bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

When do individuals maximize their inclusive fitness?

Laurent Lehmann* and François Rousset†

2019-05-01

Abstract

Adaptation is often described in behavioral ecology as individuals maximizing their inclusive fitness. Under what conditions does this hold and how does this relate to the gene-centered perspective of adaptation? We unify and extend the literature on these questions to class-structured populations. From a gene-centered perspective, we demonstrate that uninvadable traits (meaning that all deviating mutant go extinct) can be characterized as maximizing the average over classes of either class-specific average direct fitness or of class-specific inclusive fitness. These two fitness measures are defined as reproductive-value weighted averages over distributions of demographic and genetic contexts. These distributions usually depend on events in previous generations, and are thus not under individual control, which prevents, in general, from envisioning individuals as autonomous fitness maximizers. For weak selection in uninvadable population states, however, the dependence of the contextual distributions on earlier events can be neglected, and then all individuals in all classes can be envisioned as inclusive fitness maximizers (but not generally as average direct fitness maximizers). This defines an individual-centered perspective of adaptation and justifies, as a first-order approximation, the long-heralded perspective of individuals appearing to maximize their inclusive fitness.

Keywords: adaptation, inclusive fitness, game theory, social behavior, maximizing behavior.

*Department of Ecology and Evolution, University of Lausanne, Switzerland.

†University of Montpellier, CNRS, IRD, IHPE, Institut des Sciences de l'Évolution, France.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Introduction

One striking hallmark of living systems is their functional organization. From molecular, cellular, and physiological structures within individuals to behavioral interactions between them, organisms in nature display a purposefulness in form and a goal-directedness in action that has been marveled at by generations of biologists (Darwin, 1859; Fisher, 1930; Williams, 1966; Dawkins, 1996; Grafen, 2007). This outward functionality is so unequivocal that humanity has attributed purpose to animals and plants since the mists of time.

Can this purposefulness be characterized? It is well-understood that the functionality of organisms is born out of natural selection, which causes organisms to become adapted to their biotic and abiotic environments over evolutionary time. Over short time scales, mutations are limited and allele-frequency changes, resulting from differences in organismic forms and behaviors, involve selection among a limited number of alternative variants present in the population. Since to each trait combination of an organism there is an associated reproduction and survival schedule, the process of genetic adaptation is often depicted as the maximization of individual fitness. Survival and reproduction, however, also depend on the environment in which individuals reside, and in particular on the traits of conspecifics. An organism's environment thus varies in response to change in trait composition in the population. This prevents a net increase in individual fitness over evolutionary time, even supposing that at all times alleles increasing survival and reproduction are favored by evolution, since the goalposts of the survival and reproductive games of life are shifting as evolution proceeds. And even with fixed goalposts, increase in survival and reproduction may be prevented by multilocus effects in the presence of recombination, as highlighted in some classical criticisms of fitness maximization (e.g., Moran, 1964; Ewens, 2004; Bürger, 2000; Ewens, 2011).

Over long time scales, an organism can be regarded as adapted to its environment (or the totality of situations it can encounter) if no alternative trait combination or behavioral schedule can be produced by mutation, which would result in further allele frequency change (Fisher, 1930). In this long-term perspective, the maximization of the geometric growth ratio of a mutant

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

allele when rare–referred to here as *invasion fitness*–in a population where individuals express a resident allele, provides a condition of uninvadability of mutant traits (all deviating mutants go extinct). This is a defining property of an evolutionary stable population state in which the resident trait combination is a best-response to any mutant deviation (Eshel, 1983; Metz et al., 1992; Ferrière and Gatto, 1995; Eshel et al., 1998; Metz, 2011). And it is in terms of this notion of best-response that maximization of (invasion) fitness can indeed be conceived in the long-term evolutionary perspective (Eshel and Feldman, 1984; Liberman, 1988; Eshel, 1996; Hammerstein, 1996; Weissing, 1996; Eshel et al., 1998).

Invasion fitness is the per capita number of mutant copies produced by the whole mutant lineage descending from an initial mutation over a life-cycle iteration, when the mutant reproductive process has reached stationarity in a resident population. The resulting condition for uninvadability applies regardless of the underlying genetic details (Eshel and Feldman, 1984; Liberman, 1988; Eshel, 1996; Hammerstein, 1996; Weissing, 1996; Eshel et al., 1998), but shows that invasion fitness is a property of a collection of interacting individuals, and gives no reason to say that in an uninvadable population state the fitness of any of these individuals is maximized (in the best-response sense). Indeed, the gene-centered perspective of evolution (Hamilton, 1963, 1996; Dawkins, 1976, 1982; Haig, 1997b, 2012) has distanced itself from ideas of maximization of individual survival and reproduction altogether long ago, and focuses on the differential transmission of alleles to understand adaptation.

Yet Hamilton (1964) attempted to draw a bridge between the gene and the individual-centered perspective of adaptation by defining inclusive fitness, a quantity that is assigned to a representative carrier of an allele, so that natural selection proceeds as if this quantity is maximized over long-term evolutionary time scales. It is indeed attractive to think of evolution by natural selection as targeting individual behavior so that individuals themselves can be regarded as fitness-maximizing agents. This has been argued to have at least some heuristic value (e.g., Maynard Smith, 1982; Dawkins, 1978; Grafen, 1984, 2007; West and Gardner, 2013), and is a working assumption in behavioral ecology (McNamara et al., 2001; Alcock, 2005) and evolutionary psychology (Alexander, 1990; Buss, 2005). It is also a perspective often endorsed in social

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

evolution theories (Bourke, 2011; West and Gardner, 2013). For example, one may say that sterile workers maximize their inclusive fitness by helping a colony queen to raise offspring. Here, it is acknowledged that workers, being sterile, do not maximize their individual fitness, but rather the survival and expected reproduction of other individuals bearing the same allele.

Despite the attractiveness of the individual-centered perspective of adaptation, there has been few formal models supporting it and/or delineating the conditions under which individuals can be regarded as autonomous agents maximizing their own maximand. For instance, Grafen (2006a) considers that individuals maximize an inclusive fitness, which, formally, does not depend on the behavior of other conspecifics and thus appear on our reading to not cover social interactions in any broad sense. It has also be shown that, in age-structured population without social interactions and in group-structured populations with social interactions, individuals appear to maximize a reproductive-value weighted average individual fitness (respectively Grafen, 2015 and Lehmann et al., 2015), which is distinct from inclusive fitness. More generally, connections between inclusive fitness, direct fitness, and individual maximization behavior have been discussed in the literature on kin selection, evolutionary stable traits, and adaptive dynamics (e.g., Hines and Maynard Smith, 1978; Michod, 1982; Maynard Smith, 1982; Eshel, 1991; Mesterton-Gibbons, 1996; Eshel et al., 1998; Day and Taylor, 1996; Frank, 1998; Day and Taylor, 1998; Rousset, 2004; Lehmann and Rousset, 2014a; Akçay and Van Cleve, 2016; Okasha and Martens, 2016; Eshel, 2019), but not in a cohesive way, often not by emphasizing enough the distinction between the individual and the gene-centered perspective of adaptation, and generally not covering the case of class-structured populations (e.g., queen and worker, male and female, young and old individuals).

Our goal in this paper is to formalize and push forward, as far as is consistent with the gene-centered perspective, the individual-centered approach according to which individuals may maximize their own maximand in an uninvadable population state. To ground and develop this formalization, we use the common framework of evolutionary invasion analysis and proceed in three steps.

First, we present a standard model of evolution in a group-structured population with limited

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

dispersal and class-structure. Therein, we formalize both the classic gene-centered perspective of uninvadability and a corresponding individual-centered perspective. In doing so, we introduce a representation of invasion fitness for class-structured populations, which is based on an average individual fitness and provide novel results concerning the exact, gene-centered versions of inclusive fitness, which we extend to diploids with class-structure. Second, in order to highlight the conditions under which an individual-centered perspective of adaptation may or may not emerge, we discuss the properties of the component's terms of these two measures of fitness, which both involve reproductive-value weighting. Third, we provide a definition of inclusive fitness under an individual's own control in class-structured populations, and which individuals appear to maximize under weak-selection in an uninvadable population state.

By covering these three steps of analysis, this paper unifies and extends previous results on the relationship between maximizing behavior and the concept of adaptation *sensu* Reeve and Sherman (1993, p. 9), i.e., "a phenotypic variant that results in the highest fitness among a specified set of variants in a given environment". This allows us to provide a complete bridge between evolutionary invasion analysis, inclusive fitness theory, and game-theoretic approaches. We conclude with a number of take-home messages about the interpretation of fitness maximization and highlight their implications for empirical works.

The model

Assumptions

In order to formalize and compare the gene and the individual-centered approach to adaptation in a simple way but retain key biological population structural effects, we endorse two sets of well-studied assumptions.

Demographic assumptions

First, we assume that evolution occurs in a population structured into an infinite number of groups (or demes or patches), each with identical environmental conditions, and connected to each

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

other by random and uniformly distributed, but possibly limited, dispersal (i.e., the canonical demographic island model of Wright, 1931). Demographic time is discrete and during each demographic time period, reproduction, survival, and dispersal events occur in each group with exactly n individuals being censused at the end of a time period (after all relevant density-dependent events occurred). Each of the n individual in a group belongs to a class (e.g., male or female, young or old). The set of classes is denoted $\mathcal{C}$. We assume that the number of classes is finite, and that the frequencies of individuals in each class can differ among classes within a group (but not among groups).

Each individual can express a class-specific trait that affects its own survival, reproduction, and dispersal and possibly those of group neighbors. We label individuals in a focal group of the population from 1 to n and consider a focal individual i from that group. If this individual is in class a, then it is assumed to express trait $x_{a(i)}$, which is taken from the set $\mathcal{X}_a$ of feasible traits available to an individual of class a (i.e., $x_{a(i)} \in \mathcal{X}_a$). We denote by $x_i = (x_{1(i)}, x_{2(i)}, ..., x_{|\mathcal{C}|(i)})$ the vector of all traits that individual i may possibly express during its lifespan (in an age-structured population, individual i will express different traits at different ages and there may be constraints between these traits). The trait vector x_i is taken from the set $\mathcal{X}$ of all alternative traits available to an individual (i.e., $\mathcal{X} = \prod_{a\in\mathcal{C}} \mathcal{X}_a$). In terms of these notations, $w_{ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ will stand for the expected number of surviving class-u offspring *per haplogenome* produced over a demographic time period by a class-a individual i with trait $x_i \in \mathcal{X}$ when group neighbors have trait profile $\boldsymbol{x}_{-i} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_N)$ in a population where the average individual trait profile is $\bar{x} \in \mathcal{X}$ (see Appendix A for formal definitions and Table 1 for a summary of notation). We refer to w_{ua} as the *individual fitness* function as it determines the number of successful gametes per haploid set of an individual (Grafen, 1985). Individual fitness gives the average number of replicate gene copies produced by an individual per homologous gene and determines evolutionary change over a demographic time period given the traits of all individuals in the population. For simplicity of presentation, we made the assumption that the effects of individuals from different groups on a focal individual's fitness is mean-field; that is, it depends only on the population average trait. Otherwise, we need to take into account the profile of traits in each different group

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

to calculate the fitness function w_{ua} (see section "Scope of our results" for a discussion of the restrictive assumptions of our model).

Evolutionary assumptions

The above characterization of trait expression and individual fitness in the population is individual-centered. In effect, no assumptions so far have been made on the genetic composition of the population and each individual may express a different trait and is thus distinguishable from any other individual. In our to understand which traits are favored over the long term by evolution in this population, we now turn to our second set of assumptions. We place ourselves in the framework of an evolutionary invasion analysis ("ESS approach", e.g., Eshel and Feldman, 1984; Tuljapurkar, 1989; Parker and Maynard Smith, 1990; Metz et al., 1992; Charlesworth, 1994; Ferrière and Gatto, 1995; Eshel, 1996; Caswell, 2000; Otto and Day, 2007; Metz, 2011), according to which we consider a population that is monomorphic for some resident trait and aim at characterizing the conditions according to which a mutant allele changing trait expression is unable to invade the population. For this, we let resident individuals (necessarily homozygotes if diploid) have the vector $y = (y_1, y_2, ..., y_{|\mathcal{C}|}) \in \mathcal{X}$ of traits, one for each class of individuals, where y_a is the trait of a (homozygote) individual of class a. We let a heterozygote mutant individual have trait vector $x = (x_1, x_2, ..., x_{|\mathcal{C}|}) \in \mathcal{X}$ and denote by $z \in \mathcal{X}$ the trait vector of a homozygote mutant. We assume that heterozygote traits are convex combinations ("weighted averages") of homozygote traits, which means that we rule out over-, under-, and strict dominance, but otherwise allow for arbitrary gene action. This allows us to write the trait of mutant homozygotes $z(x, y)$ as a function of the traits x of heterozygote and y of resident homozygotes (Appendix eqs. A.13–A.14 for a formal definition), and this also covers the haploid case where we simply assign trait x to mutant and trait y to residents.

With these assumptions, and whether we consider a haploid population or a diploid population, the fate of a mutant allele arising as a single copy in a population can be determined by its *invasion fitness* $W(x, y)$, which is defined here as the average individual fitness w_{ua} of a randomly sampled mutant carrier over the distribution of the states in which the mutation can re-

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

side. Namely, invasion fitness is the average individual fitness w_{ua} of mutant gene copies over the different states, homozygote or heterozygote, in which this gene copies can reside ($x_i \in \{x, y\}$), and over all demographic classes, and over all combinations of mutant and resident trait profiles of all group members of different classes experienced by carriers of the mutant gene copies. Henceforth, the mutant allele cannot invade when

$$W(x, y) \leq 1. \tag{1}$$

This and all other formal arguments subtending our analysis and models are presented in the extensive Appendix, which fully develops, generalizes, and sometimes discuss more in detail the technical concepts and results presented in the main text.

The gene and the individual-centered perspectives of adaptation

Characterizing uninvadability

Suppose that a given resident trait, say $x^* = (x_1^*, x_2^*, ..., x_{|\mathcal{C}|}^*)$, is uninvadable; namely, it is resistant to invasion by any alternative trait from the set of all possible traits $\mathcal{X}$. This trait x^* must then be a best response to itself, meaning that if we can vary invasion fitness $W(x, x^*)$ by varying the mutant trait x, an uninvadable trait must be a trait maximizing invasion fitness. Formally, it follows directly from eq. (1) that an uninvadable trait x^* satisfies

$$x^* \in \arg\max_{x \in \mathcal{X}} W(x, x^*), \tag{2}$$

which means that x^* belongs to the set of traits resulting in the highest invasion fitness among all alternatives given in the set $\mathcal{X}$ of feasible traits, for the resident population at the uninvadable state (in eq. 2, x^* belongs to a set because an uninvadable trait is not necessarily unique). Hence, x^* qualifies as an adaptation in the sense of Reeve and Sherman (1993, p. 9), i.e., "A phenotypic variant that results in the highest fitness among a specified set of variants in a given environment', and it is the outcome of competition among gene lineages "attempting" to maximize their own

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

transmission across generations since "fitness" is invasion fitness.

Performing a best response in no way, however, implies an increase of mean individual fitness in the population. The key issue is that even if invasion fitness is maximized, for a given resident trait, this fitness changes with the resident trait value as a result of interactions between individuals in the population. Therefore, neither fecundity nor survival, are generally increased by evolution, a point that has repeatedly been emphasized in the literature (e.g., Maynard Smith, 1982; Parker and Maynard Smith, 1990; Kawecki, 1993; Metz et al., 2008b,a). A point that has perhaps been less emphasized, however, is that uninvadable traits can be predicted from the maximization of invasion fitness in the best-response sense is a result that does not hinge on genetic constraints, as it holds in multilocus systems as well (Eshel and Feldman, 1984; Liberman, 1988; Hammerstein, 1996; Eshel, 1996; Weissing, 1996; Eshel et al., 1998).

Characterizing individual maximizing behavior

The focus on the gene-centered maximand W that arises from the evolutionary invasion analysis raises the question: is there a fitness measure that an individual from the population will appear to be maximizing in an uninvadable population state? For individuals to be maximizing some fitness they need to be envisioned as autonomous decision-makers, each with its own class-specific maximand, which, a priori, is distinct from invasion fitness or individual fitness. We will generically denote $w_{\mathrm{I},a}$ any such individual-centered maximand, and call it the *fitness as-if* of an individual of class a. $w_{\mathrm{I},a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is the value of this function in terms of traits expressed by group and population members, where $x_i = (x_{a(i)}, x_{-a(i)})$ is individual i's trait, which is here decomposed into the trait $x_{a(i)} \in \mathcal{X}_a$ expressed when in class a and the remaining traits $x_{-a(i)} = (x_{1(i)}, x_{2(i)}, ..., x_{a-1(i)}, x_{a+1(i)}, ..., x_{|\mathcal{C}|(i)})$ available were the individual in any other class $(x_{-a(i)} \in \mathcal{X} \setminus \mathcal{X}_a)$. As for individual fitness, $\boldsymbol{x}_{-i}$ is the trait profile of the neighbors in the focal's group and $\bar{x} \in \mathcal{X}$ is the average trait of an individual in the population (the concept of fitness as-if is more formally fully developed in Appendix C.2).

The simplest and most widely used concept for the prediction of behavior under the control of the organism is that of a Nash equilibrium trait profile, compared to which no individual

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

can get a higher payoff by a unilateral deviation of behavior (see e.g., Luce and Raiffa, 1957, Fudenberg and Tirole, 1991 or Mas-Colell et al., 1995 for clear discussions of fundamental game theory concepts). A symmetric Nash equilibrium $x^* = (x_1^*, x_2^*, ..., x_{|\mathcal{C}|}^*)$, where each individual in the same class expresses the same trait when striving to maximize its fitness as-if satisfies

$$x_a^* \in \arg\max_{x_{a(i)} \in \mathcal{X}_a} w_{\mathrm{I},a}\left((x_{a(i)}, x_{-a(i)}^*), \boldsymbol{x}_{-i}^*, x^*\right) \qquad \forall a \in \mathcal{C}, \tag{3}$$

where $x_{-a(i)}^* = (x_1^*, x_2^*, ..., x_{a-1}^*, x_{a+1}^*, ..., x_{|\mathcal{C}|}^*)$ is the Nash trait profile of the focal in any class different from a, and $\boldsymbol{x}_{-i}^*$ is the trait profile of all neighbors of the focal individual at the Nash equilibrium trait profile, so that entry j of $\boldsymbol{x}_{-i}^*$ is equal to x_a^* if neighbor $j \neq i$ is of class a. Eq. (3) says that the Nash equilibrium trait x^* belongs to the set of traits maximizing the individual maximand $w_{\mathrm{I},a}$, holding the traits of all others at the Nash equilibrium.

