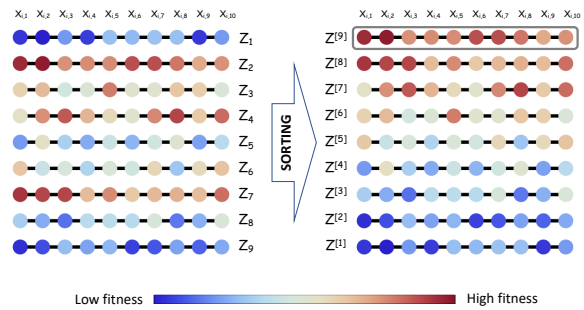

Figure 2. General setting. The population here consists of $n = 9$ genotypes represented by the 9 rows, each of which carries a genome with $m = 10$ loci represented by the 10 columns. Each dot represents a locus on an individual genome and its color indicates its genic fitness. The total fitness of the ith individual is $Z_i = \phi(X_{i1}, X_{i2}, ..., X_{im})$, where X_{ij} is the genic fitness of jth locus in the ith genotype. Strictly speaking, ϕ can be any increasing function of the genic fitnesses, X_{ij}. To give a simple and useful example, ϕ may be defined simply as the sum of its arguments. We employ this definition of ϕ 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. In an infinite population (deterministic selection), the fittest genome ($Z^{[n]}$, highlighted by a frame) always 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 interested in any statistical associations among these genic fitnesses: if that association tends to be negative, then recombination will 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 ϕ.

Density

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

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

Proof: For any continous bounded function Ψ 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 $X_1=(x_1,\ldots,x_m)$, the probability that X_1 ranks i-th is the probability that among $Z_2,\ldots,Z_n$, $i-1$ are below $\phi(x_1,\ldots,x_m)$ and $n-i$ are above. The probability for S_ℓ to be below $\phi(x_1,\ldots,x_m)$ is $H(\phi(x_1,\ldots,x_m))$. Hence the result. □

Observe that the average of the densities of A_i is the common density of all the X_i, *i.e.* $f_1(x_1),\ldots,f_m(x_m)$. This was to be expected, since choosing at random one of the A_i is equivalent to choosing at random one of the X_i. The question is whether the A_i are negatively associated in the sense of Joag-Dev and Proschan [25]; 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 this general case. We now focus on the simplest possible scenario:

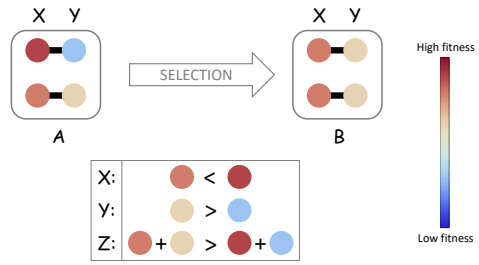

Figure 3. Two loci, two alleles. Here, a large (infinite) population consists of individuals whose genome has only two loci x and y, each of which carries one of two alleles: genotype 1 carries allele X_1 at the x locus and Y_1 at the y locus, and genotype 2 carries allele X_2 at the x locus and Y_2 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., $Z^{[2]}=\mathrm{Max}\{Z_1,Z_2\}$) 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 ϕ 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 ϕ is minimal; $(X_{(2)},Y_{(2)})$ (the *supremum*) denotes that couple (X_1,Y_1) or (X_2,Y_2) whose value by ϕ is maximal.

Proposition 1. *Let ψ 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,

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

□

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 [25]). 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 ϕ 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 ϕ 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 ϕ 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 ϕ 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 (1) and (2) 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 $i = 1, \ldots, n$, let (X_i, Y_i) be i.i.d. couples of random variables. For $i = 1, \ldots, n$, 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 $(X_i, Y_i), i = 1, \ldots, n$, uniformly distributed over $(0, 1)$. Define the random index I in $\{1, \ldots, n\}$ 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 σ-algebras, denoted by $\mathcal{F}_{2n}$ and $\mathcal{F}_n$, will be used.

