Natural selection and the advantage of recombination

Philip J Gerrish[1,2,3,*], Benjamin Galeota-Sprung[4], Fernando Cordero[5], Paul Sniegowski[4], Alexandre Colato[6], Nicholas Hengartner[2], Varun Vejalla[7], Julien Chevallier[8] & Bernard Ycart[8]

[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] *Biomathematics and Theoretical Bioinformatics, Technische Fakultät, Universität Bielefeld, Germany*
[6] *Ciências da Natureza, Matemática e Educação, Univ Fed de São Carlos, Araras SP, Brazil*
[7] *Thomas Jefferson High School for Science and Technology, Alexandria, Virginia, USA*
[8] *Mathematique Apliquée, Laboratoire Jean Kuntzmann, Université Grenoble Alpes, France*

Exchanging genetic material with another individual seems risky from an evolutionary standpoint, and yet living things across all scales and phyla do so quite regularly. The pervasiveness of such genetic exchange, or recombination, in nature has defied explanation since the time of Darwin[1–4]. Conditions that favor recombination, however, are well-understood: recombination is advantageous when the genomes of individuals in a population contain more selectively mismatched combinations of alleles than can be explained by chance alone. Recombination remedies this imbalance by shuffling alleles across individuals. The great difficulty in explaining the ubiquity of recombination in nature lies in identifying a source of this imbalance that is comparably ubiquitous. Intuitively, it would seem that natural selection should reduce the imbalance by favoring selectively matched combinations of high-fitness alleles, thereby opposing the evolution of recombination. We show, however, that this intuition is wrong; to the contrary, we find that natural selection has an encompassing tendency to assemble selectively mismatched combinations of alleles (the *products* of natural selection), thereby increasing the imbalance and promoting the evolution of recombination. We further show that population dynamics that lead to the fixation of these selectively mismatched genotypes (the *process* of natural selection) themselves produce an average imbalance that promotes the evolution of recombination. This fact is completely independent of the distribution of allelic fitness effects and is primarily due to the additive component of those effects. Our findings provide a novel vantage point from which the enormous body of established work on the evolution of sex and recombination may be viewed anew. They further suggest that recombination evolved and is maintained more as an unavoidable byproduct of natural selection than as a catalyst.

The ability to exchange genetic material through recombination (and sex) is a heritable trait[5,6] that is influenced by many different evolutionary and ecological factors, both direct and indirect, both positive and negative. Evidence from nature clearly indicates that the net effect of these factors must be positive: recombination across all levels of organismal size and complexity is undeniably the rule rather than the exception[2–4,7]. Theoretical studies, on the other hand, have revealed a variety of different mechanisms and circumstances that can promote the evolution of recombination, but each one by itself is of limited scope[2,4,8]. These studies would thus predict that the absence of recombination is the rule and its presence an exception. The sheer abundance of these exceptions, however, can be seen as amounting to a rule in its own right – a "pluralist" view that has been adopted by some authors to explain the ubiquity of recombination[3,7,9]. The necessity of this pluralist view, however, may be seen as pointing toward a fundamental shortcoming in existing theory: perhaps some very general factor that would favor recombination has been missing[3,4,8,10].

Existing theories of the evolution and maintenance of sex and recombination can be divided into those that invoke *direct* vs *indirect* selection on recombination. Theories invoking direct selection propose that recombination evolved and is maintained by some physiological effect that mechanisms of recombination themselves have on survival or on replication efficiency. Such theories might speak to the origins of sex and recombination but they falter when applied to their maintenance[1]. Most theories invoke indirect selection: they assume that any direct or immediate effect of recombination mechanisms is small compared to the trans-generational consequences of recombination.

While differing on the causal factors involved, established theoretical approaches that invoke indirect selection are unanimous in their identification of the fundamental selective environment required for sex and recombination to evolve: a population must harbor an excess of selectively mismatched combinations of alleles across loci and a deficit of selectively matched combinations. Recombination is favoured under these conditions because on average it breaks up the mismatched combinations and assembles matched combinations. Assembling selectively matched combinations increases the efficiency of natural selection: putting high-fitness alleles together can expedite their fixation[11–15], and putting low-fitness alleles together can expedite their elimination[16,17]. Under these conditions therefore, populations with recombination have an evolutionary advantage over populations without. Furthermore, competition among recombination-rate variants at a *modifier* locus under these conditions will increase recombination within a single population[18]. But it is also true that recombinants themselves formed from randomly-chosen parents within such a population have an advantage that is relatively immediate, and this will be the measure we use to assess the evolutionary potential of recombination.