Can we find a class-specific fitness as-if function $w_{\mathrm{I},a}$ such that individuals in each class can be regarded as autonomously maximizing this function in an uninvadable population state? Formally, a positive answer to this question implies that there exists a trait profile $\boldsymbol{x}^*$ satisfying simultaneously fitness as-if maximization (eq. 3) and invasion fitness maximization (eq. 2). If we can find such an as-if maximand, then long-term adaptation can be conceived at the individual level, since, in equilibrium, no unilateral change of individual behavior can be found in any class that would be favored by natural selection. In order to answer the as-if question, we discuss in a first step more in depth the gene-centered perspective of adaptation in class-structured populations and introduce two representations of invasion fitness; namely, the *average direct fitness* and the *inclusive fitness* out of which we build fitness as-if representations in a second step.

Gene-centered representations of adaptation

In order to introduce the key concepts underlying the gene-centered representations of fitness in a progressive way, we start by presenting an expression for invasion fitness for haploids with limited dispersal in the absence of within-group class structure among groups of size $N = 2$ and

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

then turn to consider class-structure without limited dispersal. In the main text we thus present the results of our analysis in a simple way, only retaining key biological features.

Average direct fitness

With two adult haploid individuals in a group without class structure, the individual fitness to a focal individual i with trait x_i when its single neighbor expresses trait x_{-i}, in a population with average trait $\bar{x}$ is simply $w(x_i, \boldsymbol{x}_{-i}, \bar{x})$ (an example of such fitness is given in Box 1). Then, the invasion fitness of a mutant allele coding for trait $x \in \mathcal{X}$ in a resident population $y \in \mathcal{X}$ is

$$W(x,y) = w(x,y,y)q_1(x,y) + w(x,x,y)q_2(x,y), \tag{4}$$

where $q_k(x,y)$ is the probability that, conditional on carrying the mutant allele, an individual finds itself in a group with k mutants. Hence, the vector $\boldsymbol{q}(x,y) = (q_1(x,y), q_2(x,y))$ gives the distribution of the genetic group states in which a carrier of the mutant allele can reside $(q_1(x,y) + q_2(x,y) = 1)$, which depends on both mutant and resident traits. Hence, for limited dispersal, invasion fitness (eq. 4) of a mutant allele is not only a function of the individual fitness of individuals bearing that allele, but also of the $\boldsymbol{q}(x,y)$ distribution of individuals bearing copies of the allele. As such, invasion fitness depends on the fitness of a collection of individuals taken over multiple generations and represents the average replication ability of a randomly sampled allele from the mutant lineage.

This distributional aspect underlying invasion fitness already appears in class-structured populations even in the absence of limited dispersal. To see this and how it affects selection, let us consider a seasonal population of diploid social insects who allocate resources to the production of three classes of individuals, reproductive males, reproductive females (queens), and workers. The life cycle is as follows. (1) At the beginning of the season each group is occupied by exactly a single mated queen that initiates a colony by producing workers that help produce sexuals. (2) At the end of the season, all reproductive individuals disperse at the same time and individuals of the parental generation die. (3) Random mating occurs, all queens mate exactly with one

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

male and then compete for vacated breeding slots to form the next generation. Under these assumptions, successful gene copies must pass through the single mated female in each group; and the number of class-u offspring produced by a class-a mutant individual per haplogenome can be written $w_{ua}(x,y)$, where trait $x = (x_{\mathrm{f}}, x_{\mathrm{m}}, x_{\mathrm{o}})$ collects, respectively, the traits of females, males, and workers (i.e., $\mathcal{C} = \{\mathrm{f}, \mathrm{m}, \mathrm{o}\}$). The trait for the resident is $y = (y_{\mathrm{f}}, y_{\mathrm{m}}, y_{\mathrm{o}})$, whereby the invasion fitness of the mutant can be written as

$$W(x,y) = \frac{1}{V(x,y)} \sum_{a \in \mathcal{C}} w_{\mathrm{T},a}(x,y)\phi_a(x,y), \tag{5}$$

where

$$\begin{aligned} w_{\mathrm{T,f}}(x,y) &= v_{\mathrm{f}}(y)w_{\mathrm{ff}}(x,y) + v_{\mathrm{m}}(y)w_{\mathrm{mf}}(x,y) \\ w_{\mathrm{T,m}}(x,y) &= v_{\mathrm{f}}(y)w_{\mathrm{fm}}(x,y) + v_{\mathrm{m}}(y)w_{\mathrm{mm}}(x,y) \\ w_{\mathrm{T,o}}(x,y) &= 0. \end{aligned} \tag{6}$$

Here, $w_{\mathrm{T},a}(x,y)$ is the total weighted expected fitness of a class-a mutant and $\phi_a(x,y)$ is the probability that, conditional on carrying the mutant allele, an individual finds itself in class a. Hence, the vector $\boldsymbol{\phi}(x,y) = (\phi_a(x,y))_{a \in \mathcal{C}}$ gives the distribution of class states in which a carrier of the mutant allele can reside and it depends on both mutant and resident traits. A worker does not reproduce and henceforth has a zero direct fitness, but its class frequency is non zero ($\phi_{\mathrm{o}}(x,y) \neq 0$; the explicit expression for $\phi_{\mathrm{f}}(x,y)$, $\phi_{\mathrm{m}}(x,y)$, and $\phi_{\mathrm{o}}(x,y)$, are given in the Appendix, eq. A.30) and it helps its parents to reproduce. Formally, the worker affects the reproduction of its male and female parents through the dependence of individual fitness on trait vector $x = (x_{\mathrm{f}}, x_{\mathrm{m}}, x_{\mathrm{o}})$ (see Box 2 for an explicit example of fitness functions affected by worker trait).

In eq. (6), an offspring of class u is weighted by the neutral reproductive value $v_u(y)$ of a gene copy in its class (i.e., the asymptotic contribution to the gene pool of a single class-u individual in a monomorphic resident population). Eq. (6) is thus the sum of the reproductive-value weighted direct-fitness components of an individual of class s, including its potentially

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

surviving self. This was called the Williams' reproductive value by Grafen (2015, p. 8), but we refer to $w_{\mathrm{T},a}$ as the *(average) direct fitness* of a class-a mutant because it counts the (expected) number of offspring of an individual, and to contrast it with the forthcoming inclusive fitness measure of a mutant (eq. 8), which also involves reproductive-value weights. So with the term "direct" in average direct fitness we here aim to emphasize the key difference between eq. (6) and the forthcoming corresponding inclusive fitness expression, eq. (8), which depends on indirect fitness effects. Eq. (5) also depends on $V(x,y)$, which is the average of the $v_u(y)$ reproductive values over the $\boldsymbol{\phi}(x,y)$ distribution, thus giving the asymptotic contribution, to the gene pool, of a randomly sampled mutant individual that is assigned the total offspring (reproductive) values of a resident individual. Hence, invasion fitness can be represented as the average direct fitness of a mutant relative to the average direct fitness this individual would have if it was assigned the individual fitness of a resident individual (see Appendix A.3.2 for more conceptual details on reproductive value).

The unit of adaptation

The examples of invasion fitness (eq. 4 for limited dispersal and eq. 5 for class structure) make explicit that invasion fitness is the average individual fitnesses over some appropriately defined distribution of the genetic-demographic states (the $\boldsymbol{q}(x,y)$ and $\boldsymbol{\phi}(x,y)$ distributions) that matters for selection (expressions taking into account both limited dispersal and class structure with arbitrary group size are given in Appendix, eq. A.19). This focus on gene replication epitomizes the gene-centered perspective of evolution, according to which it is not the individual fitness of a single individual (or a single gene copy) in a given demographic and genetic context that matters for selection, but the average of such individual fitnesses over a distribution of genetic contexts (Dawkins, 1978; Haig, 1997b, 2012). Natural selection on an allele thus depends not only on how it changes the immediate survival and reproduction of its carriers, but also the survival and reproduction of carriers through changes in the *context* in which the allele at a given locus can be found (Kirkpatrick et al., 2002, p. 1728). Indeed, an allele can reside, say in a worker or a queen, be inherited from a mother or a father, and is likely to be in a genome with many other loci with

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

different allele combinations. The importance of such contexts of alleles for their evolutionary dynamics has been much emphasized in populations genetics (e.g., Altenberg and Feldman, 1987; Kirkpatrick et al., 2002; Roze, 2009). There are even mutations that spread through selection only by way of their effects on changes of the contexts in which they are found. Typical examples are modifier alleles involved in the evolution of recombination or migration, which may spread by increasing their chance of being in a genetic context with higher fitness, despite the modifier having no direct physiological effect on reproduction and/or survival in a given context (e.g., Altenberg and Feldman, 1987; Kirkpatrick et al., 2002; Roze, 2009). From now on, we refer to $\boldsymbol{q}(x,y)$ and $\boldsymbol{\phi}(x,y)$ as the contextual distributions.

Inclusive fitness

What is missing in the previous representations for invasion fitness is twofold. First, it is a simple and intuitive quantification of the effect of spatial structure that summarizes the variation on fitness introduced by the distribution over genetic contexts (the $\boldsymbol{q}(x,y)$ distribution). Second, it is a quantification of the contribution to fitness of the different classes of individuals that really contribute to allele transmission. For instance, in the social insect example, males have fitness but their trait does not contribute to adaptation, while workers have no fitness, but their trait contributes to adaptation. This raises the question of how one can identify the force of selection on worker trait in a fitness measure?

To answer this question, we now turn to an alternative representation of uninvadability. In the presence of class structure and regardless of the number of individuals within groups and whether dispersal is limited or not, an uninvadable trait can be obtained as the best-response maximization of the inclusive fitness

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{a\in\mathcal{C}} \Delta w_{\mathrm{IF},a}(x,y) r_{\mathrm{f},a}(x,y)\phi_a(x,y), \tag{7}$$

of a rare mutant trait x in a resident diploid population. Here, $r_{\mathrm{f},a}(x,y)$ is the probability that, conditional on an individual of class a carrying the mutant allele, a randomly sampled homologous

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

gene in that individual is mutant, and

$$\Delta w_{\mathrm{IF},a}(x,y) = \sum_{u\in\mathcal{C}} v_u(y)\left[-c_{ua}(x,y) + \sum_{s\in\mathcal{C}} r_{s|a}(x,y)\, b_{us\leftarrow a}(x,y)\right] \tag{8}$$

is the inclusive fitness effect of a class-a carrier of the mutant allele. Further, $c_{ua}(x,y)$ is the additive effect on the number of class-u offspring produced by a single class-a individual when expressing the mutant instead of the resident allele and $b_{us\leftarrow a}(x,y)$ is the additive effect on the number of class-u offspring produced by all class-s neighbors and stemming from a single class-a individual expressing the mutant instead of the resident allele. These costs and benefits hold regardless of the number of group partners and was reached by using a two-predictor regression of individual fitness, as in the exact version of kin selection theory (e.g., Queller, 1992; Frank, 1997; Gardner et al., 2011; Rousset, 2015; see Appendix B for a derivation of eq. 7 and for a comparison with an alternative version of inclusive fitness based on a single-predictor regression, which may be more in line with certain empirical estimates of inclusive fitness). Finally, we have

$$r_{s|a}(x,y) = \frac{r_{\mathrm{n},s|a}(x,y)}{r_{\mathrm{f},a}(x,y)}, \tag{9}$$

which is the relatedness between a class-a actor and a class-s recipient. Here, $r_{\mathrm{n},s|a}(x,y)$ is the probability that, conditional on an individual of class a carrying the mutant allele, a randomly sampled homologous gene in a (non-self) neighbor of class s is a mutant allele. Hence, relatedness $r_{s|a}(x,y)$ can be interpreted as the ratio of the probability of indirect transmission by a class-s individual of a mutant allele taken in a class-a individual to the probability that the individual transmits itself this allele to the next generation. In the absence of selection, this is equivalent to the standard ratio of probabilities of identity-by-descent (Hamilton, 1970, p. 1219, Lehmann and Rousset, 2014b, eq. A.5). Relatedness is expressed in terms of the class-specific mutant copy number distribution (the $\boldsymbol{q}(x,y)$ distribution) and as such summarizes the statistical effect of limited dispersal on mutant-mutant interactions.

According to these definitions, $w_{\mathrm{IF},a}(x,y)$ is the total effect of an individual of class a on the

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

reproductive-value weighted number of gene copies produced by all recipients of its action(s). A fundamental difference between total offspring (reproductive) value (eq. 6) and the inclusive fitness effect of a class-a carrier (eq. 8) is that the direct fitness is non-null only for individuals who reproduce, while the inclusive fitness effect is non-null only for individuals whose trait affects individual fitnesses (theirs' or others') in the population. This fundamental difference is illustrated by our specific example of the colony of social insects, the inclusive fitness effects of males and females are nil, while the inclusive fitness effect of workers is positive in a population at the equilibrium sex-ratio (see Box 2). This contrasts with direct fitness, which is nil for workers ($w_{\mathrm{T},\mathrm{o}} = 0$) but positive for males and females ($w_{\mathrm{T},\mathrm{m}} > 0$ and $w_{\mathrm{T},\mathrm{f}} > 0$, see Box 2). More generally, however, when the sex-ratio is not at equilibrium, both the inclusive fitness effects of males and females may also be non-zero.

The subunits of adaptation

Inclusive fitness is defined at the allele level and is the average effect of individuals carrying an allele on its transmission into the gene pool. This is consistent with the original formulation of this concept (Hamilton, 1964, pp. 3-8), but our own formulation (eqs. 7-8) extends it to class-structured population, showing that it holds regardless of the complexity of social interactions at hand and the strength of selection on the mutant allele. The inclusive fitness representation of invasion fitness makes explicit that a fitness comparison is made between expressing or not the mutant allele, since this involves comparing successful number of offspring gained and lost through behavioral interactions. Hence, inclusive fitness allows to intuitively understand the adaptive significance of behavior at the gene level, as it fastens attention on the pathways determining costs and benefits (Grafen, 1988). This is particularly salient in the case of classes, where a fitness measure can be attached not only to reproductive individuals but also to sterile worker, which can thus be seen as contributing to the fitness of the gene lineage. In other words, inclusive fitness brings upfront those individuals whose traits are under selection and provides a quantification of the force of selection on it. To make this point explicit, consider a mutation $\tilde{x}_a = (x_1^*, x_2^*, ..., x_{a-1}^*, x_a, x_{a+1}^*, ..., x_{|\mathcal{C}|}^*)$ that keeps all traits at the uninvadable state, except for

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

trait x_a of class a that unilaterally deviates (e.g., we consider selection only on the worker trait in the above example). Then, from eq. (7) and eq. (B.31) of Appendix B, the inclusive fitness of this mutant is

$$W_{\mathrm{IF}}(\tilde{x}_a, x^*) = 1 + \Delta w_{\mathrm{IF},a}(\tilde{x}_a, x^*) r_{\mathrm{f},a}(\tilde{x}_a, x^*) \phi_a(\tilde{x}_a, x^*), \tag{10}$$

so that the mutant spreads if a randomly sampled mutant of class a increases its inclusive fitness effect, $\Delta w_{\mathrm{IF},a}(\tilde{x}_a, x^*) > 0$, but the inclusive fitness effect of other classes is zero and does not affect mutant spread ($\Delta w_{\mathrm{IF},v}(\tilde{x}_a, x^*) = 0$ for all $v \neq a$). The inclusive fitness formulation thus shifts attention from those individuals that are passive carriers of alleles (males in our example) to those individuals whose trait actively affects the transmission of alleles (workers).

The expressions for inclusive fitness (eqs. 7–10) also makes explicit that what is maximized in an uninvadable population state is not only the inclusive fitness effect of a particular class under focus. This is so because a mutation changing, say the level of self-sacrifice, may result in a change in the frequency distribution of workers and queens in the population, which can alter the selection pressure to which the mutant allele is exposed in subsequent generations, since it will alter the probabilities that a mutant copy resides in a queen or a worker (i.e. alter the ϕ distribution). In other words, the unit of adaptation is the gene (Dawkins, 1978, 1982; Haig, 2012), and its subunits are the collection of copies expressed differentially in particular classes of individuals. We now turn on discussing the implication of these considerations for identifying a fitness as-if that an individual appears to be maximizing through its own behavior in an uninvadable population state.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Individual-centered perspectives of adaptation

A general individual-centered maximand?

According to the expression for invasion fitness for groups of size $N = 2$ (eq. 4), one fitness as-if implementing uninvadability for haploids in the absence of class structure takes the form

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = w(x_i, x_{-i}, \bar{x}) q_1(x_i, \bar{x}) + w(x_i, x_i, \bar{x}) q_2(x_i, \bar{x}) \tag{11}$$

where $\boldsymbol{x}_{-i} = x_{-i}$. This fitness as-if is defined similarly to invasion fitness, in terms of the average fitness of an individual over a distribution of trait profiles of its group neighbors, but defined here from the probability $q_2(x_i, \bar{x}) = 1 - q_1(x_i, \bar{x})$ that an individual finds itself in a group where the neighbor expresses the same trait, x_i, as itself. Re-ordering this as

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = 1 + \underbrace{(w(x_i, x_{-i}, \bar{x}) - 1)}_{-c_{\mathrm{I}}(x_i, x_{-i}, \bar{x})} + q_2(x_i, \bar{x}) \underbrace{(w(x_i, x_i, \bar{x}) - w(x_i, x_{-i}, \bar{x}))}_{b_{\mathrm{I}}(x_i, x_{-i}, \bar{x})} \tag{12}$$

yields a representation of fitness as-if as inclusive fitness as-if, consistent with a single-predictor version of inclusive fitness (Appendix B.1.3) since $q_2(x_i, \bar{x}) = r_{\mathrm{n}}(x, y)$ where $r_{\mathrm{n}}(x, y)$ is the relatedness between two group neighbors considered in this inclusive-fitness representation.

The main structural difference between invasion fitness $W(x, y)$ (eq. 4) and fitness as-if $w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ (eq. 11) is that the trait of each individual in the group are distinguished, so that the fitness as-if of each of these individuals can be evaluated and be distinct from each other, which we take as a defining property of an individual-centered representation of trait expression. Then, each individual can be regarded as an autonomous decision maker with its own trait and maximand.

Suppose now that each individual maximizes its fitness as-if in an uninvadable population state. Then, individuals in the population will play an uninvadable trait (satisfying eq. 2), because eq. (11) reduces to the same expression as invasion fitness when both are evaluated in an uninvadable population state. This shows that it is possible to obtain a maximand that

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

individuals appear to be maximizing in the best-response sense (in fact eqs. 11–12 provides two representations for such a maximand). However, the actor does not have full control of this maximand, as this would require that whether any group neighbor plays the same trait as self or not is determined by the trait used by self. Fitness as-if can thus be interpreted as the organisms' maximand only if it *controls* the distribution $q_k(x_i, \bar{x})$ over group states. In reality, the $\boldsymbol{q}(x_i, \bar{x})$ distribution cannot be under the actor's control, as the reproduction and survival of ancestors will affect the present genetic structure in the group. Further, the individual fitness component $w(x_i, x_{-i}, \bar{x})$ is not under the individual's full control, as it depends on the trait of others, as will the cost and benefits in inclusive fitness as-if (eq. 12).

