Natural selection promotes the evolution of recombination 1: between the *products* of natural selection*

Philip J Gerrish,[1,2,3] Benjamin Galeota-Sprung,[4] Paul Sniegowski,[4]
Alexandre Colato,[5] Julien Chevallier,[6] and Bernard Ycart[6]

[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] *Departamento de Ciências da Natureza, Matemática e Educação, Univ Fed de São Carlos, Araras SP, Brazil*
[6] *Mathematique Apliquée, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France*

(Dated: 25 July 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 often stated in the literature as a self-evident premise. We show however that, upon closer examination, this premise is flawed: we find to the contrary that natural selection in fact has an encompassing tendency to assemble selectively mismatched gene combinations; 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 unavoidable byproducts of natural selection.

I. INTRODUCTION

It seems recombination should be disadvantageous most of the time. High-fitness genotypes that are amplified by natural selection (*products* of natural selection), 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 a 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 [3–12] and has largely gone unquestioned.

We define a *product* of natural selection 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.

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 tends to enhance or diminish the fitness of their offspring? In other words, will the offspring of different *products* of natural selection (as defined above) tend to be superior or inferior to their parents? The standard argument outlined in the first paragraph would 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" [13–16]) 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 [17, 18]; 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 [18, 19]. Observations of heterosis 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.

Translated to the language used in the present study, what Darwin was documenting in these early studies was that recombination between two *products* of selection tends to produce high-fitness offspring (assuming that fitness and "vigor" are correlated). This observation contradicted Darwin's theory of blending inheritance which predicts offspring fitness at the midpoint between parent fitnesses; heterosis, it would seem, had the potential at

* This article is published in concert with two companion papers referenced as PRL [1] and PRE2 [2] and Supplemental Materials referenced by the adding the prefix "S".

† pgerrish@unm.edu

least to reveal the flaw in his theory.

That these early studies of inbreeding and heterosis are relevant to the evolution of sex and recombination is not a new idea [11, 20, 21], and is subsumed under Lewontin's general proclamation that "every discovery in classical and population genetics has depended on some sort of inbreeding experiment" [20, 22, 23]. Since these early studies, several more recent studies have shown that different kinds of population structure can create conditions that make recombination across locally-evolving subpopulations favorable [10–12, 24–29]. These studies find that population structure can help to maintain the variation without which recombination would be neither advantageous nor disadvantageous, and they identify conditions under which recombination is advantageous. Spatially heterogeneous selection can, for example, create negative fitness associations if selection acts more strongly on one gene (one *locus*) in one spatial "patch" and acts more strongly on a different locus in a neighboring patch [24, 25, 29]. 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 [25].

Generally speaking, recombinants whose parents are two distinct products of natural selection will carry an immediate selective advantage, on average, when the ensemble of such products harbors an excess of selectively antagonistic (mismatched) gene combinations and a deficit of synergistic (well-matched) combinations: by randomly shuffling different gene variants (or *alleles*) across different products of selection, 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 products of selection. In doing so, we determine 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 different products of selection.

II. MEASURING SELECTIVE IMBALANCE

In much of the relevant literature, the measure of selective mismatch across loci affecting the evolution of recombination is *linkage disequilibrium* (LD) [26, 30–35], which measures the covariance in allelic *states* across two loci [36] (i.e., it measure the bias in allelic frequencies across loci) but does not retain information about the selective value of those alleles.

For the sake of presentation, we here consider an organism with just two fitness-related genes (or two *loci*) whose fitness contributions are represented by random variables X and Y. We have found (see PRE2 [2]) that the expected selective advantage of newly-formed recombinants (and the advantage of recombination over the course of a single generation) is

$$\hat{s}_r \;=\; -\sigma_{XY} \;,$$

where σ_{XY} is the covariance between X and Y. This measure of selective imbalance is superior to LD in that it retains information about both the frequencies and selective value of alleles and it directly gives the selective advantage of recombinants. Furthermore, we show in PRE2 [2] that $\hat{s}_r$ defined in this way provides a lower bound for the selective advantage of a rec^+ lineage within a single population. Our results will thus be given in terms of covariance.

III. NATURAL SELECTION: SIMULATIONS

As an introduction to how we model the selective value of recombination across different products of selection, 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 fairly 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 "products of selection". 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 between X and Y among the "products of selection" will always be negative (i.e., the mean value of recombinants across different products of selection 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.