The great challenge has been to identify an evolutionary source of the aforementioned imbalance whose prevalence in nature is comparable to the prevalence of sex and recombination in nature. One feature of living things whose prevalence approximates that of sex and recombination is right under our noses, namely, evolution by natural selection. This observation inspired the aim of the present study, which is to assess the effects of pure natural selection on the evolutionary potential of recombination.

We preface our developments with an essential technical point. In much of the relevant literature, the measure of selective mismatch across loci affecting the evolution of recombination is linkage disequilibrium (LD)[8,12,13,19–21], which measures bias in allelic frequencies across loci but does not retain information about the selective value of those alleles. Here, the objectives of our study require a slight departure from tradition: our measure of selective mismatch will be covariance between genic fitnesses. This departure is necessary 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 (Methods and Extended Data Fig. 1). Our results will thus be given in terms of covariance and we recall: negative covariance means positive selection for recombinants.

Setting

We reduce the problem to what we believe is its most essential form: we ask how the selective value of haploid recombinants is affected when natural selection simply acts on standing heritable variation. We ask this: 1) when recombination occurs across the *products* of natural selection (e.g., fixed genotypes), and 2) when recombination occurs within a population during the *process* of natural selection.

To eliminate potentially confounding factors and study the effects of natural selection in isolation, we consider large (effectively infinite) populations, each of which consists of just two competing genotypes that differ in both of two genes (or two *loci*). This simple setting permits

clean presentation and mathematical tractability and, more importantly, is biologically motivated by the observation that large clonal populations tend to be overwhelmingly dominated by one or two genotypes[22]. It further provides a connection to foundational evolution-of-sex studies[23–25]: Fisher[24] considered the case of a single beneficial mutation arising on a variable background, thereby effectively giving rise to two competing genotypes – wildtype and beneficial mutant – that differ in both the gene with the beneficial mutation (call it the x gene) and its genetic background (call it the y gene); Muller[25] considered the case of two competing genotypes, one carrying a beneficial mutation in the x gene and the other in the y gene. Both of these approaches consider two competing genotypes that differ in both of two loci, and these foundational models are thus subsumed by our framework. Simulations further confirm the adequacy of this two-genotype setting: increasing the number of genotypes only accentuates the effects we describe.

We consider a clonal haploid organism whose genome consists of just two fitness-related loci labeled x and y. Genetically-encoded phenotypes at these two loci are quantified by random variables X and Y, both of which are positively correlated with fitness. In each large population of such organisms, two genotypes exist: one encodes the phenotype (X_1, Y_1), has fitness $Z_1 = \phi(X_1, Y_1)$ and exists at some arbitrary initial frequency p; the other encodes phenotype (X_2, Y_2), has fitness $Z_2 = \phi(X_2, Y_2)$ and exists at initial frequency $1 - p$. We note that, in the absence of epistasis or dominance, the scenario we describe is formally equivalent to considering a diploid organism whose genome consists of one locus and two alleles available to each haploid copy. The question we ask is this: Does the action of natural selection, by itself, affect covariance between X and Y, denoted σ_{XY}, and if so, how?

The *products* of natural selection promote recombination

Figure 1 illustrates the problem by analogy to a set of canoe races. Figure 2 shows how the problem is posed analytically. On the surface, one might suspect that natural selection would promote well-matched combinations in which large values of X are linked to large values of Y, thereby creating a positive association between X and Y. In fact, this notion is so intuitive that it is considered self-evident, explicitly or implicitly, in much of the literature[1–3,7,9,14,26,27]. If this notion were true, recombination would break up good allelic combinations, on average, and should thus be selectively suppressed. Such allele shuffling has been called "genome dilution", a label that betrays its assumed costliness. We find, however, that the foregoing intuition is wrong. To the contrary, we find that natural selection will, on average, promote an excess of mismatched combinations in which large values of X are linked to small values of Y, or vice versa, thereby creating a negative association between X and Y. Recombination will on average break up the mismatched combinations created by natural selection, assemble well-matched combinations, and should thus be favoured.

