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

[*]To whom correspondence should be addressed; E-mail: pgerrish@unm.edu.

The ubiquity of recombination (and sex) in nature has defied explanation since the time of Darwin. Conditions that promote their evolution, however, are well-understood and arise when genomes 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. One feature of living things whose ubiquity approximates that of recombination is evolution by natural selection. Here, we find that natural selection by itself always, on average, creates the aforementioned imbalance and promotes the evolution and maintenance of recombination. Our findings suggest that recombination evolved 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 (*1*) 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 (*2–5*). 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 (*4–6*). 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 (*2, 3, 7*). 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 (*2, 5, 6*).

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 (*8*). 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 favored 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 (*9–12*), and putting low-fitness alleles together can expedite their elimination (*13, 14*). 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 tend to increase recombination rate (*1, 15*). Modifier loci can be affected by short-term and long-term indirect selection (*10*). In the short term, recombinants formed from randomly-chosen parents harboring the aforementioned imbalance have an advantage that is relatively immediate, thereby creating indirect selection for modifiers that increase the recombination rate. In the long term, such modifiers can, in theory, be favored or suppressed because of the fitness variation they augment (*10*), but this relies on an unlikely – and somewhat inconsistent – supposition that the modifier remains linked to that variation. Under the more plausible supposition that modifiers themselves can become dissociated from fitness-related loci through the very recombination they modify, it is the short-term advantage or disadvantage of recombinants that determines the fate of a modifier. If the selective imbalance is "chronic" such that recombinants are always advantageous on average, modifiers that increase the recombination rate will be indirectly favored.

The great challenge in explaining the evolution of recombination 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 evolution by natural selection. In what follows, we assess the effects of natural selection on selective imbalance and hence on the evolution of recombination.

We preface our developments with an essential technical point. In much of the relevant liter-

ature, the measure of selective mismatch across loci affecting the evolution of recombination is linkage disequilibrium (LD) (*6, 9, 10, 16–18*), 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 (Supplementary Methods, and fig. S1). Our results will thus be given in terms of covariance and we recall: negative covariance means positive selection for recombinants.

For our two-part study, 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 between two different *products* of natural selection; by "products", we mean genotypes that have come to prevail – through the action of natural selection – within their local context, e.g., their population, subpopulation, deme, or competing clone; and 2) when recombination occurs between two individuals within the same population during the *process* of natural selection.

To isolate the effects of natural selection, 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 (*19*). It further provides a connection to foundational evolution-of-sex studies: Fisher (*20*) 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 (*21*) 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 our encompassing qualitative findings thus apply to these foundational models and others. Simulations further confirm the adequacy of this two-genotype setting: increasing the number of genotypes only accentuates the effects we describe (fig. S2).

Figure 1 illustrates how the problem is posed analytically. 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?

In part one of our study, we assess how the *products* of natural selection affect the evolution of recombination. Figure 2 illustrates the problem by analogy to a set of canoe races. 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 (*2, 4, 7, 8, 11*). 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 favored.

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 (*22*)). 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, 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 (*23–26*) and may even speak to notions that out-crossing can create hybrid vigour (heterosis).

We now turn to part two of our study, where we will assess how the *process* of natural selection affects the selective value of recombinants in unstructured populations. Here, our focus is on recombinants formed from two randomly-chosen parents within the same unstructured evolving population. Here again, Fig. 1 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 ($\sigma_{XY} = \sigma_{XY}(t)$). Our measure of the net effect of natural selection on recombination is the quantity $\int_0^\infty \sigma_{XY}(t)dt$; if positive (negative), natural selection opposes (favors) recombination.

In Supplementary 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 favor recombination. Remarkably, this finding requires no assumptions about the distribution of allelic fitness contributions 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.

Our analyses further show that natural selection creates recombinant advantage even when recombinants are present in the initial variation upon which natural selection acts. Put differently, even in the presence of recombination, the effect of natural selection is to promote increased recombination. The implication is that the *process* of natural selection not only pro-

motes the evolution of recombination but also its maintenance (fig. S3).