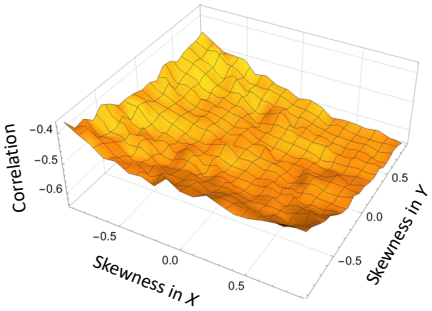

Figure 1. Correlation between genic fitness $\hat{x}_i$ and $\hat{y}_i$ among products of selection in simple simulations. A set of 20 x-values was drawn from a skew-normal distribution with mean -0.1, standard deviation 0.1 and skewness indicated by the x-axis. A set of 20 y-values was drawn from a 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 products of selection. This was repeated 5000 times. The correlation between $\hat{x}$ and $\hat{y}$ was computed and plotted for each pair of skewness values.

IV. 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 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.

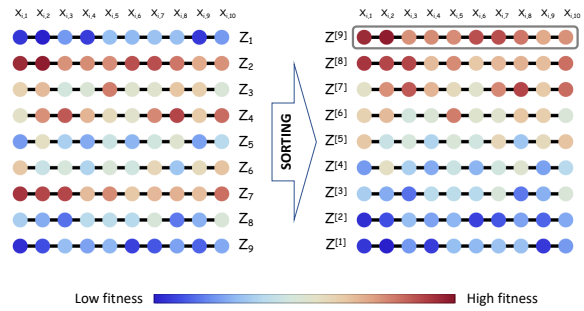

Figure 2. General setting. The population here consists of $n = 9$ genotypes represented by the 9 rows, each of which carries a genome with $m = 10$ loci represented by the 10 columns. Each dot represents a locus on an individual genome and its color indicates its genic fitness. The total fitness of 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.

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.

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 [37]; this seems a reasonable conjecture in light of Theorems 2.8 and also examples (b) and (c) of section 3.2 in that reference.

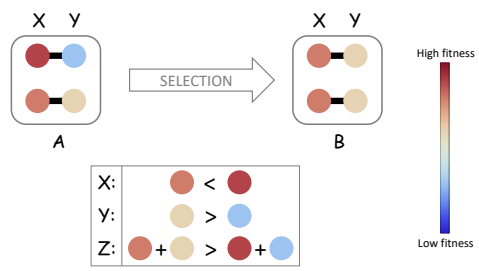

Figure 3. Two loci, two alleles. Here, a large (infinite) population consists of individuals whose genome has only two loci x and y, each of which carries one of two alleles: genotype 1 carries allele X_1 at the x locus and Y_1 at the y locus, and genotype 2 carries allele X_2 at the x locus and Y_2 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 [37]). 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

As in the $n = 2$ case developed above, we are again interested in the statistical properties of the genotypes of maximum fitness. If populations consist of n genotypes, the maximal fitness genotypes will have total fitness denoted by random variable $Z^{[n]}$, the top order statistic of total fitness Z. We are more interested, however, in the *concomitants* of $Z^{[n]}$, namely, random variables $X_{(n)}$ and $Y_{(n)}$, defined by the relation $Z^{[n]} = \phi(X_{(n)}, Y_{(n)})$. In particular, we are interested in the covariance between the concomitants, $\mathrm{cov}(X_{(n)}, Y_{(n)})$, because changing the sign of this value gives the selective advantage of recombinants (see PRE2 [2]).

Before analyzing concomitants of the top order statistic, however, the first step is to derive a general relation between a random variable Z and its concomitants X and Y when these concomitants are defined as $X + Y = Z$.

General relation between Z and its summand concomitants X and Y

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 $Z_i = X_i + Y_i$.

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:

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

The P_i thus define the discrete fitness distribution governing Z. Finally, let $(X, Y) = (X_I, Y_I)$. The goal is to derive statistical properties of concomitants X and Y of random variable $Z = X + Y$.

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:

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

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:*

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

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

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

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:

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

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:

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

□

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:*

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

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

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,

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

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:

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

□

Summand concomitants $X_{(k)}$ and $Y_{(k)}$

We define new random variable

$$S = \frac{X+Y}{\sqrt{\sigma_X^2+\sigma_Y^2+2\sigma_{XY}}},$$

which is $\mathcal{N}(0,1)$. The kth order statistic of S is denoted $S^{[k]}$ and is related to its concomitant summands as:

$$S^{[k]} = \frac{(X_{(k)}-\mu_X)+(Y_{(k)}-\mu_Y)}{\sqrt{\sigma_X^2+\sigma_Y^2+2\sigma_{XY}}}$$

[38]. Rearranging gives:

$$X_{(k)}+Y_{(k)} = \mu_X+\mu_Y+\sqrt{\sigma_X^2+\sigma_Y^2+2\sigma_{XY}}S^{[k]}$$

from which we have:

$$\mathrm{var}(X_{(k)}+Y_{(k)}) = (\sigma_X^2+\sigma_Y^2+2\sigma_{XY})\mathrm{var}(S^{[k]})$$

Plugging this expression into Eqs (6), (7) and (8) leads to the following corollary.