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

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 ϕ 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 σ-fields on Ω. Then:*

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

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

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

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:*

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

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

$$\mathbb{E}\left(X\right) = \delta + \eta_x \mathbb{E}\left(X+Y\right) \ , \tag{4}$$

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

$$\mathrm{var}\left(X\right) = \gamma + \eta_x^2 \mathrm{var}\left(X+Y\right) \ , \tag{6}$$

$$\mathrm{var}\left(Y\right) = \gamma + \eta_y^2 \mathrm{var}\left(X+Y\right) \ , \tag{7}$$

$$\mathrm{cov}\left(X, Y\right) = -\gamma + \eta_x\eta_y \mathrm{var}\left(X+Y\right) \ . \tag{8}$$

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 γ, 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,

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

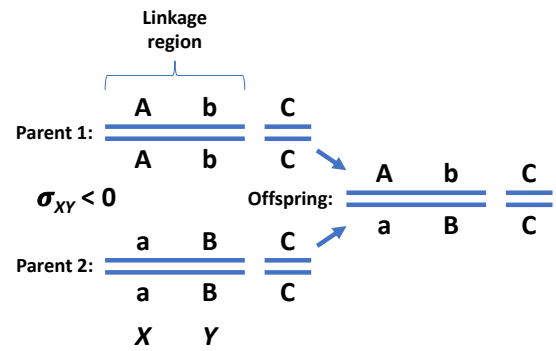

Figure 4. A novel theoretical basis for pseudo-overdominance and heterosis. Upper case (lower case) letters denote higher- (lower-) fitness alleles. Generally speaking, lower-fitness alleles tend to be recessive. Heterozygotes will therefore tend to express the higher-fitness of the two alleles at a locus. Here, parents 1 and 2 come from two different inbred populations. Inbreeding populations tend toward homozygosity and our findings show that natural selection will tend to fix alleles that are poorly matched across loci, i.e., that have negative covariance in genic fitnesses. In the simple example illustrated here, the parents expressing phenotypes (A,b,C) and (a,B,C) exhibit negative covariance in genic fitness across loci. Both parents are less fit that the offspring which expresses phenotype (A,B,C), thereby generating heterosis through an appearance of overdominance (called *pseudo-overdominance*) [26–30]. Our theory shows how the required (and hitherto mysterious) negative covariance in genic fitnesses across linked loci is an unavoidable consequence of natural selection, and thus provides a novel theoretical basis for heterosis.

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:

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

□

HETEROSIS

Our findings can be viewed as providing a theoretical basis for a kind of haploid heterosis. We will now show

that our findings also provide a theoretical basis for classical diploid heterosis as well. 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 curious phenomenon (overdominance) [4, 6, 27, 31, 32], where heterozygotes at a given locus are fitter than either homozygote. While overdominance has been observed, and there are famous examples, the genetic/mechanistic basis of overdominance is varied and nebulous.

There is a more concrete explanation for observations of overdominance that is gaining traction as increasingly detailed studies seek to untangle the genetic basis of heterosis. Increasingly, such studies reveal that observations of overdominance are not really overdominance at all but are in fact an artefact of linkage that can give the *appearance* of overdominance [33–35]. In these cases, what appears to be a single locus is in fact two or more loci in linkage with each other. If the alleles at the linked loci are selectively mismatched giving rise to negative associations in genic fitnesses across populations, outcrossing between populations can give the appearance of overdominance as fitter dominants mask less-fit recessives – a phenomenon that has been dubbed *pseudo-overdominance* [28, 33–36]. (What we are here calling "selective mismatch" has elsewhere been called "linkage repulsion" [35–37], "linkage bias" [34], or in some cases simply "linkage disequilibrium" (LD) [30].) If the outcrossed parents come from different inbred populations – common practice in agriculture – they will have high homozygosity and the heterosis effect in the offspring will be accentuated; a schematic of this scenario is presented in Fig 4.