The as-if representation of fitness, eq. (11), brings upfront that the genetic structure depends on evolving traits, and that this precludes a biologically satisfactory individual-centered representation of maximizing behavior based on the fitness components determining invasion fitness. Indeed, eq. (11) only considers the statistical rather than the causal dependence of the distribution of traits of neighbors upon the actor's behavior. The problem this raises is exacerbated in the presence of classes. In this case, both the distribution $\boldsymbol{q}(x, y)$ of the number of neighbors expressing the same trait as self, and the distribution $\boldsymbol{\phi}(x, y)$ specifying the probability that an individual finds itself in a given class, depend on evolving traits. This implies that individuals from each class, say workers and queens, cannot be envisioned as autonomously each maximizing their own inclusive fitness. By rearing the queen's offspring, the worker can be viewed as a decision-maker maximizing an inclusive fitness as-if only if the queen's behaviour is fully determined by the worker's behaviour (or vice versa).

Existence of an individual-centered maximand under weak selection

Weak selection concepts

If the genetic- and class-contextual distributions, $\boldsymbol{q}(x, y)$ and $\boldsymbol{\phi}(x, y)$, were to be independent of the mutant trait, then this would allow fitness as-if to be exogenous to the actors' own behavior and this may yield a characterization of the evolutionary equilibrium as maximization of an fitness as-if under the actor's control. There are least two ways to achieve that the state

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

distributions become independent of the mutant trait and both hinge on *weak selection* approximations implying that, to first-order, the distribution of genetic and class states will no longer be dependent on the mutant allele. Such first-order approximations are reached either by assuming that the effect of mutants is small, or that parameters determining both mutant and resident phenotypic effects are small. In the first case ("small-mutation"), one can use the approximation $\boldsymbol{q}(x,y) \sim \boldsymbol{q}(y)$ and $\boldsymbol{\phi}(x,y) \sim \boldsymbol{\phi}(y)$ so that the distributions depend only on the resident type. In the second case ("small-parameter"), the distribution of genetics and demographic states will be independent of trait values altogether $\boldsymbol{q}(x,y) \sim \boldsymbol{q}$ and $\boldsymbol{\phi}(x,y) \sim \boldsymbol{\phi}$ (see Box 3 for an example and Appendix C.4.1 for more formal details).

The key implication is that under a weak-selection approximation a mutant allele will not affect the genealogical and/or class structure to which it is exposed and this structure can thus be held constant. This was a central assumption endorsed by Hamilton (1964, p. 34), and has been used implicitly (as shown by Lessard and Soares, 2016) to obtain a representation of maximizing behavior in the absence of social interactions in age-structured population (Grafen, 2015), and explicitly in the presence of social interactions of arbitrary complexity but without class-structure (Lehmann et al., 2015). We now integrate these previous results (both obtained in the "small-parameter" case and further detailed in Appendix C) into a full game-theoretic representation of maximizing behavior with class structure, and will present a fitness as-if that takes the form of direct fitness and another one that takes the form of inclusive fitness, both being maximized in an uninvadable population state. To describe these as-if fitnesses, we denote $\tilde{w}_{ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ a weak-selection approximation of the class-specific fitness function $w_{ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$.

Average direct fitness maximization

We first reconsider the case of a haploid panmictic population with limited dispersal and no class-structure (the situation underlying eq. (11)), but now under weak selection. In this case, the fitness as-if becomes

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \tilde{w}(x_i, x_{-i}, \bar{x}) q_1(\bar{x}) + \tilde{w}(x_i, x_i, \bar{x}) q_2(\bar{x}), \tag{13}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

which individual i appears to be maximizing in un invadable population state. The fundamental difference with eq. (11) is that in eq. (13) the probability $q_2(\bar{x}) = 1 - q_1(\bar{x})$ that individual i finds itself in a group where the neighbor expresses the same trait, x_i, as itself no longer depends on one's own action. Hence, contextual events are now assigned probabilities that are independent of own trait values. This is a defining feature of the concept of an autonomous decision maker as conceived in classical decision and game theory (Savage, 1954; Luce and Raiffa, 1957; Fudenberg and Tirole, 1991).

Consider now a diploid panmictic population (no limited dispersal) with class structure given by age ($\mathcal{C} = \{0, 1, 2, ...\}$), where class "0" refers to newborns, and assume there are no effects of the traits expressed at any age on the fitness of the actor at later ages (no within-individuals inter-class trait effects). Then, let the fitness as-if of an individual of age a, expressing trait x_i when group neighbors have trait profile $\boldsymbol{x}_{-i}$, be defined as

$$w_{\mathrm{I},a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = v_0(\bar{x})\, \tilde{w}_{0a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + v_{a+1}(\bar{x})\, \tilde{w}_{(a+1)a}(x_i, \boldsymbol{x}_{-i}, \bar{x}). \tag{14}$$

This is the sum of the reproductive values of newborns and the surviving self ($\tilde{w}_{(a+1)a}$ is the probability of survival of an individual of age a). This equation generalizes to social interactions between individuals, the maximand for an asocial world derived for diploidy populations previously by Grafen (2015, eq. 38). This average fitness as-if also defines a maximand which individuals appear to be maximizing in an uninvadable population state.

These two results (eqs. 13–14), however, describe situations where there are no effects of class-specific traits on the fitness of related individuals in another class, i.e. there are no indirect fitness effects across classes (there are no classes under model eq. 13 and the fitness effects of traits are limited to the class where they are expressed in eq. 14). In the presence of indirect fitness effect of traits across classes, individuals of any class can no longer be regarded as each maximizing their own average direct fitness, since this would imply that the mutant fitness is unaffected by the effect of the trait expressed by the individual on related individuals in other classes, say for instance the spread of a mutant allele expressed by a worker is unaffected by the worker helping to rear the offspring of the queen (the worker's mother also sharing the mutant

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

allele; see also discussion after eq. C.22 in the Appendix). We next turn to describe a fitness as-if taking such effects on other classes into account.

Inclusive fitness maximization

Now we let the fitness as-if of a focal individual i, when it is of class a and expresses trait x_i in a group with neighbor trait profile $\boldsymbol{x}_{-i}$, be

$$w_{\mathrm{I},a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \sum_{u \in \mathcal{C}} v_u(\bar{x}) \left[-c_{\mathrm{I},ua}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + \sum_{s \in \mathcal{C}} r_{s|a}(\bar{x})\, b_{\mathrm{I},us \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) \right]. \tag{15}$$

Here, $c_{\mathrm{I},ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is the additive effect on the number of class-u offspring produced by a single class-a individual and $b_{\mathrm{I},us \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is the additive effect on the number of class-u offspring produced by all class-s neighbors and stemming from a single class-a individual switching behavior, and these regression coefficients depend on the behavior of all individuals in interaction. These costs and benefits are now weak-selection approximations of the exact costs and benefits obtained by performing a general regression of the individual fitness of i when in class a on the frequency in itself and its neighbors of a hypothetical allele determining the expression of trait x_i, whereby the effects of switching to expressing x_i can be assessed. This allele is taken to have the same distribution as the mutant allele in the gene-centered model, i.e. the neutral genetic contextual distribution $\boldsymbol{q}(\bar{x})$, here parameterized by the average trait $\bar{x}$ in the population. The only difference between the cost $c_{\mathrm{I},ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ in the individual-centered model and the cost $-c_{ua}(x, y)$ in the gene-centered model (recall eq. 8) is then that all individuals within groups have distinct traits in the individual-centered perspective (and likewise for the benefits $b_{\mathrm{I},us \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ versus $b_{us \leftarrow a}(x, y)$). As such, the probability $r_{s|a}(\bar{x})$ that, conditional on being in class a, a random actor and a random class-s recipient in its group share the same trait is constant with respect to the actor's trait and is equivalent to the standard relatedness in a monomorphic population with trait $\bar{x}$ (sometimes called pedigree relatedness).

Eq. (15) applies to class-structure and social interactions of arbitrary complexity, and provides an individual-centered representation of inclusive fitness. This average defines a fitness as-if which individuals appear to maximize in an uninvadable population state (see Appendix C.4.3 for a

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

proof of this result). In other words, in an uninvadable population state under weak selection, individuals from each class can be regarded as each maximizing their own inclusive fitness.

Scope of the individual-centered perspective of adaptation

We have identified two types of individual-centered maximands, average direct fitness and inclusive fitness, both of which individuals appear to be maximizing in uninvadable population states under weak selection (or if, for other reasons, the $\boldsymbol{q}$ and $\boldsymbol{\phi}$ contextual distributions are independent of selection). The class-specific average direct fitness as-if, however, is not maximized (even under weak selection) in situations where individuals affect the fitness of relatives of other classes, so that the as-if inclusive fitness representation is a more general individual-centered maximand. All these results were obtained assuming simplifying demographic assumptions; in particular, constant group size and abiotic environment, no isolation-by-distance, and discrete time. We now confront each of these the assumptions in turn. First, the number of individuals in each class and thus group size as well as abiotic environments are all likely to fluctuate. To cover these cases, it suffices to follow the recommendation of McPeek (2017) and write individual fitness w_{ua} not only as a function of an individual's trait and that of its interaction partners (group and average population members), but also as a function of relevant endogenous variables (e.g., population size, abiotic environment, cultural knowledge); namely, those variables whose distributions or values are influenced by individual traits. For weak selection, these distributions or values can then be approximated as a function of the resident traits (see Rousset and Ronce, 2004 for concrete examples of fitness functions and distributions covering both demographic and environmental fluctuations). Then, an inclusive fitness as-if under individual control can be defined; the implication being that there are now more contexts to consider relative to the case with no fluctuations (e.g., different group sizes and environments), and individuals face a maximization problem under the constraint that the endogenous distributions or values of contexts are evaluated at the uninvadable trait state. Likewise, taking isolation-by-distance into account calls for an extension of the number of contexts and relatednesses to be considered. But given the contexts and the relatednesses, their distributions or values can again be approximated as

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

function of the resident strategies under weak selection (see Rousset and Billiard, 2000 for such constructions for isolation-by-distance) and again an inclusive fitness as-if under individual control can be defined. We also assumed discrete time but comparison of our results (in particular eq. 14) to those of the continuous time model of Grafen (2015, eq. 38) suggests that here again, only a redefinition of contexts is needed to cover fitness as-if under continuous time.

Our results also relied on specific assumptions about trait expression. While traits themselves can be arbitrary complex, we assumed that they are expressed unconditionally. For instance, our model does not cover the situation where an individual helps its mother as long as she is alive and upon her death starts to help its siblings. Including such relevant cases and more generally conditional trait expression based on individual recognition again calls for an extension of the number of contexts to be considered in the definition of fitness as-if. Finally, we assumed a resident monomorphic population, but this population could be polymorphic and here previous results (Eshel and Feldman, 1984; Liberman, 1988; Eshel et al., 1998) suggest that a fitness as-if could be defined too. In conclusion, all the above scenarios involve considering more complicated $\boldsymbol{q}$ and $\boldsymbol{\phi}$ contextual distributions, so that explicit extensions of our results under these scenarios will need care for the definitions of contexts, fitnesses and relatedness. Such demographic, genetic, and behavioral extensions could be very welcome, but are unlikely to alter our conclusions, as the main requirement for our results to hold in all these cases is that the individual fitness functions and the contextual distributions are differentiable in terms of "small-mutation" and/or "small-parameter" effects.

Discussion

The evolutionary literature provides contrasting messages about the relationship between adaptation and individual behavior as the outcome of fitness maximization. We here combined core elements of evolutionary invasion analysis, inclusive fitness theory, and game theory in order to connect various strands of the literature together and get a hold on the conditions under which individuals can be envisioned as maximizing their own (individual-centered) inclusive fitness in a class-structured population that is in an uninvadable state (all deviating mutants go extinct). In

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

particular, we defined individual maximands which (a) generally differ from measures of invasion fitness (including gene-centered inclusive fitness) by being functions of the traits of an individual and of all its social partners, and thus generally differ for different individuals with the same trait but different partners' traits; but which (b) nevertheless coincide with invasion fitness at an evolutionary equilibrium and (c) appear under individual control under weak selection. We thereby defined an individual-centered inclusive fitness measure that is maximized at an evolutionary equilibrium. Our formal analysis leads to the following take-home messages, old and new.

(1) Fitness maximization obtains in the gene-centered perspective. Uninvadable traits can be characterized in terms of mutant alleles attempting to maximize their own transmission across generations. We showed that invasion fitness can be usefully expressed in terms of the average, over classes, of the class-specific inclusive fitness effect of an allele (eq. 7). This provides a decomposition of the force of selection in terms of direct and indirect transmission of replica copies of this allele, which holds for any selection strength, and allows one to fasten attention on the pathways determining costs and benefits of expressing the allele in different classes of individuals.

(2) Inclusive fitness is a gene-centered fitness measure. Selection on a mutant allele depends on both the individual fitness of its carriers and the distributions of class and genetic contexts in which these carriers reside. Since these distributions are properties of a lineage of individuals over multiple generations, inclusive fitness is not the fitness of a single individual, but that of an average carrier of a mutant allele sampled from the distributions of class and genetic contexts. As such, inclusive fitness is a gene-centered fitness measure (consistent with Hamilton's 1964 original definition) and uninvadability cannot, in general, be characterized in terms of autonomous individuals maximizing their inclusive fitness.

(3) Individual-centered inclusive fitness maximization under weak selection. For weak selection, the distributions of class and genetic contexts, and thus relatedness, can be taken to be unaffected by selection (Hamilton's 1964 original modeling assumption). In this case,

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

we showed that uninvadability can be characterized in terms of individuals, each independently of each other, maximizing an individual-centered inclusive fitness (eq. 15). In particular, this allows one to define class-specific (say, for workers versus queens) inclusive fitnesses that are each maximized at an evolutionary equilibrium. This warrants the view of individuals from different classes as autonomous decision makers, each maximizing their own inclusive fitness, and this holds regardless of the complexity of social interactions and demographic structure at hand.

(4) Individual-centered inclusive fitness is more general than average direct fitness. Invasion fitness can, in addition to gene-centered inclusive fitness, equivalently be expressed in terms of gene-centered average direct fitness; namely, in terms of the average direct fitness of a randomly sampled carrier of the mutant allele (eq. 5). Hence, the gene-centered perspective of adaptations is not unique and different maximands can be conceived (a point noted by Hamilton 1964; 1970). In the individual-centered perspective under weak selection, however, each individual does not appear in general to maximize a biologically meaningful individual-centered direct fitness. For example, sterile workers cannot be said to maximize a worker-specific direct fitness. Thus, maximization of individual-centered inclusive fitness holds in a more general sense than the maximization of individual-centered average direct fitness.

Point (1) follows directly from the fact that alleles are the information carriers of the hereditary components of organismic features and behavior. As emphasized by Dawkins (1979, p. 9), alleles do not act in isolation but in concert with all other alleles in the genome and in interaction with the environment to produce the organism. But uninvadability can be deduced from unilateral deviation of allelic effects alone, and so the inclusive fitness of a mutant allele is sufficient to characterize adaptation in the long-term evolutionary perspective.

Point (2) follows from the fact that the selection pressure on a social trait depends on what carriers and other individuals are doing. One cannot say what is the best to do for one individual, without specifying the actions of other individuals in present *and* past generations. This applies to inclusive fitness as well, and shows that it is crucial to distinguish between the gene- and the

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

individual-centered perspective of adaptation, and both cannot be assumed to be interchangeable once explicitly formalized. The gene-centered perspective is more general than the individual-centered perspective, as the effects under the control of an allele (the set of all copies of an allele) may include changes of class (demographic) and genetic contexts. The usual individual-centered characterization of Nash equilibrium in the social sciences (e.g., Luce and Raiffa, 1957; Fudenberg and Tirole, 1991; Binmore, 2007) bears similar limitations as a characterization of human (evolved) behavior.

Point (3) follows from the fact that when selection is weak, the dependence of the inclusive fitness of an allele on the frequency of this allele across generations can be simplified and the outcome of evolution can be regarded as individuals maximizing their own inclusive fitness for a given distribution of genetic-demographic contexts. The individual and the gene-centered perspective can now be interchanged. The defining feature of the individual-centered perspective is that each individual in each class is conceived as an autonomous decision-maker maximizing its own fitness measure (the "individual-as-maximizing-agent" analogy). Hence, any effect that an actor from a given class has on the individual-centered fitness of a relative in another class cannot be ascribed as an effect on that actor's own fitness. As such, the result that individuals in an age-structured population appear to be maximizing their own class-specific average direct fitness, the sum of their descendants' reproductive value including the surviving self, excludes effects on other classes (as assumed in Grafen, 2015, eq. 11), while the maximization of individual-centered inclusive fitness holds more generally (point 4).

All the individual-as-fitness maximizing-agent results derived in the paper can be connected through their relationship to allele frequency change under weak selection, under which all multi-generational effects of selection can be collapsed into a one-generation change (see Lehmann and Rousset, 2014b for a review). As the one-generation perspective is general under weak selection, the individual-as-maximizing-agent interpretation should hold under the same conditions (see also section "Scope of the individual-centered perspective of adaptation"). We finally delineate two empirical implications highlighted by our analysis.

First, in any population that is subject to density-dependent regulation and that is in an

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

uninvadable state, the average individual fitness is equal to one; thereby average inclusive fitness is necessarily equal to one. Hence, in contrast to the hypothesis considered by Bourke (2014), we do not expect a tendency for the inclusive fitness effect ("$rb - c$", the difference between the baseline fitness of "1" and inclusive fitness) to be positive even when kin selection operates through positive indirect effects ($rb > 0$). Bourke reviewed a number of studies that have attempted to test kin selection by quantifying the inclusive fitness effect, and he documented only a weak tendency for a positive bias. This meta-analysis was framed as a test of Hamilton's rule, and it could then be seen as providing little support for kin selection theory, but the inclusive fitness effect, $rb - c$, should be negative for any mutant trait in a population at an evolutionary equilibrium, since by definition it is the effect of a mutation away from an uninvadable population state on an invasion fitness maximized at this state. Hence, the results of Bourke's (2014) meta-analysis could actually be seen as evidence that populations are generally close to some evolutionary equilibrium, where the inclusive fitness effect vanishes.