Corollary 1. *Define random variable $S \sim \mathcal{N}(0,1)$ whose k^{th} order statistic from a sample of size n is denoted $S^{[k]}$. If we blindly (and wrongly) assume that order statistic distributions are normal, the first- and second-order moments of the concomitants are nevertheless exact:*

$$\mathbb{E}[X_{(k)}] = \delta+\eta_x\Gamma\mathbb{E}[S^{[k]}] \quad (9)$$

$$\mathbb{E}[Y_{(k)}] = -\delta+\eta_y\Gamma\mathbb{E}[S^{[k]}] \quad (10)$$

$$\mathrm{var}(X_{(k)}) = \gamma+\eta_x^2\Delta\mathrm{var}(S^{[k]}) \quad (11)$$

$$\mathrm{var}(X_{(k)}) = \gamma+\eta_y^2\Delta\mathrm{var}(S^{[k]}) \quad (12)$$

$$\mathrm{cov}(X_{(k)},Y_{(k)}) = -\gamma+\eta_x\eta_y\Delta\mathrm{var}(S^{[k]}) \quad (13)$$

where $\Gamma=\mu_x+\mu_y+\sqrt{\Delta}$, $\Delta=\sigma_X^2+\sigma_Y^2+2\sigma_{XY}$, and $X_{(k)}+Y_{(k)}=Z^{[k]}$, the kth order statistic in total fitness.

Comparison with simulations show the foregoing expressions to be extremely accurate. And comparison with previous studies that take a more circuitous route in different contexts [39–41] show these expressions to be exact. While our analysis is more compact than those previous studies, we suspect our approach would not be exact for higher moments.

In general, we are most interested in the top order statistic, $k=n$, because natural selection will tend to "select" the fittest genotype. In an infinite population, selection is completely deterministic and the fittest will always be fixed. We note that our findings are only weakly dependent on this assumption of deterministic selection, because the variance of top order statistics are often quite similar; hence, our findings remain relatively unchanged if suboptimal genotypes $k=n-1$ or $k=n-2$ are selected due to finite-population effects (drift). We show in [2] that the mean selective advantage of recombinants will be:

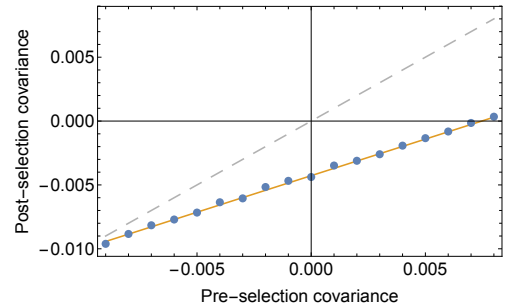

Figure 4. Covariance before and after selection. Blue dots plot covariance across simulations of 5000 subpopulations, each containing $n=20$ distinct genotypes and a bivariate normal distribution with means equal to -0.1, variances equal to 0.01, and covariance indicated on the horizontal axis. Orange line plots theoretical prediction given by Eq (14). Gray dashed line plots $y=x$ as a visual guide. Post-selection covariance is suppressed more when pre-selection covariance is strongly positive.

$$\bar{s}_R = -\mathrm{cov}(X_{(n)},Y_{(n)}) = -\sigma'_{XY}$$

where the final step is simply a change of notation. We define σ'_{XY} to be the covariance of the concomitants of the top order statistic, i.e., the covariance between X and Y after natural selection has acted locally in each of the subpopulations; σ_{XY} retains its meaning as general covariance (not covariance of concomitants) between X and Y in the initial population. More simply, σ_{XY} is pre-selection covariance and σ'_{XY} is post-selection covariance.

