Social evolution and genetic interactions in the short and long term

Jeremy Van Cleve

Department of Biology
University of Kentucky
Lexington, KY 40506 USA

National Evolutionary Synthesis Center (NESCent)
2024 W. Main Street Suite A200
Durham, NC 27705 USA

phone: 919-668-4044
fax: 919-668-9192

e-mail: vancleve@santafe.edu

Date modified: 13th October, 2014

Abstract

The evolution of social traits remains one of the most fascinating and feisty topics in evolutionary biology even after half a century of theoretical research. W. D. Hamilton shaped much of the field initially with his 1964 papers that laid out the foundation for understanding the effect of genetic relatedness on the evolution of social behavior. Early theoretical investigations revealed two critical assumptions required for Hamilton's rule to hold in dynamical models: weak selection and additive genetic interactions. However, only recently have analytical approaches from population genetics and evolutionary game theory developed sufficiently so that social evolution can be studied under the joint action of selection, mutation, and genetic drift. We review how these approaches suggest two timescales for evolution under weak mutation: (i) a short-term timescale where evolution occurs between a finite set of alleles, and (ii) a long-term timescale where a continuum of alleles are possible and populations evolve continuously from one monomorphic trait to another. We show how Hamilton's rule emerges from the short-term analysis under additivity and how non-additive genetic interactions can be accounted for more generally. This short-term approach reproduces, synthesizes, and generalizes many previous results including the one-third law from evolutionary game theory and risk dominance from economic game theory. Using the long-term approach, we illustrate how trait evolution can be described with a diffusion equation that is a stochastic analogue of the canonical equation of adaptive dynamics. Peaks in the stationary distribution of the diffusion capture classic notions of convergence stability from evolutionary game theory and generally depend on the additive genetic interactions inherent in Hamilton's rule. Surprisingly, the peaks of the long-term stationary distribution can predict the effects of simple kinds of non-additive interactions. Additionally, the peaks may capture the effect of both weak and strong selection in a manner analogous to classic diffusion approaches in population genetics. Together, the results from the short and long-term approaches suggest both how Hamilton's insight may be robust in unexpected ways and how current analytical approaches can expand our understanding of social evolution far beyond Hamilton's original work.

Key words: inclusive fitness; adaptive dynamics; fixation probability; evolutionary stability; stochastic stability; risk dominance; island model; trait substitution sequence; diffusion; cooperation

1. Introduction

The theory of evolution by natural selection as first fully elucidated by Darwin [25] is so profoundly elegant and comprehensive that truly new additions to theory have been extremely rare. In 1963, W. D. Hamilton began publishing his seminal work on how natural selection can shape social behavior [57–59], which is often either referred to as the theory of "kin selection" [100] or "inclusive fitness" [42]. It is a tribute to the importance of this work that upon his untimely death in 2000 Hamilton was called "one of the most influential Darwinian thinkers of our time" [36] and a candidate for the "most distinguished Darwinian since Darwin" [26].

In this article, we will review how the tools of population genetics and evolutionary game theory can be used to formalize Hamilton's insight. We will begin with a summary of classic analyses of Hamilton's approach and will then introduce the population genetic and game theoretic tools that currently provide a complete framework for studying social evolution under weak selection and weak mutation [92] . Using these tools, we will see how two general timescales for analysis emerge: a short-term timescale where evolution proceeds among a finite set of alleles, and a long-term timescale where populations evolve continuously among a continuum of alleles. These notions of short and long-term derive from a broader attempt to reconcile population genetic methods with evolutionary game theory [34, 62, 160].

Using the short-term approach, we show how genetic interactions between individuals [e.g. 118] can affect selection for cooperation in deme-structured populations [80]. These results extend previous analyses of stochastic evolution that have shown conditions such as "risk dominance" [13, 64, 72] and the "one-third law" [109, 113] to be important determinants of evolutionary stability. Using the substitution rate approach to long-term evolution [82, 152], we describe a diffusion equation that approximates the long-term change in monomorphic trait values. We show how peaks in the stationary distribution of this diffusion captures classic notions of evolutionary and convergence stability. Moreover, the location of these convergence stable states can be calculated using the classic direct-fitness approach of kin selection [123, 125, 142]. Applying this long-term approach to a simple non-additive social interaction, we find surprisingly that the long-term analysis can capture these non-additive effects even though the diffusion integrates over only additive interactions. Moreover, the long-term approach appears to reproduce results from some strong selection models, which suggests an unexpected robustness of the long-term diffusion. Together, the results from the short and long-term approaches reveal the usefulness of the current framework for integrating Hamilton's original insight with recent results from population genetics and evolutionary game theory.

1.1. Hamilton's rule

The core insight in Hamilton's work is often summarized with his eponymous rule: an allele for a social behavior increases in frequency when the "inclusive fitness effect" is positive, namely

$$-c + b\,r > 0\,. \tag{1}$$

In Hamilton's rule (1), b is the increase in fitness (benefit) of a social partner from the behavior of a focal individual, c is the decrease in fitness (cost) of a focal individual that performs the behavior, and r measures genetic relatedness between focal and recipient individuals [41]. More generally, $-c$ is called the "direct fitness effect" and b the "indirect fitness effect". Hamilton [58] initially emphasized that genetic relatedness is generated by a genealogical process that produces alleles identical by descent (IBD) among a group of socially interacting individuals. Another general definition of genetic relatedness says that it is the regression of the genotypes of social partners on the genotype of the focal individual [54, 60]. Hamilton's rule crystalized the notion that natural selection depends not only on how genes within an individual effect that individual's fitness but also on the indirect effect of those genes on the fitness of social partners. Although Darwin [25], Fisher [39], and Haldane [56], among others, had expressed this idea in relation to the how evolution would lead one individual to sacrifice its fitness for another, Hamilton was the first to present a compelling framework applicable to social evolution more generally.

Within Hamilton's inclusive fitness framework, behaviors that decrease the fitness of a focal individual ($c > 0$) but increase the fitness of social partners ($b > 0$) are "altruistic". Well-known examples of altruism include worker sterility in eusocial insects [8], stalk cells that give up reproduction to disperse spore cells in *Dictyostelium discoideum* [134], and costly human warfare [61, 84]. Other behaviors can also be classified in Hamilton's framework [58, and Table 2]: (i) behaviors are "mutualistic" when they increase the fitness of the focal individual and its social partners, (ii) "selfish" when they increase the fitness of the focal at the expense of the fitness of social partners, and (iii) "spiteful" when they decrease the fitness of both the focal individual and its social partners. Although there are other potential definitions of altruism and other behaviors [see 14, 76], Hamilton's classification based on direct and indirect effects has proven useful for distinguishing different kinds of helping behaviors (mutualisms and altruisms) and for showing how different biological mechanisms can promote or inhibit the evolution of these behaviors [85, 161].

Though Hamilton's approach was initially accepted among empiricists [167] and some theorists [100, 115],

other theorists were concerned about the generality of the approach due to its emphasis on fitness maximization and optimality modeling [16, 73, 165]. Fitness maximization was viewed as untenable because examples where it is violated are well known [102]. Optimality models were additionally viewed with skepticism because, by neglecting gene frequency dynamics, they cannot study genetic polymorphisms; in effect, such models must assume that mutant alleles that invade a population also reach fixation. An initial wave of population genetic studies in response to these concerns showed that Hamilton's rule was generally a correct mutant invasion condition so long as selection is weak and fitness interactions between individuals are additive [1, 16, 149, 150, 153]. However, these models were family structured where cooperation occurs between close relatives and could not address the applicability of Hamilton's rule in populations with more generic structure, such as deme structure in island [168] and lattice models [78, 98, 99].

1.2. The Price equation and the individually-based approach

Part of the difficulty with the population genetic methods used to analyze family-structured models is that they use genotypes as state variables. This quickly increases the dimensionality of the model as the number of loci, family size, or demes increases and makes approximation difficult. An important alternative approach was introduced to population genetics by George Price with his eponymous equation [116, 117]. The Price equation uses the distribution of allele frequencies in each individual in the population as the set of state variables and tracks the first population-level moment of this distribution, which is the mean allele frequency. If $\mathbf{p} = (p_1, \ldots, p_{N_T})$ represents the allele frequency distribution for N_T haploid individuals ($p_i = 0$ or 1 for individual i), the Price equation yields

$$\mathrm{E}[w\Delta p|\mathbf{p}] = \mathrm{Cov}[w_i, p_i] + \mathrm{E}[w_i \Delta p_i] \tag{2}$$

where $\mathrm{E}[w\Delta p|\mathbf{p}]$ is the expected change in mean allele frequency weighted by mean fitness $\overline{w}$ and conditional on $\mathbf{p}$ in the parental generation. The first term on the right hand side, the covariance between individual fitness w_i and allele frequency p_i, measures the effect of selection on the change in mean allele frequency in the population. The second term, $\mathrm{E}[w_i \Delta p_i]$, measures the effect of non-selective transmission forces, such as mutation and migration (and recombination for changes in genotype frequencies), on the change in mean allele frequency. When selection is the only force on allele frequencies and the population size remains fixed ($w = 1$), the Price equation simplifies to

$$\mathrm{E}[\Delta p|\mathbf{p}] = \mathrm{Cov}[w_i, p_i]\,. \tag{3}$$

Calculating higher-order moments of the allele frequency distribution is necessary to measure the exact dynamics of the distribution over time; thus, moment-based approaches like the Price equation are not necessarily more tractable than directly tracking genotype frequencies. However, an important observation about moment-based approaches is that they are readily amenable to approximation. When selection is weak relative to other forces such as recombination and migration, a kind of separation of timescales occurs where allele frequency dynamics converge very slowly and associations between alleles, linkage disequilibrium between loci [10, 79, 105] and F_{ST} between individuals in a deme [124, 127, 129, 156], converge much more quickly. Because of this "separation of timescales", linkage disequilibrium, F_{ST}, and other associations will converge to "quasi-equilibrium" (QE) values that are a function of mean allele frequencies. This means that the mean allele frequency dynamics can be expressed as a closed system of equations, which considerably simplifies analysis of multilocus systems or structured populations.

With respect to social evolution, the QE results for structured populations are particularly useful as they have helped to establish a rigorous basis for kin selection and Hamilton's rule in populations with finite size, localized dispersal, or both [123, 125]. Initiated by the seminal work of François Rousset [122, 125], the mean allele frequency dynamics in this approach are calculated under weak selection and can be expressed as functions of F_{ST} and other between individual genetic associations evaluated under selective neutrality. In the simplest cases, this approach shows that Hamilton's rule holds for weak selection and additive genetic interactions in populations with island-type structure [125] and family-structured populations [128]. More generally, this approach produces analogues of Hamilton's rule where the direction of selection is given by a sum of relatedness coefficients and fitness effects indexed by the spatial distance between a focal individual and its social partners [92, 123, 125] or by the demographic class (e.g., juvenile vs. adult or worker vs. queen in social insects) of the focal and its partners [126, 128]. It is this weak selection and QE approach that we will use to study genetic interactions and their affect on cooperation in deme structured populations below.

1.3. Genetic drift, adaptive dynamics, and evolution in the short and long term

Another difficulty with the early analyses of kin selection and Hamilton's rule in family structured populations was that those population genetic models could easily produce stable polymorphic equilibria [147, 149, 150], which made general predictions concerning the level of altruism or other social behaviors difficult. Generally, such equilibria are of intrinsic biological and mathematical interest since they illuminate stabilizing selection that can maintain genetic variation in levels of cooperation. In finite populations however, even alle-

les under stabilizing selection either eventually go extinct or reach fixation due to genetic drift. If genetic drift is sufficiently strong relative to the rate of mutation μ, then the population will spend most of its time fixed for one of a set of possible alleles generated by mutation. This occurs for large N_T and small μ when

$$N_T \, \mu \log N_T \ll 1, \tag{4}$$

which can be arrived at heuristically by using the expected number of alleles in the population in an infinite alleles model under neutrality [74, 158, 159, 169]. Moreover, this condition ensures that any mutation that can arise will either fix or go extinct before another mutation arrives [17, 19][1]; in other words, there are at most two alleles in the population at one time: a "resident" and a "mutant" that either goes extinct or fixes and becomes the new resident. This process of sequential substitution of alleles is called the "trait substitution sequence" [TSS; 18, 30] and is the fundamental biological model of adaptive dynamics [28, 50, 101] when population size goes to infinity, $N_T \to \infty$. In addition, the TSS is often the implicit dynamical process behind many phenotypic models of kin selection and evolutionary game theory [27, 37, 92, 125, 142, 160, 162] that use an optimality criterion (i.e., fitness maximization) in search of an evolutionarily stable strategy (ESS).

On timescales short compared those required for generating phenotypic novelty, alleles generated by mutation in the TSS constitute a finite set, and the population "jumps" between alleles in this set as each allele invades and fixes. Assuming that it is possible to mutate from one allele to any other in the set through a sequence of zero or more intermediates (i.e., the mutation process is irreducible), the short-term process equilibrates to a stationary distribution $\boldsymbol{\lambda}$ among the fixation or monomorphic states. This short-term TSS corresponds to "short-term evolution" as defined by Eshel [33, 34] where a fixed set of genotypes are allowed to change frequency but new mutations outside this set do not occur. The length of time the population spends fixed for each allele is primarily determined by the likelihood each allele arises via mutation (μ) and fixes (fixation probability, π) in populations monomorphic for the other possible alleles [44]. Assuming weak selection, the QE approach discussed above can be used to calculate fixation probabilities even in spatially or demographically structured populations [123]. Together with the TSS condition (4), this allows a complete description of the stationary distribution of allelic states under the forces of selection, mutation, and genetic drift in the short-term.

For longer timescales, novel phenotypes are possible due to the invasion of mutations outside of a given finite set. For example, processes such gene [96] and genome duplication [146], transposons [114], and lateral

[1] This condition is very similar to the one obtained in the "strong-selection weak-mutation" limit of Gillespie [p. 221; 52] and the successional-mutations regime of Desai and Fisher [eq. 1; 29]: $N_T \mu \log N_T \omega \ll 1$ where ω is the strength of selection.

gene transfer [70] can generate novel physiological and ecological functions not possible with small changes in single genes. These processes suggest that the set of possible phenotypes may have a continuum of values over the long term. Suppose that for phenotype z the probability density of generating a mutant allele of phenotype $z+\delta$ is $u(\delta,z)$. If the support of $u(\delta,z)$ covers a fitness peak (i.e., it is possible to generate a mutant that resides exactly at the peak), then it is possible that the population will not only approach the peak, but it will spend most of its time fixed for a phenotype within a small neighborhood of the peak. This long-term TSS corresponds to the definition of "long-term evolution" by Eshel [33, 34] where invasion of new genotypes allow the population to approach phenotypic equilibria defined by the classic evolutionarily stable strategy (ESS) condition [35, 63, 95]. Without any assumptions on the distribution of mutational effects $u(\delta,z)$, the long-term TSS is described mathematically as a Markov jump process and is given by an integro-differential (master) equation [17–19, 82]. Often for the purpose of tractability, only small mutants are allowed in small intervals of time, which means that $u(\delta,z)$ is narrowly peaked around z and the population cannot make large jumps quickly. This assumption turns the jump process into a diffusion process [20] that is the stochastic analogue of the deterministic canonical equation of adaptive dynamics [18, 30]. The long-term TSS diffusion also has a stationary probability density, $\rho(z)$, and the phenotypes located at peaks in that density correspond to equilibria obtained from classic ESS or adaptive dynamics analyses [82, 152]. Exactly as in the short-term TSS, weak selection can be used to calculate the fixation probabilities that determine $\rho(z)$. This leads to a long-term stationary distribution of phenotypes that captures selection, mutation, and drift in spatially or demographically structured populations.

1.4. Putting it all together: Hamilton and social evolution in the short and long term

The two main assumptions above, weak selection and the TSS condition (4), allow us to describe the stationary density of transitions between a discrete set of phenotypes in the short term or among a continuum of phenotypes in the long-term. When there are only two types in the short term, cooperative and noncooperative, and the population is spatially structured or family structured, Hamilton's rule in equation (1) is readily recovered when genetic interactions are additive [125, 128]. As we will see below, this is a result of comparing the stationary density of cooperative versus noncooperative types. Moreover, much of the recent work in evolutionary game theory that focuses on finite populations uses this same short-term TSS model to calculate a stationary distribution of types [e.g., 45, 66, 71, 109, 113, 131]. When there is a continuum of levels of cooperation in the long term, $br-c$ in Hamilton's rule becomes the gradient of the potential function used to solve for the stationary density of the TSS diffusion [82]. Since $br-c$ can be thought to measure the change in inclusive

fitness for additive genetic interactions [92, 125, 142], phenotypes at peaks in inclusive fitness ($br - c = 0$) correspond to peaks in the stationary density. Thus, the long-term action of natural selection, *assuming additivity*, leads to a kind maximization of inclusive fitness, which supports the use of classic inclusive fitness analyses [however, see refs 91, 92, for how difficulties in interpreting this result as broadly justifying "inclusive fitness maximization"].

1.5. Genetic interactions and non-additivity

If one is willing to assume weak selection, weak mutation relative to genetic drift (i.e., the TSS), and additive genetic interactions, then a direct application of Hamilton's rule can be justified using the theoretical work discussed above. However, the ability to predict short and long term distributions of types under weak selection and the TSS is possible even when genetic interactions are non-additive. Non-additivity at the genetic level allows for interaction among alleles, within or between individuals. Within individuals, such interactions produce dominance and epistasis and between individuals they produce scenarios analogous to classic two-player games with pure strategies, such as the Hawk-Dove or Stag-Hunt games. Non-additive interactions are important because they produce frequency dependence in the sign of the change in allele frequency (eq. 3) even for weak selection [86, 112, 151]. In the case of social behavior, this implies that Hamilton's rule becomes frequency dependent and no longer provides an unambiguous prediction of the effect of selection in either the short or the long term. Rather, applying the tools of QE and the TSS for non-additive interactions requires additional terms to account for higher-order genetic associations.

Once we calculate these additional terms, we determine the effect of non-additive interactions on the short-term stationary distribution of types in a given demographic context. Here, we apply the theory to Wright's island model of population structure where there are n demes or groups each containing N haploid individuals [168] ($N_T = nN$). All groups are connected equally by migration at rate m. One of the important features of the Price equation approach is that allows us to expression the genetic associations (e.g., F_{ST}) in terms of mean times to coalescence. Using results from coalescent theory to calculate the genetic associations, we replicate and generalize well known short-term results from inclusive fitness (the Taylor cancellation result [112, 139, 140]) and evolutionary game theory (the one-third law [109] and risk dominance [13, 64, 72]). Moreover, we show how changing the competitive environment (hard vs. soft vs. group selection) changes these well known results, particularly in the presence of non-additive interactions.

The long-term approach, in contrast, only accounts for additive genetic interactions. Nevertheless, at least

in social interactions with simple non-additive payoffs, the long-term approach remarkably reproduces results from the short-term approach that explicitly includes non-additive genetic interactions. We discuss a potential explanation for this power of the long-term approach, which suggests that three-way genetic interactions may be uniquely analytically tractable among possible non-additive interactions.