By contrast to the net inclusive fitness effect $rb - c$, the indirect fitness effect, rb will be non-zero when kin selection operates at an evolutionary equilibrium. In testing kin selection theory, a difference should thus be made between attempting to measure the inclusive fitness of an allele, which in itself is not informative about the importance of kin selection, and the indirect fitness effect, which quantifies how the force of selection on a mutant depends on relatedness. Nevertheless, the inclusive fitness of a particular class of individuals (eq. 8) is itself informative about the importance of kin selection, since it assigns fitness contributions even to individuals that do not reproduce.

The second and more significant implication of our results is to support the common conception in behavioral ecology and evolutionary psychology, of adaptation as the result of interacting individuals maximizing their own inclusive fitness (e.g., Alexander, 1990; Alcock, 2005; Buss, 2005; Grafen, 2007, 2008; Davies et al., 2012; West and Gardner, 2013; Crespi, 2014). Insofar as evolutionists think about adaptation in terms of individuals maximizing their (inclusive) fitness, they should, however, keep in mind the underlying conditions (points (2-3) above), and the definition of inclusive fitness for which it holds, namely a function of the traits of an individual and

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

of its social partners, yet which coincides with invasion fitness at an evolutionary equilibrium. Previous work has been able to justify that individuals appear to be to maximize their inclusive fitness only for behaviors that involve additive effects on fitness, i.e., the realm of optimization (Grafen, 2006a)[1]. Our analysis has formally established this conception, as a weak-selection approximation, for social behavior of arbitrary complexity and for class-structured populations.

While Hamilton (1996, p.27–28) has emphasized the importance of weak selection for the evolutionary process, weak-selection approximations are still sometimes vilified in evolutionary biology (as reviewed by Birch, 2017). The value of approximations, however, can only be assessed by their impact on a field. Humans were landed on the Moon using Newtonian mechanics (Wakker, 2015)–a first-order approximation to the real (relativistic) mechanics of the solar system (Okun, 2012). The more recent observation of gravitational waves has also crucially relied on various approximations of the relativistic two-body dynamics (Damour, 2016). Thus, technological and scientific achievements regarded as paradigmatic are as dependent on approximations as is the individual-centered version of inclusive fitness. A number of unique predictions about social behavior have been made by focusing on individual inclusive fitness-maximizing behavior, from conflicts over sex-ratios and resources within families to inbreeding tolerance and genomic imprinting (e.g., Trivers and Hare, 1976; Haig, 1997a; Alcock, 2005; Macke et al., 2011; Davies et al., 2012; Szulkin et al., 2013). Our analysis justifies formally this long-heralded view.

Acknowledgments

We thank R. Bshary for having initially stimulated the writing of this paper and T. Priklopil for useful discussions. We also thank M. Chapuisat, N. Raihani, and D. Bolnick for useful comments on previous version of the manuscript, and S. Lion, A. Grafen, and an anonymous reviewer, for extensive and very useful comments.

[1]Strictly speaking, Grafen (2006a) did not formally prove any form of inclusive fitness maximization, since as discussed in Lehmann and Rousset (2014a), he considered selection on a mutant allele over a single generation under the assumption that there is a single copy of this allele in the population, but his results about individual-centered inclusive fitness maximization can be shown to hold by considering multigenerational effects and are implied by the present analysis.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Notations	**Meaning and references to the Appendix**
$x_i, \boldsymbol{x}_{-i}, \bar{x}$	Respectively, trait of individual i, its group neighbor's trait profile, and the average trait over all individuals in the population.
x, y	Trait of a mutant and a resident individual in a haploid population. In a diploid population, y is the trait of a homozygote resident, x the trait of heterozygote, and $z(x,y)$ that of homozygote mutant (eq. A.14).
$w(x_i, \boldsymbol{x}_{-i}, \bar{x})$	Individual fitness in the absence of class structure. This is the expected number of surviving offspring per haplogenome produced by an individual (possibly including self) over one demographic time period.
$w_{us}(x_i, \boldsymbol{x}_{-i}, \bar{x})$	Individual fitness through class-u offspring of a class-s parent per haplogenome (eq. A.2).
$\tilde{w}_{us}(x_i, \boldsymbol{x}_{-i}, \bar{x})$	Weak-selection approximation of class-specific individual fitness.
$v_s(y)$	Neutral reproductive value of single gene copy in class s (eq. A.18).
$W(x,y)$	Invasion fitness of a mutant gene copy (eqs. A.6,A.15, A.23).
$w_{\mathrm{T},a}(x,y)$	Reproductive value-weighted individual fitness of a class-a mutant gene copy (eq. A.24).
$V(x,y)$	Average of the $v_u(y)$ reproductive values (eq. A.20).
$W_{\mathrm{IF}}(x,y)$	Inclusive fitness of a mutant gene copy (eq. B.28), which is related to invasion fitness as $W_{\mathrm{IF}}(x,y) = 1 + V(x,y)\,(W(x,y) - 1)$, whereby $W_{\mathrm{IF}}(x,y) = W(x,y)$ if $V(x,y) = 1$.
$\Delta w_{\mathrm{IF},a}(x,y)$	Inclusive fitness effect of a class-a mutant gene copy (eq. B.29).
$w_{\mathrm{I},a}(x, \boldsymbol{x}_{-i}, \bar{x})$	fitness as-if of a class a individual (eq. C.10).
$q_k(x,y)$	Conditional probability of identity in the absence of class structure (eq. A.7).
$\boldsymbol{q}(x,y)$	Conditional distribution of identity.
$\phi_s(x,y)$	Probability that a mutant gene copy resides in a class s individual (eq. A.21).
$\boldsymbol{\phi}(x,y)$	Distribution for $\phi_s(x,y)$.
$r_{s\|a}(x,y)$	Conditional relatedness, with a class-s individual, of a mutant gene copy taken in a class-a individual (eq. B.18). Under weak selection this is written $r_{s\|a}(y)$; namely, as a function of only the resident population.
$r_{\mathrm{n},s\|a}(x,y)$	Conditional relatedness with a class-s individual, of a mutant gene copy taken in a class-a individual (eq. B.18).
$r_{\mathrm{f},a}(x,y)$	Conditional relatedness with itself of a mutant gene copy taken in a class-a individual (eq. B.18).
$c_{ua}(x,y)$	Average effect on its own fitness through class-u offspring of a gene substitution in a class-a individual (eq. B.10).
$b_{us\leftarrow a}(x,y)$	Average effect of a gene substitution in a single class-a individual on the fitness of all class-s recipients (eq. B.27) through class-u offspring.
$c_{\mathrm{I},ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$	Average effect of a class-a individual on its own fitness as-if (eq. C.23) through class-u offspring.
$b_{\mathrm{I},us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$	Average effect of a single class-s individual on the fitness as-if of all class-a recipients (eq. C.23) through class-u offspring.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Box 1: individual fitness example for haploid case without class structure. As a concrete example of an individual fitness $w(x_i, x_{-i}, \bar{x})$ in eq. (4), let us assume that x_i represents the dispersal probability of an individual, ($\mathcal{X} \in [0,1]$) and that exactly one individual in a group dies per demographic time step (Moran process, e.g., Ewens, 2004) in a group of size $N = 2$. Then, because we have only two individuals, we can set $\boldsymbol{x}_{-i} = x_{-i} \in \mathcal{X}$ and

$$w(x_i, x_{-i}, \bar{x}) = \frac{1}{2} + \frac{1}{2}\left[\frac{1 - x_i}{((1 - x_i) + (1 - x_{-i}))/2 + s\bar{x}} + \frac{x_i s}{(1 - \bar{x}) + s\bar{x}}\right], \tag{B.1}$$

where s is the survival probability during dispersal. In this expression, $1/2$ is the probability that an individual survives to the next demographic time period, whereby there is fraction $1/2$ of open breeding spots in each group. Locally, in the focal group, a relative number $1 - x_i$ of offspring of the focal individual compete for that breeding spot against other offspring produced locally [relative number $((1 - x_i) + (1 - x_{-i}))/2$] and a relative number $s\bar{x}$ of immigrants. In other groups, a relative number $x_i s$ of offspring of the focal compete against local and immigrants, in total relative number $(1 - \bar{x}) + s\bar{x}$. Now consider a population where residents do not disperse. This might, at first glance, be thought to be an equilibrium trait since dispersal reduces viability. Actually, it is well-known that some level of dispersal is selected as soon as $s > 0$, as successful emigrants avoid kin competition (Hamilton and May, 1977; Frank, 1998). To assess this conclusion, we need to evaluate the invasion fitness of a mutant allele which affects the dispersal probability. Substituting eq. B.1 into eq. (4) it then found that there is a unique uninvadable trait:

$$x^* = \frac{1}{1 + 2(1 - s)} \tag{B.2}$$

(Mullon et al., 2016, p. 188 and further details therein).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Box 2: individual fitness for social insect example. As a concrete example of the individual fitnesses in eq. (4), making the (evolutionary) role of workers more explicit, let us assume that each female produces exactly one worker, which increases colony productivity according to its trait x_{o}. We also consider that the female trait x_{f} determines the sex-ratio, and nothing else. Finally, we consider that the male trait does not affect any fitness component. Since the worker is heterozygote with probability $1/2$ and homozygote for the resident with probability $1/2$, the expected number of (reproductive) daughters of a mutant female can be written as

$$w_{\mathrm{ff}}(x,y) = \frac{(1+P(x_{\mathrm{o}}))\,x_{\mathrm{f}}}{2\,(1+P(y_{\mathrm{o}}))\,y_{\mathrm{f}}} \times \frac{1}{2} + \frac{(1+P(y_{\mathrm{o}}))\,x_{\mathrm{f}}}{2\,(1+P(y_{\mathrm{o}}))\,y_{\mathrm{f}}} \times \frac{1}{2}. \tag{B.3}$$

Here, x_{f} is the proportion of offspring that become female. The worker affects the relative fecundity of a female, which is assumed to be given by $1+P(\cdot)$, where $P(\cdot)$ is some function of worker trait. In other words, the worker trait increases offspring production of the queen relative to some baseline. The first term in eq. (B.3) is for the case where the worker is heterozygote and the second when it is homozygote resident. The denominators in eq. (B.3) reflects female production by other colonies that are monomorphic for the resident allele and the 2 reflects the fact that we measure fitness per haplogenome. Likewise, the number of sons produced by a female is

$$w_{\mathrm{mf}}(x,y) = \frac{(1+P(x_{\mathrm{o}}))\,(1-x_{\mathrm{f}})}{2\,(1+P(y_{\mathrm{o}}))\,(1-y_{\mathrm{f}})} \times \frac{1}{2} + \frac{(1+P(y_{\mathrm{o}}))\,(1-x_{\mathrm{f}})}{2\,(1+P(y_{\mathrm{o}}))\,(1-y_{\mathrm{f}})} \times \frac{1}{2}. \tag{B.4}$$

While male trait does not impact fitness, the mutant allele may still occur in a male and the mutant male fitness components will depend on the worker trait, which affects offspring production by the male's mate(s), whereby

$$w_{\mathrm{fm}}(x,y) = \frac{(1+P(x_{\mathrm{o}}))}{2\,(1+P(y_{\mathrm{o}}))} \times \frac{1}{2} + \frac{1}{4} \quad \text{and} \quad w_{\mathrm{mm}}(x,y) = \frac{(1+P(x_{\mathrm{o}}))}{2\,(1+P(y_{\mathrm{o}}))} \times \frac{1}{2} + \frac{1}{4}. \tag{B.5}$$

For this model, the inclusive fitness effects of a female, male, and worker carrying the mutant in a population at the equilibrium sex-ratio of $x_{\mathrm{f}}^{*} = 1/2$ are, respectively,

$$\begin{aligned} \Delta w_{\mathrm{IF,f}}(x,y) &= 0 \\ \Delta w_{\mathrm{IF,m}}(x,y) &= 0 \\ \Delta w_{\mathrm{IF,o}}(x,y) &= v_{\mathrm{f}}(y)\left(\frac{P(x_{\mathrm{o}})-P(y_{\mathrm{o}})}{1+P(y_{\mathrm{o}})}\right) + v_{\mathrm{m}}(y)\left(\frac{P(x_{\mathrm{o}})-P(y_{\mathrm{o}})}{1+P(y_{\mathrm{o}})}\right) \end{aligned} \tag{B.6}$$

(see Appendix B.3 for a proof).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Box 3: weak selection concepts. As concrete example of both "small-mutation" and "small-parameter" weak selection, we can use the social-insects scenario and corresponding fitnesses given in Box 2. Then, we can first Taylor-expand the fitness components, say the number of daughters produced by queens (eq. B.3), in mutant trait around the resident trait and neglect higher-order terms to obtain

$$w_{\mathrm{ff}}(x,y) \sim \underbrace{\frac{1}{2} + \frac{\partial w_{\mathrm{ff}}(x,y)}{\partial x_{\mathrm{o}}}\bigg|_{x=y}(x_{\mathrm{o}} - y_{\mathrm{o}}) + \frac{\partial w_{\mathrm{ff}}(x,y)}{\partial x_{\mathrm{f}}}\bigg|_{x=y}(x_{\mathrm{f}} - y_{\mathrm{f}})}_{\tilde{w}_{\mathrm{ff}}(x,y)}, \tag{B.7}$$

where the right-hand side gives a small-mutation approximation to fitness as the fitness of a resident individual in a monomorphic resident population plus the marginal changes in fitness weighted by their phenotypic differences. Alternatively, we can linearize fitness in terms of the effect $P(x_{\mathrm{o}})$ of workers on female fecundity

$$w_{\mathrm{ff}}(x,y) \sim \underbrace{\frac{x_{\mathrm{f}}}{2y_{\mathrm{f}}} + \frac{x_{\mathrm{f}}\Big(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}})\Big)}{4y_{\mathrm{f}}}}_{\tilde{w}_{\mathrm{ff}}(x,y)}, \tag{B.8}$$

where the right-hand side represents small-parameter approximation to fitness.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Appendix

Contents

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Appendix A: Invasion fitness

In this Appendix, we derive the expressions for invasion fitness given in the main text, eq. (4)–(5), and their generalizations. The novel results in this Appendix not appearing previously in the literature are eq. (A.15) and eqs. (A.23)–(A.27) for the average direct fitness of an allele. As explained in the main text, we limit or discussion to a population that is divided into an infinite number of groups that are all of constant size N and connected by random dispersal with reproduction occurring in discrete time periods (i.e., Wright's 1931 canonical island model of dispersal). Since even under this assumption notations and concepts becomes rapidly complicated in the presence of class-structure and diploidy, we will progressively introduce the different cases, concepts, and notations. We start by defining the central building block of our analysis, which is individual fitness.

A.1 Building blocks

A.1.1 Individual fitness

Denote by $\mathcal{X}$ the set of feasible traits (or phenotypes) that can be expressed by an individual.[2] In the absence of class-structure, we define the *individual fitness function* as

$$w : \mathcal{X} \times \mathcal{X}^{N-1} \times \mathcal{X} \to \mathbb{R}_+, \tag{A.1}$$

such that $w(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is the expected number of successful offspring produced (*per haplogenome*) by an individual i with trait $x_i \in \mathcal{X}$ in a group where neighbors have trait profile $\boldsymbol{x}_{-i} \in \mathcal{X}^{N-1}$ in a population where the average trait over all individuals is $\bar{x} \in \mathcal{X}$. The expectation is over all within-generation stochastic effects on settled offspring number in the descendant generation and conditional on realized trait profile $(x_i, \boldsymbol{x}_{-i}, \bar{x})$ in the parental generation.

In the presence of class-structure, we assume that there is a finite number of classes within

[2] We assume that $\mathcal{X}$ is a locally convex Hausdorff space; namely, it is a nonempty, compact, and convex set in a topological vector space (Alipantris and Border, 2006, p. 55). We are not aware of any applications in evolutionary biology that is not covered by this case (e.g., it covers discrete finite trait sets, infinite-dimensional reaction norms (or function value traits) taking values in the reals, combination of these two, etc.), and is the space for which general results concerning function maximization exists (Alipantris and Border, 2006, pp. 581-585).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

each group, and use the following notations (see also section "Demographic assumptions" of the main text): n_a denotes the number of individuals in class a, $\mathcal{C}$ denotes the set of classes (e.g., workers and queens, males and females; $N = \sum_{a\in\mathcal{C}} n_a$), and $\mathcal{X}_a$ denotes the feasible trait set of an individual of class $a \in \mathcal{C}$ with the total trait set being $\mathcal{X} = \prod_{a\in\mathcal{C}} \mathcal{X}_a$ (products of sets are taken as Cartesian products throughout). With this, we define the individual fitness function

$$w_{us} : \mathcal{X} \times \prod_{a\in\mathcal{C}} \mathcal{X}_a^{n_a-\delta_{as}} \times \mathcal{X} \to \mathbb{R}_+ \quad \forall (u,s) \in \mathcal{C}^2, \tag{A.2}$$

such that $w_{us}(x_i, \boldsymbol{x}_{-i}, \bar{\boldsymbol{x}})$ is the expected number of successful class-u offspring produced over a demographic time step by a class-s individual (*per haplogenome*) that has trait $x_i \in \mathcal{X}$ in a group where neighbors have trait profile $\boldsymbol{x}_{-i} \in \prod_{a\in\mathcal{C}} \mathcal{X}_a^{n_a-\delta_{as}}$ (which has dimension $n_1 n_2 \cdots n_{s-1}(n_s - 1)n_{s+1} \cdots n_{|\mathcal{C}|}$ owing to the fact that δ_{as} is the Kronecker delta) in a population where the vector of average traits is $\bar{\boldsymbol{x}} \in \mathcal{X}$.

A.1.2 Distinct and indistinct individuals

The formulation of the fitness functions (eqs. A.1–A.2) allows for a characterization of the population where each individual in a group can be distinguished from each other. This means that the trait profile $(x_i, \boldsymbol{x}_{-i})$ in a focal group, i.e., its state, belongs to the set $\mathcal{X}^N$ of all ordered trait profiles (i.e., all ordered groups states are considered). In an evolutionary invasion analysis, however, we consider that only two alleles –mutant and resident– segregate in the population and so there can be a maximum number of only two types of individuals in each class in a haploid population (or three types in diploids: one heterozygote and the two homozygotes). Hence, we have group states with N individuals, where each member belongs only to one among a finite number of genotypic types. This allows for an alternative, simpler, characterization of the population, where one just counts the number of individuals bearing identical traits in a group and thus individuals are no longer distinguished (i.e., only unordered groups states are considered).