We further define the variance of the top order statistic of a standard normal random variable: $\sigma_n^2=\mathrm{var}(S^{[n]})$, which has the property $0<\sigma_n^2\le 1$ [38, 42]. Full general expressions are given in the SM. Two simplified cases are illuminating and are discussed here:

The first illuminating case is when X and Y are independent. In this case, Eq (13) becomes:

$$\sigma'_{XY} = \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}(\sigma_n^2-1) < 0$$

because two or more genotypes ($n\ge 2$) are required for recombination to make a difference, and $\sigma_n^2<1$ for $n\ge 2$. In words, after natural selection has run its course in local subpopulations, recombination across those local subpopulations will be advantageous.

The second illuminating case is when $\sigma_X=\sigma_Y=\sigma$. In this case, we have the following equivalent expressions:

$$\sigma'_{XY} = \frac{1}{2}(\sigma_{XY}(1+\sigma_n^2)-\sigma^2(1-\sigma_n^2)) \quad (14)$$

$$\sigma'_{XY} = \frac{1}{2}\sigma^2(\rho(1+\sigma_n^2)-(1-\sigma_n^2)) \quad (15)$$

where $\rho=\sigma_{XY}(\sigma_X\sigma_Y)^{-1}$, the pre-selection correlation coefficient.

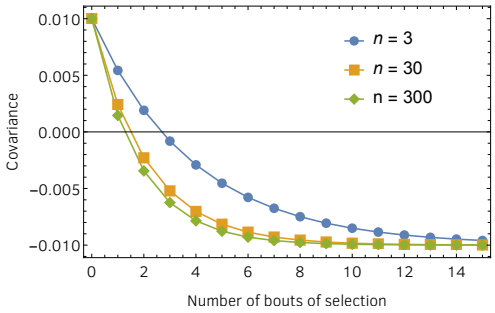

Figure 5. Covariance dynamics over several bouts of selection. Initially, we assign covariance its maximal possible value of $\sigma_X\sigma_Y = 0.01$ in order to illustrate the fact that, at least in theory (under an extreme condition), it may take more than one bout of selection for covariance to become negative. Covariance does become negative rather quickly, however, and converges to the predicted value of $-\sigma_X\sigma_Y = -0.01$, which is the minimal value for covariance.

The first thing to notice is that the effect of natural selection is always to reduce covariance by an amount whose lower bound depends only on the number of competing genotypes:

$$\begin{aligned}\sigma'_{XY} &= \frac{1}{2}(\sigma_{XY}(1+\sigma_n^2) - \sigma^2(1-\sigma_n^2)) \\ &\leq \frac{1}{2}(\sigma_{XY}(1+\sigma_n^2) - \sigma_{XY}(1-\sigma_n^2)) \\ &= \sigma_{XY}\sigma_n^2\end{aligned}$$

In the above expressions, it is apparent that post-selection covariance can in theory be positive if pre-selection covariance is strongly positive. (In other words, post-selection recombinant advantage can in theory be negative if pre-selection recombinant advantage is strongly negative.) The condition for covariance to be negative after a single bout of selection is best expressed as a condition on pre-selection correlation; post-selection covariance will be negative when:

$$\rho < \frac{1-\sigma_n^2}{1+\sigma_n^2} > 0 \ .$$

In theory, at least, one bout of selection may not result in negative covariance. Several bouts of selection, however, are guaranteed to result in negative covariance. The equilibrium covariance achieved after many bouts of selection can be determined by setting $\sigma'_{XY} = \sigma_{XY}$ in Eq (14), giving:

$$\sigma_{XY} \xrightarrow{t} -\sigma^2 < 0$$

For the general case where the variances are not equal, $\sigma_{XY} \xrightarrow{t} -\sigma_X^2$ and $\sigma_{XY} \xrightarrow{t} -\sigma_Y^2$ are both stable equilibria.

V. HETEROSIS

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

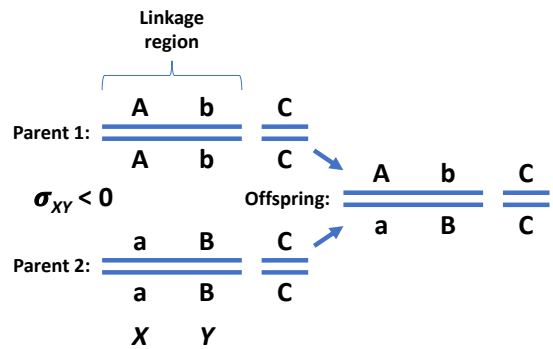

Figure 6. A novel theoretical basis for pseudo-overdominance and heterosis. Upper case (lower case) letters denote higher- (lower-) fitness alleles. Generally speaking, lower-fitness alleles tend to be recessive. Heterozygotes will therefore tend to express the higher-fitness of the two alleles at a 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*) [43–47]. Our theory shows how the required negative covariance in genic fitnesses across linked loci is an unavoidable consequence of natural selection, and thus provides a novel theoretical basis for heterosis.