Figure 3 illustrates why our initial intuition was wrong and why natural selection instead tends to create negative fitness associations among genes. For simplicity of presentation, we assume here that an individual's fitness is $Z = \phi(X, Y) = X + Y$, i.e., that X and Y are simply additive genic fitness contributions, and that X and Y are independent. In the absence of recombination, selection does not act independently on X and Y but on their sum, $Z = X + Y$. Perhaps counter-intuitively, this fact alone creates negative associations. To illustrate, we suppose that we

know the fitness of successful genotypes – the *products* of natural selection – to be some constant, z, such that $X + Y = z$; here, we have the situation illustrated in Fig 3a and we see that X and Y are negatively associated; indeed, covariance is immediate: $\sigma_{XY} = -\sigma_X \sigma_Y \leq 0$. Of course, in reality the fitnesses of successful genotypes will not be known *a priori* nor will they be equal to a constant; instead, they will follow a distribution of maxima of Z as illustrated in Fig 3b. If populations consist of n contending genotypes, then $X_{(n)} + Y_{(n)} = Z^{[n]}$, the nth order statistic of Z with genic components $X_{(n)}$ and $Y_{(n)}$ (called *concomitants* in the probability literature[28]). In general, $Z^{[n]}$ will have smaller variance than Z. Components $X_{(n)}$ and $Y_{(n)}$, therefore, while not exactly following a line as in Fig 3a, will instead be constrained to a comparatively narrow distribution about that straight line, illustrated by Fig 3b, thereby creating a negative association. Figure 3c plots ten thousand simulated populations evolving from their initial (green dots) to final (black dots) mean fitness components; this panel confirms the predicted negative association.

What we have shown so far is that, if recombination occurs across the *products* of natural selection, such as fixed genotypes in different populations, subpopulations, demes or competing clones, the resulting offspring should be more fit than their parents, on average. This effect provides novel insight into established observations that population structure can favor recombination[29–33] and may even speak to notions that out-crossing can create hybrid vigour (heterosis).

The *process* of natural selection promotes recombination

Here, our focus is on recombinants formed from two randomly-chosen parents within the same unstructured evolving population. Here again, Fig 2 shows how the problem is posed analytically. Natural selection will cause the two competing genotypes to change in frequency, causing σ_{XY} to change over time ($\sigma_{XY} = \sigma_{XY}(t)$). If $\int_0^\infty \sigma_{XY}(t)dt$ is positive (negative) on average, this indicates that natural selection creates conditions that oppose (favour) recombination on average.

In the Methods, we show that, in expectation, covariance over the long run is unconditionally non-positive, $\mathbb{E}[\int_0^\infty \sigma_{XY}(t)dt] \leq 0$, implying that the *process* of natural selection, on average, always creates conditions that favour recombination. Remarkably, this finding requires no assumptions about the distribution of X and Y; in fact, a smooth density is not required. Indeed, this distribution can have strongly positive covariance, and yet the net effect of natural selection is still to create negative time-integrated covariance. Montecarlo integration shows that time-integrated covariance is indeed negative under a wide range of different distributions (Extended Data Fig. 2), and recombinant advantage increases as the number of alleles increases (Extended Data Fig. 3).

We further show that it is primarily the additive component of fitness that causes time-integrated covariance to be negative. This fact stands in contrast with previous notions that non-additive effects, specifically negative or fluctuating epistasis, are an essential ingredient in the evolution of recombination[19,21,34–37].

Discussion

Some authors[2,38] have argued that negative associations build up within a population because positive associations, in which alleles at different loci are selectively matched, are either removed efficiently (when they are both similarly deleterious), or fixed efficiently (when they are both similarly beneficial), thereby contributing little to overall within-population associations. Genotypes that are selectively mismatched, on the other hand, have longer sojourn times, as the less-fit loci effectively shield linked higher-fitness loci from selection. The net effect, it is argued, should be that alleles across loci will on average be selectively mismatched within a population. Our findings differ from these arguments in one respect, namely, we find that even genotypes that are ultimately fixed carry selectively mismatched alleles. In another respect, however, these arguments are entirely consistent with our findings: Equation (1) in Proposition 6 gives time-integrated covariance; it is intuitively more likely that the numerator of that equation will be negative when the denominator is small, i.e., when Z_1 and Z_2 are close to each other. Negative values are thus amplified because they tend to occur when total fitness of the two genotypes are close to each other and thus coexist for a longer period of time before one displaces the other. Indeed this is the intuitive way to understand Proposition 7.

We have identified a phenomenon that is an inherent consequence of natural selection and gives rise to selectively mismatched combinations of alleles across loci. Generally speaking, this pervasive phenomenon is an example of counter-intuitive effects caused by probabilistic conditioning. For example, "Berkson's bias"[39,40] arises when a biased observational procedure produces

spurious negative correlations. In the original context, among those admitted to hospital due to illness, a negative correlation among potentially causative factors was observed because those with no illness (who tended to have no causative factors) were not admitted to the hospital and hence not observed. Similarly, negative correlations arise across genic fitnesses in part because genotypes in which both loci have low genic fitness are purged by selection; here, however, the bias is not observational but actual, as these low-fitness genotypes no longer exist in the population.