To illustrate these concepts, consider a haploid population without class structure with individuals either expressing a mutant trait $x \in \mathcal{X}$ or expressing a resident trait $y \in \mathcal{X}$. Since each individual in a group is either mutant or resident, there is a total number of 2^N ordered groups

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

states. But to evaluate the fitness of an individual, one typically does not distinguish all these states, and could simply count the number of individuals carrying the mutant allele and write the individual fitness of an individual i with mutant trait $x_i = x$ when $k-1$ of its neighbors express the mutant x as

$$w(x_i, \boldsymbol{x}_{-i}, \bar{x}) = w(x, \boldsymbol{x}_k, \bar{x}) \qquad \forall \boldsymbol{x}_{-i} \in \mathcal{S}_k. \tag{A.3}$$

Here, $\boldsymbol{x}_k$ is a vector of dimension $N-1$ with $k-1$ entries equal to x and $N-k$ entries equal to y and $\mathcal{S}_k$ is the set of all subsets of the set $\mathcal{S} = \{x, y\}^{N-1}$ of ordered neighbor trait profiles such that exactly $k-1$ individuals have trait x and $N-k$ individuals having trait y (note that $\mathcal{S} = \cup_{k=1}^{N} \mathcal{S}_k$). The number of such profiles that characterize group states that are indistinct is given by the Binomial coefficient

$$\mathcal{B}(N,k) = \binom{N-1}{k-1}, \tag{A.4}$$

where the sum over all indistinct cases gives the total number of trait profiles that could be distinguished ($\sum_{k=1}^{N} \mathcal{B}(N,k) = 2^{N-1}$). In eq. (A.3), individual fitness is thus permutation-invariant on the trait profile of its neighbors. In the class-structured case, permutation-invariance is on the trait profile of neighbors belonging to the same class. If this permutation-invariance (or symmetry) did not hold, then individuals would belong to different classes, hence permutation-invariance is not an assumption.

These considerations show that one can characterize a group state in a class structured population from the perspective of an individual i either by distinguishing all individuals (ordered group states) or by not distinguishing individuals in identical states (unordered group states). While in evolutionary analysis individuals are usually not distinguished because this is often mathematically simpler (an exception being the Price equation, Price, 1970; Frank, 1998), distinguishing them is fundamental to the individual-centered perspective of adaptation. As such, we develop the invasion fitness by distinguishing individuals when this will be needed for the analysis of the indivdual-centered perspective, but start by not distinguishing individuals to

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

frame the model into the classical approach and to introduce concepts in a progressive way.

A.2 Invasion fitness without classes

A.2.1 Haploids

Indistinct individuals. In the absence of within-group class structure (homogeneous individuals), a mutant allele with trait $x \in \mathcal{X}$ introduced as a single copy in a resident haploid population otherwise monomorphic for a resident allele with trait $y \in \mathcal{X}$ goes extinct with probability one if

$$W(x,y) \leq 1, \tag{A.5}$$

where

$$W(x,y) = \sum_{k=1}^{N} w(x,\boldsymbol{x}_k,y) q_k(x,y) \tag{A.6}$$

is the invasion fitness of the mutant x in a resident y population (formally, the invasion fitness function is $W : \mathcal{X}^2 \to \mathbb{R}_+$ with $W(y,y) = 1$ for all $y \in \mathcal{X}$). Here, the individual fitness $w(x,\boldsymbol{x}_k,y)$ is given by eq. (A.3) and $q_k(x,y)$ is the probability that a randomly sampled mutant individual from the mutant lineage descending from the initial mutant resides in a group with k mutants ($\sum_{k=1}^{N} q_k(x,y) = 1$). When $N = 2$, invasion fitness (eq. A.6) reduces to eq. (4) of the main text.

The $q_k(x,y)$ probability is evaluated under the assumption that the mutant is overall rare in the population, and that the growth of the mutant lineage descending from a single initial copy has reached stationarity. That is,

$$q_k(x,y) = \frac{k u_k(x,y)}{\sum_{i=1}^{N} i u_i(x,y)}, \tag{A.7}$$

where $\boldsymbol{u} = (u_1, u_2, ..., u_N)$ is the right eigenvector associated to the leading eigenvalue $W(x,y)$ of the matrix $\boldsymbol{A}(x,y)$ describing the growth of the mutant when it is overall rare in the population

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

(multitype branching process):

$$\boldsymbol{A}(x,y)\boldsymbol{u}(x,y) = W(x,y)\boldsymbol{u}(x,y). \tag{A.8}$$

The ijth entry of $\boldsymbol{A}(x,y)$ gives the expected number of groups with $i > 0$ mutants descending over one time step from a group with $j > 0$ mutant, and $u_i(x,y)$ is the stationary probability that there are i mutants in a group, conditional on there being at least one mutant (see Lehmann et al., 2016 for a proof of eq. A.6 and a more detailed characterization of the reproductive process underlying mutant dynamics and derivation of $q_k(x,y)$).

Distinct individuals. We now make the link to characterizing invasion fitness by considering all ordered groups states (as this will be useful in the individual-centered perspective). To obtain this representation, we note that from eq. (A.3), we can write

$$w(x,\boldsymbol{x}_k,y) = \frac{1}{\mathcal{B}(N,k)} \sum_{\boldsymbol{x}_{-i}\in\mathcal{S}_k} w(x,\boldsymbol{x}_{-i},y), \tag{A.9}$$

where on the right-hand side we have distinguished all trait profiles in the focal group. Let us now further define

$$q_k^{\mathrm{D}}(x,y) = \frac{q_k(x,y)}{\mathcal{B}(N,k)}, \tag{A.10}$$

which is the probability that, conditional on an individual carrying the mutant allele, an ordered neighbor trait profile $\boldsymbol{x}_{-i} \in \mathcal{S}$ contains exactly $k-1$ individuals also carrying the mutant (hence $\sum_{k=1}^{N}\sum_{\boldsymbol{x}_{-i}\in\mathcal{S}_k} q_k^{\mathrm{D}}(x,y) = 1$). On substituting eqs. (A.9)-(A.10) into eq. (A.6), we can write the invasion fitness of mutant allele with trait x introduced into a haploid resident population with trait y as

$$W(x,y) = \sum_{k=1}^{N}\sum_{\boldsymbol{x}_{-i}\in\mathcal{S}_k} w(x,\boldsymbol{x}_{-i},y)q_k^{\mathrm{D}}(x,y) \quad \forall (x,y)\in\mathcal{X}^2, \tag{A.11}$$

which is the average fitness over all ordered trait profiles in a group.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Writing explicitly the sums appearing in eq. (A.11) and detailing the permutation under the more general diploid and class-structured model will be cumbersome and we now present an alternative and more compact representation of invasion fitness. To that end, let us collect all $q_k^{\mathrm{D}}(x,y)$ probabilities into the vector $\boldsymbol{q}^{\mathrm{D}}(x,y)$, which is the distribution of ordered group states experienced by an individual with trait x and that has support[3] in $\mathcal{S}$. With this, we can write invasion fitness as

$$W(x,y) = \mathrm{E}_{\boldsymbol{x}_{-i}\sim\boldsymbol{q}^{\mathrm{D}}(x,y)}[w(x,\boldsymbol{x}_{-i},y)]\,, \tag{A.12}$$

where the notation $\sim$ specifies that variable $\boldsymbol{x}_{-i}$ follows distribution $\boldsymbol{q}^{\mathrm{D}}(x,y)$.

A.2.2 Diploids

When individuals are diploid, we need to take into account that they can be homozygote for the mutant allele. To do this, it will be convenient to build on our notations for mutant and resident traits introduced for a haploid population. For a diploid population, we let $y \in \mathcal{X}$ be the trait of an individual that is homozygote for the resident allele and $x \in \mathcal{X}$ be the trait of an individual that is heterozygote for the mutant allele. We then denote by $z \in \mathcal{X}$ the trait of a homozygote mutant and assume that the trait of an heterozygote is obtained as the following convex combination of the trait of the two homozygotes:

$$x = \alpha y + (1-\alpha)z \tag{A.13}$$

for the scalar $\alpha \in (0,1)$. Hence, we rule out over-, under-, and strict dominance, but otherwise allow for arbitrary gene action. Eq. (A.13) guarantees that for all $y \in \mathcal{X}$ and $z \in \mathcal{X}$, we have $x \in \mathcal{X}$ and it allows us to express conveniently the trait of a homozygote mutant as a function $z : \mathcal{X}^2 \to \mathcal{X}$ of heterozygote and resident homozygote traits, where

$$z(x,y) = y + \frac{x-y}{1-\alpha}. \tag{A.14}$$

[3]The support of a distribution is the set of possible values of a random variable having that distribution.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

For arbitrary group size N, the invasion fitness of a mutant allele with heterozygote trait x introduced into a resident diploid population with homozygote trait y can be written as

$$W(x,y) = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim \boldsymbol{q}^{\mathrm{D}}(x,y)}[w(x_i,\boldsymbol{x}_{-i},y)]\,, \tag{A.15}$$

where $x_i \in \{z(x,y),x\}$. Each component x_j of the neighbor trait profile $\boldsymbol{x}_{-i} = (x_1,,...,x_{i-1},x_{i+1},...x_N)$ takes values in the set $\{z(x,y),x,y\}$. The expectation in eq. (A.15) is over the distribution $\boldsymbol{q}^{\mathrm{D}}(x,y)$, conditional on an individual carrying at least one copy of the mutant allele, of all possible group ordered profiles of strategies with support in $\mathcal{S} = \{z(x,y),x\}\times\{z(x,y),x,y\}^{N-1}$.

Eq. (A.15) shows that invasion fitness can, as in the haploid case, be expressed as an average of fitness components $w(\cdot,\cdot,\cdot)$ over a distribution $\boldsymbol{q}^{\mathrm{D}}(x,y)$, but which is generally more involved than in the haploid case. For instance, for $N=2$, we have

$$\begin{aligned} W(x,y) &= w(x,y,y)q_{0,\mathrm{he}}(x,y) + w(x,x,y)q_{1,\mathrm{he}}(x,y) + w(x,z(x,y),y)q_{2,\mathrm{he}}(x,y) \\ &+ w(z(x,y),y,y)q_{0,\mathrm{ho}}(x,y) + w(z(x,y),x,y)q_{1,\mathrm{ho}}(x,y) + w(z(x,y),z(x,y),y)q_{2,\mathrm{ho}}(x,y), \end{aligned} \tag{A.16}$$

where $q_{j,\mathrm{he}}(x,y)$ is the probability that, conditional on an individual carrying the mutant allele, it is heterozygote and its group neighbor has j copies of the mutant allele ($j=0,1,2$, then stem, respectively, for the neighbor to be homozygote resident, heterozygote, and homozygote mutant), while $q_{j,\mathrm{ho}}(x,y)$ is the probability that, conditional on an individual carrying the mutant allele, it is an homozygote and its group neighbor has j copies of the mutant allele. For this case, the distribution over group configurations is given by

$$\boldsymbol{q}^{\mathrm{D}}(x,y) = (q_{0,\mathrm{he}}(x,y),q_{1,\mathrm{he}}(x,y),q_{2,\mathrm{he}}(x,y),q_{0,\mathrm{ho}}(x,y),q_{1,\mathrm{ho}}(x,y),q_{2,\mathrm{ho}}(x,y)), \tag{A.17}$$

whose elements sum up to one and could be expressed in terms of probabilities of identity in state of alleles in pairs of individuals (Michod, 1982, Fig. 1).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

A.3 Invasion fitness with classes

A.3.1 Haploids

In the presence of classes, the trait x of the mutant in a haploid population is taken as a vector of actions (or stream of actions), one for each class the individual may belong to, so we write $x = (x_1, x_2, ..., x_{|\mathcal{C}|}) \in \mathcal{X}$, where x_a is the trait of a mutant individual when of class a, that is assumed to be under its own control. Likewise, we have $y = (y_1, y_2, ..., y_{|\mathcal{C}|}) \in \mathcal{X}$. Using eq. (A.2), we let $w_{us}(x, \boldsymbol{x}_{\mathbf{k}}, y)$ be the expected number of class-u offspring produced by a class-s mutant when in a group in state $\mathbf{k} = (k_1, ..., k_{|\mathcal{C}|})$, which is the vector of the number of individuals carrying the mutant allele in each class, with k_a being the number of mutants in class a, whereby $\boldsymbol{x}_{\mathbf{k}}$ is a vector that has $(k_s - 1)$ entries with trait x_s, k_a entries with trait x_a for each $a \neq s$, while all remaining entries are for the corresponding element of the resident trait vector y.

A central quantity in our analysis is the reproductive value $v_s(y)$ of a single gene copy residing in an individual of class s in a monomorphic resident population (neutral reproductive value), which satisfies

$$v_s(y) = \sum_{u \in \mathcal{C}} v_u(y) w_{us}(y, \boldsymbol{x}_{\mathbf{0}}, y) \tag{A.18}$$

(e.g., Taylor, 1990; Frank, 1998; Rousset, 2004; Grafen, 2006b; Lehmann et al., 2016). With these definitions, the invasion fitness of a mutant allele with trait x in a resident population with trait y can be written as a sum over, respectively, possible group states, offspring classes, and parent classes:

$$W(x, y) = \frac{1}{V(x, y)} \sum_{\mathbf{k} \in I} \sum_{u \in \mathcal{C}} \sum_{s \in \mathcal{C}} v_u(y) w_{us}(x, \boldsymbol{x}_{\mathbf{k}}, y) q_{\mathbf{k},s}(x, y), \tag{A.19}$$

where $q_{\mathbf{k},s}(x, y)$ is the probability that a randomly sampled member of the mutant lineage finds itself in class s and in a group in state $\mathbf{k}$; $I = (I_1 \times \cdots \times I_{n_c}) \setminus \mathbf{0}$ is the set of possible group

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

states with $I_u = \{0, 1, ..., n_u\}$ being the set of the number of mutant alleles in class u; and

$$V(x,y) = \sum_{s\in\mathcal{C}} v_s(y)\phi_s(x,y), \tag{A.20}$$

where

$$\phi_s(x,y) = \sum_{\mathbf{k}\in I} q_{\mathbf{k},s}(x,y) \tag{A.21}$$

is the probability that a randomly sampled gene copy from the mutant lineage is a class-s individual. Hence, $V(x,y)$ is the total (neutral) reproductive value of a randomly sampled mutant from its lineage. Owing to eq. (A.18), $V(x,y)$ can be seen as the average reproductive value of a mutant that would have its fitness components assigned those of a resident individual (instead of expressing mutant fitness components, the $w_{us}(x,\boldsymbol{x}_{\mathbf{k}},y)$'s, it expresses resident fitness components, the $w_{us}(y,\boldsymbol{x}_{\mathbf{0}},y)$'s).

In eq. (A.19) we have not distinguished identical individuals within classes and for such a class-structured population invasion fitness $W(x,y)$ still satisfies eq. (A.8), but matrix $\boldsymbol{A}(x,y)$ describing the growth of the mutant lineage when rare in the population has now elements giving the expected number of mutant copies in context $\mathbf{i}$ that descend from a mutant copy in context $\mathbf{j}$, and the $q_{\mathbf{k},s}(x,y)$ distribution is then expressed in terms of the leading right eigenvector of this matrix; namely,

$$q_{\mathbf{k},s}(x,y) = \frac{k_s u_{\mathbf{k}}(x,y)}{\sum_{\mathbf{k}\in I}\sum_{s\in\mathcal{C}} k_s u_{\mathbf{k}}(x,y)} \tag{A.22}$$

(see Lehmann et al., 2016, Appendix F for more details and a proof of eq. A.19).

A.3.2 Invasion fitness in terms of average direct fitness

It will be useful to write eq. (A.19) as

$$W(x,y) = \frac{1}{V(x,y)} \sum_{s\in\mathcal{C}} w_{\mathrm{T},s}(x,y)\phi_s(x,y), \tag{A.23}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

where

$$w_{\mathrm{T},s}(x,y) = \sum_{u\in\mathcal{C}} \sum_{\mathbf{k}\in I} v_u(y) w_{us}(x,\boldsymbol{x}_{\mathbf{k}},y) q_{\mathbf{k}|s}(x,y), \tag{A.24}$$

is the expected reproductive value-weighted fitness of a class s individual and

$$q_{\mathbf{k}|s}(x,y) = \frac{q_{\mathbf{k},s}(x,y)}{\phi_s(x,y)} \tag{A.25}$$

is the probability that, conditional on an individual being mutant and of class s, the individual resides in a group in state $\mathbf{k}$. Eq. (A.24) is the sum of the reproductive values of the descendants of an individual of class s, including its potentially surviving self. We thus refer to $w_{\mathrm{T},s}(x,y)$ as the average direct fitness of a class s individual, and, for a panmictic population, this quantity was previously called Williams' reproductive value (Grafen, 2015, p. 8). Hence, invasion fitness eq. (A.23) is the total average direct fitness of a mutant relative to the reproductive value that individual would have if it expressed the resident trait.

However, any non-null vector of weights could have been chosen in eq. (A.19) and eq. (A.23) to compute the geometric growth rate, which is so because the right-hand side eq. (A.19) is obtained by rearranging the leading eigenvalue-eigenvector equation, where the leading eigenvector can be normalized by any non-null vector (see Lehmann et al., 2016, Appendix B and C for more details). We can in particular choose the unit vector $(1,1,...,1)$, whereby invasion fitness becomes the average of the individual fitnesses of a randomly sampled mutant from its lineage. In eq. (A.19) (and eq. (6) of the main text), we here choose reproductive-value weights for two reasons. First, average direct fitness is then expressed with the same weights as is inclusive fitness (see next section "Inclusive fitness"), given that for inclusive fitness there is no choice but to use the reproductive-value weights. Second, and more importantly, the reproductive-value weights play a pivotal role in the forthcoming weak selection analysis (section "Individual maximands under weak selection"), where they allow to obtain meaningful expressions for the different average fitnesses, a feature that follows from the well-established fact that the reproductive-value weights are also the unique weights that would allow to apply eqs. (A.23) when the mutant is no longer

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

rare to predict the direction of average allele frequency change by a scalar fitness measure at all allele frequencies under weak selection (e.g., Taylor, 1990; Rousset, 2004; Grafen, 2006b).

Finally, we note that we could normalize the reproductive values such that $V(x,y) = 1$, however this would induce the $v_u(y)$'s to become a function of the mutant, since the $\phi_s(\mu,y)$ probabilities in eq. (A.20) depend on the mutant. We would like to avoid this here, otherwise differentiation of $W(x,y)$ requires differentiating the reproductive values, and so we need a notation distinguishing the case where reproductive values depend on the mutant from the case of weak selection (investigated below), where this dependence drops out. In order to have a uniform notation throughout the main text, we normalize the $v_u(y)$'s such that the average neutral reproductive value of a randomly sampled resident individual is one:

$$V(y,y) = \sum_{s\in\mathcal{C}} v_s(y)\phi_s(y,y) = 1, \tag{A.26}$$

where $v_s(y)\phi_s(y,y)$ can be recognized as the reproductive value of class s in a monomorphic resident population (e.g., Taylor, 1990; Rousset, 2004) and so eq. A.26 is the standard normalization of the reproductive values.