2. Theory: short-term evolution

2.1. Weak mutation,the TSS, and evolutionary success

Consider evolution in a population with total size, N_T, where the population can be group structured (n groups of size N) or otherwise spatially structured with some pattern of migration between spatial locations. Recall from section 1.3 that the short-term TSS requires considering only two alleles, which we label A and a where p_i measures the frequency of A in individual i (see Table 1 for a description of symbols used throughout this paper). Suppose that the mutation rate from A to a is $\mu_{a|A}$ and $\mu_{A|a}$ is the rate from a to A. We assume $\mu = \max(\mu_{A|a}, \mu_{a|A})$ measures the strength of mutation. The weak mutation condition that defines the TSS, condition (4), is derived under the limit as $N_T \to \infty$ and $\mu \to 0$. In this limit, the TSS is consists of the population jumping between states fixed for allele A and fixed for allele a. To represent the jump process between these two fixation states, we create a Markov chain with the following matrix

$$\Lambda(\mu) = \begin{bmatrix} 1 - \frac{\mu_{a|A}}{\mu}\pi_{a|A} & \frac{\mu_{a|A}}{\mu}\pi_{a|A} \\ \frac{\mu_{A|a}}{\mu}\pi_{A|a} & 1 - \frac{\mu_{A|a}}{\mu}\pi_{A|a} \end{bmatrix} \tag{5}$$

where $\pi_{A|a}$ and $\pi_{a|A}$ are the probabilities that alleles A and a, respectively, reach fixation starting from an initial frequency of $1/N_T$ in a population where the other allele has frequency $1 - 1/N_T$. Rescaling the mutation rates by the overall rate μ allows a nontrivial stationary distribution of the Markov chain (i.e., the left eigenvector of $\Lambda(\mu)$) as $\mu \to 0$. Inspired by ideas in large deviations theory [43], Fudenberg and Imhof [44] show that the stationary distribution $\boldsymbol{\lambda}$ of the TSS as $\mu \to 0$ is simply the stationary distribution of the Markov chain in (5) in the limit as $\mu \to 0$. Thus, instead of having to calculate the stationary distribution of the complex stochastic process with many different possible population states, we need only calculate the stationary distribution of the much simpler "embedded" chain composed of fixation states. Calculating the stationary distribution using the embedded chain yields

$$\boldsymbol{\lambda} = \left(\frac{\mu_{A|a}\pi_{A|a}}{\mu_{A|a}\pi_{A|a} + \mu_{a|A}\pi_{a|A}}, \frac{\mu_{a|A}\pi_{a|A}}{\mu_{A|a}\pi_{A|a} + \mu_{a|A}\pi_{a|A}} \right). \tag{6}$$

If we are interested only in the effect of selection on the stationary distribution, we can assume that the

mutation rates are symmetric, $\mu_{A|a} = \mu_{a|A}$. In this case, the expected frequency of allele A in the population at stationarity, which we write as $\mathrm{E}[p]$, becomes in the limit as the mutation rate goes to zero

$$\lim_{\mu \to 0} \mathrm{E}[p] = \frac{\pi_{A|a}}{\pi_{A|a} + \pi_{a|A}} \,. \tag{7}$$

An intuitive condition for the evolutionary success of allele A relative to allele a is that A is more common at stationarity, or

$$\mathrm{E}[p] > \frac{1}{2} \,. \tag{8}$$

Using equation (7), condition (8) is equivalent to

$$\pi_{A|a} > \pi_{a|A} \tag{9}$$

when $\mu \to 0$, which means we need only compare complementary fixation probabilities in order to determine which allele is "favored" by natural selection [7, 45]. This condition on fixation probabilities is the evolutionary success condition that we will use to derive Hamilton's rule (eq. 1).

2.2. Fixation probability and the Price equation

Calculating the fixation probabilities in a model with arbitrarily complex demography or spatial structure can be daunting if not impossible. Thus, our next aim is to show how to connect fixation probabilities to the Price equation, which will make it straightforward to use weak selection and QE results. Suppose that $p(t)$ is the mean frequency of allele A at time t. Following recent methods [89, 94, 122], we can write the fixation probability as

$$\begin{aligned} \pi_{A|a} &= \mathrm{E}[p(\infty)|p(0)] \\ &= p(0) + \mathrm{E}\left[\sum_{t=0}^{\infty} \Delta p(t) \middle| p(0) \right] \\ &= p(0) + \sum_{t=0}^{\infty} \mathrm{E}[\, \Delta p(t) | p(0)] \end{aligned} \tag{10}$$

where $\Delta p(t) = p(t+1) - p(t)$ and we can exchange the expectation and the infinite sum in the last line because the Markov chain converges in mean [94]. Expanding the sum in (10) by conditioning on all possible

population states $\mathbf{p}(t)$ yields

$$\pi_{A|a} = \pi^{\circ} + \sum_{t=0}^{\infty} \sum_{\mathbf{p}(t)} \Pr[\mathbf{p}(t)|p(0)] \, \mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,] \tag{11}$$

where we have used the fact that the fixation probability of a neutral allele, π°, is its initial frequency $p(0)$. The second term in the sum, $\mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,]$, is exactly the left-hand side of the Price equation (3).

2.3. Fixation probability under weak selection

The most straightforward way to calculate the probability of fixation $\pi_{A|a}$ assuming weak selection is to Taylor expand $\pi_{A|a}$ in terms of a parameter that measures the strength of selection [see: 89, 94, 122], which we call ω. This expansion is simply

$$\pi_{A|a} = \pi^{\circ} + \frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega}\omega + O(\omega^2) \tag{12}$$

where the derivative $\frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega}$ is evaluated under neutrality ($\omega = 0$). Using equation (11), we can calculate the derivative of the fixation probability under neutrality as

$$\frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega} = \sum_{t=0}^{\infty} \sum_{\mathbf{p}} \frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\Pr[\mathbf{p}(t)|p(0)] \, \mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,]\Big] \tag{13}$$

where the exchange of derivative and the limit is justified provided the derivatives converge uniformly [see Appendix of 94, for such a proof]. Expanding the derivative in the sum in (13) using the chain rule yields

$$\frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\Pr[\mathbf{p}(t)|p(0)] \, \mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,]\Big] = \frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\Pr[\mathbf{p}(t)|p(0)]\Big]\mathrm{E}^{\circ}[\, \Delta p(t)|\mathbf{p}(t)\,] + \Pr{}^{\circ}[\mathbf{p}(t)|p(0)]\frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,]\Big] \tag{14}$$

with the symbol $^{\circ}$ indicating evaluation of an expectation or probability in the neutral case when $\omega = 0$. The first term on the right hand side of (14) is zero since the expected change in allele frequency under neutrality is zero. Simplifying equation (13) with this fact yields

$$\frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega} = \sum_{t=0}^{\infty} \sum_{\mathbf{p}} \Pr{}^{\circ}[\mathbf{p}(t)|p(0)]\frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,]\Big] \tag{15}$$

which we can write as

$$\frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega} = \sum_{t=0}^{\infty} \mathrm{E}^{\circ}\left[\frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\mathrm{E}[\, \Delta p(t)|\mathbf{p}(t)\,]\Big]\right] \tag{16}$$

where E° implies expectation over the neutral realizations of $\mathbf{p}$ given an initial frequency of A of $p(0)$.

In order to evaluate the derivative of the fixation probability in equation (16), we need the derivative of the expected change in mean allele frequency. This is a quantity that is relatively simple to calculate since all ones needs is the first-order term in an expansion of $\mathrm{E}[\,\Delta p(t)|\mathbf{p}(t)\,]$ in terms of selection strength ω. To obtain this expansion, we first expand the fitness of the focal individual i in terms of selection strength. Without loss of generality, the fitness of individual i is

$$\begin{aligned} w_i &= 1 + \omega \sum_{d=1}^{N_\mathrm{T}} \sum_{k_1<\ldots<k_d}^{N_\mathrm{T}} s_{i,k_1\cdots k_d}\, p_{k_1} \cdots p_{k_d} + O(\omega^2) \\ &= 1 + \omega S_i(\mathbf{p}) + O(\omega^2) \end{aligned} \tag{17}$$

where $s_{i,1}, \ldots, s_{i,1\cdots N_\mathrm{T}}$ are often called "selection coefficients" [10, 79] and $S_i(\mathbf{p})$ is the polynomial given by the summations in the first line. The $d = 1$ term in the summation yields the "additive" fitness components where selection coefficients $s_{i,j}$ are multiplied p_j. Terms in the summation $d > 1$ are "non-additive" fitness components since the selection coefficients there are multiplied by products of allele frequencies. Given that mean fitness is equal to one (which is true for populations of fixed size N_T), $S_i(\mathbf{p})$ must satisfy to first order in ω

$$\sum_i S_i(\mathbf{p}) = 0\,, \tag{18}$$

which also implies that the sum of the selection coefficients must be zero by setting $\mathbf{p} = \mathbf{1}$ in (18); this constraint on the sum of the selection coefficients is a common feature of weak selection models with a fixed demography [e.g.: 89, 125].

Using the expression for fitness in (17), the change in mean allele frequency from the Price equation (3) becomes

$$\begin{aligned} \mathrm{E}[\,\Delta p|\mathbf{p}] &= \mathrm{Cov}[1 + \omega S_i(\mathbf{p}), p_i] \\ &= \omega \frac{\mathbf{S}(\mathbf{p}) \cdot \mathbf{p}}{N_\mathrm{T}} + O(\omega^2) \end{aligned} \tag{19}$$

where $\mathbf{S}(\mathbf{p}) = (S_1(\mathbf{p}), \ldots, S_{N_\mathrm{T}}(\mathbf{p}))$ and $\mathbf{S}(\mathbf{p}) \cdot \mathbf{p}$ is the scalar product of the two vectors. For example, if in a single panmictic population allele A confers a fitness advantage of ω for every individual that has it, then $w_i = 1 + \omega(p_i - p)$ and

$$\mathrm{E}[\,\Delta p|\mathbf{p}] = \omega p(1-p)$$

which is the standard result for an advantageous allele under weak selection [38]. Taking the first-order term from (19) and inserting it in equation (16) produces

$$\frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega} = \frac{1}{N_\mathrm{T}} \sum_{t=0}^{\infty} \mathrm{E}^\circ[\mathbf{S}(\mathbf{p}) \cdot \mathbf{p}] \,. \tag{20}$$

Expanding (20) into the first-order Taylor series for the probability of fixation in (12) yields

$$\pi_{A|a} = \pi^\circ + \frac{\omega}{N_\mathrm{T}} \sum_{t=0}^{\infty} \sum_{d=1}^{N_\mathrm{T}} \sum_{i=1}^{N_\mathrm{T}} \sum_{k_1<\ldots<k_d}^{N_\mathrm{T}} s_{i,k_1\cdots k_d} \, \mathrm{E}^\circ\left[p_i p_{k_1} \cdots p_{k_d}\right] + O(\omega^2) \,. \tag{21}$$

At this point, even though we have not explicitly specified the effect of gene expression on fitness or the population structure, we can see how additive and non-additive effects of selection affect fixation probability. The additive terms, $d = 1$ in the above sum, depend on expected pairs of allele frequency, $\mathrm{E}\left[p_i p_j\right]$. These expected pairs are essentially probabilities of genetic identity between two different individuals in the population. Thus, measures of average pairwise genetic identity within a structured population, such as Wright's F_{ST}, are natural statistics to use when considering the effect of selection due to additive genetic interactions. The non-additive terms, $d > 1$, contribute expected d-order products of allele frequency and thus require higher order statistics that F_{ST}.

In order to better interpret the expression for fixation probability in (21), which contains a difficult infinite sum over time, we follow the argument given in Rousset [122] and expanded in Lessard and Ladret [94] and Lehmann and Rousset [89] that interprets the expected allele frequency products in terms of coalescence probabilities. Recall from the TSS that in a population composed of allele a, a single A mutation will arise and either fix or go extinct. In this case, the expected allele frequency product, $\mathrm{E}^\circ\left[p_i p_{k_1} \cdots p_{k_d}\right]$, is the probability that individuals i and k_1 through k_d all have allele A at some future time t. Going backwards in time, this probability is equivalent to the probability that those lineages coalesce before time t, $\Pr^\circ\left[T_{ik_1\cdots k_d} \leq t\right]$, times the probability that the ancestral lineage is allele A, which is the initial frequency $p(0) = \pi^\circ = 1/N_\mathrm{T}$. Writing $\Pr^\circ\left[T_{ik_1\cdots k_d} \leq t\right]$ as $1 - \Pr^\circ\left[T_{ik_1\cdots k_d} > t\right]$ and using the fact that the selection coefficients sum to zero (eq. 18),

equation (21) becomes

$$\begin{aligned}\pi_{A|a} &= \frac{1}{N_{\mathrm{T}}} - \frac{\omega}{N_{\mathrm{T}}} \sum_{d=1}^{N_{\mathrm{T}}} \sum_{i=1}^{N_{\mathrm{T}}} \sum_{k_1<\ldots<k_d}^{N_{\mathrm{T}}} s_{i,k_1\cdots k_d} \sum_{t=0}^{\infty} \Pr^{\circ}\left[T_{ik_1\cdots k_d} > t\right] / N_{\mathrm{T}} + O(\omega^2) \\ &= \frac{1}{N_{\mathrm{T}}} - \frac{\omega}{N_{\mathrm{T}}} \sum_{d=1}^{N_{\mathrm{T}}} \sum_{i=1}^{N_{\mathrm{T}}} \sum_{k_1<\ldots<k_d}^{N_{\mathrm{T}}} s_{i,k_1\cdots k_d} \mathrm{E}^{\circ}\left[T_{ik_1\cdots k_d}\right] / N_{\mathrm{T}} + O(\omega^2) \end{aligned} \tag{22}$$

which matches previous results [eq. 15 in 122; eqs. 59 and 61 in 94]. Equation (22) says that the effect of selection on the fixation probability of allele A is simply a sum of selection coefficients and expected coalescence times under neutrality, $\mathrm{E}^{\circ}\left[T_{ik_1\cdots k_d}\right]$. One advantage of expressing the fixation probability in terms of coalescence times is that results from coalescence theory [157] can be used, which is the approach we utilize when we apply these methods to a population with island-type structure.

The condition for allele A to be more common at stationarity than allele a, condition (9), requires both fixation probabilities $\pi_{A|a}$ and $\pi_{a|A}$. Expanding $\pi_{a|A}$ under weak selection can be accomplished using the same reasoning above for equation (22). First, observe that that the expected change in the frequency of the a allele is $\mathrm{E}[\Delta q(t)] = -\mathrm{E}[\Delta p(t)]$. This implies that $\frac{\mathrm{d}\pi_{a|A}}{\mathrm{d}\omega}$ is simply the negative of the first-order term in equation (21) *except* that $\mathrm{E}^{\circ}\left[p_i p_{k_1} \cdots p_{k_d}\right]$ is evaluated under a neutral process where the initial frequency of a is $1/N_{\mathrm{T}}$ (rather than $1 - 1/N_{\mathrm{T}}$ as was the case for $\pi_{A|a}$). In calculating $\pi_{A|a}$, $\mathrm{E}^{\circ}\left[p_i p_{k_1} \cdots p_{k_d}\right]$ was interpreted as a coalescence probability times the initial frequency of A; analogously, we can write each p_i as $1 - q_i$ and interpret the products $q_i q_{k_1} \cdots q_{k_d}$ in $\mathrm{E}^{\circ}\left[(1-q_i)(1-q_{k_1})\cdots(1-q_{k_d})\right]$ as a coalescence probabilities, $\Pr^{\circ}\left[T_{ik_1\cdots k_d} \leq t\right]$, times the initial frequency of a. Calculating $\pi_{a|A}$ using the analogue of (21), expanding the expected products of q_i, and summing over time yields

$$\pi_{a|A} = \frac{1}{N_{\mathrm{T}}} + \frac{\omega}{N_{\mathrm{T}}} \sum_{d=1}^{N_{\mathrm{T}}} \sum_{i=1}^{N_{\mathrm{T}}} \sum_{k_1<\ldots<k_d}^{N_{\mathrm{T}}} s_{i,k_1\cdots k_d} \sum_{l=1}^{d} (-1)^l \sum_{m_1<\cdots<m_l}^{\{k_1,\ldots,k_d\}} \left(\mathrm{E}^{\circ}\left[T_{m_1\cdots m_l}\right] - \mathrm{E}^{\circ}\left[T_{im_1\cdots m_l}\right]\right) / N_{\mathrm{T}} + O(\omega^2) \tag{23}$$

where the last sum on the right hand is over all $m_1 < \cdots < m_l$ drawn from the set $\{k_1, \ldots, k_d\}$ and $\mathrm{E}^{\circ}\left[T_{k_j}\right] = 0$ for any single lineage k_j by definition.

With expressions from both fixation probabilities, $\pi_{A|a}$ from equation (22) and $\pi_{a|A}$ from equation (23), we can combine them to complete the condition for allele A to be more common than a at stationarity, $\pi_{A|a} > \pi_{a|A}$,

which becomes

$$\sum_{d=1}^{N_T}\sum_{i=1}^{N_T}\sum_{k_1<\ldots<k_d}^{N_T} s_{i,k_1\cdots k_d}\left(\mathrm{E}^{\circ}\left[T_{ik_1\cdots k_d}\right]+\sum_{l=1}^{d}(-1)^l\sum_{m_1<\cdots<m_l}^{\{k_1,\ldots,k_d\}}\left(\mathrm{E}^{\circ}\left[T_{m_1\cdots m_l}\right]-\mathrm{E}^{\circ}\left[T_{im_1\cdots m_l}\right]\right)\right)<0 \tag{24}$$

to first order in ω. This is the main condition for evolutionary success for the short-term TSS. Given a population where the demography and the relationship between gene expression and fitness are known (i.e., the selection coefficients $s_{i,k_1\cdots k_d}$ are known), all that remains is to calculate expected coalescence times under neutrality, which will be a function of the demography. For models of cooperative behavior, condition (24) generates both Hamilton's rule when genetic interactions are additive ($d=1$) and the risk-dominance condition when interactions are non-additive ($d>1$).