Finally, we show that it is primarily the additive component of fitness that causes time-integrated covariance to be negative. This fact stands in contrast to some previous indications that non-additive effects, specifically negative or fluctuating epistasis, are an essential ingredient in the evolution of recombination (*27, 28*).

Some authors (*4, 29*) 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. The findings from part one of our study differ from these arguments: we find that even genotypes that are ultimately fixed carry selectively mismatched alleles. The findings from part two of our study, however, are entirely consistent with these arguments which provide an intuitive way to understand our remarkable Proposition S7.

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 paradox" (*30*) 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.

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.

References and Notes

1. M. Nei, *Genetics* **57**, 625 (1967).

2. S. P. Otto, T. Lenormand, *Nat. Rev. Genet.* **3**, 252 (2002).

3. M. Hartfield, P. D. Keightley, *Integr. Zool.* **7**, 192 (2012).

4. S. P. Otto, *Am. Nat.* **174 Suppl 1**, S1 (2009).

5. N. H. Barton, B. Charlesworth, *Science* **281**, 1986 (1998).

6. J. Felsenstein, *Genetics* **78**, 737 (1974).

7. S. A. West, C. M. Lively, A. F. Read, *J. Evol. Biol.* (1999).

8. J. A. G. M. de Visser, S. F. Elena, *Nat. Rev. Genet.* **8**, 139 (2007).

9. S. P. Otto, N. H. Barton, *Genetics* **147**, 879 (1997).

10. N. H. Barton, *Genetics* **140**, 821 (1995).

11. A. F. Agrawal, *Curr. Biol.* **16**, R696 (2006).

12. J. A. Arjan, *et al.*, *Science* **283**, 404 (1999).

13. A. S. Kondrashov, *Nature* **336**, 435 (1988).

14. P. D. Keightley, S. P. Otto, *Nature* **443**, 89 (2006).

15. J. Felsenstein, S. Yokoyama, *Genetics* **83**, 845 (1976).

16. N. H. Barton, *Genet. Res.* **65**, 123 (1995).

17. N. H. Barton, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 2559 (2010).

18. S. P. Otto, M. W. Feldman, *Theor. Popul. Biol.* **51**, 134 (1997).

19. E. Baake, A. González Casanova, S. Probst, A. Wakolbinger, *Theor. Popul. Biol.* **127**, 58 (2019).

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

21. H. J. Muller, *Am. Nat.* **66**, 118 (1932).

22. S. S. Yang, *Ann. Stat.* **5**, 996 (1977).

23. G. Martin, S. P. Otto, T. Lenormand, *Genetics* **172**, 593 (2006).

24. L. Becks, A. F. Agrawal, *Nature* **468**, 89 (2010).

25. A. O. B. Whitlock, R. B. R. Azevedo, C. L. Burch, *Evolution* **73**, 1089 (2019).

26. T. Lenormand, S. P. Otto, *Genetics* **156**, 423 (2000).

27. R. D. Kouyos, S. P. Otto, S. Bonhoeffer, *Genetics* **173**, 589 (2006).

28. S. Gandon, S. P. Otto, *Genetics* **175**, 1835 (2007).

29. N. H. Barton, S. P. Otto, *Genetics* **169**, 2353 (2005).

30. J. B. Miller, A. Sanjurjo, *J. Econ. Perspect.* **33**, 144 (2019).

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.

Supplementary Materials

Methods

Supplementary Text

Figs S1 to S12

Tables S1 to S3

References (31 to 40)

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

Competing interests The authors declare no competing interests.