A.3.3 Diploids and social insects

In order to generalize eq. (A.24) to diploidy, we let $z_a(x_a,y_a) \in \mathcal{X}_a$ be the trait of an homozygote mutant of class a when the profile of heterozygote mutant traits across classes is $x = (x_1,x_2,...,x_{|\mathcal{C}|}) \in \mathcal{X}$ and the trait profile of a homozygote resident individual is $y = (y_1,y_2,...,y_{|\mathcal{C}|}) \in \mathcal{X}$ (following eqs. A.13–A.14, we assume that for each a, $z_a(x_a,y_a)$ is obtained by assuming that heterozygotes are a convex combination of the homozygotes). With this, we denote by $z(x,y) = (z_1,z_2,...,z_{|\mathcal{C}|}) \in \mathcal{X}$ the profile of homozygote mutants. Then, the invasion fitness of a mutant allele with heterozygote (multidimensional) trait x introduced into a resident diploid population with homozygote trait y can be written as

$$w_{\mathrm{T},s}(x,y) = \sum_{u\in\mathcal{C}} v_u(y)\mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim q_s^{\mathrm{D}}(x,y)}\Big[w_{us}(x_i,\boldsymbol{x}_{-i},y)\Big]. \tag{A.27}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Here, $x_i \in \{z(x,y), x\}$ and each component x_j of the neighbor trait profile $\boldsymbol{x}_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ...x_N)$ takes values in $\{z_a(x_a, y_a), x_a, y_a\}$ if the corresponding individual j is of class a. In eq. (A.27), the couple $(x_i, \boldsymbol{x}_{-i})$ follows the distribution $\boldsymbol{q}_s^{\mathrm{D}}(x,y)$ of ordered focal group trait profiles, conditional on a individual carrying at least one copy of the mutant allele and being of class s. This distribution has support in $\mathcal{S}_s = \{z(x,y), x\} \times \prod_{a\in\mathcal{C}} \{z_a(x_a, y_a), x_a, y_a\}^{(n_a - \delta_{sa})}$, since among the neigbors of an individual of class s we have n_a individuals of class $a \neq s$ and $n_s - 1$ class-s individuals.

A special case of eq. (A.27) is when there is only a single individual per group under complete dispersal and random mating. In that case, a mutant individual can only be heterozygote (as long as the mutant is rare) and there is no dependence on group members in the fitness function w_{ua}. Then, eq. (A.27) reduces to

$$w_{\mathrm{T},s}(x,y) = \sum_{u\in\mathcal{C}} v_u(y) w_{us}(x,y). \tag{A.28}$$

As a concrete example, we work out the model of a seasonal population of social insects presented in the main text. Since we are interested in considering the three classes of individuals demographically, the census stage of fitness is taken right before dispersal (end of stage (1) of the life cycle). When the mutant allele is rare, the dynamics of the number of mutant allele copies in females, males, and workers in the population between successive census stages can be described by the matrix

$$\mathbf{A}(x,y) = \begin{bmatrix} w_{\mathrm{ff}}(x,y) & w_{\mathrm{fm}}(x,y) & 0 \\ w_{\mathrm{mf}}(x,y) & w_{\mathrm{mm}}(x,y) & 0 \\ w_{\mathrm{of}}(x,y) & w_{\mathrm{om}}(x,y) & 0 \end{bmatrix}, \tag{A.29}$$

From this matrix, the probabilities that a randomly sampled copy of the mutant allele is in a female, male, or worker, are respectively

$$\phi_{\mathrm{f}}(x,y) = \frac{w_{\mathrm{ff}}\left(w_{\mathrm{ff}} + w_{\mathrm{mm}}\right)}{X},\ \phi_{\mathrm{f}}(x,y) = \frac{w_{\mathrm{mm}}\left(w_{\mathrm{ff}} + w_{\mathrm{mm}}\right)}{X} \text{ and } \phi_{\mathrm{o}}(x,y) = \frac{w_{\mathrm{ff}} w_{\mathrm{of}}}{X}, \tag{A.30}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

where $X = (w_{\mathrm{ff}} + w_{\mathrm{mm}})^2 + w_{\mathrm{ff}} w_{\mathrm{of}}$. The reproductive values are $v_{\mathrm{o}}(y) = 0$, $v_{\mathrm{f}}(y) > 0$ and $v_{\mathrm{m}}(y) > 0$, and the invasion fitness is given by

$$W(x,y) = \frac{1}{V(x,y)} \left[w_{\mathrm{T,f}}(x,y)\phi_{\mathrm{f}}(x,y) + w_{\mathrm{T,m}}(x,y)\phi_{\mathrm{m}}(x,y) \right]. \tag{A.31}$$

The two direct fitnesses appearing in this equation are given by eq. (6) of the main text. Supposing there is only one worker in the colony (e.g., assumptions in the main text), then, in a monomorphic population, we have $\phi_{\mathrm{f}}(y,y) = \phi_{\mathrm{f}}(y,y) = \phi_{\mathrm{o}}(y,y) = 1/3$ and the reproductive values, normalized so as to satisfy eq. (A.26), are

$$v_{\mathrm{o}}(y) = 0, \quad v_{\mathrm{f}}(y) = \frac{3}{2}, \quad v_{\mathrm{m}}(y) = \frac{3}{2}. \tag{A.32}$$

Appendix B: Inclusive fitness

In this Appendix, we derive from invasion fitness the expression for inclusive fitness given in the main text (eq. 7). In so doing, we explain how inclusive fitness can accomodate one-predictor and two-predictor regression interpretations for the costs and benefits. And as we did in Appendix A, we progressively introduce the different concepts, starting with haploids, and all results pertaining to class structure and diploidy are novel (the key result being eq. B.28).

B.1 Inclusive fitness for haploids without classes

We start by deriving inclusive fitness from invasion fitness (eq. A.6) for the haploid case and without class structure, so as to present it for the simplest case. To that end, we use the relatedness coefficient defined as

$$r(x,y) = \sum_{k=1}^{N} \left(\frac{k-1}{N-1} \right) q_k(x,y), \tag{B.1}$$

which is the probability that a randomly sampled neighbor of a mutant (itself randomly sampled from its lineage when rare) also carries the mutant allele (when $N = 2$, we have $r(x,y) = q_2(x,y)$).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Inclusive fitness is a representation of the fitness of an allele (Hamilton, 1964, p. 6, i.e., number of replica copies of an allele produced by a focal copy), as a partition of this fitness in terms of direct and indirect changes in transmission of replica copies (the "cost" and "benefit" of expressing the allele), and where the indirect effects are weighted by relatedness coefficient(s). For a model with arbitrary strength of selection, a general expression for inclusive fitness has been reached for the case $N = 2$ by performing a *two-predictor* regression of the fitness of a representative individual from the population, on the mutant allele frequency it carries (zero or one for haploids) and on the frequency of the mutant in its neighbors (e.g., Queller, 1992; Frank, 1997; Gardner et al., 2011; Rousset, 2015).

Alternatively, one may perform a single-predictor regression of the individual fitness of a carrier of the mutant on the frequency of the mutant allele among its neighbors, which may be more in line with certain empirical estimates of inclusive fitness where only the social neigborhood of an individual expressing a particular behavior is varied (Krakauer, 2005; Dobson et al., 2012). A single-predictor regression was also used in Lehmann et al. (2016, Box.1) as a justification to derive an exact version of inclusive fitness for haploid class-structured populations, although a two-predictor interpretation was retained for the resulting cost and benefits. In order to confirm that such a two-predictor interpretation holds for these previous results, avoid further confusions, and delineate the differences between the partitions of fitness by single and two-predictor regressions, we will show that both interpretations apply to an arbitrary number of interacting individuals with or without class structure and in the presence and absence of diploidy.

B.1.1 Regression with respect to neighbors

For the one-predictor regression version of inclusive fitness for haploids, we aim to write the individual fitness of a mutant x in a group with trait profile $\boldsymbol{x}_k$ as

$$w(x, \boldsymbol{x}_k, y) = 1 - \gamma(x, y) + \beta(x, y) \left(\frac{k-1}{N-1} \right) + \text{residual}, \tag{B.2}$$

where $1 - \gamma(x, y)$ is the intercept of the regression, $\beta(x, y)$ is the additive effect of allele frequency in neighbors, and $(k-1)/(N-1)$ is the frequency of the mutant allele among neighbors of a

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

mutant. The "cost" (γ) and "benefit" (β) of this single predictor are determined by minimizing over the $q_k(x,y)$ distribution the expected mean-square difference between individual fitness $w(x,\boldsymbol{x}_k,y)$ and the regression. Thus, for all $(x,y)\in\mathcal{X}^2$, we minimize the sum of squares

$$Q(\gamma,\beta,x,y)=\sum_{k=1}^{N}\left[1-\gamma+\beta\left(\frac{k-1}{N-1}\right)-w(x,\boldsymbol{x}_k,y)\right]^2 q_k(x,y), \tag{B.3}$$

with respect to γ and β, which are practically obtained by setting $\partial Q(\gamma,\beta,x,y)/\partial\gamma$=0 and $\partial Q(\gamma,\beta,x,y)/\partial\beta$=0, and solving for γ and β, which are thus obtained as functions of x and y (i.e., $\gamma=\gamma(x,y)$ and $\beta=\beta(x,y)$). It follows directly by averaging the regression over the $q_k(x,y)$ distribution, that we can write invasion fitness in terms of the so-obtained coefficient as

$$W(x,y)=1-\gamma(x,y)+r(x,y)\beta(x,y) \tag{B.4}$$

for relatedness defined in eq. (B.1).

B.1.2 Regression with respect to focal and neighbors

For the two-predictor regression version of inclusive fitness, the additional predictor variable for the fitness of an individual is its own allelic type. To take this into account in a least-squares regression framework, we need to consider a population where the average mutant frequency is no longer rare. We denote p this frequency, and by a slight abuse of notation, we denote $w(x,\boldsymbol{x}_k,p)$ the individual fitness of a mutant in a group with a total number k of mutant neighbors, in a population where the mutant frequency is p. More generally, whenever we will consider fitness at all mutant frequencies, we will replace the last argument of the fitness function with the mutant frequency in the population). Fitness $w(y,\boldsymbol{x}_{k+1},p)$ likewise stands for the fitness of an individual carrying the resident allele in the same context of a group including k mutants (hence $\boldsymbol{x}_{k+1}$ is any vector of dimension $N-1$ with k entries equal to x and $N-k$ entries equal to y). The sum of squares characterizing the regression of the expected number of offspring of a mutant x with

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

frequency p in a resident y population is:

$$Q(c,b,x,y,p) = \left(\sum_{k=1}^{N}\left[1-c+b\left(\frac{k-1}{N-1}\right)-w(x,\boldsymbol{x}_k,p)\right]^2 q_k(x,y,p)\right)p + \left(\sum_{k=0}^{N-1}\left[1+b\frac{k}{N-1}-w(y,\boldsymbol{x}_{k+1},p)\right]^2 \tilde{q}_k(x,y,p)\right)(1-p), \quad \text{(B.5)}$$

where c and b are regression coefficients, $q_k(x,y,p)$ is the probability that, given an individual is a mutant with trait x in a population where the frequency of mutants is p and residents play trait y, it will reside in a group where there are k mutants. Likewise, $\tilde{q}_k(x,y,p)$ is the probability that, given an individual is a resident with trait y in a population where the frequency of mutants with trait x is p, it will reside in a group with k mutants. Minimizing the quadratic form $Q(c,b,x,y,p)$ (by solving $\partial Q(c,b,x,y,p)/\partial c = 0$ and $\partial Q(c,b,x,y,p)/\partial b = 0$) we then obtain the regression coefficients $c = c(x,y,p)$ and $b = b(x,y,p)$, which depend on the population state.

When the mutant is rare ($p \to 0$), the fitness of a mutant is $w(x,\boldsymbol{x}_k,p) \to w(x,\boldsymbol{x}_k,y)$ (same as in eq. A.6) and the regression thus predicts this fitness as

$$w(x,\boldsymbol{x}_k,y) = 1 - c(x,y) + b(x,y)\left(\frac{k-1}{N-1}\right) + \text{residual}, \quad \text{(B.6)}$$

where the residual and the cost and benefit will depend on mutant trait, resident trait, and mutant frequency, but are evaluated as $p \to 0$; i.e., $c(x,y) = \lim_{p\to 0} c(x,y,p)$ and $b(x,y) = \lim_{p\to 0} b(x,y,p)$. When the mutant is rare we also have that $q_k(x,y,p) \to q_k(x,y)$ because in that case the mutant frequency dynamics within patches is described by the mean matrix $\boldsymbol{A}$, which is also the matrix of the linearized dynamical system around $p = 0$, and so $q_k(x,y,p)$ can be expressed in terms of the same leading right eigenvector as that subtending $q_k(x,y)$ (eq. A.8). Averaging eq. (B.6) over $q_k(x,y)$ (which cancels the expected residuals since they are uncorrelated with regressors when regression coefficients minimize the quadratic form; Cox and Wermuth, 1996, section 3.3.2), we see that we can then write invasion fitness as

$$W(x,y) = 1 - c(x,y) + r(x,y)b(x,y) \quad \text{(B.7)}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

for the relatedness coefficient defined in eq. (B.1) and which provides the two-predictor representation of invasion fitness by inclusive fitness.

B.1.3 Comparing single- and two-predictor regression

The main difference between the single and two-predictor regression version of inclusive fitness (eq. B.3 and eq. B.5) is that only mutant fitness in different contexts configurations (the set of $w(x, \boldsymbol{x}_k, p)$) are taken into account into the single-predictor regression (eq. B.3), while all contexts for mutant and residents (the set of $w(x, \boldsymbol{x}_k, p)$ and $w(y, \boldsymbol{x}_{k+1}, p)$ values) are taken into account in the two-predictor version (eq. B.5). Technically, this implies that one has to consider explicitly the average mutant allele frequency p in the total population to derive the two-predictor version. Biologically, this implies that the interpretation of costs and benefits differ. Indeed, while the variable β in eq. B.4 and b in eq. B.7 and are both regression coefficients of fitness to mutant frequency in neighbors, in general $\beta \neq b$, since the value of a regression coefficient depends on the other predictor variables considered. Likewise, γ and c differ. This is best seen in the case where $N = 2$, where the single-predictor regression line exactly describes the fitness for $k = 1$ and $k = 2$, hence $1 - \gamma$ is the fitness of a single mutant in a group. Indeed, in this case

$$\begin{aligned} \gamma &= 1 - w(x, y, y) \\ \beta &= w(x, x, y) - w(x, y, y). \end{aligned} \tag{B.8}$$

By contrast, in the case of non-additive interactions between group members, it is known that $1 - c$, as given by the two-predictor regression, is not the fitness of a single mutant (e.g., Gardner et al., 2011, eq. 7). Further, in this case already for $N = 2$, both c and b will depend on relatedness coefficients (e.g., Gardner et al., 2011) and are given explicitly by

$$\begin{aligned} c &= \frac{1}{1 + r(x, y)} \left(1 - w(x, y, y)\right) + \frac{r(x, y)}{1 + r(x, y)} \left(w(y, x, y) - w(x, x, y)\right) \\ b &= \frac{1}{1 + r(x, y)} \left(w(y, x, y) - 1\right) + \frac{r(x, y)}{1 + r(x, y)} \left(w(x, x, y) - w(x, y, y)\right), \end{aligned} \tag{B.9}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

which is different from γ and β that depend only on differences in individual fitness.

B.2 Inclusive fitness for diploids with classes

B.2.1 General regression approach

We now turn to the case with diploidy and classes and derive a gene-centered expression for inclusive fitness by performing an extension of the two-predictor regression of the fitness of a representative gene copy from the population. To construct this fitness measure, we consider all possible group trait profiles, and write for each class s of parents and class u of descendants, the fitness (per haplogenome) of an individual i with trait x_i in a group with trait profile $\boldsymbol{x}_{-i}$ as

$$w_{us}(x_i, \boldsymbol{x}_{-i}, y) = w_{us}(y, \boldsymbol{x}_{\mathbf{0}}, y) - c_{us}(x, y)p_i + \sum_{a\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x, y)p_{a(i)} + \text{residual}. \tag{B.10}$$

Here, p_i is the frequency of the mutant allele in individual i (zero, one-half, or one, so that the individual expresses, respectively, trait y, x or $z(x, y)$), $p_{a(i)}$ stands for the mutant allele frequency in neighbors of class a of individual i, and $c_{us}(x, y)$ and $b^{\mathrm{N}}_{us\leftarrow a}(x, y)$ are regression coefficients depending on mutant and resident trait values. Eq. (B.10) must hold for all group trait profiles $(x_i, \boldsymbol{x}_{-i}) \in \{z(x, y), x, y\} \times \prod_{a\in\mathcal{C}} \{z_a(x_a, y_a), x_a, y_a\}^{(n_a-\delta_{sa})}$. The superscript in $b^{\mathrm{N}}_{us\leftarrow a}(x, y)$ emphasizes the fact that this regression coefficient is suitable in neighbor-modulated representations of fitness, such as eq. (B.10), where fitness effects are grouped as effects of neighbors on a single recipient, while we will later introduce regression coefficients suitable for inclusive fitness representations of fitness, such as eq. (B.27), where fitness effects are grouped as effects of a single actor. The regression coefficients are determined as follows.

First, note that eq. (B.10) says that we seek to obtain a predictor of class-u fitness by an s parent as a linear regression of the fitness of each gene copy on the mutant allele frequency carried by the actor and its neighbors in each class, where the regression is given by

$$\hat{w}_{us}\left(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, p_i, p_{a(i)}, y\right) = w_{us}(y, \boldsymbol{x}_{\mathbf{0}}, y) - c_{us}p_i + \sum_{a\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}p_{a(i)}, \tag{B.11}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

where $\boldsymbol{b}^{\mathrm{N}}_{us} = (b^{\mathrm{N}}_{us\leftarrow 1},...,b^{\mathrm{N}}_{us\leftarrow|\mathcal{C}|})$. Second, we denote $w_{us}(x_i,\boldsymbol{x}_{-i},\boldsymbol{p})$ the fitness of an individual in a population where the mutant frequencies in the different classes are no longer rare, and are collected in the vector $\boldsymbol{p} = (p_1,...,p_{|\mathcal{C}|})$. We further denote $\boldsymbol{q}^{\mathrm{D}}_s(x,y,\boldsymbol{p})$ the ordered distribution of group traits, conditional on an individual carrying the mutant allele and being of class s.