2.4. Additive genetic interactions and evolutionary success

The effect of selection on fixation probability given in (22) can be arbitrary complex if individual fitness depends on any combination of genotypes (i.e., products of allele frequency) of individuals in the population. In the simplest social interactions, fitness only depends additively on the genotype of focal individual and the genotypes of other individuals in the population, which means $d=1$ in equation (17). Using $d=1$ in the expression for $\pi_{A|a}$ in (22), we find that

$$\pi_{A|a}=\frac{1}{N_T}-\frac{\omega}{N_T}\sum_{i=1}^{N_T}\sum_{j=1}^{N_T}s_{i,j}\,\mathrm{E}^{\circ}\left[T_{ij}\right]/N_T+O(\omega^2) \tag{25}$$

Similarly, applying $d=1$ for the expression for $\pi_{a|A}$ produces

$$\pi_{a|A}=\frac{1}{N_T}+\frac{\omega}{N_T}\sum_{i=1}^{N_T}\sum_{j=1}^{N_T}s_{i,j}\,\mathrm{E}^{\circ}\left[T_{ij}\right]/N_T+O(\omega^2)\,, \tag{26}$$

and this immediately yields

$$\pi_{A|a}+\pi_{a|A}=\frac{2}{N_T}+O(\omega^2)\,, \tag{27}$$

which transforms the evolutionary success condition $\pi_{A|a}>\pi_{a|A}$ to (after neglecting $O(\omega^2)$ terms)

$$\pi_{A|a}>\frac{1}{N_T}\,, \tag{28}$$

which is the standard condition for type A to be advantageous compared to a neutral allele. Inserting the first-order expansion for $\pi_{A|a}$ in equation (12) in condition (28) produces another evolutionary success condition for allele A:

$$\frac{\mathrm{d}\pi_{A|a}}{\mathrm{d}\omega} > 0\,. \tag{29}$$

Thus, additive genetic interactions simply the analysis of evolutionary success considerably since we have to evaluate only one fixation probability, $\pi_{A|a}$, or its derivative with respect to the strength of selection. Conversely, analyses that only use condition (28) and compare the fixation probability of one type versus fixation under neutrality are correct predictors evolutionary success *only* when interactions are additive. For example, the one-third law from evolutionary game theory is an evaluation of condition (28) in a finite population where individuals interact in a pairwise manner and the payoffs from their interactions can produce non-additive genetic interactions [109, 113]. In effect, the one-third law then yields a measure of the relative stability of a population fixed for allele a, but does not provide information about the stability of the population when fixed for allele A [see Fig. 2b in 109] unless genetic interactions are additive.

2.5. Weak effect mutations and continuous phenotypes

An important case where genetic interactions are additive occurs when the difference between the phenotypes produced by alleles A and a is small and phenotypes are allowed to take a continuum of values. Suppose that the phenotype of type a is z and that of type A is $z+\delta$ where δ is called the phenotypic deviation. Further, let the phenotype of individual i be $z_i = z + \delta p_i$, and the vector $\mathbf{z} = \mathbf{z}(\mathbf{p}) = (z_1, \ldots, z_{N_\mathrm{T}})$ contain the phenotypes for the whole population. Since phenotype is a continuous variable, we assume that the fitness of each individual i, $w_i(\mathbf{z})$, is a differentiable function of phenotype [p. 41 in ref 24]. This also implies, using equations (3) and (11), that the fixation probability a single mutant with trait $z+\delta$ in a population with resident trait z, denoted $\pi(\delta, z)$, is differentiable. When the phenotypic deviation δ is small (weak effect mutations), we can ignore terms $O(\delta^2)$ and individual fitness can be written as a Taylor series in δ:

$$\begin{aligned} w_i(\mathbf{z}) &= 1 + \delta\frac{\mathrm{d}w_i(z)}{\mathrm{d}\delta} + O(\delta^2) \\ &= 1 + \delta\sum_{i}^{N_\mathrm{T}}\frac{\partial w_i(z)}{\partial z_j}p_j + O(\delta^2) \end{aligned} \tag{30}$$

where $\frac{\partial w_i}{\partial z_j}$ are evaluated at $\delta = 0$ (written as $\frac{\partial w_i(z)}{\partial z_j}$) when the population is fixed for allele a and $w_i(z) = 1$. Comparing equation (30) to the expression for fitness when phenotypes are discrete (eq. 17) reveals that the

phenotypic deviation δ is analogous to the selection strength ω and that the derivatives of fitness with respect to phenotype, $\frac{\partial w_i(z)}{\partial z_j}$, are equivalent to additive selection coefficients. Thus, so-called "δ-weak" selection [163] implies additivity of fitness effects and allele frequency, which is well known in the literature [e.g.: 123, 138].

Finally, using the fitness function in (30) in the fixation probability equation (25) and taking the first-order term, we obtain the the derivative of the fixation probability $\pi(\delta, z)$ with respect to δ (since it measures selection strength) evaluated at $\delta = 0$,

$$\left(\frac{\mathrm{d}\pi}{\mathrm{d}\delta}\right)_{\delta=0} = S(z) = -\frac{1}{N_\mathrm{T}} \sum_{i=1}^{N_\mathrm{T}} \sum_{j=1}^{N_\mathrm{T}} \frac{\partial w_i(z)}{\partial z_j} \mathrm{E}^\circ\left[T_{ij}\right] / N_\mathrm{T}\,, \tag{31}$$

which is often called the phenotypic "selection gradient", $S(z)$, in adaptive dynamics [50, 93], inclusive fitness theory [90]. and quantitative genetics [81]. The "gradient" terminology suggests that the zeros of the selection gradient, which correspond to extrema of the fixation probability, will be candidate evolutionary equilibria. We will show this to be the case though only once we have described evolution under the long-term TSS in section 4.

2.6. Coalescence time and identity by descent

So far, we have shown how the fixation probability of an allele with effects on social behavior depends on mean coalescence times (eq. 22). However, the relatedness term in Hamilton's rule is often expressed as a function of probabilities of identity by descent [58, 60]. Translating between mean coalescence times and probabilities of identity by descent is possible using an argument first presented by Slatkin [133]. Suppose that Q_{ij} represents the expected probability of identity by descent between alleles i and j. Since mutations are distributed independently on a neutral genealogy [157] and IBD requires that the alleles not mutate before coalescing, the IBD probability can be expressed as

$$Q_{ij} = \sum_{t=1}^{\infty} (1-\mu)^{2t} \Pr^\circ\left[T_{ij} = t\right]. \tag{32}$$

Since the TSS assumes weak mutation, we can ignore terms $O(\mu^2)$ and rewrite equation (32) as

$$\mathrm{E}^\circ\left[T_{ij}\right] = \lim_{\mu \to 0} \frac{1 - Q_{ij}}{2\mu}\,. \tag{33}$$

We use the limit as $\mu \to 0$ in equation (33) since we derived the fixation probabilities and their dependence on coalescence time under the TSS assumption that new mutation is not possible until the old mutation either

fixes or goes extinct. Equation (33) only pertains to pairwise coalescence times and IBD probabilities, so we can only apply it to additive genetic interactions. The relationship between three-way (and more generally d-way) coalescence times and IBD probabilities is more complex and deserves further study.

Applying the relationship between pairwise coalescence times and IBD probabilities in (33) to the selection gradient in (31) in the case of δ-weak selection yields

$$S(z) = \lim_{\mu\to 0} \frac{1}{2N_{\mathrm{T}}^2\mu} \sum_{i=1}^{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \frac{\partial w_i(z)}{\partial z_j} Q_{ij} \,. \tag{34}$$

This expression was first obtained by Rousset and Billiard [125], and an analogous derivation was presented by Rousset [122]. For models with simple population structure (homogenous structures [112, 143] like the island [168] or stepping-stone [78] models) and simple demography, the IBD probabilities Q_{ij} are relatively easy to obtain in the low mutation limit. The fitness function $w_i(\mathbf{z})$ depends on nature of the social interactions as well as on the demography and population structure.

2.7. Inclusive fitness effect and Hamilton's rule

Essentially, the right hand side of (34) is a measure of how expression of allele A affects inclusive fitness [123, 125]; fitness effects are given by the derivative of the fitness of individual i with respective to the phenotype of individual j, and each effect is weighted by likely individuals i and j are to share alleles IBD. Applying the evolutionary success condition $S(z) > 0$ to (34) yields a Hamilton-type rule,

$$\lim_{\mu\to 0} \frac{1}{\mu} \sum_{i=1}^{N_{\mathrm{T}}} \sum_{j=1}^{N_{\mathrm{T}}} \frac{\partial w_i(z)}{\partial z_j} Q_{ij} > 0 \,, \tag{35}$$

where the population structure and fitness functions haven't been specified. In order to obtain the classic form of Hamilton's rule from condition (1), we make some simplifying assumptions about the social interaction and population structure.

Suppose that there is a homogeneous structure with n groups each containing N haploid individuals ($N_{\mathrm{T}} = nN$). This implies that we need only to track two kinds of IBD probabilities, Q_0, which measures the chance that two alleles drawn from different individuals in the same group are IBD, and Q_1, which is the probability that two alleles from different groups are IBD. Individuals socially interact within their group, but social effects between groups also occur due to differential productivity of groups (i.e., "hard selection" [21]). Since the fitness derivatives $\frac{\partial w_i(z)}{\partial z_j}$ are evaluated at $\delta = 0$ where all individual behave the same (as if they express the

a allele), there are only three different fitness derivatives: (i) individuals i and j are the same individual, and $\frac{\partial w_i(z)}{\partial z_i}$ is the effect of the individual's behavior on itself, which we call the "cost" or $-c$; (ii) individuals i and j live in the same group, and $\frac{\partial w_i(z)}{\partial z_j}$ is the effect on the individual due to its group mate's behavior, which we call the "benefit" or $b/(N-1)$ (for each of the $N-1$ group mates); and (iii) individuals i and j live in different groups where we can set $\frac{\partial w_i(z)}{\partial z_j} = -(b-c)/((n-1)N)$ since the selection coefficients must sum to zero (eq. 18). Putting these expression into equation (35) and simplifying produces

$$\lim_{\mu\to 0} \frac{1-Q_1}{\mu}\left(-c + b\,\frac{Q_0-Q_1}{1-Q_1}\right) > 0\,. \tag{36}$$

Typically, in finite populations ($N_T < \infty$), the IBD probabilities Q_0 and Q_1 will go to one as the mutation rate goes to zero since one lineage will eventually fix in the population. In these cases (such as the island model), the IBD probabilities can often be expressed as $1-\mu\,O(1)$ [123], which suggests that the first term in (36) has a positive limit as $\mu \to 0$. The ratio multiplying the benefit b turns out to be Wright's F_{ST}, which will also have a positive limit under low mutation. Setting the relatedness to $r = F_{ST} = \frac{Q_0-Q_1}{1-Q_1}$ and simplifying, we then obtain Hamilton's rule for this population

$$-c + b\,r > 0\,.$$

where the left-hand side of the inequality is the inclusive fitness effect.

There are two important points about Hamilton's rule that are illuminated by the above derivation. The first is that the cost and benefit terms are effects on *fitness* as measured over the whole lifecycle. This implies that fitness effects will be functions of demographic parameters (such as population size and migration rate) and of social effects on both survival and fertility. In order to determine the signs of $-c$ and b, which are required for classifying a behavior as altruistic or not (see Table 2), the demography and the effect of social behavior on both fertility and survival must be specified. Second, relatedness is a measure of relative genetic identity that depends on population structure. In some simple cases [e.g., family structured populations; 128], this will simplify to classic pedigree measures of relatedness, but generally r will depend explicitly on the demography and population structure.

3. Application: social games in island-type populations

To recap a bit, we have shown above how weak mutation and the TSS allow a simple criterion for evolutionary success, $\pi_{A|a} > \pi_{a|A}$, in the short term. When selection is weak and genetic interactions are additive,

this simplifies the condition for evolutionary success considerably and we recover a measure of inclusive fitness and Hamilton's rule in (34). When non-additive genetic interactions are important and selection is still weak, the evolutionary success condition is given generally by condition (24).

There are many biological scenarios where non-additive interactions are important though the simplest one that is often invoked is a two-player game. Each individual plays one of two pure strategies, "cooperation" or "noncooperation", with a social partner where the payoffs for the game are given in Table 3. When two individuals are noncooperators, they receive no payoff. If one does not cooperate and the other cooperates, the cooperator receives $-C$ and the noncooperator receives B. When two individuals cooperate, they each receive payoff $B - C + D$ where D is a measure of non-additivity or "synergy". The strategy names and payoffs are inspired by the Prisoner's Dilemma game [11] where B and C are positive and $D \geq 0$, though we will allow the parameters to take negative values as well in order to study other games like the Stag-Hunt [132]. In the simplest case, the strategies are fixed by the genotype of the individual where individuals bearing the A allele cooperate and individuals bearing a do not cooperate. This means there is no phenotypic plasticity that might result from changing strategies over repeated interactions (e.g., reciprocity [9, 148] or responsiveness [2, 3]).

We assume an island-type population structure where n groups of N haploids ($N_{\mathrm{T}} = nN$) are connected each by migration rate m. Generations are non-overlapping. The frequency of the cooperation allele A in individual i in group g is p_{gi}, the mean frequency in group g is p_g including individual i and $p_{g\backslash i}$ excluding i, and p is the mean frequency in the whole population. Individuals choose social partners at random and the mean payoff from the social interactions in the group determines how each individual's fertility differs from the baseline of one. Given these assumptions, the fertility of individual i in group g is $1 + \omega f_{gi}$ where

$$f_{gi} = 1 + B\, p_{g\backslash i} - C\, p_{gi} + D\, p_{gi} p_{g\backslash i} \tag{37}$$

and B and C are clearly additive and D non-additive in allele frequency. This fertility function represents not only the two-player case but also the n-player case within the group when payoff is an additive function of the frequency of other player types in the group [152, Appendix C].

3.1. Baseline demography: hard selection

We begin with a model of hard selection (see section 2.7) with groups producing different numbers of migrants depending on their composition. To first order in ω, the fitness of individual i in group g can be

written implicitly as

$$w_{gi} = \frac{(1-m)\big(1+\omega f_{gi}\big)}{(1-m)\big(1+\omega f_g\big) + m\big(1+\omega f_{\setminus g}\big)} + \frac{1}{n-1}\sum_{k\neq g}^{n} \frac{m\big(1+\omega f_{gi}\big)}{(1-m)(1+\omega f_k) + m\big(1+\omega f_{\setminus k}\big)}\,, \tag{38}$$

where $f_{\setminus g}$ is the mean fertility in the population excluding group g, or explicitly as

$$w_{gi} = 1 + \omega\Big(f_{gi} - \Big[(1-m)^2 f_g + m(2-m) f_{\setminus g} + \frac{m}{n-1}\big(f_g - f_{\setminus g}\big)\Big]\Big)\,, \tag{39}$$

which takes the same form as equation (17). All that remains in order to calculate the fixation probabilities in equations (22) and (22) are the expected coalescence times under neutrality.

Fortunately, coalescence times in structure populations are a well-studied topic [67, 108, 155, 164] in coalescent theory where the process often considered is called the structured coalescent [106, 107]. As applied to the island model, the structured coalescent usually assumes that the migration rate m is $O(1/N)$ and $nNm/(n-1) \to M$ as $N \to \infty$. This impies that during a small interval of time, either two lineages within a group can coalesce or one lineage can migrate from one group to another, but more than one such event does not occur. For two lineages, either both can be in the same group with configuration $(0,1)$ or each can be in different groups with configuration $(2,0) = (2)$ where the first element of the configuration is the number of groups with a single lineage, the second element is the number of groups with two lineages, etc (e.g., see [155]). We denote the expected coalescence times of these configurations by $\mathrm{E}^{\circ}\big[T_{(0,1)}\big]$ and $\mathrm{E}^{\circ}\big[T_{(2)}\big]$, respectively. With synergistic payoffs ($D \neq 0$) that create non-additive fitness effects, we also have to track three lineage samples. For three lineages, there are three possible configurations: (3), $(1,1)$, and $(0,0,1)$. Using the master equation for the continuous-time Markov process that describes coalescence (eq. 2.8 in [108]), a system of equations for the five expected coalescence times can be constructed (Wakeley [155] describes the method nicely); these equations are given in section 5.1 of [80], and we only provide the solutions here (time in absolute units):

$$\begin{aligned}
\mathrm{E}^{\circ}\big[T_{(0,1)}\big] &= N_{\mathrm{T}}, & \mathrm{E}^{\circ}\big[T_{(2)}\big] &= N_{\mathrm{T}}\Big(1 + \frac{1}{2M}\Big), \\
\mathrm{E}^{\circ}\big[T_{(0,0,1)}\big] &= N_{\mathrm{T}}\Big(\frac{4}{3} + \frac{n-1}{6n(1+M)}\Big), & \mathrm{E}^{\circ}\big[T_{(1,1)}\big] &= N_{\mathrm{T}}\Big(\frac{4}{3} + \frac{1}{2M} - \frac{1}{6n(1+M)}\Big), \\
& & \mathrm{E}^{\circ}\big[T_{(3)}\big] &= N_{\mathrm{T}}\Big(\frac{4}{3} + \frac{2}{3M} - \frac{1}{6n(1+M)}\Big).
\end{aligned} \tag{40}$$

Note that evaluating a more complex fitness function with higher-order frequency dependence would require

calculating expected coalescence times for four lineage or larger samples. The number of configurations and equations grows quickly as the lineage sample size increases, which makes this method cumbersome for complex fitness functions. Working on the related problem of calculating the total length of the coalescent genealogy, Wakeley [155] shows expected coalescence times can be calculated for arbitrarily large samples so long as n is large; this suggests, an analogous method might work for coalescence times that could be used to calculate fixation probabilities.

Applying the fitness function from (39) and the coalescence times in (40) to equation (22), we calculate the probability allele A (the cooperation allele) fixes in a population of a (the noncooperation allele) as

$$\pi_{A|a} = \frac{1}{N_{\mathrm{T}}} + \omega\left(-C + D\left(\frac{1}{3} + \frac{1 - \frac{1}{n}}{6(1+M)}\right)\right) \tag{41}$$

and the fixation probability of a in a population of A as

$$\pi_{a|A} = \frac{1}{N_{\mathrm{T}}} + \omega\left(C - D\left(\frac{2}{3} - \frac{1 - \frac{1}{n}}{6(1+M)}\right)\right). \tag{42}$$

First derived by Ladret and Lessard [80, eq. 29], these expressions are correct to first order in selection strength ω and zeroth order in $O(1/N)$ since our coalescence times assume that $N \to \infty$. These two fixation probabilities are the first main result of this section and produce a few important observations. First, they reproduce the classic cancellation result from Taylor [139, 140, 143] for additive genetic interactions ($D = 0$). Taylor's result says that the benefits of cooperation are exactly balanced out by the effect of competition between related individuals within a group (so-called "kin competition" or "local competition") when population structure is "homogenous" [143] and generations do not overlap. In our model notation, the cancellation result implies that the benefit to others B will cancel out of the fixation probability expressions; equations (41) and (42) show that this indeed does occur. Thus, when interactions are additive ($D = 0$), the cooperation allele A fixes with a probability greater than the neutral probability $\pi^\circ = 1/N_{\mathrm{T}}$ (eq. 28) only when there is negative direct cost (i.e., cooperation is directly beneficial). Likewise, $\pi_{a|A}$ in equation (42) shows that the noncooperation allele a is advantageous when the cost is positive. In an extension of the cancellation result, Ohtsuki [112] shows (for $n \to \infty$) that positive synergy that does not change the structure of the game, $0 < D < C$, cannot not result in positive selection for cooperation. Equations (41) and (42) reproduce this result since even the strongest population structure, $M \to 0$, results cooperation fixing more likely than chance only when $D > 2C$.