Several relevant issues are beyond the scope of this study. For example, our approach implicitly assumes that recombination will evolve as a result of persistent recombinant advantage, but we have not explicitly shown this. In simulations (Extended Data Figs. 4-7), we show that our reductionist approach does indeed cause recombination modifiers to increase in frequency. These simulations are perhaps superfluous, however, given our finding that recombinants themselves are chronically advantageous. Also, we have omitted mutation, which was necessary in order to study the effects of natural selection in isolation.

Many previous studies, in one way or another, point to the increase in agility and efficiency of adaptation that recombination confers as the primary cause of its evolution. Here, we invert the perspective of those earlier studies, asking not whether recombination speeds adaptation, but whether adaptation via natural selection generally creates selective conditions that make recombinants directly and immediately advantageous. If so, as our findings indicate, then: 1) the ubiquity of recombination in nature might be less enigmatic than previously thought, and 2) perhaps recombination arose and is maintained more as an unavoidable byproduct than as a catalyst of natural

selection.

Methods

Notes. In the main text, we employ the shorthand σ_{XY} to denote covariance. In what follows, however, we use σ_{XY} and $\mathrm{Cov}(X, Y)$ (for clarity) interchangeably. Several of the proofs here are abridged; full proofs are in the SI, as well as alternative and supplemental proofs.

Covariance and recombinant advantage. Much work on the evolution of recombination employs linkage disequilibrium (LD) as the measure of across-loci associations. It is straight-forward to estimate LD from genomic sequence data, which likely explains the popularity of this measure. LD, however, contains no information about the selective cost of such associations. Covariance, on the other hand, retains all of the information regarding both the prevalence of linkage and its selective cost (i.e., recombinant advantage), and is thus the measure we employ. We note that when the fitness function is a bivariate Bernoulli distribution ($\phi(X, Y) = \mathbb{P}\{X = i, Y = j\} = p_{i,j},\ i, j \in \{0, 1\}$) then covariance and disequilibrium are equivalent ($\sigma_{XY} = D = p_{1,1} - p_{1,\bullet}p_{\bullet,1}$). Recombinants are formed from two randomly-chosen contemporaneous parents such that their genetic makeup is simply an unbiased random sampling of the pool of available alleles at the x and y loci. As such, their instantaneous advantage is zero on average: $\mathbb{E}_R[X + Y] - \mathbb{E}[X + Y] = 0$, where subscript R denotes recombinant and no subscript denotes wildtype. Recombinants and wildtype, however, gain fitness at different rates: $\partial_t \mathbb{E}_R[X+Y] = \sigma_X^2 + \sigma_Y^2$ and $\partial_t \mathbb{E}[X+Y] = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$. A first order expansion thus reveals that the selective advantage of recombinants after a single generation of growth is $\partial_t \mathbb{E}_R[X+Y] - \partial_t \mathbb{E}[X+Y] = -2\sigma_{XY}$. A single-generation Moran

model (Extended Data Fig. 1) shows this prediction to be accurate and that covariance increases linearly in the first generation, implying that the mean selective advantage of recombinants over that first generation is $-\sigma_{XY}$. We note that a generalized linear relationship between fitness and phenotypes X and Y, i.e., $Z = k_0 + k_X X + k_Y Y$, yields a recombinant advantage of $-k_X k_Y \sigma_{XY}$. A full treatment of the relation between covariance and recombinant advantage is found in the SI, as well as the relation between our approach and classical population genetics.

The *products* of natural selection promote recombination. The setting for this problem is shown in Fig 2. No hypothesis on the fitness function ϕ is made at this point, apart from being measurable. For the sake compact presentation we assume here (relaxed in SI) that (X_1, Y_1, X_2, Y_2) are i.i.d.; departures from this and other simplifying assumptions are dealt with in the SI. As defined in Fig 2, $Z_i = \phi(X_i, Y_i)$, $Z^{[i]} = \phi(X_{(i)}, Y_{(i)})$, and $Z^{[2]} > Z^{[1]}$.