The $\boldsymbol{q}^{\mathrm{D}}_s(x,y,\boldsymbol{p})$ distribution has the same support as $\boldsymbol{q}^{\mathrm{D}}_s(x,y)$ and generalizes it to arbitrary allele frequency. Likewise, we denote $\tilde{\boldsymbol{q}}_s(x,y,\boldsymbol{p})$ the ordered distribution of group traits for non-rare mutant frequency, conditional on an individual carrying the resident allele and being of class s (this distribution has support in $\{x,y\}\times\prod_{a\in\mathcal{C}}\{z_a(x_a,y_a),x_a,y_a\}^{(n_a-\delta_{sa})}$). With these notations, the expected sum of squares for the regression to minimize can be written

$$Q_{u,s}(c_{us},\boldsymbol{b}^{\mathrm{N}}_{us},x,y,\boldsymbol{p}) = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim\boldsymbol{q}^{\mathrm{D}}_s(x,y,\boldsymbol{p})}\left[\left(\hat{w}_{us}\left(c_{us},\boldsymbol{b}^{\mathrm{N}}_{us},p_i,p_{a(i)},y\right)-w_{us}(x_i,\boldsymbol{x}_{-i},\boldsymbol{p})\right)^2\right]p_s + \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim\tilde{\boldsymbol{q}}^{\mathrm{D}}_s(x,y,\boldsymbol{p})}\left[\left(\hat{w}_{us}\left(c_{us},\boldsymbol{b}^{\mathrm{N}}_{us},p_i,p_{a(i)},y\right)-w_{us}(x_i,\boldsymbol{x}_{-i},\boldsymbol{p})\right)^2\right](1-p_s). \quad \text{(B.12)}$$

By solving $\partial Q_{u,s}(c_{us},\boldsymbol{b}^{\mathrm{N}}_{us},x,y,\boldsymbol{p})/\partial c_{us} = 0$ and $\partial Q_{u,s}(c_{us},\boldsymbol{b}^{\mathrm{N}}_{us},x,y,\boldsymbol{p})/\partial b^{\mathrm{N}}_{us\leftarrow a} = 0$ for all $a\in\mathcal{C}$ we then obtain the coefficients $c_{us}(x,y,\boldsymbol{p})$ and $b^{\mathrm{N}}_{us\leftarrow a}(x,y,\boldsymbol{p})$, which depend on the population state.

Our aim is now to evaluate the so-obtained regression coefficients under vanishing mutant allele frequency. To do this we need a single (scalar) measure of allele frequency such that allele frequencies in all classes vanish simultaneously when this measure vanishes. As such a measure, we use the weighted average allele frequency $p = \sum_{a\in\mathcal{C}}\alpha_a(y)p_a$ in the population, where the weights are the neutral class reproductive values (the $\alpha_a(y) = v_a(y)\phi_a(y,y)$ elements in eq. A.26). To evaluate the regression coefficients, we then need to be able to express each class-specific frequency p_a in terms of p and $\phi_a(x,y)$, at least when the mutant allele is rare. For this purpose, we recall that as long as the mutant allele is rare, its growth is characterized by the leading eigenvalue (invasion fitness) and by the associated right eigenvector (quasi-stationary distribution) $\boldsymbol{u}(x,y)$ of the transition matrix $\boldsymbol{A}(x,y)$ [i.e., eq. A.8]. Eigenvectors are defined up to a constant factor, so the relationship between allele frequencies p_a in each class a and the

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

eigenvector can be specified up to a constant, here denoted L_1. We write this relationship as

$$p_a = L_1 u_a(x,y) \tag{B.13}$$

where $u_a(x,y) = \sum_{\mathbf{k}\in I} k_a u_{\mathbf{k}}(x,y)$ is (up to a constant factor) the frequency of the mutant allele in class a under the quasi-stationary distribution. The average allele frequency is then $p = L_1 \sum_{a\in\mathcal{C}} \alpha_a(y) u_a(x,y)$, whereby $L_1 = p/\left[\sum_{a\in\mathcal{C}} \alpha_a(y) u_a(x,y)\right]$ and

$$p_a = p\frac{u_a(x,y)}{\sum_{a\in\mathcal{C}} \alpha_a(y) u_a(x,y)} = p\frac{u_a(x,y)}{\sum_{a\in\mathcal{C}} u_a(x,y)}\frac{\sum_{a\in\mathcal{C}} u_a(x,y)}{\sum_{a\in\mathcal{C}} \alpha_a(y) u_a(x,y)}. \tag{B.14}$$

From eq. (A.22), the middle fraction on the right-hand side is the probability $\phi_a(x,y)$ that a randomly sampled gene copy from the mutant lineage is in class a, introduced in eq. (A.21): $\phi_a(x,y) = \sum_{\mathbf{k}\in I} k_a u_{\mathbf{k}}(x,y)/\left[\sum_{\mathbf{k}\in I}\sum_{a\in\mathcal{C}} k_a u_{\mathbf{k}}(x,y)\right] = u_a(x,y)/\sum_{a\in\mathcal{C}} u_a(x,y)$. The last fraction is then the inverse of $\sum_{a\in\mathcal{C}} \alpha_a(y)\left[u_a(x,y)/\sum_{a\in\mathcal{C}} u_a(x,y)\right] = \sum_{a\in\mathcal{C}} \alpha_a(y)\phi_a(x,y)$, and

$$p_a = p\frac{\phi_a(x,y)}{\sum_{a\in\mathcal{C}} \alpha_a(y)\phi_a(x,y)}. \tag{B.15}$$

Substituting eq. (B.15) into $c_{us}(x,y,\boldsymbol{p})$ and $b^{\mathrm{N}}_{us\leftarrow a}(x,y,\boldsymbol{p})$, we compute the regression coefficients of eq. (B.10) as $c_{us}(x,y) = \lim_{p\to 0} c_{us}(x,y,\boldsymbol{p})$ and $b^{\mathrm{N}}_{us\leftarrow a}(x,y) = \lim_{p\to 0} b^{\mathrm{N}}_{us\leftarrow a}(x,y,\boldsymbol{p})$ (see section B.3 for a concrete application). We further note that, by construction, $\boldsymbol{q}_s(x,y,\boldsymbol{p}) \to \boldsymbol{q}_s(x,y)$ as $p \to 0$. This then allows us to define

$$r_{\mathrm{f},s}(x,y) = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim\boldsymbol{q}^{\mathrm{D}}_s(x,y)}[p_i]\,, \tag{B.16}$$

which is the probability that, conditional on an individual of class s carrying the mutant allele, a randomly sampled homologous gene in that individual is a mutant, and

$$r_{\mathrm{n},a|s}(x,y) = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim\boldsymbol{q}^{\mathrm{D}}_s(x,y)}\left[p_{a(i)}\right], \tag{B.17}$$

is the probability that, conditional on an individual of class s carrying the mutant allele, a

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

randomly sampled homologous gene in a neighbor of class a is a mutant allele. In terms of the $r_{\mathrm{f},s}(x,y)$ and $r_{\mathrm{n},a|s}(x,y)$ probabilities, we define the relatedness coefficient between a class-s actor and a class-a recipient as

$$r_{a|s}(x,y) = \frac{r_{\mathrm{n},a|s}(x,y)}{r_{\mathrm{f},s}(x,y)}. \tag{B.18}$$

Now substitute eq. (B.10) into direct fitness (eq. A.27) and then into invasion fitness (eq. A.19). Then, by dint of the reproductive values recursion (eq. A.18), relatedness coefficients (eqs. B.16–B.18), the reproductive values normalizer (eq. A.20), and recalling that the residual term in eq. (B.10) cancels when averaged over the $\boldsymbol{q}_s^{\mathrm{D}}(x,y)$ distribution, since they are uncorrelated with regressors, the invasion fitness of a mutant allele x introduced as a single copy in a resident population otherwise monomorphic for allele y can be put under the form

$$W(x,y) = 1 + \frac{1}{V(x,y)}\left[W_{\mathrm{IF}}(x,y) - 1\right], \tag{B.19}$$

where

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{u\in\mathcal{C}}\sum_{s\in\mathcal{C}} v_u(y)\left[-c_{us}(x,y) + \sum_{a\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y) r_{a|s}(x,y)\right] r_{\mathrm{f},s}(x,y)\phi_s(x,y). \tag{B.20}$$

is the average inclusive fitness of an allele. These two equations were previously derived for the haploid case ($r_{\mathrm{f},s}(x,y) = 1$ for all $s \in \mathcal{C}$) in Lehmann et al. (2016, eqs. C.1-C.6) assuming that $1 - c_{u,s}$ was the intercept and only $b^{\mathrm{N}}_{u,s}$ were the regression coefficients of fitness, thus performing a multiple-neighbors extension of the single-predictor regression to obtain inclusive fitness. To obtain such coefficients it suffices to set $p_s = 1$ in eq. (B.12) and otherwise follow the same line of argument.

Since $V(x,y) > 0$, we have from eq. (B.19) that

$$W(x,y) \leq 1 \iff W_{\mathrm{IF}}(x,y) \leq 1. \tag{B.21}$$

Hence, trait x^* is uninvadable if it solves $\max_{x\in\mathcal{X}} W_{\mathrm{IF}}(x,x^*)$. Finally, we note that if one chooses

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

to normalize the (neutral) reproductive such that $V(x,y) = 1$, then one would have $W(x,y) = W_{\mathrm{IF}}(x,y)$.

B.2.2 Class-specific inclusive fitness

In eq. (B.20), social interactions among all individuals in the population are grouped by recipients since

$$-c_{us}(x,y) + \sum_{a\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y) r_{a|s}(x,y) \tag{B.22}$$

is the total effect of all actors in a group on class-u offspring produced by a representative recipient of class s recipient expressing a copy of the mutant allele. We now rearrange eq. (B.20) in order to obtain an inclusive fitness perspective, by grouping actions by actor (e.g., Hamilton, 1970, Frank, 1998, Rousset, 2004, Fig. 7.1). To reach this perspective, we note that for $s \neq a$,

$$r_{\mathrm{n},a|s}(x,y)\phi_s(x,y) = r_{\mathrm{n},s|a}(x,y)\phi_a(x,y)\frac{n_s}{n_a}. \tag{B.23}$$

To check this result, we highlight that it considers two ways of sampling gene copies: either we sample gene copies uniformly from the mutant lineage (by definition, $\phi_s(x,y)$ is the probability that a gene sampled in this way is in a class-s individual), or we sample gene copies uniformly among class-a individuals ($r_{\mathrm{n},a|s}(x,y)$ is the probability that, when a given gene copy from a class-s individual is mutant, a given gene copy from a class-a individual in the same group is mutant). Thus, the expected number of pairs of gene copies in class-s and class-a individuals within a group per copy of the mutant allele is $\phi_s(x,y)2n_a r_{\mathrm{n},a|s}(x,y)$. By the same logic (but considering the case where the "first" copy is sampled in a class-a individual), this is also $\phi_a(x,y)2n_s r_{\mathrm{n},s|a}(x,y)$, from which the above result follows.

Substituting eq. (B.23) into eq. (B.20), making the change of dummy variable

$$\sum_{s\in\mathcal{C}} c_{us}(x,y) r_{\mathrm{f},s}(x,y) = \sum_{a\in\mathcal{C}} c_{ua}(x,y) r_{\mathrm{f},a}(x,y), \tag{B.24}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

and rearranging we obtain

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{u\in\mathcal{C}}\sum_{a\in\mathcal{C}} v_u(y)\left[-c_{ua}(x,y) + \sum_{s\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y)\frac{n_s}{n_a}\, r_{s|a}(x,y)\right] r_{\mathrm{f},a}(x,y)\phi_a(x,y). \tag{B.25}$$

Here

$$-c_{ua} + \sum_{s\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y)\frac{n_s}{n_a}\, r_{s|a}(x,y) \tag{B.26}$$

is the average effect of a single individual of class a when switching from expressing zero to one copy of the mutant allele on the number of mutant gene copies in class u produced by all recipients of the action; that is, the recipients of each class. Eq. (B.26) is consistent with eq. (8) of Grafen, 2006a who assumed (a) additive separable fitness effects and (b) relatedness independent of evolving trait values. To further simplify expression (B.25) for inclusive fitness, we let

$$b_{us\leftarrow a}(x,y) = \frac{b^{\mathrm{N}}_{us\leftarrow a}(x,y)n_s}{n_a}, \tag{B.27}$$

which is the average effect on the number of class-u offspring produced per haplogenome by all class-s individuals in a group and stemming from a single gene copy in a class-a individual switching to expressing the mutant instead of the resident allele. Hence, the coefficient $b_{us\leftarrow a}(x,y)$ groups fitness effects by a single actor on all recipients in class s. Substituting eq. (B.27) into eq. (B.25), we can write invasion fitness as an average inclusive fitness effect:

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{a\in\mathcal{C}} \Delta w_{\mathrm{IF},a}(x,y) r_{\mathrm{f},a}(x,y)\phi_a(x,y), \tag{B.28}$$

where

$$\Delta w_{\mathrm{IF},a}(x,y) = \sum_{u\in\mathcal{C}} v_u(y)\left[-c_{ua}(x,y) + \sum_{s\in\mathcal{C}} b_{us\leftarrow a}(x,y)\, r_{s|a}(x,y)\right] \tag{B.29}$$

is the inclusive fitness effect of an average class-a carrier of the mutant allele.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Finally, let us denote by $\tilde{x}_a = (x_1^*, x_2^*, ..., x_{a-1}^*, x_a, x_{a+1}^*, ..., x_{|\mathcal{C}|}^*)$ the trait profile of a mutation that holds all traits at the uninvadable state, except for trait x_a of class a that can unilaterally deviate. Then, if the mutant $\tilde{x}_a$ appears in a population at the uninvadable state x^*, we have

$$\Delta w_{\mathrm{IF},v}(\tilde{x}_a, x^*) = 0 \quad \forall v \neq a, \tag{B.30}$$

which, from eq. (B.28), implies

$$W_{\mathrm{IF}}(\tilde{x}_a, x^*) = 1 + \Delta w_{\mathrm{IF},a}(\tilde{x}_a, x^*) r_{\mathrm{f},a}(\tilde{x}_a, x^*) \phi_a(\tilde{x}_a, x^*). \tag{B.31}$$

Eq. (B.30) says that if a mutant allele changes only the trait expression of individuals of class a, then the inclusive fitness effect of any other class is nil, which is so since $\Delta w_{\mathrm{IF},a}(\tilde{x}_a, x^*)$ captures all effects of individuals of class a expressing the mutant trait x_a (the "actors") on mutant allele transmission. A more formal proof follows from the fact that $c_{uv}(\tilde{x}_a, x^*) = 0$ and $b_{us\leftarrow v}(\tilde{x}_a, x^*) = 0$ for all $v \neq a$ and all u and s because the u type fitness w_{us} of an individual of class s is a constant with respect to the traits of individuals in any class $v \neq a$, since all individuals in any class $v \neq a$ express the same trait value x_v^*. Hence, all such regression coefficients on class-v individuals will be nil, since there is no variation in individual fitness to be explained by any such regressor.

B.3 Inclusive fitness for social insects example

We here derive the inclusive fitness effects for the social insect model (eq. B.6 in Box 2) from the fitness functions defined in the main text (see eqs. B.3–B.5 in Box 2) and assuming the population has reached the uninvadable sex ratio of $1/2$ for this model. From eqs. (B.3), we can

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

write the fitness components of a female i whose worker offspring has trait $x_{\mathrm{o}(i)} \in \{y_{\mathrm{o}}, x_{\mathrm{o}}, z_{\mathrm{o}}\}$ as

$$\begin{aligned} w_{\mathrm{ff}}\Big((x_{\mathrm{o}(i)}, y_{\mathrm{f}}), y\Big) &= \frac{\Big(1 + P(x_{\mathrm{o}(i)})\Big)}{2\,(1 + P(y_{\mathrm{o}}))} \\ w_{\mathrm{mf}}\Big((x_{\mathrm{o}(i)}, y_{\mathrm{f}}), y\Big) &= \frac{\Big(1 + P(x_{\mathrm{o}(i)})\Big)}{2\,(1 + P(y_{\mathrm{o}}))}, \end{aligned} \tag{B.32}$$

while, from eq. (B.5) the fitness components of a male i whose worker offspring has trait $x_{\mathrm{o}(i)} \in \{y_{\mathrm{o}}, x_{\mathrm{o}}, z_{\mathrm{o}}\}$ is

$$\begin{aligned} w_{\mathrm{fm}}\Big((x_{\mathrm{o}(i)}, y_{\mathrm{f}}), y\Big) &= w_{\mathrm{ff}}\Big((x_{\mathrm{o}(i)}, y_{\mathrm{f}}), y\Big) \\ w_{\mathrm{mm}}\Big((x_{\mathrm{o}(i)}, y_{\mathrm{f}}), y\Big) &= w_{\mathrm{mf}}\Big((x_{\mathrm{o}(i)}, y_{\mathrm{f}}), y\Big). \end{aligned} \tag{B.33}$$

In order to evaluate the sum of squares for the regression coefficients, we need to take into account all possible matings as this determines the number of mutant allele copies in the worker offspring. There is a total number of 9 matings, since a female can be homozygote mutant (probability denoted $p_{\mathrm{ho,f}}$), heterozygote (probability denoted $p_{\mathrm{he,f}}$), or homozygote resident (probability $(1 - p_{\mathrm{ho,f}} - p_{\mathrm{he,f}})$), and her mate can be of the same respective types (with respective probabilities, $p_{\mathrm{ho,m}}$, $p_{\mathrm{he,m}}$, and $(1 - p_{\mathrm{ho,m}} - p_{\mathrm{he,m}})$). The assumption that we consider a population with random mating at the uninvadable sex-ratio, implies that the fitness functions for both males and females are equivalent (e.g., eq. B.33), and that the frequency of the mutant allele will be the same in males and females, $p_{\mathrm{m}} = p_{\mathrm{f}} = p$. Henceforth, we can evaluate the genotype frequencies in terms of allele frequencies at Hardy-Weinberg equlibrium:

$$p_{\mathrm{ho,m}} = p_{\mathrm{ho,f}} = p^2 \qquad \text{and} \qquad p_{\mathrm{he,f}} = p_{\mathrm{he,m}} = 2p(1-p). \tag{B.34}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

B.3.1 Regressions for female fitness components

Taking into account all matings, we write the sum of squares for female fitness through offspring of type $j \in \{\mathrm{f},\mathrm{m}\}$ as

$$Q_{jf}(c_{jf}, b^{\mathrm{N}}_{jf\leftarrow \mathrm{o}}, b^{\mathrm{N}}_{jf\leftarrow \mathrm{m}}) = Q_{jf|\mathrm{ho}(x)} p_{\mathrm{ho,f}} + Q_{jf|\mathrm{he}} p_{\mathrm{he,f}} + Q_{jf|\mathrm{ho}(y)}(1 - p_{\mathrm{ho,f}} - p_{\mathrm{he,f}}), \tag{B.35}$$

where $Q_{jf|\mathrm{ho}(x)}$, $Q_{jf|\mathrm{he}}$, and $Q_{jf|\mathrm{ho}(y)}$ are, respectively, the sum of squares when the female is homozygote mutant, heterozygote, and homozygote resident. Application of eqs. (B.11)–(B.12) shows that when the female is homozygote

$$\begin{aligned} Q_{jf|\mathrm{ho}(x)} &= \left(\frac{1}{2} - c_{jf} + b^{\mathrm{N}}_{jf\leftarrow \mathrm{o}} + b^{\mathrm{N}}_{jf\leftarrow \mathrm{m}} - w_{jf}((z_{\mathrm{o}}, y_{\mathrm{f}}), y)\right)^2 p_{\mathrm{ho,m}} \\ &+ \left[\left(\frac{1}{2} - c_{jf} + b^{\mathrm{N}}_{jf\leftarrow \mathrm{o}} + \frac{b^{\mathrm{N}}_{jf\leftarrow \mathrm{m}}}{2} - w_{jf}((z_{\mathrm{o}}, y_{\mathrm{f}}), y)\right)^2 \frac{1}{2}\right. \\ &+ \left.\left(\frac{1}{2} - c_{jf} + \frac{b^{\mathrm{N}}_{jf\leftarrow \mathrm{o}}}{2} + \frac{b^{\mathrm{N}}_{jf\leftarrow \mathrm{m}}}{2} - w_{jf}((x_{\mathrm{o}}, y_{\mathrm{f}}), y)\right)^2 \frac{1}{2}\right] p_{\mathrm{he,m}} \\ &+ \left(\frac{1}{2} - c_{jf} + \frac{b^{\mathrm{N}}_{jf\leftarrow \mathrm{o}}}{2} - w_{jf}((x_{\mathrm{o}}, y_{\mathrm{f}}), y)\right)^2 (1 - p_{\mathrm{ho,m}} - p_{\mathrm{he,m}}) \end{aligned} \tag{B.36}$$

where in the present example w_{jf} is given by eq. (B.32). The first, second, and third summand, stand, respectively, for the case where the male mate of the focal female is homozygote mutant, heterozygote, or homozygote resident. When the male is heterozygote, then with probability $1/2$ the worker inherits a copy of his mutant allele and will be homozygote (first term in the second summand), while with probability $1/2$ the worker does not inherit a copy of the mutant allele from its father and will be heterozygote (second term in the second summand).