The expression for $\pi_{A|a}$ in equation (41) also produces the one-third law from evolutionary game theory [109]. The one-third law says that as $N \to \infty$ in a single panmictic population, the cooperation allele A fixes with a probability greater than chance when the mixed strategy equilibrium of the game in Table 3 is less than one third. If an opponent cooperates with probability x and does not cooperate with probability $1 - x$, the mixed strategy equilibrium is the value of x where an individual does equally well against the opponent by either cooperating or not cooperating [69]. Using the payoffs from Table 3, the mixed strategy equilibrium equation is $x^*(B - C + D) - (1 - x^*)C = x^*B$, which yields $x^* = C/D$. Thus, the one-third law translates to

$$D > 3C\,. \tag{43}$$

We can immediately recover the one-third law from $\pi_{A|a} > 1/N_{\mathrm{T}}$ by taking the high migration limit $M \to \infty$ in (41), which results in an unstructured population. The complementary condition for fixation of the noncooperation allele a, $\pi_{a|A} > 1/N_{\mathrm{T}}$, becomes $D < 3C/2$ in the high migration limit. In contrast, population structure is at its strongest in the low migration limit when $M \to 0$ and when the number of groups is large, $n \to \infty$. Fixation of the cooperation allele becomes easier in this case as $\pi_{A|a} > 1/N_{\mathrm{T}}$ translates to $D > 2C$. Conversely, fixation of the noncooperation allele also becomes easier when population structure is strong since the fixation condition becomes $D < 2C$. These conditions are summarized in Table 4.

As discussed in section 2.4, each fixation condition alone (and, consequently, the one-third law, $\pi_{A|a} > 1/N_{\mathrm{T}}$) is sufficient as a measure of evolutionary success only when genetic interactions are additive. When non-additive or synergistic interactions are included, the condition $\pi_{A|a} > \pi_{a|A}$ (eq. 9) should be used. This condition is

$$D > 2C \tag{44}$$

using the fixation probabilities in (41) and (42) (see Table 4). Interestingly, this implies that whether the cooperation allele A or the noncooperation allele a is more common at stationarity is independent of the strength of population structure, M, and depends only on the payoffs in the social game. This is a generalization of the Taylor cancellation result in the sense that the kind of simple population structure considered here (non-overlapping generations, homogenous island-type migration, hard selection) is not sufficient for the benefits of cooperation B to affect selection for cooperation. Rather, synergistic effects are important, but they must be significantly outweigh the costs in order for cooperation to be more prevalent than noncooperation.

In fact, once synergistic effects outweigh the costs at all, $D > C$, they change the structure of the social game from a Prisoner's Dilemma where noncooperation is the strictly dominant strategy to a Stag-Hunt or coordination game where both cooperation and noncooperation are Nash equilibria [11]. In coordination games in unstructured populations, resident populations of cooperators and noncooperators are both resistant to invasion by the complementary type when evolution is deterministic (i.e., there is no genetic drift). This implies that whether allele A or a becomes fixed in the population depends on the initial frequency of A. When the initial frequency is greater than the mixed strategy equilibrium $x^* = C/D$, selection leads to fixation of the cooperation allele A, and fixation of the noncooperation allele a occurs for initial frequencies less than C/D. In effect, if the phenotypic space is the probability of cooperation x, then the mixed strategy equilibrium is a fitness valley and pure cooperation and noncooperation are fitness peaks [152]. The basin of attraction for the cooperation peak in this case would be $(x^*, 1)$, and for the noncooperation peak the basin would be $(0, x^*)$. An intuitive condition for the cooperation peak to be more likely to evolve is that is has the larger basin of attraction under a model of simple deterministic evolution. This condition, which is called "risk dominance" in game theory [64, 72], is equivalent to $1 - x^* > x^*$ or

$$x^* = \frac{C}{D} < 1/2 \,. \tag{45}$$

However, this is exactly the same condition we obtained from $\pi_{A|a} > \pi_{a|A}$ in (44). The fact that we can derive the evolutionary success condition for cooperation in a structured population with selection, drift, and mutation from the risk dominance condition in a purely deterministic model is another way of describing the generalized cancellation result introduced above.

Even though the equivalence between $\pi_{A|a} > \pi_{a|A}$ and risk dominance is proven here for the infinite island model ($n \to \infty$ and $N \to \infty$), it approximately holds for the finite island model as well. In order to show this, we make two observations. First, examining the general condition for $\pi_{A|a} > \pi_{a|A}$ in equation (24), we observe that so long as fitness depends on at most three way genetic interactions ($d = 2$), which is true for the fitness function in (39), all three-way coalescence times cancel out. Thus, we only need exactly pairwise coalescence times (we elaborate on further ramifications of this fact in the Discussion). Second, as suggested by Ladret and Lessard [80, p. 416], exact expected coalescence times can be calculated using a discrete-time Markov process that produces the same linear equations in Notohara [108, p. 66] that were used to generate the structured coalescent results in (40). These equations, given by (I) in Ladret and Lessard [80], produce the following exact

expected coalescence times:

$$\mathrm{E}^{\circ}\left[T_{(0,1)}\right] = N_{\mathrm{T}}, \quad \mathrm{E}^{\circ}\left[T_{(2)}\right] = N_{\mathrm{T}}\left(1 - \frac{1}{N} + \frac{1}{M(2 - M/N)}\right). \tag{46}$$

Combining the above expressions and the fitness function from (39) with the evolutionary success condition $\pi_{A|a} > \pi_{a|A}$ in equation (24), we get

$$D > 2C + \frac{2(B - C + D)}{N_{\mathrm{T}}}. \tag{47}$$

Condition (47) contains a single correction to the risk dominance condition, $2(B - C + D)/(N_{\mathrm{T}})$, which is due to local competition in a finite population. Compared to the infinite island model, the cooperation allele A is only slightly less prevalent in this case. If the number of groups is infinite or if each group if infinitely large, condition (47) simplifies to risk dominance.

3.2. Demography and the scale of competition

Under the baseline demography of hard selection, we obtain a general cancellation result where the additive benefits of cooperation are canceled. This cancellation is a function of how demography shapes both the competitive environment and genetic identity within and between groups. Thus, demographies that create a different competitive environment may be more or less conducive towards the evolutionary success of the cooperation allele.

One of the most common alternative demographies to hard selection is soft selection [21] where individuals compete for resources or breeding spots within their group before the migration stage. Thus, each group contributes the same number of individuals to the next generation. Intuitively, this should make the it more difficult for the cooperation allele A to succeed since groups with a higher frequency of A will not be more productive than groups with a lower frequency. In this case, the fitness function is

$$w_{gi} = \frac{1 + \omega f_{gi}}{1 + \omega f_{g}}, \tag{48}$$

[123][p. 125] when the number of groups is large ($n \to \infty$). For additive genetic interactions ($D = 0$) and using the exact pairwise expected coalescence times in (46), the evolutionary success condition $\pi_{A|a} > 1/N_{\mathrm{T}}$ is

$$-C > \frac{B - C}{N}, \tag{49}$$

which agrees with previous analyses [90, 123]. In contrast to the case of hard selection where the evolutionary success condition is $-C > 0$, soft selection increases the strength of local competition so that cooperation is actually selected against in finite groups with a strength proportional to the net benefits, $B - C$. Allowing for non-additive interactions ($D \neq 0$), we calculate the evolutionary success condition $\pi_{A|a} > \pi_{a|A}$ to be

$$D > 2C + \frac{2(B - C + D)}{N}, \tag{50}$$

which again has a stronger local competition correction to the risk dominance condition than the hard selection case (eq. 47). In fact, since it doesn't depend on the number of groups n, the soft selection correction is not simply a finite population size ($N_{\mathrm{T}} < \infty$) effect and is due to competition occurring exclusively within groups.

A strong contrast to the soft selection is where the social interaction still occurs within the group but the scale of competition is the whole population. This might occur when groups directly compete for resources and successful groups produce propagules whose genotype frequency is proportional to their frequency within the group [e.g., 48, 88, 90]. If migration occurs at the adult stage, then the fitness function is given by

$$w_{gi} = \frac{1 + \omega f_{gi}}{1 + \omega f}, \tag{51}$$

and the neutral demography is still represented by an island model whose expected coalescence times are given in equations (40). The evolutionary success condition $\pi_{A|a} > \pi_{a|A}$ for this case is

$$D + \frac{B - C + D}{M}\left(1 - \frac{1}{n}\right) > 2C \tag{52}$$

for large groups ($N \to \infty$). When migration is high and the population is unstructured ($M \to \infty$), we recover the risk dominance condition from (52). Strong population structure from weak migration however strongly selects for cooperation. In fact, for any level of cost, there is a low enough population migration rate M that results in more cooperation alleles A at stationarity. This strong selection for cooperation is a direct result of the lack of local competition, which allows the benefits of cooperation to accrue to individuals who cooperate through the other individuals in their group who share their genes IBD.

4. Theory: long-term evolution

We showed above how the forces of selection, mutation, and genetic drift generate the stationary distribution between two alleles when mutation is weak and evolutionary change follows the TSS. This stationary distribution gave us a measure of evolutionary success of one allele relative to another, which was simply $\pi_{A|a} > \pi_{a|A}$. Under the additional assumption of weak selection, we showed how to calculate this condition (eq. 24) for arbitrary non-additive genetic interactions. From this condition, we obtained Hamilton's rule, the one-third law, risk dominance, and generalizations of these conditions. However, this condition only gives the stationary distribution among a fixed set of alleles and thus does not explicitly make predictions for longer timescales when the evolutionary process can sample a continuum of alleles.

Studying long-term evolution among a continuum of possible alleles requires specifying how those alleles are generated by mutation and how mutations are fixed or lost over the long-term due to selection and drift. Just as with the short-term model among a finite set of alleles, we will assume weak mutation so that the population can be described by the TSS and we only need to track the evolution of a population from one fixed, or monomorphic, state to another. Our approach to modeling the long-term process uses the "substitution rate" approach of Lehmann Lehmann [82], Van Cleve and Lehmann [152], which derives from population genetic approaches to adaptation [51, 52] and to kin selection in finite populations [123, 125, 141, 145] and from the adaptive dynamics approach [18, 30, 101].

4.1. Substitution rate approach and the TSS diffusion

The essence of this approach is that, over the long term, the evolutionary process at each point in time can be fully characterized by a substitution or transition rate that measures how likely the population is to move from one monomorphic state to another. We assume that the substitution rate is an instantaneous measure of change, which is justified since organismal life cycles and generation times become very short on the scale of long-term evolution. If $\rho(z_1, t_1|z_0, t_0)$ is the probability density of that a population is monomorphic for trait z_1 at time t_1 given it was monomorphic for z_o at time t_0, then we define the substitution rate as

$$\lim_{\Delta t \to 0} \frac{\rho(z+\delta, t+\Delta t|z,t)}{\Delta t} = k(\delta, z)\ ,$$

which is simply the rate at which mutations of type $z+\delta$ are produced and fixed in the population of type z. The substitution rate is a function of both the mutation rate, μ, and the distribution of mutational effects, $u(\delta, z)$, which represents the probability density that a mutant offspring is of type $z+\delta$ given that its parent is of type z.

For simplicity, we assume that μ does not depend on the resident trait in the population, z. Using weak mutation and the TSS condition in (4), Champagnat [17, 20] showed that the long-term TSS can be characterized as a Markov jump process (continuous in time and phenotypic space) with instantaneous jump rate

$$k(\delta, z) = N_{\mathrm{T}}\, \mu\, u(\delta, z)\, \pi(\delta, z) \tag{53}$$

(see also eq. 2 in [20] and eq. 2 in [82]). This rate is conceptually analogous to the classic long-term neutral substitution rate $k = N_{\mathrm{T}}\mu(1/N_{\mathrm{T}}) = \mu$ from molecular evolution [77].

Following standard methods in stochastic processes, the Markov jump process representing the TSS with jump rate $k(\delta, z)$ can be represented with the following (forward) master equation [30, 46, 82]

$$\frac{\partial \rho(z,t)}{\partial t} = \int k(\delta, z-\delta)\rho(z-\delta, t) - k(\delta, z)\rho(z,t)\, \mathrm{d}\delta\,, \tag{54}$$

where we write $\rho(z,t) = \rho(z,t|z_0,t_0)$ for simplicity. The master equation captures the intuition that the change in probability density for trait z at time t is equal to the sum of the jumps towards that trait minus the sum of the jumps away from that trait. Without further assumptions about the size of jumps, the master equation is difficult to analyze. A common way to approximate Markov jump processes is to assume that the jumps are small, which turns the discontinuous jump process into a continuous process and turns the master equation into a diffusion equation. Specifically, standard methods (e.g., a Kramers-Moyal expansion [46]) generate a diffusion equation by ignoring third-order and higher moments of the jump process. Biologically, we can justify this approximation by assuming that the mutational effects δ cluster tightly enough around the mean trait value z so that the evolutionary dynamic is affected only by the variance of the distribution of mutational effects and higher order moments can be neglected. The (forward) diffusion equation obtained with these methods for the jump process in (54) is

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial z}[a(z)\rho(z,t)] + \frac{1}{2}\frac{\partial^2}{\partial z^2}[b(z)\rho(z,t)] \tag{55}$$

where

$$a(z) = \int \delta k(\delta, z)\, \mathrm{d}\delta \tag{56}$$

is the "drift" term and measures the mean jump away from the trait z and

$$b(z) = \int \delta^2 k(\delta, z)\, \mathrm{d}\delta \tag{57}$$

is the "diffusion" term and measures the variance of the jumps away from z. The drift and diffusion terms can be simplified by assuming that the fixation probability is differentiable (which is true if fitness is differentiable; see section 2.5) and approximating the substitution or jump rate by the first-order Taylor series

$$k(\delta, z) = N_\mathrm{T}\, \mu\, u(\delta, z)\big(\pi^\circ(z) + S(z)\delta + O\big(\delta^2\big)\big) \tag{58}$$

where $\pi^\circ(z) = \pi(0, z)$ and $S(z) = \left(\frac{\mathrm{d}\pi(\delta,z)}{\mathrm{d}\delta}\right)_{\delta=0}$ is the selection gradient. Using the expansion in (58) and assuming that the mutational distribution is symmetric in δ, the drift and diffusion terms become, respectively,

$$\begin{aligned} a(z) &= N_\mathrm{T}\, \mu \int \delta u(\delta, z)\big(\pi^\circ(z) + S(z)\delta + O\big(\delta^2\big)\big)\, \mathrm{d}\delta \\ &= N_\mathrm{T}\, \mu\, S(z) \int \big(\delta^2 + O\big(\delta^3\big)\big) u(\delta, z)\, \mathrm{d}\delta \\ &\approx N_\mathrm{T}\, \mu\, \sigma^2(z) S(z) \end{aligned} \tag{59}$$

and

$$\begin{aligned} b(z) &= N_\mathrm{T}\, \mu \int \delta^2 u(\delta, z)\big(\pi^\circ(z) + S(z)\delta + O\big(\delta^2\big)\big)\, \mathrm{d}\delta \\ &\approx N_\mathrm{T}\, \mu\, \sigma^2(z) \pi^\circ(z) = \mu\, \sigma^2(z) \end{aligned} \tag{60}$$

where $\pi^\circ(z) = 1/N_\mathrm{T}$, $\sigma^2(z) = \int \delta^2 u(\delta, z)\, \mathrm{d}\delta$ is the second moment (or raw variance) of the mutational effects distribution and third-order and higher moments are neglected. Substituting these expressions for $a(z)$ and $b(z)$ into the diffusion equation (57) yields

$$\frac{\partial \rho}{\partial t} = -N_\mathrm{T}\, \mu\, \frac{\partial}{\partial z}\big[S(z)\sigma^2(z)\rho(z, t)\big] + \frac{\mu}{2}\frac{\partial^2}{\partial z^2}\big[\sigma^2(z)\rho(z, t)\big]\,, \tag{61}$$

which was derived by Lehmann [82, eq. 4] and is analogous to the stochastic differential equation derived by Champagnat and Lambert [20, eq. 3]. The diffusion equation in (61) is our main mathematical description of how the TSS evolves over the long-term. In a sense, this diffusion equation is a stochastic version of the

deterministic canonical equation of adaptive dynamics [30]. The first term in (61) measures the deterministic effect of selection on the trait and is the counterpart of the canonical equation of adaptive dynamics [20]. The second term measures the stochastic effect of genetic drift through the neutral fixation rate (eq. 60).

4.2. Evolutionary success in the long term TSS

Just as in the short-term TSS analysis, we will define the evolutionary success for a trait z in the long-term TSS as its stationary probability or $\rho(z) = \lim_{t\to\infty} \rho(z,t)$. Using the long-term TSS diffusion in (61) and standard methods [38, 46, 75], we find that

$$\rho(z) = \frac{1}{K\sigma^2(z)} \exp\left[2N_{\mathrm{T}} \int^z S(y)\, \mathrm{d}y\right] \tag{62}$$

where K is a normalizing constant that ensures $\rho(z)$ integrates to one over its support. The most successful traits wil be those that reside at peaks of the stationary distribution, and the least successful will reside at troughs. Obtaining the peaks and troughs, when they do not reside at the boundaries of trait space, requires calculating the extrema of $\rho(z)$, which must satisfy

$$S(z) = \frac{1}{2N_{\mathrm{T}}} \frac{\mathrm{d}\log\sigma^2(z)}{\mathrm{d}z} \tag{63}$$

evaluated at a candidate extremum $z = z^*$. So long as either population size N_{T} is very large or the mutational variance $\sigma^2(z)$ does not depend on the resident trait z, equation (63) becomes

$$S(z^*) = 0 \tag{64}$$

Consequently, in these two cases, the extrema of the stationary distribution are given the zeros of the selection gradient $S(z)$, which are the extrema of the fixation probability. For the remainder of this analysis, we will assume that $\mathrm{d}\sigma^2(z)/\,\mathrm{d}z = 0$ and the extrema of the stationary density are zeros of the selection gradient.

The zeros of the selection gradient are the candidate evolutionary equilibria obtained using evolutionary game theory, adaptive dynamics, and inclusive fitness theory. A candidate evolutionary equilibrium z^* is called "convergence stable" when $S'(z^*) < 0$ [32, 34, 138], which means that for resident trait values close to z^*, mutants invade such resident populations only when those mutants are closer to z^* than the resident. Convergence stability is a natural way to characterize long-term evolutionary attractors (which may or may not be "branching points"; see refs [154] for the relevance of branching in the TSS). The condition for an extremum z^* of the

stationary density $\rho(z)$ to be a local maximum, $\mathrm{d}^2\rho/\,\mathrm{d}z^2 < 0$, turns out to be precisely the convergence stability condition. Convergence stable traits may also reside on the boundaries of the trait space. In this case, the lower boundary is convergence stable when $S(z) < 0$ and the upper boundary is convergence stable when $S(z) > 0$. These two condition correspond, respectively, to boundary maxima for the stationary density, $\mathrm{d}\rho/\,\mathrm{d}z < 0$ for the lower boundary and $\mathrm{d}\rho/\,\mathrm{d}z > 0$ for the upper boundary. Thus, long-term evolutionary attractors given by convergence-stable equilibria obtained from the selection gradient are generally local maxima of the stationary density of the long-term TSS diffusion [82].