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) [14, 16, 44, 48, 49], 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.

Increasingly detailed studies of heterosis 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 [50–52]. 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 neg-

ative associations in genic fitnesses across populations, out-crossing between populations can give the appearance of overdominance as fitter dominants mask less-fit recessives – a phenomenon that has been dubbed *pseudo-overdominance* [45, 50–53]. (What we are here calling "selective mismatch" has elsewhere been called "linkage repulsion" [52–54], "linkage bias" [51], or "linkage disequilibrium" (LD) [47].) If the out-crossed parents come from different inbred populations – a 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 6.

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) [51, 52]. Some authors have made verbal arguments invoking a combination of mutation, weak mutational effect and small effective population size to meet this requirement [43]. In a very recent paper, Waller [52] 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].

VI. CONCLUDING REMARKS

To summarize what has been modeled in this paper, we revisit our definition of "products of selection". These are genotypes that are locally prevalent, due to natural selection. Products of selection 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.

In simulations (SM), we placed such recombinant offspring in head-to-head competition with non-recombinant offspring, with no further recombination occurring during the competition. We found that the recombinant offspring displaced the non-recombinant offspring $> 95\%$ of the time under a wide range of conditions.

The mathematical analyses in this study is a bit more restrictive than our analyses in companion paper [2], most of which has zero dependence on the initial parent distribution of genic fitnesses. Here, our mathematical analyses eventually require: 1) an assumption that the initial distribution is symmetric, for the 2-locus, 2-genotype case, and 2) an assumption that the initial distribution is normal, for the 2-locus, n-genotype case. Our finding that the lower central moments of concomitants $X_{(k)}$ and $Y_{(k)}$ are exact despite a bold assumption of normality is at least suggestive that our findings might be robust more generally, i.e., to non-normal parent distributions. Furthermore, our qualitative results are corroborated by simulations with a wide variety of divergent parent distributions (SM).

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.

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

[4] Redfield, A truly pluralistic view of sex and recombination, J. Evol. Biol. **12**, 1043 (1999).

[5] S. A. West, C. M. Lively, and A. F. Read, A pluralist approach to sex and recombination, J. Evol. Biol. **12**, 1003 (1999).

[6] J. A. G. M. de Visser and S. F. Elena, The evolution of sex: empirical insights into the roles of epistasis and drift, Nat. Rev. Genet. **8**, 139 (2007).

[7] S. P. Otto, The evolutionary enigma of sex, Am. Nat. **174 Suppl 1**, S1 (2009).

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

[9] M. Hartfield and P. D. Keightley, Current hypotheses for the evolution of sex and recombination, Integr. Zool. **7**, 192 (2012).

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

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

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

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

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

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

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

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

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

[19] P. Vorzimmer, Charles darwin and blending inheritance, Isis **54**, 371 (1963).

[20] D. M. Waller and L. F. Keller, Inbreeding and inbreeding depression, in *Evolutionary Biology* (Oxford University Press, 2020).

[21] R. W. Siegel, Hybrid vigor, heterosis and evolution in paramecium aurelia, Evolution **12**, 402 (1958).

[22] R. C. Lewontin, Genetics. (book reviews: The theory of inbreeding), Science **150**, 1800 (1965).

[23] R. C. Lewontin, The theory of inbreeding. sir ronald a. fisher. academic press, new york, ed. 2, 1965. viii + 150 pp. illus. $6, Science **150**, 1800 (1965).

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

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

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

[27] 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).

[28] 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).

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

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

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

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

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

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

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

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

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

[38] H. A. David, *Order Statistics* (Wiley, 1970).

[39] N. Balakrishnan, Multivariate normal distribution and multivariate order statistics induced by ordering linear combinations, Stat. Probab. Lett. **17**, 343 (1993).

[40] R. Song and J. A. Deddens, A note on moments of variables summing to normal order statistics (1993).

[41] R. Song, S. G. Buchberger, and J. A. Deddens, Moments of variables summing to normal order statistics, Stat. Probab. Lett. **15**, 203 (1992).

[42] H. J. Godwin, Some low moments of order statistics (1949).

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

[44] 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).

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

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

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

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

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

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

[51] 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.

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

[53] X. Li, X. Li, E. Fridman, T. T. Tesso, and J. Yu, Dissecting 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).

[54] 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).