The weak link in the pseudo-overdominance theory is the requirement that linkage repulsion (selective mismatch) somehow develop within blocks of linked loci – dubbed pseudo-overdominance blocks (or PODs) [34, 35]. Some authors have made verbal arguments invoking a combination of mutation, weak mutational effect and small effective population size to meet this requirement [26]. In a very recent paper, Waller [35] makes an elegant argument showing how slightly-deleterious mutations can mask strongly-deleterious mutations, thereby maintaining inbreeding depression over long periods of time – an observation that confounded Darwin. The theory we have developed in the present study speaks directly to the requirement of linkage repulsion and provides a general theoretical foundation for the pseudo-overdominance theory of heterosis. Mutation, weak mutational effects and small population sizes are not required. The required linkage repulsion is produced across populations of any size simply by natural selection acting on heritable variation. Conceptually, heterosis due to pseudo-overdominance can ultimately be a product of the counter-intuitive phenomenon outlined in Figs 2 and 3 of our companion paper [1].

NATURAL SELECTION PROMOTES SEGREGATION

Our focus in the present study has been on recombination, which is one consequence of sexual reproduction. Another consequence of sexual reproduction is segregation, whereby different alleles at a single locus diverge in frequency owing to a combination of allelic fitness and dominance. When dominance effects are present, it is usually the fittest allele that is dominant; i.e., deleterious alleles tend to be recessive. If the locus in question is actually more than one loci in linkage, similar to the PODs mentioned above, then the processes described in this paper can give rise to an evolutionary advantage for segregation in structured populations.

Pseudo-overdominance is essentially a *segregant advantage* that arises when some loci are in linkage repulsion and when some degree of dominance exists. The theory we have developed shows that the linkage repulsion giving rise to segregant advantage emerges as an inevitable consequence of natural selection.

Simulations of natural selection acting in local populations can give rise to a substantial segregant advantage when the two parents come from different local populations (Fig 5). The segregant advantage is equivalent to the contribution of PODs to heterosis. These observations are consistent with some previous studies [9, 38–41] which have found that the advantage of sexual reproduction can derive at least as much from segregation as from recombination.

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.

Specifically, employing the notation of the *Simulations* section above, a second bout of selection is simulated

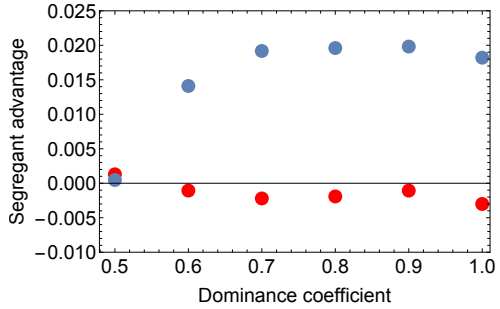

Figure 5. Segregant advantage before (red dots) and after (blue dots) natural selection has acted on local populations, plotted as a function of the dominance coefficient h defined as follows: Let w_A and w_a denote fitnesses of alleles A and a such that $w_A > w_a$. Genic fitness at this locus is $w = hw_A + (1-h)w_a$, so that $h = 1$ models complete dominance and $h = 0.5$ models no dominance. Initial genic fitnesses in local populations were drawn from the same bivariate distribution with zero correlation. Before and after selection, two local populations were sampled with replacement from which the two parents were chosen. The fitness of the segregant offspring was recorded and the processes repeated many time. Mean parent fitness was subtracted from mean offspring fitness and this quantity was reported as the *segregant advantage* (vertical axis).