4.3. Additive genetic interactions, weak selection, and the long-term TSS diffusion

In sum, the long-term TSS diffusion in (61) describes how selection, mutation, and drift interact to shape the evolution of a continuous trait assuming that the mutation rate is weak and mutational effects are small enough to be sufficiently described by the mutational variance. It is important to note that by neglecting third order and higher moments of the mutational effects distribution, we only need a first-order approximation of the fixation probability $\pi(\delta, z)$ in terms of δ when calculating the substitution or jump rate. In section 2.5, we showed that a first-order approximation of fitness with respect to δ, which is necessary for a first-order approximation of fixation probability, results both in an assumption of weak selection and additive genetic interactions. This could suggest that the diffusion equation for the long-term TSS only represents evolution under weak selection and additive genetic interactions. Intriguingly, this is far from true as we will show in section 5. By integrating over a continuum of mixed strategies, the diffusion model of the long-term TSS can reproduce all the results of the short-term TSS with three-way genetic interactions in a group structured population. The diffusion approach also appears to reproduce some *strong selection* results generated from other TSS approaches; specifically, these approaches show that the condition for evolutionary success in an unstructured population under strong selection is no longer simply risk dominance [45] and the long-term TSS can reproduce this effect. We will explore these results now.

5. Application: social games in structured populations

In this section, we will apply the long-term TSS diffusion to the same social game as in section 3 with a similar group-structured population. Our goal is to track the evolution of the continuous trait z that measures the fraction of time that an individual cooperates with social partners living in its own group; the complementary fraction $1 - z$ is the fraction of time the individual does not cooperate. Using the payoffs from Table 3, the

fertility of individual i in group g is

$$f_{gi} = 1 + B\, z_{gi} - C\, z_{g\backslash i} + D\, z_{gi} z_{g\backslash i}\,, \tag{65}$$

which is analogous to the fertility function in equation (37). Following the weak effect mutation model for the short-term TSS, we assume that the phenotype of individual i in group g is $z_{gi} = z + p_{gi}\delta$ where p_{gi} is the frequency of the mutant allele in that individual.

5.1. Selection gradient

Instead of calculating fitness for the specific case of an island model as in section 3, we follow Van Cleve and Lehmann [152] and write the selection gradient for a more general group-structured case where the demographic process that translates fertility and survival into fitness is not explicitly specified though dispersal is still potentially local so that genetic identity or relatedness can build up. These are in fact the same assumptions we made in section 2.7 where we reproduced Hamilton's rule from the selection gradient in equation (34). Briefly, these assumptions ensure that the derivatives of fitness with respect to the trait values of different individuals at neutrality take only three possible values: $\frac{\partial w_{gi}(z)}{\partial z_{gi}}$ for the effect of the focal individual's trait on its own fitness, $\frac{\partial w_{gi}(z)}{\partial z_{g\backslash i}}$ for the effect of the average group member's trait on the focal's fitness, and $\frac{\partial w_{gi}(z)}{\partial z_{g'j}}$ for the effect of an individual in another group on the focal's fitness. The latter derivative, $\frac{\partial w_{gi}(z)}{\partial z_{g'j}}$, can written in terms of the former two since the selection coefficients must sum to zero. The IBD probabilities also collapse to three categories: identity with self, which is one, identity with another individual in the group or Q_0, and identity with an individual in another group or Q_1. These facts together allow us to rewrite the selection gradient in equation (34) as

$$S(z) = \lim_{\mu\to 0} \frac{1-Q_1}{1-Q_0}\left(\frac{\partial w_{gi}(z)}{\partial z_{gi}} + \frac{\partial w_{gi}(z)}{\partial z_{g\backslash i}}\frac{Q_0-Q_1}{1-Q_1}\right) \propto -c + b\,r\,. \tag{66}$$

where we have used the fact that $1 - Q_0 = 2N_{\mathrm{T}}\mu + O(\mu^2)$ (eqs. 26 and 46 in [104] and eq. 3.68 in [123]). The selection gradient in (66) is the same as that derived by Rousset [123, 125] for group structured populations. Using the definitions for terms in Hamilton's rule in section 2.7 where $-c = \frac{\partial w_{gi}(z)}{\partial z_{gi}}$, $b = \frac{\partial w_{gi}(z)}{\partial z_{g\backslash i}}$ and $r = F_{\mathrm{ST}} = \frac{Q_0-Q_1}{1-Q_1}$, it is clear that selection gradient $S(z)$ is proportional to the inclusive fitness effect, which is a function of the phenotypic trait z and all of the demographic effects that shape b, c, and r.

Even though we do not specify precisely how fitness is a function of fertility, we can expand $\frac{\partial w_{gi}(z)}{\partial z_{gi}}$ and $\frac{\partial w_{gi}(z)}{\partial z_{g\backslash i}}$ each into sums of products of two components, a derivative of fitness with respect to fertility and a

derivative of fertility with respect to trait values. Additionally, since the total population size is fixed, individual fitness is a function of relative fertility. More specifically, we assume that fitness is a ratio of linear functions of fertility (e.g., equation 38), which is the case when competition for resources or breeding patches is linear with respect the number of offspring produced (i.e., the "contest success function" is a linear ratio [15, 68]). The above assumptions allow us to rewrite the selection gradient in (66) as

$$S(z) = \frac{k}{f_{gi}(z)}\left(\frac{\partial f_{gi}(z)}{\partial z_{gi}} + \kappa \frac{\partial f_{gi}(z)}{\partial z_{g\backslash i}}\right). \tag{67}$$

The first term in the parentheses is the effect of the trait in the focal individual on its own fertility. The second term is the effect of the trait in the group (excluding the focal) on the fertility of the focal weighted by κ, which is a "scaled relatedness coefficient" that accounts for both relative genetic identity due to genetic relatedness and competitive effects due to demography and finite population size [90, 151, 152]. In general, the scaled relatedness κ can take a value between -1 and 1 depending on the demography [90]. The coefficient k also captures demographic and competitive effects and scales the magnitude of the selection gradient. We assume that both κ and k do not depend on z, which is true if the demographic variables, such survival, migration, population size, etc, do not depend on the phenotypic trait. Additionally, fertility must also be generally large or Poisson distributed in order to neglect the effect of the trait on demographic stochasticity [83].

Compared to the fitness effects and genetic identity formulation of the selection gradient in (66), the selection gradient in (67) partitions terms into fertility effects and scaled relatedness. The former partition is that used to define b, c, and r in Hamilton's rule, which leads to the definitions for different social behaviors based on fitness effects in Table 2. In the latter partition, all the effects of demography and local competition are encompassed by scaled relatedness κ and the coefficient k and the effects of the phenotypic trait on the social interaction are isolated in the fertility effects. In so far as we are interested in understanding how the immediate payoffs from social interactions and demography *independently* contribute to selection on social behavior, the latter partition is advantageous. Previous models have used this partition where κ is also called scaled relatedness [2, 90, 151] or "compensated relatedness" [55],"potential for altruism" [47], "potential for helping" [121], "index of assortativity σ_0" [4], or the "structure coefficient σ" [5, 135, 136]. Of particular relevance here, the structure coefficient σ introduced by Tarnita and collaborators [5, 135, 136] has been used to analyze the evolution of discrete strategies in finite and structured populations where non-additive genetic interactions are possible due non-additive payoffs (i.e., $D \neq 0$). As we will see below, our long-term TSS diffusion can

reproduce a central result from the work using σ even though the selection gradient $S(z)$ only captures additive genetic interactions at any one point in time.

Applying the fertility function in (65) to the selection gradient in (67) yields

$$S(z) = k\left(\frac{B\kappa - C + Dz(1+\kappa)}{1 + z(B - C + Dz)}\right). \tag{68}$$

Recalling that $S(z^*) = 0$ define candidate evolutionary equilibria, we find one internal equilibrium at

$$z^* = \frac{C - \kappa B}{D(1+\kappa)}, \tag{69}$$

which is viable mixed strategy only when $D(1+\kappa) > C - \kappa B > 0$. When there is no effect of population structure and $\kappa = 0$, this simplifies to $D > C$, which implies that the social game is one of coordination where both the boundary equilibria $z = 0$ and $z = 1$ are convergence stable (i.e., $S(0) < 0$ and $S(1) > 0$) and the internal equilibrium z^* is unstable ($S'(z^*) > 0$). If $D < C$, then the game is Prisoner's Dilemma and only $z = 0$ is stable. When population structure is at its strongest, $\kappa = 1$, $B > C$ ensures that the internal equilibrium does not exist (not a valid mixed strategy) and only $z = 1$ is stable, which implies the social interaction is a mutualism game. Thus, non-additive payoffs, or synergy, can generate the possibility of cooperation by changing the game from a Prisoner's Dilemma to a coordination game [2]. Population structure can also generate cooperation, but in a stronger way since when cooperation is stable, it is the only equilibrium.

5.2. Stationary distribution

Substituting the selection gradient in (68) into the stationary distribution in (62) yields

$$\rho(z) = \frac{1}{K}\exp(2N_T\phi(z)) \tag{70}$$

where

$$\phi(z) = \frac{k}{2}((1+\kappa)\log(1 + z(B - C + Dz)) + (1-\kappa)Y(z)) \tag{71}$$

is the "potential function" [46],

$$Y(z) = \frac{2(B+C)}{V}\left[\tan^{-1}\left(\frac{B-C}{V}\right) - \tan^{-1}\left(\frac{B-C+2Dz}{V}\right)\right], \tag{72}$$

and $V = \sqrt{4D - (B-C)^2}$. In the left panel of Figure 1, we plot the stationary density $\rho(z)$ in (71) for $N_T = 20$, $B = 1$, $C = 0.5$, and $\kappa = 0$. As expected for $D = 0$ when the game is a Prisoner's Dilemma, we see a peak in the density at $z = 0$ and the mean value of z, $\mathrm{E}[z]$, is close to $z = 0$. When $D > C = 0.5$, the social interaction becomes a coordination game and both full noncooperation, $z = 0$, and full cooperation, $z = 1$, become convergence stable. The full cooperation peak is initially very small, and the population spends most of its time with a trait value close to $z = 0$ until D increases to at least $2C = 1$. Recall from condition (45) that $D > 2C$ is the risk dominance condition that ensures the basin of attraction of full cooperation is larger than that of full noncooperation. The mean value of the trait $\mathrm{E}[z]$ crosses $z = 1/2$ at $D \approx 1.15 > 2C$, which implies that cooperation under the long-term TSS model is more difficult to obtain than the risk dominance condition suggests. However, risk dominance appears to be the correct condition when the payoffs are much weaker. This is shown in the right panel of Figure 1 where $B = 1 \times 10^{-2}$, and $C = 0.5 \times 10^{-2}$. The lower values of the payoffs induce much weaker selection, which leads to a mean trait value much closer to $1/2$ for all values of D, but the crossing point is almost exactly at $D = 2C$ as risk dominance predicts. In fact, we will show that the risk dominance condition can be recovered from the stationary density $\rho(z)$ as part of a more general condition for evolutionary success under weak payoffs.

The scaled relatedness κ has a strong effect on the stationary distribution of trait values [152], which we show in Figure 2 by plotting the mean trait value $\mathrm{E}[z]$. In the upper left panel, the parameters are the same as Figure 1 except we vary κ from -0.5 to 0.5. When population structure increases genetic relatedness and does not induce much local competition, κ takes positive values. The plot shows that positive values of κ produce a shift in the $\mathrm{E}[z]$ curve so that lower values of the synergistic payoff D are required to obtain a high level of cooperation in the population compared to $\kappa = 0$ when there no effect of population structure. The converse is true for negative values of κ, which require higher levels of D for any significant amount of cooperation. In these cases, IBD within groups is low and competition locally between relatives is strong. These results are directly analogous to those obtained from the short-term TSS model in section 3. Instead of varying population structure continuously in the short-term TSS model, we analyzed different demographic scenarios and found that strong local competition led to a more stringent condition for the cooperation allele to be evolutionarily successful (eq. 50) and no local competition lead to a more relaxed condition (eq. 52).

5.3. Stochastic stability

The remaining panels in Figure 2 demonstrated the effect of increasing total population size N_T. As N_T grows, the $E[z]$ curves approach step functions that appear in the $N_T \to \infty$ panel; for values of D below a threshold, $E[z] = 0$ and the population is fully noncooperative, and above the threshold, $E[z] = 1$ and the population is fully cooperative. When the total population size is small, genetic drift is strong and the peaks in the stationary density are small. This implies significant probability density for trait values distant from the peaks (see the right panel in Figure 1 where drift is strong relative to selection). Using the metaphor of a fitness valley, the population frequently crosses the valleys between small peaks due to small N_T. Increasing N_T reduces the effect of genetic drift and increases the height of the peaks in the stationary density, which means most of the probability density is located at the peaks. In this case, the population rarely crosses fitness valleys. In the limit as $N_T \to \infty$, the peaks of the stationary density get infinitely tall and the population only visits the peaks. The amount of time the population spends in fitness valleys goes to zero but crossing still occur "enough" so that the population can escape a lower peak and visit a higher peak. Thus, only a small set of peaks are visited by the population in the long-run. Such peaks are called "stochastically stable states" and were first studied in game theory in models of agents with simple learning rules [12, 40, 72, 111, 130]. A trait z is stochastically stable when

$$\lim_{N_T \to \infty} \rho(z) > 0 \,. \tag{73}$$

Van Cleve and Lehmann [152] prove that only the highest peak, as measured by the potential function $\phi(z)$ whose peaks are those of $\rho(z)$, is stochastically stable. Since peaks in the stationary density correspond to convergence stable equilibria, only the convergence stable equilibrium associated with the highest peak is stochastically stable. In our current model of social interaction in a group structured population, the lower right panel of Figure 2 shows the stochastically stable value of z as a function of D and κ.

The advantage of identifying the stochastically stable state is that it is often unique as a function of the demography, κ, and the payoffs of the social game, B, C, and D. Using the stationary density in (70), we can show [152] that full cooperation ($z = 1$) is stochastically stable when

$$\left(\frac{1+\kappa}{1-\kappa}\right) \log(1 + B - C + D) + Y(1) > 0 \tag{74}$$

and full noncooperation ($z = 0$) is stochastically stable when the opposite condition holds. Condition (74)

reveals immediately the positive effect that population structure has on the stochastic stability of cooperation since the left hand side is an increasing function of κ. Holding κ constant however, the relationship between condition (74) and the risk dominance condition is still difficult to discern. The results in Figure 1 suggest that stochastic stability in the long-term TSS diffusion and risk dominance may coincide when the payoffs are weak. Figure 3 shows more evidence for this by plotting $\mathrm{E}[z]$ for the same values of N_{T} and κ as Figure 2 except the payoffs are multiplied by 10^{-2}. As population grows large, the $\kappa = 0$ curve crosses $\mathrm{E}[z] = 1/2$ at exactly the risk dominance prediction of $D = 2C = 1$. In fact, we can show this analytically by assuming that B, C, and D are small and ignoring quadratic and higher terms in condition (74), which produces

$$-C + \kappa B + \frac{1+\kappa}{2} D > 0 \,. \tag{75}$$

This is a "weak payoff" stochastic stability condition and is exactly the risk dominance condition when $\kappa = 0$ and there is no effect of population structure. Recall that risk dominance is the same condition as $\pi_{A|a} > \pi_{a|A}$ for the short-term TSS under hard selection when local competition exactly cancels the benefits of cooperating with relatives. In fact, we can rederive other results from the short-term TSS model by inserting appropriate values of κ into equation (75). For example, using $\kappa = -1/(N-1)$ for the soft selection demography [90, eq. A-8] in condition (75) reproduces condition (50). The scaled relatedness for the group competition demography is F_{ST} under the infinite island model, $\kappa = F_{\mathrm{ST}} = 1/(1+2M)$ [90, limit as $Nm \to M$ and $N \to \infty$ of eq. A-3], and this value of κ reproduces condition (52) when the number of groups is infinite. Additionally, condition (75) is the same as the one derived by Tarnita et al. [135, eq. 4] if one exchanges our scaled relatedness κ for their structure coefficient σ. It is notable that the condition of Tarnita et al. [135] is derived from a discrete strategy model under weak mutation and weak selection that is analogous to our short-term TSS whereas our result is obtained from the long-term TSS diffusion under weak payoff.

The ability of the weak payoff stochastic stability condition of the long-term TSS to reproduce the weak selection results for the short-term TSS suggests that strong payoffs might reproduce strong selection results from the short-term TSS. Our evidence for this conjecture is less conclusive since there is no known approximation of the evolutionary success condition $\pi_{A|a} > \pi_{a|A}$ under strong selection and generic population structure. However, we can gain some intuition from a result from Fudenberg et al. [45] who analyze social interactions in a finite Moran model with no population structure. Assuming that $N_{\mathrm{T}} = N \to \infty$ and that selection strength

is arbitrary ($\omega = 1$ and arbitrary B, C, and D) in the short-term TSS, they find that $\pi_{A|a} > \pi_{a|A}$ when

$$\frac{(1+B-C+D)\log(1+B-C+D)-(1-C)\log(1-C)}{B+D}-\frac{(1+B)\log(1+B)}{B} \tag{76}$$

[45, Theorem 2 part b.3]. If we assume weak selection in condition (76) by replacing each parameter with its value times the selection strength ω and ignoring terms $O(\omega^2)$, we recover the risk dominance condition in (45). For strong selection, condition (76) and the long-term TSS stochastic stability condition (75) are not equivalent, but yields similar numerical results. Evidence of this match is in Figure 4, which plots the value of D at which the stochastic stability condition (75) is an equality as a function of scaled relatedness κ. The black curve is the strong payoff case ($\omega = 1$) and the blue and red curves are weak payoffs ($\omega = 10^{-2}$ in blue and $\omega = 10^{-4}$ in red). The equivalent point from condition (76) where $\kappa = 0$ is plotted with the circled dots using equivalent selection strengths plotted in the same colors. It is evident from the figure that the short-term TSS results from Fudenberg et al. match very closely to the stochastic stability results from the long-term TSS diffusion. Both results show that increased selection strength requires a higher value of synergy D for full cooperation to be evolutionary successful. The correspondence between the short-term TSS model under weak selection and the long-term TSS diffusion under weak payoffs and their numerical match under strong selection and strong payoffs, respectively, suggests that the long-term process might capture the full potential of selection to shape social interactions under the TSS.

6. Discussion

Due in part to its simplicity, Hamilton's rule has proved to be a remarkably useful tool for deriving insight into the effect of natural selection on the evolution of social behavior. By emphasizing the role of genetic correlations between individuals, Hamilton's rule catalyzed interested in the effect of genetic population structure on all kinds of social behavior from parental care and cooperative breeding to cooperative hunting and colony defense. Hamilton's rule also suggested a simple way to categorize social behaviors based on how they affect the fitness of focal actors and social partners (Table 2).

Early theoretical investigations of Hamilton's rule and inclusive fitness quickly identified the source of this simplicity in two key assumptions of the rule, namely weak selection and additive genetic interactions with respect to fitness. The general utility and limitations of these assumptions for populations with arbitrary group or class structure became clearer in approaches that use an individually centered approach like the Price equation that describes evolutionary change via the statistical moments of the allele frequency distribution. Such

approaches usually assume weak selection, which allows the calculation of higher-order moments via a quasi-equilibrium (QE) approximation. For the case of evolution at a single locus in a group-structured population, the QE approximation entails simply calculating F_{ST} or some other measure of population structure under neutral evolution. Assuming genetic additivity in addition to weak selection guarantees that Hamilton's rule is independent of the allele frequency [125], which means that it predicts both invasion and fixation of a mutant allele in a monomorphic population.