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 $I \in \{1, 2\}$, and for now $\mathbb{P}(I = 1) = \mathbb{P}(I = 2) = 1/2$, and I is independent of (X_1, Y_1, X_2, Y_2). The couple (X_I, Y_I) is distributed as (X_1, Y_1). Hence, $\mathbb{E}(\psi(X_I, Y_I)) = \mathbb{E}(\psi(X_1, Y_1))$, however, $\mathbb{E}(\psi(X_I, Y_I)) = \mathbb{E}(\mathbb{E}(\psi(X_I, Y_I) \,|\, I)) = \frac{1}{2}\mathbb{E}(\psi(X_{(1)}, Y_{(1)})) + \frac{1}{2}\mathbb{E}(\psi(X_{(2)}, Y_{(2)}))$. □

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

PROOF: The couples $(X_{(I)}, Y_{(I)})$ and $(X_{(I)}, Y_{(3-I)})$ are both distributed as (X_1, Y_1). There-

fore their covariances are null. These covariances can also be computed by conditioning on I (see *e.g.* formula (1.1) in [41]). For $(X_{(I)}, Y_{(I)})$ we have: $\mathrm{Cov}(X_{(I)}, Y_{(I)}) = \mathbb{E}(\mathrm{Cov}(X_{(I)}, Y_{(I)}|I)) + \mathrm{Cov}(\mathbb{E}(X_{(I)}|I), \mathbb{E}(Y_{(I)}|I))$. 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 fitness 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 (Z_1, Z_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 i and s of minimal and maximal sum. □

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

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

The *process* of natural selection promotes recombination. We recall that recombinant advantage is $-\sigma_{XY}$. Here, we study how the selection-driven changes in types (X_1, Y_1) and (X_2, Y_2) *within a single unstructured population* change $\sigma_{XY} = \sigma_{XY}(t)$ over time. We are interested in the net effect of these changes, given by $\int_0^\infty \sigma_{XY}(t)dt$; in particular, we are interested in knowing whether this quantity is positive (net recombinant disadvantage) or negative (net recombinant advantage).

PROPOSITION 6. *Within-population covariance integrated over time is:*

$$\int_0^\infty \sigma_{XY}(t)dt \ = \ q\mathbb{E}[\frac{(X_2 - X_1)(Y_2 - Y_1)}{|Z_2 - Z_1|}] \tag{1}$$

where q is the initial frequency of the inferior genotype. No assumption about the distribution of (X, Y) is required. And $Z_i = \phi(X_i, Y_i)$ where fitness function ϕ can be any function.

PROOF: We let p denote initial frequency of the superior of the two genotypes, and we let $q = 1-p$ denote initial frequency of the inferior genotype. Time-integrated covariance is:

$$\int_0^\infty \sigma_{X,Y}(t)dt = \mathbb{E}[(X_{(2)} - X_{(1)})(Y_{(2)} - Y_{(1)}) \int_0^\infty \frac{pq\mathrm{e}^{\left(Z^{[1]}+Z^{[2]}\right)t}}{\left(p\mathrm{e}^{Z^{[2]}t} + q\mathrm{e}^{Z^{[1]}t}\right)^2} dt]$$

Integration by parts yields:

$$\int_0^\infty \sigma_{XY}(t)dt = q\mathbb{E}[\frac{(X_{(2)} - X_{(1)})(Y_{(2)} - Y_{(1)})}{Z^{[2]} - Z^{[1]}}]$$

where q in Prop 6 is written as $1 - p_0$. We observe that:

$$(X_{(2)} - X_{(1)})(Y_{(2)} - Y_{(1)}) = (X_{(1)} - X_{(2)})(Y_{(1)} - Y_{(2)}) = (X_2 - X_1)(Y_2 - Y_1)$$

and that

$$Z^{[2]} - Z^{[1]} = |Z_2 - Z_1|$$

from which we have:

$$\mathbb{E}[\frac{(X_{(2)} - X_{(1)})(Y_{(2)} - Y_{(1)})}{Z^{[2]} - Z^{[1]}}] = \mathbb{E}[\frac{(X_2 - X_1)(Y_2 - Y_1)}{|Z_2 - Z_1|}]$$

□

PROPOSITION 7. *We define spacings $\Delta X = X_2 - X_1$, $\Delta Y = Y_2 - Y_1$, and $\Delta Z = Z_2 - Z_1 = \Delta X + \Delta Y$. If the pairs (X_i, Y_i) are independently drawn from any distribution, then ΔX and ΔY are symmetric about zero, and time-integrated covariance is unconditionally non-positive:*

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

PROOF: There is no need to assume that $(\Delta X, \Delta Y)$ has a density. This proof also reveals that the result also holds for discrete random variables. Let ΔX, ΔY be two real-valued random variables

such that: $(-\Delta X, \Delta Y)$ has the same distribution as $(\Delta X, \Delta Y)$. We have:

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

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

COROLLARY 1. *Proposition 7 holds for divergent expectations.*

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

$$\begin{aligned}
\mathbb{E}[UV\{1/(U+V) - 1/(|U-V|)\}] &= \mathbb{E}[mM\{-2m/(M^2 - m^2)\}] \\
&= -2\mathbb{E}[Mm^2/(M^2 - m^2)] \leq 0
\end{aligned}$$

Indeed, if the expectation is divergent, then it is always $-\infty$. This approach removes the need to make the argument that $U + V > |U - V|$ and avoids the need to take a difference of expectations. An alternative approach is given in an expanded statement and proof of Proposition 7 in the SI. □

1. de Visser, J. A. G. M. & Elena, S. F. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat. Rev. Genet.* **8**, 139–149 .

2. Otto, S. P. The evolutionary enigma of sex. *Am. Nat.* **174 Suppl 1**, S1–S14 .

3. Otto, S. P. & Lenormand, T. Resolving the paradox of sex and recombination. *Nat. Rev. Genet.* **3**, 252–261 .

4. Barton, N. H. & Charlesworth, B. Why sex and recombination? *Science* **281**, 1986–1990 .

5. Bodmer, W. F. & Parsons, P. A. Linkage and recombination in evolution. In Caspari, E. W. & Thoday, J. M. (eds.) *Advances in Genetics*, vol. 11, 1–100 .

6. Nei, M. Modification of linkage intensity by natural selection. *Genetics* **57**, 625–641 .

7. Hartfield, M. & Keightley, P. D. Current hypotheses for the evolution of sex and recombination. *Integr. Zool.* **7**, 192–209 .

8. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 .

9. West, S. A., Lively, C. M. & Read, A. F. A pluralist approach to sex and recombination. *J. Evol. Biol.* .

10. Otto, S. P. & Barton, N. H. Selection for recombination in small populations. *Evolution* **55**, 1921–1931 .

11. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102-103**, 127–144 .

12. Otto, S. P. & Barton, N. H. The evolution of recombination: removing the limits to natural selection. *Genetics* **147**, 879–906 .

13. Barton, N. H. Linkage and the limits to natural selection. *Genetics* **140**, 821–841 .

14. Agrawal, A. F. Evolution of sex: Why do organisms shuffle their genotypes? *Curr. Biol.* **16**, R696–R704 .

15. Arjan, J. A. *et al.* Diminishing returns from mutation supply rate in asexual populations. *Science* **283**, 404–406 .

16. Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 .

17. Keightley, P. D. & Otto, S. P. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* **443**, 89–92 .

18. Felsenstein, J. & Yokoyama, S. The evolutionary advantage of recombination. II. individual selection for recombination. *Genetics* **83**, 845–859 .

19. Barton, N. H. A general model for the evolution of recombination. *Genet. Res.* **65**, 123–145 .

20. Barton, N. H. Genetic linkage and natural selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 2559–2569 .

21. Otto, S. P. & Feldman, M. W. Deleterious mutations, variable epistatic interactions, and the evolution of recombination. *Theor. Popul. Biol.* **51**, 134–147 .

22. Baake, E., González Casanova, A., Probst, S. & Wakolbinger, A. Modelling and simulating lenski's long-term evolution experiment. *Theor. Popul. Biol.* **127**, 58–74 .

23. Crow, J. F. & Kimura, M. Evolution in sexual and asexual populations. *Am. Nat.* **99**, 439–450 .

24. Fisher, R. A. *The genetical theory of natural selection* (OxfordClarendon Press, 1930).

25. Muller, H. J. Some genetic aspects of sex. *Am. Nat.* **66**, 118–138 .

26. Jaffe, K. Emergence and maintenance of sex among diploid organisms aided by assortative mating. *Acta Biotheor.* **48**, 137–147 .

27. Redfield. A truly pluralistic view of sex and recombination. *J. Evol. Biol.* **12**, 1043–1046 .

28. Yang, S. S. General distribution theory of the concomitants of order statistics. *Ann. Stat.* **5**, 996–1002 .

29. Martin, G., Otto, S. P. & Lenormand, T. Selection for recombination in structured populations. *Genetics* **172**, 593–609 .

30. Becks, L. & Agrawal, A. F. Higher rates of sex evolve in spatially heterogeneous environments. *Nature* **468**, 89–92 .