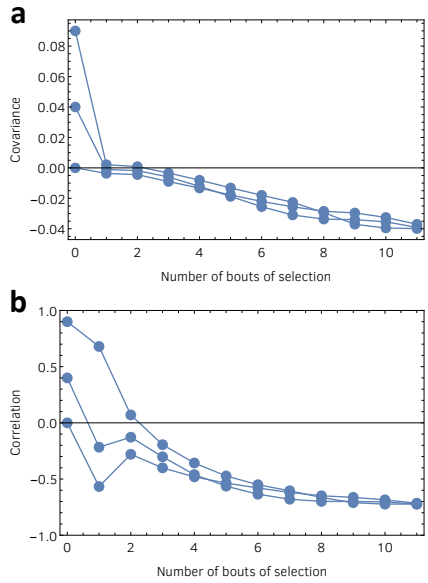

Figure 6. Covariance and correlation as functions of the number of bouts of selection. Employing the notation of the *Simulations* section above, genic fitness pairs (x_i, y_i), $i = 1, 2, ..., n$ are chosen by drawing at random from the $(\hat{x}_j, \hat{y}_j)$ array, and a new selection process produces 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.

by drawing at random a new set of genic fitness pairs (x_i, y_i), $i = 1, 2, ..., n$ from the set of selected genotypes from the first bout of selection, $(\hat{x}_j, \hat{y}_j)$, 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. This procedure is repeated to generate more bouts of 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 usually becomes negative within two or three bouts of selection (Fig 6).

Concluding remarks

To summarize what 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.

Finally, our findings correct the straw-man argument, outlined in the first paragraph of this paper, 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. We find that, when natural selection is operating in isolation, the opposite is true quite generally. We find 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 its evolution is thus promoted.

If all alleles are accessible by mutation, incorporating mutation can lead eventually to the absence of selective imbalance, where all local populations fix the highest-fitness allele at both loci. Incorporating mutation may reveal that the selective imbalance we find here to be a matter of scale in evolutionary time. This consideration is in agreement with observations that out-crossing confers a fitness benefit only when the genetic distance is not too great; beyond a certain distance, out-crossing is detrimental (consistent with the hybrid incompatibility underlying speciation) [42].

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] G. H. Shull, The genotypes of maize, Am. Nat. **45**, 234 (1911).

[4] J. F. Crow, The rise and fall of overdominance, Plant Breed. Rev. (2000).

[5] J. F. Crow, Mid-century controversies in population genetics, Annu. Rev. Genet. **42**, 1 (2008).

[6] J. F. Crow, Alternative hypotheses of hybrid vigor, Genetics **33**, 477 (1948).

[7] C. Darwin, *The effects of cross and self fertilization in the vegetable kingdom* (John Murray, London, 1876).

[8] B. Charlesworth and D. Charlesworth, Darwin and genetics, Genetics **183**, 757 (2009).

[9] A. F. Agrawal, Spatial heterogeneity and the evolution of sex in diploids, Am. Nat. **174 Suppl 1**, S54 (2009).

[10] L. Becks and A. F. Agrawal, Higher rates of sex evolve in spatially heterogeneous environments, Nature **468**, 89 (2010).

[11] A. F. Agrawal, Evolution of sex: Why do organisms shuffle their genotypes?, Curr. Biol. **16**, R696 (2006).

[12] S. P. Otto and T. Lenormand, Resolving the paradox of sex and recombination, Nat. Rev. Genet. **3**, 252 (2002).

[13] T. Lenormand and S. P. Otto, The evolution of recombination in a heterogeneous environment, Genetics **156**, 423 (2000).

[14] N. H. Barton, Genetic linkage and natural selection, Philos. Trans. R. Soc. Lond. B Biol. Sci. **365**, 2559 (2010).

[15] A. O. B. Whitlock, R. B. R. Azevedo, and C. L. Burch, Population structure promotes the evolution of costly sex in artificial gene networks, Evolution **73**, 1089 (2019).

[16] R. A. Neher and B. I. Shraiman, Competition between recombination and epistasis can cause a transition from allele to genotype selection, Proc. Natl. Acad. Sci. U. S. A. **106**, 6866 (2009).

[17] G. Martin, S. P. Otto, and T. Lenormand, Selection for recombination in structured populations, Genetics **172**, 593 (2006).

[18] N. H. Barton, A general model for the evolution of recombination, Genet. Res. **65**, 123 (1995).

[19] N. H. Barton, Linkage and the limits to natural selection, Genetics **140**, 821 (1995).