6.1. The trait substitution sequence and short and long-term evolution

Even with weak selection, polymorphisms are still possible since non-additive interaction can generate stabilizing selection and high mutation rates can maintain heterozygosity. If mutation rates are low enough relative to population size so that fixation or extinction occurs more quickly than a mutation, then only a single mutation will segregate in the population. This low mutation assumption, given in condition (4), generates the trait substitution sequence (TSS) whose short and long-term dynamics can be described in a remarkably complete fashion. Using weak selection and weak mutation, we showed above how to analyze evolutionary change under the TSS both in the short term when the set of possible alleles is finite (e.g. single nucleotide polymorphisms) and in the long term when a continuum of alleles is possible (e.g. morphological traits shaped by multiple cis-regulatory elements).

This difference between short versus long-term evolution with respect to the TSS is related to the broader conception of short and long-term evolution by Eshel [33, 34], Hammerstein [62, 63, "streetcar theory"], and others [160]. Growing out of an attempt to reconcile explicit population genetics approaches with phenotypic approaches from evolutionary game theory, Eshel characterizes short-term evolution as where, within a set of fixed genotypes, "natural selection will operate, in the short run, to change the genotype frequencies toward a new internally stable equilibrium" [34, p. 489]. Such an equilibrium can be monomorphic or polymorphic due to any process that could generate stabilizing selection such as heterosis, epistasis, or local adaptation in a structured population. Eshel defines long-term evolution as "characterized by the repeated introduction of new mutations into the population and in between periods of changes of genotype frequencies (say, short-term evolution) within the new simplex of genotypes" [34, p. 489]. Determining whether or not a new mutation will invade a generic internally-stable equilibrium is very difficult though extremely suggestive results were obtained by Eshel and Feldman [35], Liberman [95], and Hammerstein and Selten [63]; broadly, these authors found that new mutations cannot invade when the internally-stable equilibrium generates an ESS phenotype

(under a linear population game) and that invading mutants shift a population towards an ESS if the population is already within a neighborhood of the ESS [35, 95]. This is a more general result than those obtained from our long-term TSS in that there are no restrictions on mutation rate or selection strength, which allows for complex polymorphisms. However, the assumptions of the TSS allows us to characterize both the short and long-term dynamics completely, which is not possible in the more general case of long-term evolution.

6.2. The short-term TSS

Beginning with the short term TSS, we described how a natural condition for evolutionary success, the expected frequency of the mutant allele being greater than the resident or $\mathrm{E}[p] > 1/2$, is equivalent under weak mutation to the fixation probability of the mutant being larger than the resident or $\pi_{A|a} > \pi_{a|A}$ [7, 44, 125]. Under weak selection, fixation probabilities can be written in terms of selection coefficients and expected coalescence times (eq. 22), which yields the general expression for $\pi_{A|a} > \pi_{a|A}$ given in (24). This general condition, which is new to the literature, reveals how non-additive genetic interactions quickly increase the sensitivity of evolutionary success to pairwise, triplet, and higher-order coalescence times. When genetic interactions are additive, the condition $\pi_{A|a} > \pi_{a|A}$ is equivalent to the simpler condition that the derivative of the fixation probability, or selection gradient, be positive: $\frac{\mathrm{d}\pi}{\mathrm{d}\delta} = S(z) > 0$. This latter condition readily reproduces Hamilton's rule.

The appearance of expected coalescence times in the approximation of fixation probabilities [89, 94, 122] is useful both from conceptual and practical perspectives. Conceptually, expected coalescence times arise directly out of the fact that we calculate fixation probabilities after the generation of a new mutation in a monomorphic population and before another mutation enters the population [122]. Moving backwards in time, coalescence times express the effect of population structure and demography on the genealogy of the invading mutant. How the genealogy interacts with selection on the mutant depends on the fitness function; fitness functions linear in mutant allele frequency generate dependence on pairwise coalescence times, quadratic fitness functions generate coalescence times between three lineages, and so on. Practically, using expected coalescence times allow us to leverage the extensive results in coalescence theory [106, 157] that detail the effect of demography [e.g., 120] and population structure [e.g., 108, 155] on coalescence times.

Using Slatkin's formula (eq. 33) for relating expected pairwise coalescence times to pairwise IBD probabilities [133], we showed how for additive genetic interactions the evolutionary success condition $\frac{\mathrm{d}\pi}{\mathrm{d}\delta} > 0$ can be written in terms of IBD probabilities (eq. 34). This approximation in terms of IBD probabilities can also be

obtained by calculating the evolutionary success condition

$$\lim_{\mu \to 0} \mathrm{E}[\Delta p | \mathbf{p} \notin \{\mathbf{0}, \mathbf{1}\}] > 0 \tag{77}$$

for additive genetic interactions [125]. Allen and Tarnita [7] showed under weak selection that condition (77) is equivalent to $\pi_{A|a} > \pi_{a|A}$, which says that calculating condition (77) can yield the equivalent of condition (24) except with IBD probabilities instead of expected coalescence times. Our application of the Slatkin formula effectively allowed us to show this equivalency for additive genetic interactions; showing this equivalency for non-additive interactions requires relating coalescence times and IBD probabilities among an arbitrary number of lineages, which is a task for future work.

Using the success condition $\pi_{A|a} > \pi_{a|A}$, we reproduced and generalized several important previous results from group-structured (island-type) populations with the non-additive social interaction in Table 3. First, we reproduced the expressions for fixation probability $\pi_{A|a}$ given by Ladret and Lessard [80], which easily generate the one-third law of evolutionary game theory [109] by evaluating $\pi_{A|a} > 1/N_{\mathrm{T}}$ as population structure disappears or $M \to \infty$. Applying the evolutionary success condition $\pi_{A|a} > \pi_{a|A}$ yielded the well-known risk dominance condition regardless of the level of population structure as measured by M. While risk dominance was initially proposed as a condition for determining which Nash equilibrium is optimal in two-player games in economics [64], it was subsequently shown to predict the stochastically stable strategy in population games where agents update their strategy by learning the best response [72] even when agents can only learn from local neighbors [13, 31]. Thus, these results confirm an important connection between the process of strategy selection in economic models and the process of natural selection with a certain type of population structure.

Risk dominance also can be seen as a generalization of the classic cancellation result of Taylor [139, 140, 143] and others [119, 166] that says localized dispersal or population viscosity alone is not enough to create selection for cooperation or altruism. The cancellation results shows that the benefits of cooperation are exactly canceled by the costs of competing locally with kin when the population is homogeneously structured, density-dependent regulation occurs after dispersal (hard selection), and generations are non-overlapping. Our risk dominance result generalizes the cancellation result because it shows that the condition for cooperation to be evolutionarily successful is still independent of the amount of population structure even with non-additive or synergistic payoffs.

Moreover, our results go further because we showed how the risk dominance condition holds only in the

limit of large total population size for the baseline demography of hard selection. When the total population size is small, the cancellation results no longer holds and local competition with kin degrades some of the benefit cooperators obtain when interacting with one another. More generally, the scale of competition can either increase or decrease the strength of local competition relative to relatedness, which makes selection for cooperation more or less stringent. Demographies with soft selection, where density-dependent regulation occurs before dispersal and competition occurs within the group, have greatly increased local competition that depends not only the total population size, but on the local group size N. In contrast, when the scale of competition is the total population, local competition is negligible and selection for cooperation depends primarily on the degree of genetic relatedness. If relatedness is strong enough (M is small), then the evolutionary success of the cooperation allele can be guaranteed. We can also relate these results back to the results in economics that suggest risk dominance is equivalent to stochastic stability in structured populations [13, 31, 72] (see also Sandholm [130] for a refinement of risk dominance). Our results suggest that models of learning and best-response dynamics in economics make strong implicit biological assumptions that eliminate the potential of population structure to affect strategic evolution; thus, these biological assumptions should be clearly specified and studied in economic models.

6.3. The long-term TSS

Moving to the long-term TSS and a continuum of possible traits, we applied a substitution rate approach that assumes the long-term process can be fully characterized by the rate a population jumps from one monomorphic state to another. This substitution or jump rate to a particular mutant trait $z+\delta$ from a resident trait z is the number of mutants of type $z+\delta$ times the probability these mutations fix. The substation rate in the long-term TSS is analogous to the neural substitution rate of molecular evolution [77] except that the substitution rate is an explicit function of trait value, population size, and other parameters. By assuming that mutational effects are tightly clustered around the resident value, the substitution rates leads directly to a diffusion process [82] that characterizes trait change under the long term TSS (eq. 61), which is analogous to the canonical diffusion of adaptive dynamics [20].

From the long-term TSS diffusion, we obtained the stationary long-term trait density $\rho(z)$ that can be used to asses evolutionary success. Specifically, local peaks in the stationary density correspond to classic evolutionary stable states [100] in linear or discrete games. In continuous trait games, these peaks correspond to convergence stable states [22, 32], which are attracting evolutionarily stable states where the selection gradient

crosses zero from above [82, 152]. Being located at a peak in the stationary density is then a natural criterion for evolutionary success. Moreover, as the total population size goes to infinity and genetic drift becomes extremely weak compared to selection, only the highest peaks retain any positive probability at stationarity; all other peaks, even if locally convergence stable, are visited by the population with zero probability. The traits at these highest peaks are called stochastically stable. Since stochastically stable states are often unique, they represent one of the simplest possible predictions for trait evolution in the long term.

Using the stochastic stability criterion (eq. 73), we showed that the long-term TSS diffusion can a reproduce a range of results from the short-term TSS model presented here as well as results from other work using TSS-type assumptions. To show this, we first transformed the selection gradient from a Hamilton's rule form with fitness effects and genetic relatedness to a form with fertility effects and "scaled relatedness" or κ. The scaled relatedness combines the effects of demography on both genetic identity and local competition whereas the fertility effects are strictly functions of the payoffs from the social interaction. Assuming "weak payoffs", the condition for stochastic stability using this expression for the selection gradient matches exactly with the previous condition from Tarnita et al. [135] for the cooperation allele to be evolutionarily successful ($\pi_{A|a} > \pi_{a|A}$) where scaled relatedness κ corresponds to their structure coefficient σ. Moreover, by using the values of scaled relatedness κ that correspond to the soft selection and group competition demographies, we reproduced the evolutionary success conditions derived from the short-term TSS. This is despite the fact that the selection gradient used in the long-term TSS depends only on additive genetic interactions, whereas the results from the short-term TSS and from Tarnita et al. [135] include non-additive interactions. We suggest below why the additive selection gradient can recover the results of the non-additive analyses.

An intriguing aspect of the long-term TSS diffusion is that it only reproduced the short-term TSS results after assuming weak payoffs in the selection gradient. This suggested that strong payoffs in the selection gradient might produce results analogous to strong selection in the short-term TSS. In fact, we found some numerical evidence for this by comparing the results of the long-term TSS diffusion with the strong selection results of Fudenberg et al. [45] (see Figure 4) who derived a condition for stochastic stability in a panmictic population. This behavior is potentially analogous to how the classical diffusion approach for allele frequencies [38] has a "strong selection" regime for $N_T\omega > 1$ even though the diffusion requires that $N_T \to \infty$ and $\omega \to 0$. Comparisons between the long-term TSS diffusion and strong selection in structured populations remain to be done due to the lack of analytical results in that area (however see [103]) and the difficulty in defining a scaled relatedness κ for strong selection. If such comparisons find additional support for the analogy between

the long-term TSS diffusion and the classical diffusion from population genetics, then the long-term TSS may be a powerful tool for the analysis of social evolution even under strong selection.

6.4. Non-additivity and the power of the additive long-term approach

One of the crucial points of this article is that our derivation of Hamilton's rule in the context of a dynamical model with selection, mutation, and genetic drift relies on three assumptions: (i) weak selection, (ii) weak mutation to ensure the TSS, and (iii) additive genetic interactions. The impact of breaking the last assumption and allowing non-additivity or synergy has been the subject of much interest [e.g., 49, 65, 87, 112, 118, 137, 144], and our analysis of the short-term TSS reinforces the importance of non-additivity in shaping the evolution of social outcomes. Nevertheless, our analysis of the long-term TSS also shows that even additive interactions, which generate Hamilton's rule and the selection gradient used in the long-term TSS, can still generate useful and even powerful results. This is in contrast to some suggestions that analyses based on inclusive fitness methods are severely limited due to their assumption of additivity [6, 110], which does not appear to hold in some empirical systems [23, 53, 97]. In order to better understand precisely the extent to which additivity is limiting, we need to clarify the different levels at which additivity can hold and understand why additivity of genetic interactions is more powerful than commonly assumed.

Throughout this article, we have used additivity to refer specifically to additive genetic interactions with respect to fitness; this is equivalent to additivity of fitness effects with respect to allele frequency ($d = 1$ in eq. 17). With weak selection and the short-term TSS, we showed that additive genetic interactions lead directly to Hamilton's rule since evolutionary success depends directly on the derivative of the fixation probability with respect to selection strength (i.e., the selection gradient), which is a sum of selection coefficients times pairwise expected coalescence times or IBD probabilities. Analyzing the social interaction in Table 3 where individuals carry either a cooperation (A) or noncooperation allele (a), we showed how assuming additivity of payoffs (i.e., the synergistic payoff $D = 0$) implies additive fitness effects. However, the converse is not true; additive fitness effects do not imply that payoffs in the social interaction are additive.

To see this, consider an arbitrary fitness function for individual i, $w_i(\mathbf{z})$, which is a function of the (continuous) trait values of all other individuals in the population. The fitness function depends on the traits values of other individuals in part because individual i obtains payoff from a social interaction with some set of social partners. This payoff can be highly nonlinear as a function of the trait values of the social partners, which results in nonlinearity of the fitness function. However, so long as the fitness function is differential with respect

to the traits values, a weak selection expansion of the fixation probability with respect to the mutant deviation δ can still be calculated, and the resulting δ-weak expansion has additive fitness effects. Thus, assuming a differentiable fitness function and δ-weak selection imply that the selection gradient $S(z)$ predicts the direction of selection and that Hamilton's rule holds, even when payoffs are non-additive.

A simple example of a nonlinear payoff function is equation (65), which comes from the social interaction in Table 3. We showed this payoff function leads to the selection gradient in equation (68). We can apply the short-term TSS δ-weak condition, $S(z) > 0$, to this selection gradient and determine for any probability z of choosing the cooperation strategy whether a mutant trait $z + \delta$ is more common under the stationary distribution, even for a synergistic payoff $D \neq 0$. In general, the nonlinearities of the payoff function will have important effects which traits are evolutionarily successful under the short-term TSS or convergence stable under the long-term TSS; for example, synergistic effects as measure by mixed partial derivatives (e.g., $\frac{\partial^2 f_i}{\partial z_i \partial z_j}$) can have a powerful effect on the evolution of cooperative behavior [2, 3]. Such effects are important even in the absence of population structure [3] when such behavior would be a mutualism according to Hamilton's rule (see Table 2). Cooperation is generated in these cases by reciprocity or behavioral responses [3], which can also enhance cooperation in structured populations where relatedness is important [2].

The weakness of the short-term TSS with respect to additive genetic interactions is that it cannot directly compare the relative evolutionary success of traits that differ by more than δ, such as full cooperation ($z = 1$) and full noncooperation ($z = 0$). Comparing the evolutionary success of full cooperation and no cooperation with the short-term TSS requires non-additive genetic interactions. In contrast, the long-term TSS measures the relative evolutionary success of full cooperation and noncooperation using the diffusion process to integrate population jumps over the whole trait space. Under weak payoffs for the long-term TSS, we showed that the diffusion approach produces the same results as the short-term TSS *even though the short-term TSS condition includes non-additive genetic interactions and the long-term diffusion does not*.

The likely explanation for this unexpected feature of the long-term TSS diffusion is that the social interaction we study (eqs. 37 and 65) has at most triplet genetic interactions and the triplet expected coalescence times cancel out of the short-term TSS condition (eq. 24); in other words, the short-term evolutionary success of cooperation versus noncooperation even with synergy ($D \neq 0$) *depends only on pairwise measures of genetic identity*. This fact is easily missed when analyzing the conditions $\pi_{A|a} > 1/N_{\mathrm{T}}$ and $\pi_{a|A} > 1/N_{\mathrm{T}}$ independently, as is done in analyses based on the one-third law, since the triplet coalescence time terms cancel only in the full evolutionary success condition $\pi_{A|a} > \pi_{a|A}$. For more complex non-additive interactions, condition (24)

reveals that higher-order expected coalescence times may not cancel, which suggests that the long-term TSS diffusion may have more difficulty approximating these interactions. The relationship between the long-term approach based on the selection gradient and arbitrary non-additive genetic interactions remains a topic ripe for further development. Nonetheless, the long-term TSS may be a powerful approach for modeling the evolution of social traits in structured populations under simple forms of non-additivity and potentially under strong selection.

6.5. Conclusion

Many of the analytical tools we have used and referenced in this review are much more sophisticated than those readily available to Hamilton when he first discussed inclusive fitness and his eponymous rule [58]. Regardless, the power of the inclusive fitness effect persists both in its ability to suggest broad qualitative patterns connecting natural selection and sociality and in its recurrence as a quantity in complex models of social evolution.

Acknowledgements

J.V. was supported by the National Evolutionary Synthesis Center (NESCent) under NSF grant #EF-0423641.

References

[1] Abugov, R., and R. E. Michod. 1981. On the relation of family structured models and inclusive fitness models for kin selection. J. Theor. Biol. 88:743–754.

[2] Akçay, E., and J. Van Cleve. 2012. Behavioral responses in structured populations pave the way to group optimality. Am. Nat. 179:257–269.

[3] Akçay, E., J. Van Cleve, M. W. Feldman, and J. Roughgarden. 2009. A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proc. Natl. Acad. Sci. U. S. A. 106:19061–19066.

[4] Alger, I., and J. W. Weibull. 2012. A generalization of hamilton's rule—love others how much? J. Theor. Biol. 299:42 – 54.

[5] Allen, B., M. A. Nowak, and U. Dieckmann. 2013. Adaptive dynamics with interaction structure. Am. Nat. 181:E139–E163.

[6] Allen, B., M. A. Nowak, and E. O. Wilson. 2013. Limitations of inclusive fitness. Proc. Natl. Acad. Sci. U. S. A. 110:20135–20139.

[7] Allen, B., and C. E. Tarnita. 2014. Measures of success in a class of evolutionary models with fixed population size and structure. J. Math. Biol. 68:109–143.

[8] Andersson, M. 1984. The evolution of eusociality. Annu. Rev. Ecol. Syst. 15:165–189.

[9] Axelrod, R., and W. D. Hamilton. 1981. The evolution of cooperation. Science 211:1390–1396.

[10] Barton, N. H., and M. Turelli. 1991. Natural and sexual selection on many loci. Genetics 127:229–255.

[11] Binmore, K. G. 2007. Playing for Real: A Text on Game Theory. Oxford University Press, Oxford.

[12] Binmore, K. G., L. Samuelson, and R. Vaughan. 1995. Musical chairs: Modeling noisy evolution. Games Econ. Behav. 11:1 – 35.