31. Hartfield, M., Otto, S. P. & Keightley, P. D. The maintenance of obligate sex in finite, structured populations subject to recurrent beneficial and deleterious mutation. *Evolution* **66**, 3658–3669 .

32. Whitlock, A. O. B., Azevedo, R. B. R. & Burch, C. L. Population structure promotes the evolution of costly sex in artificial gene networks. *Evolution* **73**, 1089–1100 .

33. Lenormand, T. & Otto, S. P. The evolution of recombination in a heterogeneous environment. *Genetics* **156**, 423–438 .

34. Kouyos, R. D., Otto, S. P. & Bonhoeffer, S. Effect of varying epistasis on the evolution of recombination. *Genetics* **173**, 589–597 .

35. Otto, S. P. & Michalakis, Y. The evolution of recombination in changing environments. *N. Jb. Geol. Paläont. Mh.* **486**, 516 .

36. Gandon, S. & Otto, S. P. The evolution of sex and recombination in response to abiotic or coevolutionary fluctuations in epistasis. *Genetics* **175**, 1835–1853 .

37. Peters, A. D. & Lively, C. M. The red queen and fluctuating epistasis: A population genetic analysis of antagonistic coevolution. *Am. Nat.* **154**, 393–405 .

38. Barton, N. H. & Otto, S. P. Evolution of recombination due to random drift. *Genetics* **169**, 2353–2370 .

39. Miller, J. B. & Sanjurjo, A. A bridge from monty hall to the hot hand: The principle of restricted choice. *J. Econ. Perspect.* **33**, 144–162 .

40. Berkson, J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics* **2**, 47–53 .

41. Joag-Dev, K. & Proschan, F. Negative association of random variables with applications. *Ann. Statist.* **11**, 286–295 .

Acknowledgements 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. 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. The authors thank D. Chencha, J. Streelman, R. Rosenzweig and the Biology Department at Georgia Institute of Technology for critical infrastructure and computational support. 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.

Author contributions P.G. conceived the theory conceptually; P.G., P.S., B.S. and A.C. developed the theory verbally and with simulation; P.G, B.Y. and J.C. developed the theory mathematically; B.Y. and J.C. provided mathematical proofs for the across-population part; P.G., V.V., F.C. and N.H. provided mathematical proofs for the within-population part. P.G. wrote the paper with critical help and guidance from B.S., P.S. and B.Y.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s....

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

Paddler strength

0.88	0.57	Race 1
0.59	1.23	
1.74	0.92	Race 2
0.87	0.53	
1.16	0.73	Race 3
1.67	0.43	

Fig 1 | Canoe race analogy. Each canoe contains two paddlers. The strength of each paddler is measured and reported in the table. In any given canoe race, there is no correlation between paddler strengths A and B. In each race, paddler strengths are recorded (tables on right), and the winning canoe is that in which the sum of the strengths of the two paddlers is the greatest (highlighted). Three such canoe races are conducted. We ask: what is the covariance between the strengths of paddlers A and B among winning canoes only? While it seems reasonable to suppose that winning canoes would carry two strong paddlers thereby resulting in positive covariance, the counter-intuitive answer we find is that the covariance is, for all practical purposes, unconditionally negative in expectation. By analogy, paddlers are genes, paddler strength is genic fitness, and canoes are genotypes. Natural selection picks the winner.