[20] J. Felsenstein, The evolutionary advantage of recombination, Genetics **78**, 737 (1974).

[21] S. P. Otto and M. W. Feldman, Deleterious mutations, variable epistatic interactions, and the evolution of recombination, Theor. Popul. Biol. **51**, 134 (1997).

[22] S. P. Otto and N. H. Barton, The evolution of recombination: removing the limits to natural selection, Genetics **147**, 879 (1997).

[23] M. Slatkin, Linkage disequilibrium—understanding the evolutionary past and mapping the medical future, Nat. Rev. Genet. **9**, 477 (2008).

[24] S. P. Otto, Selective interference and the evolution of sex, J. Hered. **112**, 9 (2021).

[25] K. Joag-Dev and F. Proschan, Negative association of random variables with applications, Ann. Statist. **11**, 286 (1983).

[26] D. Charlesworth and J. H. Willis, The genetics of inbreeding depression, Nat. Rev. Genet. **10**, 783 (2009).

[27] 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).

[28] J. A. Birchler, H. Yao, S. Chudalayandi, D. Vaiman, and R. A. Veitia, Heterosis, Plant Cell **22**, 2105 (2010).

[29] J. A. Birchler, D. L. Auger, and N. C. Riddle, In search of the molecular basis of heterosis, Plant Cell **15**, 2236 (2003).

[30] L. V. Khotyleva, A. V. Kilchevsky, and M. N. Shapturenko, Theoretical aspects of heterosis, Russian Journal of Genetics: Applied Research **7**, 428 (2017).

[31] M. R. Labroo, A. J. Studer, and J. E. Rutkoski, Heterosis and hybrid crop breeding: A multidisciplinary review, Front. Genet. **12**, 234 (2021).

[32] Z. B. Lippman and D. Zamir, Heterosis: revisiting the magic, Trends Genet. **23**, 60 (2007).

[33] D. F. Jones, Dominance of linked factors as a means of accounting for heterosis, Genetics **2**, 466 (1917).

[34] E. T. Bingham, Role of chromosome blocks in heterosis and estimates of dominance and overdominance, in *Concepts and Breeding of Heterosis in Crop Plants* (Crop Science Society of America, Madison, WI, USA, 2015) pp. 71–87.

[35] D. M. Waller, Addressing darwin's dilemma: Can pseudo-overdominance explain persistent inbreeding depression and load?, Evolution **75**, 779 (2021).

[36] X. Li, X. Li, E. Fridman, T. T. Tesso, and J. Yu, Dis-

secting repulsion linkage in the dwarfing gene dw3 region for sorghum plant height provides insights into heterosis, Proc. Natl. Acad. Sci. U. S. A. **112**, 11823 (2015).

[37] Y. Semel, J. Nissenbaum, N. Menda, M. Zinder, U. Krieger, N. Issman, T. Pleban, Z. Lippman, A. Gur, and D. Zamir, Overdominant quantitative trait loci for yield and fitness in tomato, Proc. Natl. Acad. Sci. U. S. A. **103**, 12981 (2006).

[38] S. P. Otto, The advantages of segregation and the evolution of sex, Genetics **164**, 1099 (2003).

[39] M. Kirkpatrick and C. D. Jenkins, Genetic segregation and the maintenance of sexual reproduction, Nature **339**, 300 (1989).

[40] K. Jaffe, Emergence and maintenance of sex among diploid organisms aided by assortative mating, Acta Biotheor. **48**, 137 (2000).

[41] H. Bernstein, H. C. Byerly, F. A. Hopf, and R. E. Michod, Origin of sex, J. Theor. Biol. **110**, 323 (1984).

[42] M. Lynch, THE GENETIC INTERPRETATION OF INBREEDING DEPRESSION AND OUTBREEDING DEPRESSION, Evolution **45**, 622 (1991).

[43] 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).

[44] W. J. Ewens, *Mathematical Population Genetics: I. Theoretical Introduction* (Springer, New York, NY, 2004).