[13] Blume, L. E. 1993. The statistical mechanics of strategic interaction. Games Econ. Behav. 5:387 – 424.

[14] Bshary, R., and R. Bergmuller. 2008. Distinguishing four fundamental approaches to the evolution of helping. J. Evol. Biol. 21:405–420.

[15] Cant, M. A. 2012. Suppression of social conflict and evolutionary transitions to cooperation. Am. Nat. 179:pp. 293–301.

[16] Cavalli-Sforza, L. L., and M. W. Feldman. 1978. Darwinian selection and "altruism". Theor. Popul. Biol. 14:268–280.

[17] Champagnat, N. 2006. A microscopic interpretation for adaptive dynamics trait substitution sequence models. Stochastic Process. Appl. 116:1127 – 1160.

[18] Champagnat, N., R. Ferrière, and G. Ben Arous. 2001. The canonical equation of adaptive dynamics: A mathematical view. Selection 2:73–83.

[19] Champagnat, N., R. Ferrière, and S. Méléard. 2006. Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models. Theor. Popul. Biol. 69:297–321.

[20] Champagnat, N., and A. Lambert. 2007. Evolution of discrete populations and the canonical diffusion of adaptive dynamics. Ann. Appl. Probab. 17:102–155.

[21] Christiansen, F. B. 1975. Hard and soft selection in a subdivided population. Am. Nat. 109:11–16.

[22] ———. 1991. On conditions for evolutionary stability for a continuously varying character. Am. Nat. 138:37 – 50.

[23] Chuang, J. S., O. Rivoire, and S. Leibler. 2010. Cooperation and Hamilton's rule in a simple synthetic microbial system. Mol. Syst. Biol. 6:398.

[24] Courant, R., and F. John. 2000. Introduction to calculus and analysis, vol. II. Springer, Berlin.

[25] Darwin, C. 1859. On the origin of species by means of natural selection. John Murray, London.

[26] Dawkins, R. 2000. Obituary: Professor W. D. Hamilton. Independent (Lond.) Mar. 10:6.

[27] Day, T., and P. D. Taylor. 1998. Unifying genetic and game theoretic models of kin selection for continuous traits. J. Theor. Biol. 194:391–407.

[28] Dercole, F., and S. Rinaldi. 2008. Analysis of evolutionary processes: the adaptive dynamics approach and its applications. Princeton University Press, Princeton, N.J.

[29] Desai, M. M., and D. S. Fisher. 2007. Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176:1759–98.

[30] Dieckmann, U., and R. Law. 1996. The dynamical theory of coevolution: a derivation from stochastic ecological processes. J. Math. Biol. 34:579–612.

[31] Ellison, G. 1993. Learning, local interaction, and coordination. Econometrica 61:pp. 1047–1071.

[32] Eshel, I. 1983. Evolutionary and continuous stability. J. Theor. Biol. 103:99 – 111.

[33] ———. 1991. Game theory and population dynamics in complex genetical systems: the role of sex in short-term and in long-term evolution. *In* R. Selten, ed., Game Equilibrium Models, vol. I, pages 6–28. Springer Verlag, Berlin.

[34] ———. 1996. On the changing concept of evolutionary population stability as a reflection of a changing point of view in the quantitative theory of evolution. J. Math. Biol. 34:485–510.

[35] Eshel, I., and M. W. Feldman. 1984. Initial increase of new mutants and some continuity properties of ESS in 2-locus systems. Am. Nat. 124:631 – 640.

[36] ———. 2001. Individual selection and altruistic relationships: The legacy of W. D. Hamilton. Theor. Popul. Biol. 59:15–20.

[37] Eshel, I., U. Motro, and E. Sansone. 1997. Continuous stability and evolutionary convergence. J. Theor. Biol. 185:333–343.

[38] Ewens, W. J. 2004. Mathematical Population Genetics. Springer, New York.

[39] Fisher, R. A. 1930. The genetical theory of natural selection. The Clarendon Press, Oxford.

[40] Foster, D., and H. P. Young. 1990. Stochastic evolutionary game dynamics. Theor. Popul. Biol. 38:219–232.

[41] Frank, S. A. 1998. Foundations of Social Evolution. Princeton University Press, Princeton, NJ.

[42] ———. 2013. Natural selection. VII. History and interpretation of kin selection theory. J. Evol. Biol. 26:1151–1184.

[43] Freidlin, M. I., and A. D. Wentzell. 1984. Random perturbations of dynamical systems. Springer-Verlag, New York.

[44] Fudenberg, D., and L. A. Imhof. 2006. Imitation processes with small mutations. J. Econ. Theory 131:251–262.

[45] Fudenberg, D., M. A. Nowak, C. Taylor, and L. A. Imhof. 2006. Evolutionary game dynamics in finite populations with strong selection and weak mutation. Theor. Popul. Biol. 70:352–363.

[46] Gardiner, C. W. 2009. Stochastic methods: a handbook for the natural and social sciences. Springer series in synergetics, 4th ed. Springer, Berlin.

[47] Gardner, A. 2010. Sex-biased dispersal of adults mediates the evolution of altruism among juveniles. J. Theor. Biol. 262:339–45.

[48] Gardner, A., and S. A. West. 2006. Demography, altruism, and the benefits of budding. J. Evol. Biol. 19:1707–1716.

[49] Gardner, A., S. A. West, and N. H. Barton. 2007. The relation between multilocus population genetics and social evolution theory. Am. Nat. 169:207–226.

[50] Geritz, S. A. H., E. Kisdi, G. Meszena, and J. A. J. Metz. 1998. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. Evol. Ecol. 12:35 – 57.

[51] Gillespie, J. H. 1983. A simple stochastic gene substitution model. Theor. Popul. Biol. 23:202–215.

[52] ———. 1991. The causes of molecular evolution. Oxford University Press, New York.

[53] Gore, J., H. Youk, and A. van Oudenaarden. 2009. Snowdrift game dynamics and facultative cheating in yeast. Nature 459:253–6.

[54] Grafen, A. 1985. A geometric view of relatedness. Oxf. Surv. Evol. Biol. 2:28–89.

[55] Grafen, A., and M. Archetti. 2008. Natural selection of altruism in inelastic viscous homogeneous populations. J. Theor. Biol. 252:694–710.

[56] Haldane, J. B. S. 1955. Population genetics. New Biol. 18:34–51.

[57] Hamilton, W. D. 1963. The evolution of altruistic behavior. Am. Nat. 97:pp. 354–356.

[58] ———. 1964. The genetical evolution of social behaviour. I. J. Theor. Biol. 7:1–16.

[59] ———. 1964. The genetical evolution of social behaviour. II. J. Theor. Biol. 7:17–52.

[60] ———. 1970. Selfish and spiteful behaviour in an evolutionary model. Nature 228:1218–1220.

[61] ———. 1975. Innate social aptitudes of man: an approach from evolutionary genetics. *In* R. Fox, ed., Biosocial Anthropology, pages 133–155. Malaby Press, London.

[62] Hammerstein, P. 1996. Darwinian adaptation, population genetics and the streetcar theory of evolution. J. Math. Biol. 34:511–532.

[63] Hammerstein, P., and R. Selten. 1994. Game theory and evolutionary biology. *In* R. Aumann and S. Hart, eds., Handbook of Game Theory with Economic Applications, vol. 2, chap. 28, pages 929–993. Elsevier.

[64] Harsanyi, J. C., and R. Selten. 1988. A general theory of equilibrium selection in games. MIT Press, Cambridge, Mass.

[65] Hauert, C., F. Michor, M. A. Nowak, and M. Doebeli. 2006. Synergy and discounting of cooperation in social dilemmas. J. Theor. Biol. 239:195–202.

[66] Hauert, C., A. Traulsen, H. Brandt, M. A. Nowak, and K. Sigmund. 2007. Via freedom to coercion: the emergence of costly punishment. Science 316:1905–1907.

[67] Herbots, H. M. 1997. The structured coalescent. *In* P. J. Donnelly and S. Tavare, eds., Progress in Population Genetics and Human Evolution, vol. 87 of *The IMA Volumes in Mathematics and its Applications*. Springer, New York.

[68] Hirshleifer, J. 1989. Conflict and rent-seeking success functions: Ratio vs. difference models of relative success. Public Choice 63:101–112.

[69] Hofbauer, J., and K. Sigmund. 2003. Evolutionary game dynamics. Bull. Amer. Math. Soc. 40:479–519.

[70] Husnik, F., N. Nikoh, R. Koga, L. Ross, R. P. Duncan, M. Fujie, M. Tanaka, et al. 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. Cell 153:1567–1578.

[71] Imhof, L. A., D. Fudenberg, and M. A. Nowak. 2005. Evolutionary cycles of cooperation and defection. Proc. Natl. Acad. Sci. U. S. A. 102:10797–10800.

[72] Kandori, M., G. Mailath, and R. Rob. 1993. Learning, mutation, and long-run equilibria in games. Econometrica 61:29–56.

[73] Karlin, S., and C. Matessi. 1983. The eleventh R. A. Fisher Memorial Lecture: Kin selection and altruism. Proc. R. Soc. B 219:327–353.

[74] Karlin, S., and J. McGregor. 1967. The number of mutant forms maintained in a population. *In* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, pages 415–438. University of California Press, Berkeley, CA.

[75] Karlin, S., and H. M. Taylor. 1981. A second course in stochastic processes. Academic Press, New York.

[76] Kerr, B., P. Godfrey-Smith, and M. W. Feldman. 2004. What is altruism? Trends Ecol. Evol. 19:135–140.

[77] Kimura, M., and T. Ohta. 1971. Protein polymorphism as a phase of molecular evolution. Nature 229:467–469.

[78] Kimura, M., and G. H. Weiss. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

[79] Kirkpatrick, M., T. Johnson, and N. H. Barton. 2002. General models of multilocus evolution. Genetics 161:1727–1750.

[80] Ladret, V., and S. Lessard. 2007. Fixation probability for a beneficial allele and a mutant strategy in a linear game under weak selection in a finite island model. Theor. Popul. Biol. 72:409–425.

[81] Lande, R. 1979. Quantitative genetic-analysis of multivariate evolution, applied to brain - body size allometry. Evolution 33:402–416.

[82] Lehmann, L. 2012. The stationary distribution of a continuously varying strategy in a class-structured population under mutation-selection-drift balance. J. Evol. Biol. 25:770–87.

[83] Lehmann, L., and F. Balloux. 2007. Natural selection on fecundity variance in subdivided populations: kin selection meets bet hedging. Genetics 176:361–377.

[84] Lehmann, L., and M. W. Feldman. 2008. War and the evolution of belligerence and bravery. Proc. R. Soc. B 275:2877–2885.

[85] Lehmann, L., and L. Keller. 2006. The evolution of cooperation and altruism–a general framework and a classification of models. J. Evol. Biol. 19:1365–1376.

[86] ———. 2006. Synergy, partner choice and frequency dependence: their integration into inclusive fitness theory and their interpretation in terms of direct and indirect fitness effects. J. Evol. Biol. 19:1426–1436.

[87] ———. 2006. Synergy, partner choice and frequency dependence: their integration into inclusive fitness theory and their interpretation in terms of direct and indirect fitness effects. J. Evol. Biol. 19:1426–1436.

[88] Lehmann, L., N. Perrin, and F. Rousset. 2006. Population demography and the evolution of helping behaviors. Evolution 60:1137–1151.

[89] Lehmann, L., and F. Rousset. 2009. Perturbation expansions of multilocus fixation probabilities for frequency-dependent selection with applications to the Hill-Robertson effect and to the joint evolution of helping and punishment. Theor. Popul. Biol. 76:35–51.

[90] ———. 2010. How life history and demography promote or inhibit the evolution of helping behaviours. Philos. Trans. R. Soc. B 365:2599–2617.

[91] ———. 2014. Fitness, inclusive fitness, and optimization. Biol. Philos. 29:181–195.

[92] ———. 2014. The genetical theory of social behaviour. Philos. Trans. R. Soc. B 369:20130357.

[93] Leimar, O. 2001. Evolutionary change and darwinian demons. Selection 2:65–72.

[94] Lessard, S., and V. Ladret. 2007. The probability of fixation of a single mutant in an exchangeable selection model. J. Math. Biol. 54:721–744.

[95] Liberman, U. 1988. External stability and ESS: criteria for initial increase of new mutant allele. J. Math. Biol. 26:477–485.

[96] Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

[97] MaClean, R. C., A. Fuentes-Hernandez, D. Greig, L. D. Hurst, and I. Gudelj. 2010. A mixture of "cheats" and "co-operators" can enable maximal group benefit. PLoS Biol. 8.

[98] Malecot, G. 1967. Identical loci and relationship. *In* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, pages 317–332. University of California Press, Berkeley, CA.

[99] Malécot, G. 1948. Les mathématiques de l'hérédité. Masson, Paris.

[100] Maynard Smith, J. 1964. Group selection and kin selection. Nature 201:1145–1147.

[101] Metz, J. A. J., S. A. H. Geritz, G. Meszéna, F. J. A. Jacobs, and J. S. van Heerwaarden. 1996. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. *In* S. J. van Strien and S. M. Verduyn Lunel, eds., Stochastic and spatial structures of dynamical systems, vol. 45 of *Konink. Nederl. Akad. Wetensch. Verh. Afd. Natuurk. Eerste Reeks*, pages 183–231. North-Holland, Amsterdam.

[102] Moran, P. A. P. 1964. On the nonexistence of adaptive topographies. Ann. Hum. Genet. 27:383–393.

[103] Mullon, C., and L. Lehmann. 2014. The robustness of the weak selection approximation for the evolution of altruism against strong selection. J. Evol. Biol. .

[104] Nagylaki, T. 1983. The robustness of neutral models of geographical variation. Theor. Popul. Biol. 24:268–294.

[105] ———. 1993. The evolution of multilocus systems under weak selection. Genetics 134:627–647.

[106] Nordborg, M. 2007. Coalescent theory. *In* D. J. Balding, M. Bishop, and C. Cannings, eds., Handbook of Statistical Genetics, 3rd ed., chap. 25, pages 843–877. John Wiley & Sons, Ltd.

[107] Nordborg, M., and S. M. Krone. 2002. Separation of time scales and convergence to the coalescent in structured populations. *In* M. Slatkin and M. Veuille, eds., Modern Developments in Theoretical Population Genetics, pages 194–232. Oxford University Press, Oxford.

[108] Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29:59–75.

[109] Nowak, M. A., A. Sasaki, C. Taylor, and D. Fudenberg. 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428:646–650.

[110] Nowak, M. A., C. E. Tarnita, and E. O. Wilson. 2010. The evolution of eusociality. Nature 466:1057–1062.

[111] Nöldeke, G., and L. Samuelson. 1993. An evolutionary analysis of backward and forward induction. Games Econ. Behav. 5:425–454.

[112] Ohtsuki, H. 2012. Does synergy rescue the evolution of cooperation? an analysis for homogeneous populations with non-overlapping generations. J. Theor. Biol. 307:20 – 28.

[113] Ohtsuki, H., P. Bordalo, and M. A. Nowak. 2007. The one-third law of evolutionary dynamics. J. Theor. Biol. 249:289–295.

[114] Oliver, K. R., and W. K. Greene. 2009. Transposable elements: powerful facilitators of evolution. BioEssays 31:703–714.

[115] Oster, G., I. Eshel, and D. Cohen. 1977. Worker-queen conflict and the evolution of social insects. Theor. Popul. Biol. 12:49–85.

[116] Price, G. R. 1970. Selection and covariance. Nature 227:520 – 521.

[117] ———. 1972. Extension of covariance selection mathematics. Ann. Hum. Genet. 35:485–490.

[118] Queller, D. C. 1985. Kinship, reciprocity and synergism in the evolution of social behavior. Nature 318:366–367.

[119] ———. 1994. Genetic relatedness in viscous populations. Evol. Ecol. 8:70–73.

[120] Ramachandran, S., N. A. Rosenberg, M. W. Feldman, and J. Wakeley. 2008 Dec. Population differentiation and migration: Coalescence times in a two-sex island model for autosomal and x-linked loci. Theor. Popul. Biol. 74:291–301.

[121] Rodrigues, A. M. M., and A. Gardner. 2012. Evolution of helping and harming in heterogeneous populations. Evolution 66:2065–2079.

[122] Rousset, F. 2003. A minimal derivation of convergence stability measures. J. Theor. Biol. 221:665–668.

[123] ———. 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton, New Jersey.

[124] ———. 2006. Separation of time scales, fixation probabilities and convergence to evolutionarily stable states under isolation by distance. Theor. Popul. Biol. 69:165–179.

[125] Rousset, F., and S. Billiard. 2000. A theoretical basis for measures of kin selection in subdivided populations: Finite populations and localized dispersal. J. Evol. Biol. 13:814 – 825.

[126] Rousset, F., and O. Ronce. 2004. Inclusive fitness for traits affecting metapopulation demography. Theor. Popul. Biol. 65:127–141.

[127] Roze, D., and F. Rousset. 2003. Selection and drift in subdivided populations: a straightforward method for deriving diffusion approximations and applications involving dominance, selfing and local extinctions. Genetics 165:2153–2166.

[128] ———. 2004. The robustness of Hamilton's rule with inbreeding and dominance: kin selection and fixation probabilities under partial sib mating. Am. Nat. 164:214–231.

[129] ———. 2008. Multilocus models in the infinite island model of population structure. Theor. Popul. Biol. 73:529–542.

[130] Sandholm, W. H. 2010. Orders of limits for stationary distributions, stochastic dominance, and stochastic stability. Theor. Econ. 5:1–26.

[131] Sigmund, K., H. De Silva, A. Traulsen, and C. Hauert. 2010. Social learning promotes institutions for governing the commons. Nature .

[132] Skyrms, B. 2004. The stag hunt and the evolution of social structure. Cambridge University Press, Cambridge, UK.

[133] Slatkin, M. 1991. Inbreeding coefficients and coalescence times. Genet. Res. 58:167–175.

[134] Strassmann, J. E., Y. Zhu, and D. C. Queller. 2000. Altruism and social cheating in the social amoeba dictyostelium discoideum. Nature 408:965–967.

[135] Tarnita, C. E., H. Ohtsuki, T. Antal, F. Fu, and M. A. Nowak. 2009. Strategy selection in structured populations. J. Theor. Biol. 259:570–81.

[136] Tarnita, C. E., N. Wage, and M. A. Nowak. 2011. Multiple strategies in structured populations. Proc. Natl. Acad. Sci. U. S. A. 108:2334–7.

[137] Taylor, C., and M. A. Nowak. 2007. Transforming the dilemma. Evolution 61:2281–2292.

[138] Taylor, P. D. 1989. Evolutionary stability in one-parameter models under weak selection. Theor. Popul. Biol. 36:125–143.

[139] ———. 1992. Altruism in viscous populations – an inclusive fitness model. Evol. Ecol. 6:352 – 356.

[140] ———. 1992. Inclusive fitness in a homogeneous environment. Proc. R. Soc. B 249:299 – 302.

[141] Taylor, P. D., T. Day, and G. Wild. 2007. From inclusive fitness to fixation probability in homogeneous structured populations. J. Theor. Biol. 249:101–110.