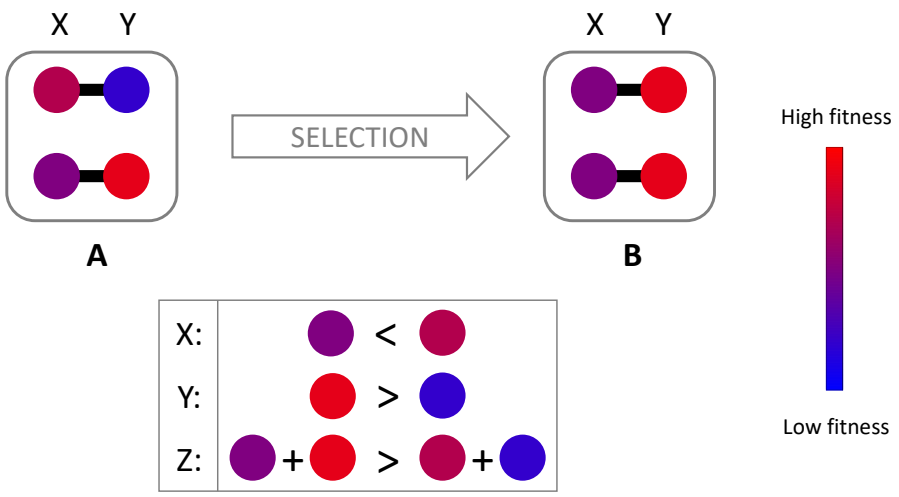

Fig 2 | Two loci, two alleles. Here, a large (infinite) population consists of individuals whose genome has only two loci x and y, each of which carries one of two alleles: genotype 1 encodes quantified phenotype X_1 at the x locus and Y_1 at the y locus, and genotype 2 carries quantified phenotype X_2 at the x locus and Y_2 at the y locus. Fitness is indicated by color. An individual's fitness is a function of the two phenotypes: $Z = \phi(X, Y)$; here we make the simplifying assumption that $\phi(X, Y) = X + Y$, so that the fitnesses of genotypes 1 and 2 are $Z_1 = X_1 + Y_1$ and $Z_2 = X_2 + Y_2$, respectively. The fitter of these two genotypes has total fitness denoted $Z^{[2]}$ (i.e., $Z^{[2]} = \text{Max}\{Z_1, Z_2\}$) and genic fitnesses $X_{(2)}$ and $Y_{(2)}$ (i.e., $Z^{[2]} = X_{(2)} + Y_{(2)}$). Similarly, the less-fit of these two genotypes has total fitness $Z^{[1]} = X_{(1)} + Y_{(1)}$. We note: $Z^{[2]} > Z^{[1]}$ by definition, but this does *not* guarantee that $X_{(2)} > X_{(1)}$ or that $Y_{(2)} > Y_{(1)}$, as illustrated in the lower box. The population labeled A consists of two distinct genotypes but selection acts to remove the inferior genotype leaving a homogeneous population in which individuals are all genetically identical (with fitness $Z^{[2]}$) as illustrated in the population labeled B. We derive selective mismatch measured by covariance σ_{XY}: 1) among the *products* of natural selection (among different B), given by $\sigma_{X_{(2)}Y_{(2)}}$, and 2) during the *process* of natural selection (going from A to B), given by $\int_0^\infty \sigma_{XY}(t)dt$. We find both to be unconditionally negative, in expectation, indicating encompassing recombinant advantage caused by natural selection.

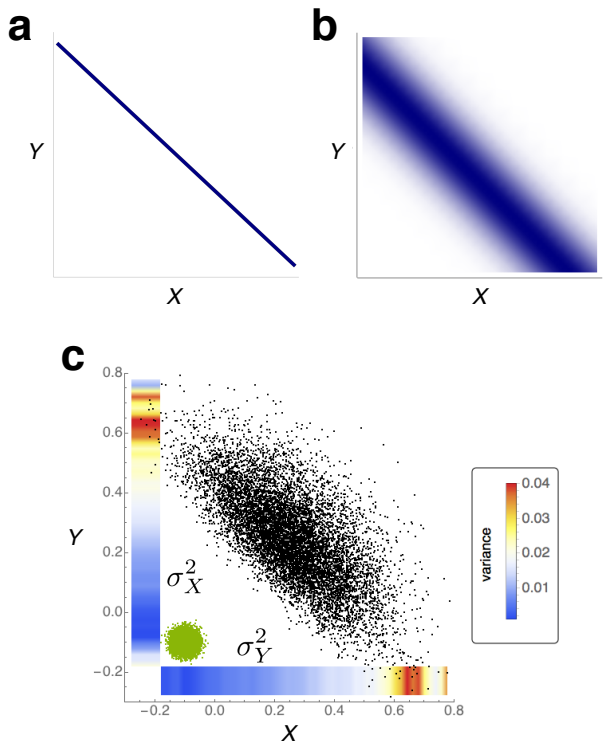

Fig 3 | Natural selection promotes negative associations. In the absence of recombination, selection does not act independently on X and Y but organismal fitness which, for simplicity, we here assume to be their sum, $Z = \phi(X, Y) = X + Y$. Perhaps counterintuitively, this fact alone creates negative associations. As discussed in the main text, this fact gives rise to a correlation of exactly negative one when the sum is a constant (**a**) and something intuitively negative when the sum is distributed as expected (**b**), i.e., as an order statistic. **c**, Ten thousand simulated populations move from their initial (green dots) to final (black dots) mean fitnesses. Here, the predicted negative covariance in the final state is apparent. The heatmap bars indicate variance in Y along the x-axis and variance in X along the y-axis, a manifestation of Hill-Robertson interference.