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

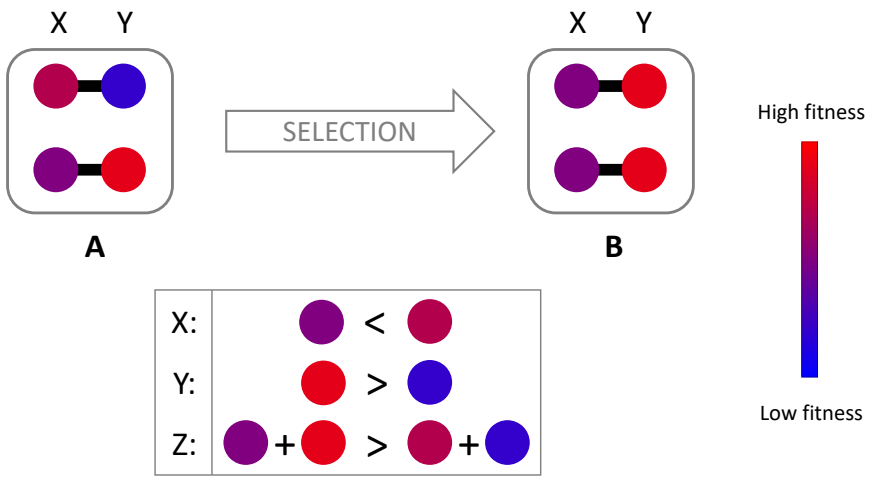

Fig 1. Two loci, two alleles. Here, a large (infinite) population consists of individuals whose genome has only two loci x and y, each of which carries one of two alleles: genotype 1 encodes quantified phenotype X_1 at the x locus and Y_1 at the y locus, and genotype 2 carries quantified phenotype X_2 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 $\int_0^\infty \sigma_{XY}(t)dt$.

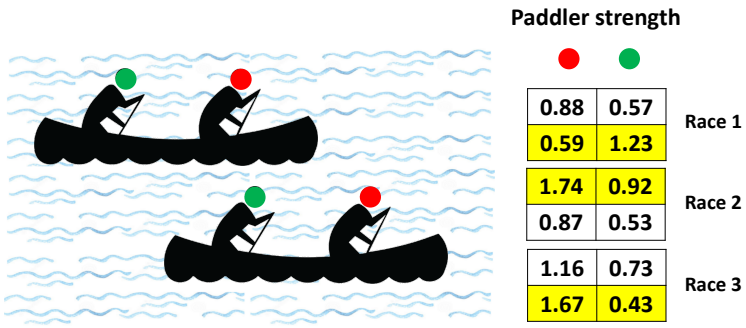

Fig 2. Canoe race analogy. Each canoe contains two paddlers. The strength of each paddler is measured and reported in the table. In any given canoe race, there is no correlation between paddler strengths A and B. In each race, paddler strengths are recorded (tables on right), and the winning canoe is 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.

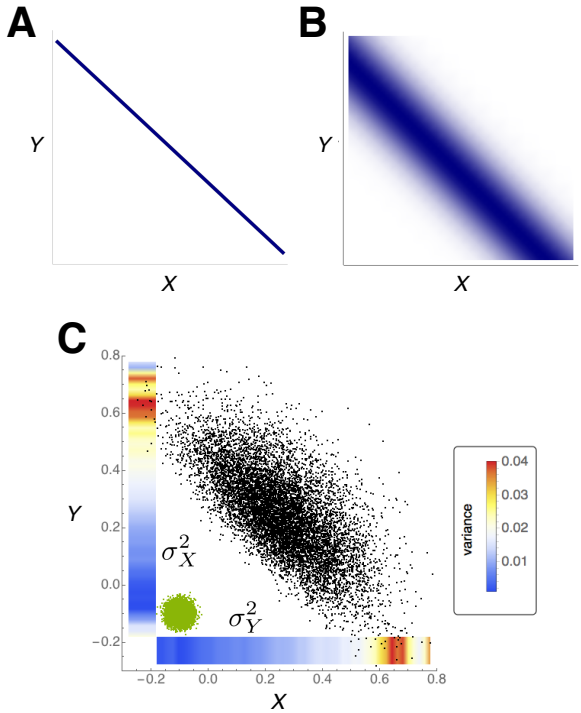

Fig 3. Natural selection promotes negative associations. In the absence of recombination, selection does not act independently on X and Y but organismal fitness which, for simplicity, we here assume to be their sum, $Z = \phi(X, Y) = X + Y$. Perhaps counterintuitively, this fact alone creates negative 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 (*14, 23*).