[142] Taylor, P. D., and S. A. Frank. 1996. How to make a kin selection model. J. Theor. Biol. 180:27–37.

[143] Taylor, P. D., T. Lillicrap, and D. Cownden. 2011. Inclusive fitness analysis on mathematical groups. Evolution 65:849–859.

[144] Taylor, P. D., and W. Maciejewski. 2012. An inclusive fitness analysis of synergistic interactions in structured populations. Proc. R. Soc. B 279:4596–4603.

[145] ———. 2014. Hamilton's inclusive fitness in finite-structured populations. Philos. Trans. R. Soc. B 369:20130360.

[146] Thornton, J. W. 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. Proc. Natl. Acad. Sci. U. S. A. 98:5671–5676.

[147] Toro, M., R. Abugov, B. Charlesworth, and R. E. Michod. 1982. Exact versus heuristic models of kin selection. J. Theor. Biol. 97:699–713.

[148] Trivers, R. L. 1971. The evolution of reciprocal altruism. Q. Rev. Biol. 46:35–57.

[149] Uyenoyama, M. K., and M. W. Feldman. 1981. On relatedness and adaptive topography in kin selection. Theor. Popul. Biol. 19:87 – 123.

[150] Uyenoyama, M. K., M. W. Feldman, and L. D. Mueller. 1981. Population genetic theory of kin selection - multiple alleles at one locus. Proc. Natl. Acad. Sci. U. S. A. 78:5036 – 5040.

[151] Van Cleve, J., and E. Akçay. 2014. Pathways to social evolution: Reciprocity, relatedness, and synergy. Evolution 68:2245–2258.

[152] Van Cleve, J., and L. Lehmann. 2013. Stochastic stability and the evolution of coordination in spatially structured populations. Theor. Popul. Biol. 89:75–87.

[153] Wade, M. J. 1979. The evolution of social interactions by family selection. Am. Nat. 113:399–417.

[154] Wakano, J. Y., and L. Lehmann. 2012. Evolutionary and convergence stability for continuous phenotypes in finite populations derived from two-allele models. J. Theor. Biol. 310:206 – 215.

[155] Wakeley, J. 1998. Segregating sites in wright's island model. Theor. Popul. Biol. 53:166–174.

[156] ———. 2003. Polymorphism and divergence for island-model species. Genetics 163:411–420.

[157] ———. 2009. Coalescent theory: an introduction. Roberts & Co. Publishers, Greenwood Village, CO.

[158] Watterson, G. A. 1974. The sampling theory of selectively neutral alleles. Adv. Appl. Prob. 6:463–488.

[159] ———. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–76.

[160] Weissing, F. J. 1996. Genetic versus phenotypic models of selection: can genetics be neglected in a long-term perspective? J. Math. Biol. 34:533–555.

[161] West, S. A., A. S. Griffin, and A. Gardner. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J. Evol. Biol. 20:415–432.

[162] Wild, G., and P. D. Taylor. 2004. Fitness and evolutionary stability in game theoretic models of finite populations. Proc. R. Soc. B 271:2345–2349.

[163] Wild, G., and A. Traulsen. 2007. The different limits of weak selection and the evolutionary dynamics of finite populations. J. Theor. Biol. 247:382–390.

[164] Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. 37:535–585.

[165] Williams, B. J. 1981. A critical review of models in sociobiology. Annu. Rev. Anthropol. 10:163–192.

[166] Wilson, D. S., G. B. Pollock, and L. A. Dugatkin. 1992. Can altruism evolve in purely viscous populations? Evol. Ecol. 6:331–341.

[167] Wilson, E. O. 1975. Sociobiology: the new synthesis. Harvard University Press, Cambridge, Mass.

[168] Wright, S. 1931. Evolution in mendelian populations. Genetics 16:97–159.

[169] ———. 1969. Evolution and the genetics of populations, vol. 2. University of Chicago Press, Chicago.

Symbol	Description
N_{T}	Total population size
N	Group (or deme) size
n	Number of groups (or demes)
m	Migration rate
M	Population migration rate, $M = nNm/(n-1)$
p_i (p_{gi})	Frequency of allele A in haploid individual i (living in group g)
p_g	Mean frequency of allele A in group g
p (q)	Mean frequency of allele A (a) in the population
$\mathbf{p} = (p_1, \ldots, p_{N_{\mathrm{T}}})$	Vector of allele frequencies for the population
μ	Probability an offspring carries a mutant allele
$\mu_{A\|a}$ ($\mu_{a\|A}$)	Probability an offspring carries allele a (A) with a parent carrying allele A (a)
z_i (z_{gi})	Phenotype of individual i (living in group g)
$\mathbf{z} = (z_1, \ldots, z_{N_{\mathrm{T}}})$	Vector of phenotypes for the population
δ	Phenotypic deviation of mutant phenotype from resident phenotype z
$u(\delta, z)$	Distribution of mutant deviations δ (mutational effects) given a mutation from resident phenotype z
$\sigma^2(z)$	Second moment (raw variance) of the distribution of mutant deviations
$k(\delta, z)$	Instantaneous rate of substitution of population monomorphic for trait z with one monomorphic for trait $z+\delta$
$\rho(z,t)$ ($\rho(z)$)	Probability density that population is monomorphic for trait z at time t ($t \to \infty$)
w_i (w_{gi})	Fitness of individual i (living in group g)
w	Mean fitness in the population
b	Fitness benefit to the focal individual of the expression of allele A in a social partner
c	Fitness cost of the expression of allele A in the focal individual
f_{ig}, f_g, f	Fertility of individual i in group g, mean fertility in group g, mean fertility in the population
$\pi_{A\|a}$ ($\pi_{a\|A}$)	Probability of a single mutant A (a) allele fixing in a population of $N_{\mathrm{T}}-1$ a (A) alleles
π° ($\pi^\circ(z)$)	Probability of a neutral allele (in a population resident for phenotype z) fixing in the population
$S(z)$	Derivative of fixation probability with respect to δ and evaluated at $\delta = 0$; also called the "selection gradient"
ω	Selection strength
$s_{i,k_1 \cdots k_d}$	Selection coefficient for individual i of the frequency of A in individuals $k_1, \ldots, k_d$
$S_i(\mathbf{p})$	Sum of selection coefficients for individual i times their respective allele frequency products
$\mathbf{S}(\mathbf{p}) = (S_1(\mathbf{p}), \ldots, S_{N_{\mathrm{T}}}(\mathbf{p}))$	Vector of sums of selection coefficients and allele frequency products
$\mathrm{E}^\circ\left[T_{ik_1 \cdots k_d}\right]$	Expected coalescence time of alleles in individuals i and k_1 through k_d under neutrality
Q_{ij}	Probability of identity by descent between alleles in individual i and j
r	Genetic relatedness
κ	Scaled relatedness coefficient
k	Selection gradient proportionality constant

Table 1: Description of symbols.

	Effect on social partner (b)	
Effect on focal ($-c$)	+	−
+	Mutualism	Selfishness
−	Altruism	Spite

Table 2: Definitions of social behavior using Hamilton's rule (eq. 1)

Focal individual	Social partner	
	A | coop	a | noncoop
A | coop	$B - C + D$	$-C$
a | noncoop	B	0

Table 3: Payoffs for the cooperation (coop) allele, A, and noncooperation (noncoop) allele, a, in the social game.

Success condition	Population migration rate		
	$0 < M < \infty$	$M \to \infty$	$M \to 0$
$\pi_{A\|a} > 1/N_{\mathrm{T}}$	$D > 3\left(1 - \frac{1}{3+2M}\right)C$	$\boxed{D > 3C}$	$D > 2C$
$\pi_{a\|A} > 1/N_{\mathrm{T}}$	$D < \frac{3}{2}\left(1 + \frac{1}{3+4M}\right)C$	$D < \frac{3}{2}C$	$D < 2C$
$\pi_{A\|a} > \pi_{a\|A}$	$D > 2C$		

Table 4: Evolutionary success conditions calculated in a infinite island model ($N \to \infty$ and $n \to \infty$) with hard selection. The solid boxed condition is the one-third law [109] and the dashed box condition is the risk dominance condition [31, 64, 72] that holds for all values of M.

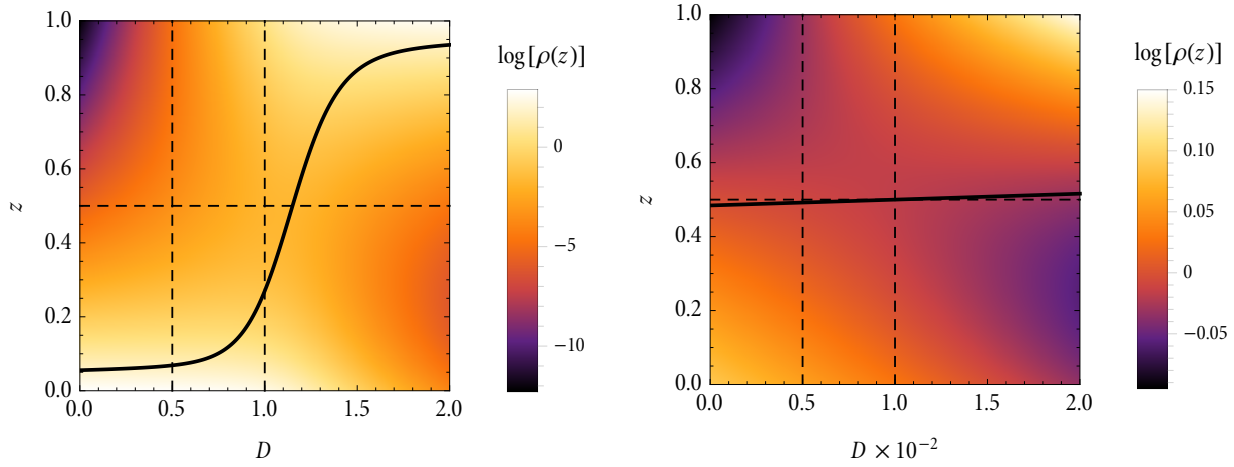

Figure 1: Log stationary density, $\log[\rho(z)]$, for $N_T = 20$ and $\kappa = 0$ plotted as a function of the synergy parameter D. The solid black line represents the mean trait value, $E[z]$. The first vertical dashed line represents the boundary between a Prisoner's Dilemma for $D < C$ and a coordination game $D > C$. For $D < 2C$, full noncooperation ($z = 0$) is risk dominant (see condition 45), and full cooperation is risk dominant for $D > 2C$. Left panel: $B = 1$, and $C = 0.5$. Right panel: $B = 1 \times 10^{-2}$, and $C = 0.5 \times 10^{-2}$.

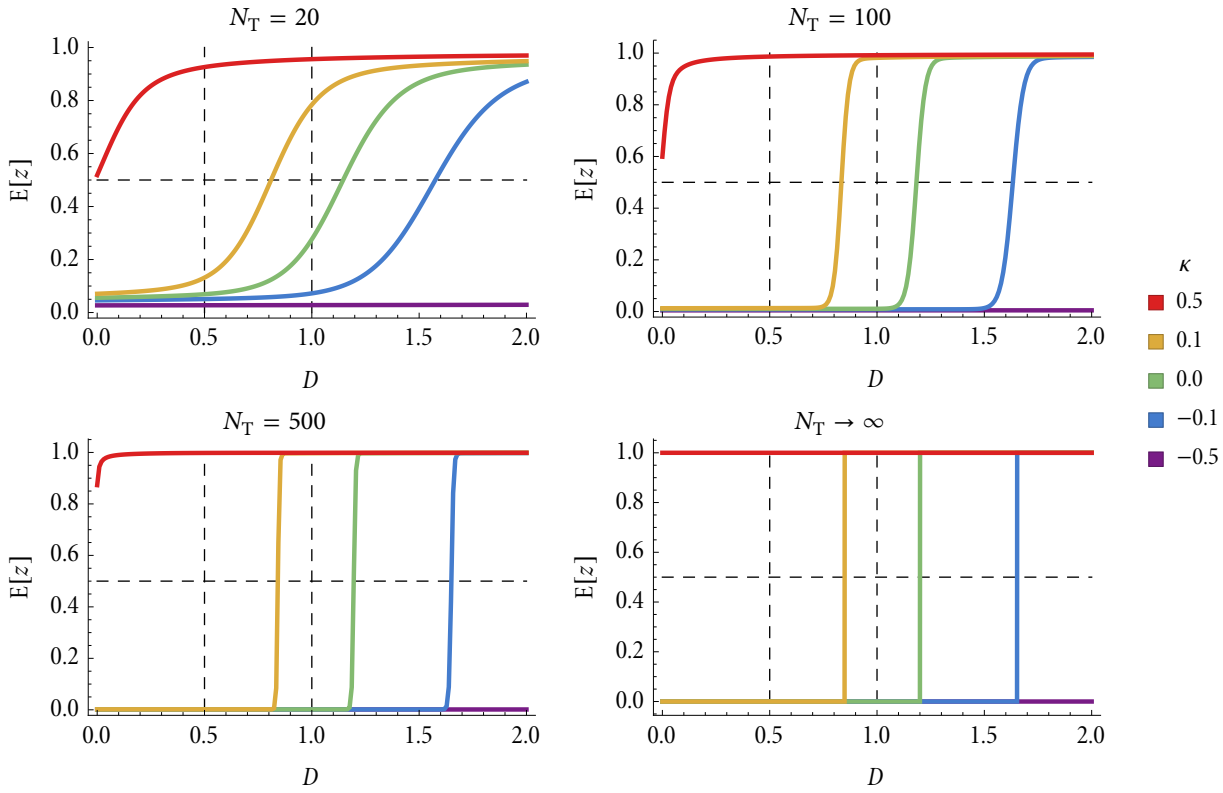

Figure 2: Expected trait value $\mathrm{E}[z]$ at stationarity under the long-term TSS diffusion (eq. 70) as a function of the synergistic payoff D. Total population size is given above each plot and scaled relatedness coefficients are given by the line colors in the legend varying from $\kappa = -0.5$ in purple to $\kappa = 0.5$ in red. The remaining payoffs are set at $B = 1$ and $C = 0.5$, which implies the game is a Prisoner's Dilemma for $D < 0.5$ and a coordination game for $D > 0.5$. For $D < 1$, the trait $z = 0$, full noncooperation, is risk dominant, and full cooperation or $z = 1$ is risk dominant for $D > 1$.

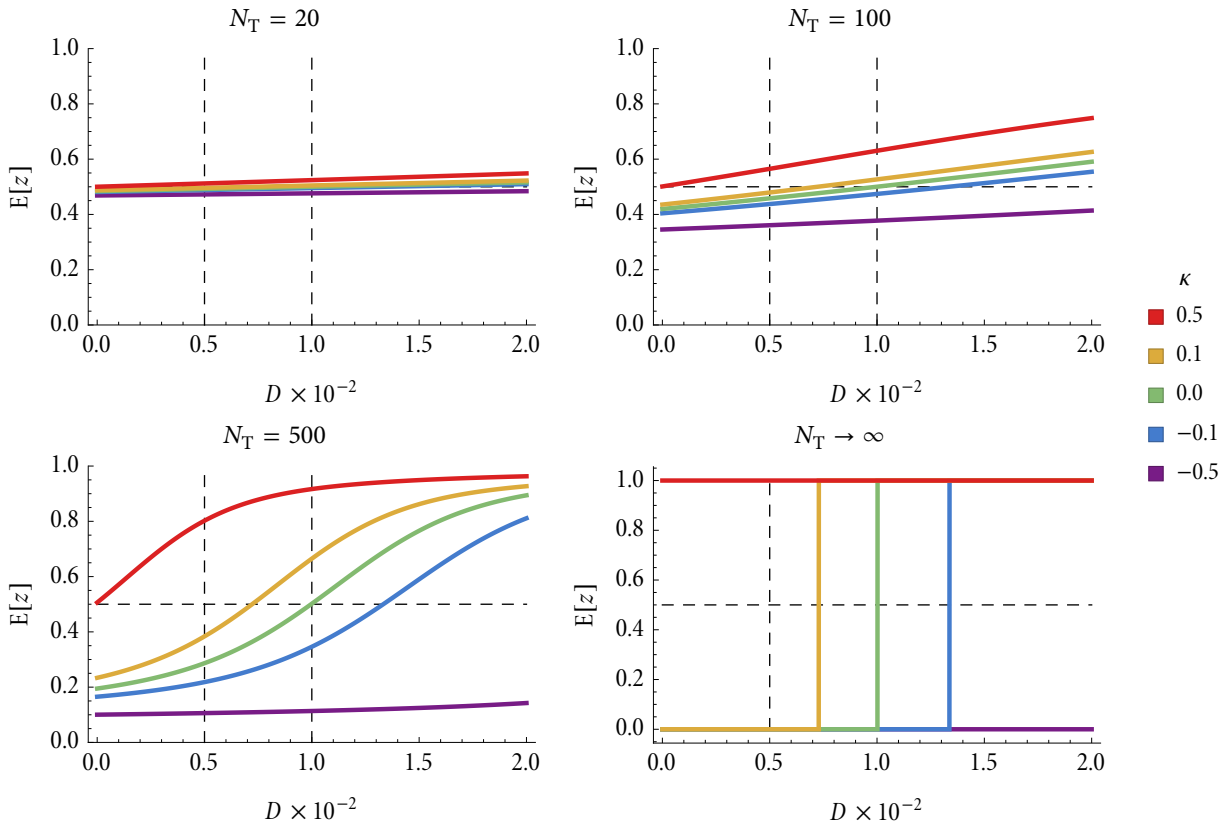

Figure 3: Expected trait value $\mathrm{E}[z]$ at stationarity under the long-term TSS diffusion as a function of D for weak payoffs where all payoffs are scaled by 10^{-2}. Plots are otherwise identical to Figure 2.

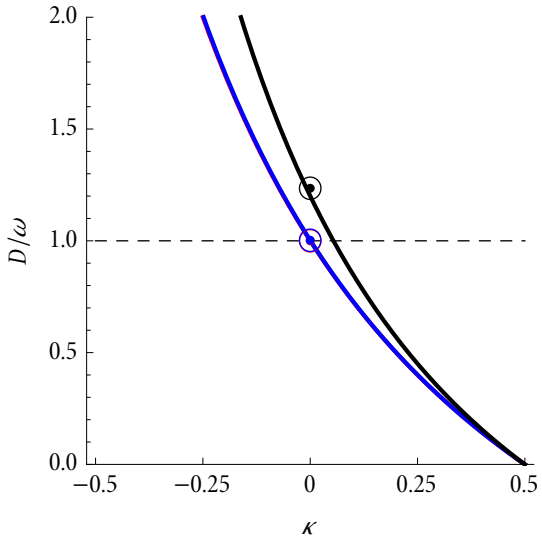

Figure 4: Threshold value of D according to condition (74) above which $z = 1$ and below which $z = 0$ are stochastically stable ($N_T \to \infty$) as a function of scaled relatedness κ. Payoffs are scaled by selection intensity ω where the black curve has $\omega = 1$, blue $\omega = 10^{-2}$, and red $\omega = 10^{-4}$. The circled dots represent the analogous threshold calculated for a panmictic population using condition (76) from Fudenberg et al. [45]. Payoffs are set at $B = 1$ and $C = 0.5$.