When the female is heterozygote, we write the sum of squares as $Q_{jf|\mathrm{he}} = (1/2)Q_{jf|\mathrm{he},1} + (1/2)Q_{jf|\mathrm{he},0}$, where $Q_{jf|\mathrm{he},1}$ represents the case where the worker inherits the mutant allele from its mother and $Q_{jf|\mathrm{he},0}$ for the case the worker does not inherit the mutant from its mother. We

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

find that

$$\begin{aligned}Q_{jf|he,1} &= \left(\frac{1}{2} - \frac{c_{jf}}{2} + b^{N}_{jf\leftarrow o} + b^{N}_{jf\leftarrow m} - w_{jf}((z_o, y_f), y)\right)^2 p_{ho,m} \\ &+ \left[\left(\frac{1}{2} - \frac{c_{jf}}{2} + b^{N}_{jf\leftarrow o} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}((z_o, y_f), y)\right)^2 \frac{1}{2}\right. \\ &+ \left.\left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}((x_o, y_f), y)\right)^2 \frac{1}{2}\right] p_{he,m} \\ &+ \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} - w_{jf}((x_o, y_f), y)\right)^2 (1 - p_{ho,m} - p_{he,m}) \quad \text{(B.37)}\end{aligned}$$

and

$$\begin{aligned}Q_{jf|he,0} &= \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + b^{N}_{jf\leftarrow m} - w_{jf}((x_o, y_f), y)\right)^2 p_{ho,m} \\ &+ \left[\left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}((x_o, y_f), y)\right)^2 \frac{1}{2} + \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}((y_o, y_f), y)\right)^2 \frac{1}{2}\right] p_{he,m} \\ &+ \left(\frac{1}{2} - \frac{c_{jf}}{2} - w_{jf}((y_o, y_f), y)\right)^2 (1 - p_{ho,m} - p_{he,m}). \quad \text{(B.38)}\end{aligned}$$

Finally, when the female is homozygote resident, we have that

$$\begin{aligned}Q_{jf|ho(y)} &= \left(\frac{1}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + b^{N}_{jf\leftarrow m} - w_{jf}((x_o, y_f), y)\right)^2 p_{ho,m} \\ &\left[\left(\frac{1}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}((x_o, y_f), y)\right)^2 \frac{1}{2} + \left(\frac{1}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}((x_o, y_f), y)\right)^2 \frac{1}{2}\right] p_{he,m},\end{aligned} \quad \text{(B.39)}$$

since a worker from a homozygote resident mother can inherit the mutant allele only from its father, and when the father is heterozygote the worker inherits the mutant with probability $1/2$.

We now minimize the sum of squares $Q_{jf}(c_{jf}, b^{N}_{jf\leftarrow o}, b^{N}_{jf\leftarrow m})$ with respect to the relevant re-

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

gression coefficients, which requires that, for $j \in \{\mathrm{f},\mathrm{m}\}$, we solve

$$\frac{\partial Q_{j\mathrm{f}}(c_{j\mathrm{f}}, b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}, b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}})}{\partial c_{j\mathrm{f}}} = 0\ , \ \frac{\partial Q_{j\mathrm{f}}(c_{j\mathrm{f}}, b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}, b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}})}{\partial b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}} = 0 \text{ and } \frac{\partial Q_{j\mathrm{f}}(c_{j\mathrm{f}}, b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}, b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}})}{\partial b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}} = 0 \quad \text{(B.40)}$$

for $c_{j\mathrm{f}}$, $b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}$ and $b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}}$. Substituting eq. (B.34) into the so-obtained regression coefficients and letting $p \to 0$, we finally obtain that $c_{j\mathrm{f}} = 0$, $b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}} = 0$ for $j \in \{\mathrm{f},\mathrm{m}\}$, and

$$\begin{aligned} b^{\mathrm{N}}_{\mathrm{ff}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})} \\ b^{\mathrm{N}}_{\mathrm{mf}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})}. \end{aligned} \quad \text{(B.41)}$$

B.3.2 Regressions for male fitness components

We now derive the regression coefficients for the male fitness components. The model for the male side is exactly symmetric to that of the female side and to compute the corresponding sum of squares $Q_{j\mathrm{m}}(c_{j\mathrm{m}}, b^{\mathrm{N}}_{j\mathrm{m}\leftarrow\mathrm{o}}, b^{\mathrm{N}}_{j\mathrm{m}\leftarrow\mathrm{f}})$ for $j \in \{\mathrm{f},\mathrm{m}\}$ we only interchange m and f subscripts in all equations of the previous section. Otherwise, the calculations carry over *mutatis mutandis* to give $c_{j\mathrm{m}} = 0$, $b^{\mathrm{N}}_{j\mathrm{m}\leftarrow\mathrm{f}} = 0$ for $j \in \{\mathrm{f},\mathrm{m}\}$, and

$$\begin{aligned} b^{\mathrm{N}}_{\mathrm{fm}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})} \\ b^{\mathrm{N}}_{\mathrm{mm}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})}. \end{aligned} \quad \text{(B.42)}$$

B.3.3 Inclusive fitness effects

Using the regression coefficients computed in the last two sections, we are now in the position to compute the inclusive fitness effects. First, the inclusive fitness effects of females and males is null

$$\begin{aligned} \Delta w_{\mathrm{IF},\mathrm{f}}(x,y) &= 0 \\ \Delta w_{\mathrm{IF},\mathrm{m}}(x,y) &= 0. \end{aligned} \quad \text{(B.43)}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

To obtain the inclusive fitness effect for a worker, we note that from eq. B.27,

$$b_{us\leftarrow \mathrm{o}}(x,y) = b^{\mathrm{N}}_{us\leftarrow \mathrm{o}}(x,y) \tag{B.44}$$

for $u \in \{\mathrm{f},\mathrm{m}\}$ and $s \in \{\mathrm{f},\mathrm{m}\}$ since the number of individuals of each class $n_{\mathrm{f}} = n_{\mathrm{m}} = n_{\mathrm{o}} = 1$. With this, eq. (B.29), and eqs. (B.41)–(B.42), we obtain

$$\Delta w_{\mathrm{IF},\mathrm{o}}(x,y) = v_{\mathrm{f}}(y)\left(\frac{P(x_{\mathrm{o}}) - P(y_{\mathrm{o}})}{1+P(y_{\mathrm{o}})}\right) + v_{\mathrm{m}}(y)\left(\frac{P(x_{\mathrm{o}}) - P(y_{\mathrm{o}})}{1+P(y_{\mathrm{o}})}\right). \tag{B.45}$$

Appendix C: Individual-centered perspective of adaptation

In this Appendix, we explain how to derive expressions for fitness as-if from the invasion fitness (Appendix A) and inclusive fitness (Appendix B), and we prove eq. (13), eq. (14), and eq. (15) of the main text. As detailed in the section "Characterizing individual maximizing behavior" of the main text, we aim at identifying an individual-centered maximand that individuals appear to be maximizing in an uninvadable population state. More formally, in the absence of class structure and diploidy, we aim to identify a meaningful fitness as-if function w_{I} satisfying

$$x^* \in \arg\max_{x\in\mathcal{X}} w_{\mathrm{I}}\left(x, \boldsymbol{x}^*_{-i}, x^*\right) \iff x^* \in \arg\max_{x\in\mathcal{X}} W(x,x^*), \tag{C.1}$$

where x^* is a symmetric Nash equilibrium and invasion fitness is given by eq. (A.6).

In the presence of class structure and diploidy, we aim to identify, for a class-a individual, a fitness as-if function $w_{\mathrm{I},a}$ satisfying

$$x^*_a \in \arg\max_{x_{a(i)}\in\mathcal{X}_a} w_{\mathrm{I},a}\left((x_{a(i)}, x^*_{-a(i)}), \boldsymbol{x}^*_{-i}, x^*\right)\ \forall a\in\mathcal{C} \iff x^* \in \arg\max_{x\in\mathcal{X}} W(x,x^*), \tag{C.2}$$

where $x^*_{-a(i)} = (x^*_1, x^*_2, ..., x^*_{a-1}, x^*_{a+1}, ..., x^*_{|\mathcal{C}|})$, entry j of the vector of neighbor trait profile $\boldsymbol{x}^*_{-i}$ is equal to x^*_a if neighbour j is of class a, $x^* = (x^*_1, x^*_2, ..., x^*_{|\mathcal{C}|})$ is a Nash equilibrium (symmetric in each class), and invasion fitness is given by eq. (A.23) with eq. (A.27).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

C.1 Sufficient conditions for fitness as-if maximization

We start by presenting sufficient conditions for fitness as-if to satisfy eq. (C.2), which is useful to identify the type of functions that can and cannot be taken to represent fitness as-if. To do this, recall that $\tilde{x}_a = (x_1^*, x_2^*, ..., x_{a-1}^*, x_a, x_{a+1}^*, ..., x_{|\mathcal{C}|}^*)$ stands for the trait profile of a mutation that has all traits at the uninvadable state (e.g., eq. B.30), except for trait x_a of class a that can unilaterally deviate. Then, uninvadability can be characterized as

$$x_a^* \in \arg\max_{x_a \in \mathcal{X}_a} W(\tilde{x}_a, x^*) \ \forall a \in \mathcal{C} \iff x^* \in \arg\max_{x \in \mathcal{X}} W(x, x^*), \tag{C.3}$$

since x^* must be a best response to any unilateral deviation in any class a with $W(\tilde{x}_a, x^*)$ being the invasion fitness of the mutant $\tilde{x}_a$ in a resident population at x^*. Eq. (C.3) shows that if fitness as-if is equal to $W(\tilde{x}_a, x^*)$ at x^*; namely, if

$$w_{\mathrm{I},a}\left((x_a, x_{-a(i)}^*), \boldsymbol{x}_{-i}^*, x^*\right) = W(\tilde{x}_a, x^*), \tag{C.4}$$

then eq. (C.2) is satisfied. If $w_{\mathrm{I},a}\left((x_a, x_{-a(i)}^*), \boldsymbol{x}_{-i}^*, x^*\right)$ is an affine function of $W(\tilde{x}_a, x^*)$ or a monotonically increasing function thereoff, then eq. (C.2) is also satisfied.

By the same token and in force of eq. (B.21), we have

$$x_a^* \in \arg\max_{x_a \in \mathcal{X}_a} W_{\mathrm{IF}}(\tilde{x}_a, x^*) \ \forall a \in \mathcal{C} \iff x^* \in \arg\max_{x \in \mathcal{X}} W(x, x^*), \tag{C.5}$$

so if

$$w_{\mathrm{I},a}\left((x_{a(i)}, x_{-a(i)}^*), \boldsymbol{x}_{-i}^*, x^*\right) = \Delta w_{\mathrm{IF},a}(\tilde{x}_a, x^*)\, r_{\mathrm{f},a}(\tilde{x}_a, x^*) \phi_a(\tilde{x}_a, x^*), \tag{C.6}$$

then we see from eq. (B.31) that eq. (C.5) is satisfied. Likewise, if $w_{\mathrm{I},a}\left((x_a, x_{-a(i)}^*), \boldsymbol{x}_{-i}^*, x^*\right)$ is an affine function of $W_{\mathrm{IF}}(\tilde{x}_a, x^*)$ or a monotonically increasing function thereof, then eq. (C.5) is again satisfied.

Eq. (C.4) and eq. (C.6) put constraints on the representation of fitness as-if and make clear

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

that it should be close to invasion fitness and consist of the same fitness components, w or w_{us}. But fitness as-if $w_{\mathrm{I},a}\left((x_{a(i)}, x_{-a(i)}), \boldsymbol{x}_{-i}, \bar{x}\right)$ needs also to be well defined outside the uninvadable population state x^*, so that the trait profile $\boldsymbol{x}_{-i}$ of neighbors should cover the case where the traits of neighbors are all distinct from each other, otherwise the concept of autonomous decision maker does make full biological sense.

C.2 The instrumental distribution

We now present a way to construct fitness as-if for the case where the traits of neighbors can be distinct from each other. This departs from the population genetic models of the previous sections where invasion fitness was depending only on heterozygote and homozygote mutant and resident traits, with the distribution $\boldsymbol{q}^{\mathrm{D}}(x, y)$ of group states describing correlated trait expression within groups. In order to take this difference into account, we consider that, while fitness as-if should in general consist of the same fitness components, w or w_{us}, as invasion fitness, it should be averaged over a different distribution of correlated trait expression within groups (in particular, a distribution with a different support allowing for each individual expressing a different trait). We refer to this new distribution as the *instrumental* distribution, and it will be reminiscent of the so-called subjective probability distribution that neighbors play a given profile of traits as considered in the construction of an individual's utility function in game theory (e.g., Fudenberg and Tirole, 1991, Mas-Colell et al., 1995). To describe how we obtain the instrumental distribution, we first define its support, beginning with a haploid population without class structure and using average direct fitness as-if as an example.

C.2.1 Haploids without classes

For the haploid case, where $w_{\mathrm{I}}(x, \boldsymbol{x}_{-i}, \bar{x})$ is the fitness as-if of an individual with trait x_i in a group with neighbor trait profile $\boldsymbol{x}_{-i} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_N)$ in a population with average group trait $\bar{x}$, the instrumental distribution is constructed as follows. We first consider the support (sample space) defined from the actual neighbor trait profile $\boldsymbol{x}_{-i}$, defined by replacing any number of the elements of $\boldsymbol{x}_{-i}$ by i's trait. Thus, for any $k \in \{1, \ldots, N\}$, we consider the

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

set $\mathcal{P}_k(\boldsymbol{x}_{-i})$ of hypothetical neighbor trait profiles $\tilde{\boldsymbol{x}}_{-i}$ such that exactly $k-1$ components of the (true) profile $\boldsymbol{x}_{-i}$ are replaced by i's trait x_i, while the remaining $N-k$ components of $\tilde{\boldsymbol{x}}_{-i}$ are identical to those in $\boldsymbol{x}_{-i}$ (this operation will capture correlated trait expression within groups). The set of all such profiles is $\mathcal{S}_i = \cup_{k=1}^{N} \mathcal{P}_k = \prod_{j\neq i}^{N-1} \{x_i, x_j\}$. From the perspective of individual i, we can think of $\tilde{\boldsymbol{x}}_{-i}$ as a hypothetical profile where neighbors' traits have been replaced with traits similar to self, and if such a profile were to obtain in individual i's group, then its fitness would be $w(x_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x})$.

Any probability distribution $\sigma(x_i, \boldsymbol{x}_{-i}, \bar{x})$ on the support $\mathcal{S}_i$ takes values in the simplex $\Delta(\mathcal{S}_i)$ induced by $\mathcal{S}_i$, and assigns probabilities $\sigma_k(\tilde{\boldsymbol{x}}_{-i}; x_i, \boldsymbol{x}_{-i}, \bar{x})$ such that these probabilities satisfy

$$\sum_{k=1}^{N} \sum_{\tilde{\boldsymbol{x}}_{-i} \in \mathcal{P}_k(\boldsymbol{x}_{-i})} \sigma_k(\tilde{\boldsymbol{x}}_{-i}; x_i, \boldsymbol{x}_{-i}, \bar{x}) = 1. \tag{C.7}$$

The instrumental distribution $\sigma(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is yet undefined beyond its support. In particular, this distribution has yet no imposed relation to the probabilities of events that occur in the actual reproductive process in the population under consideration (the $q^{\mathrm{D}}(x, y)$ distribution), but it retains the ability to describe within-group correlated trait expression. The exact connection between the instrumental distribution $\sigma(x_i, \boldsymbol{x}_{-i}, \bar{x})$ and $q^{\mathrm{D}}(x, y)$ will be developed in section C.3.

Given the instrumental distribution on support $\mathcal{S}_i$, we can define the average direct fitness as-if of an individual with trait x_i as

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \sum_{k=1}^{N} \sum_{\tilde{\boldsymbol{x}}_{-i} \in \mathcal{P}_k(\boldsymbol{x}_{-i})} w(x_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x}) \sigma_k(\tilde{\boldsymbol{x}}_{-i}; x_i, \boldsymbol{x}_{-i}, \bar{x}), \tag{C.8}$$

which is the average of individual fitness over the distribution $\sigma(x_i, \boldsymbol{x}_{-i}, \bar{x})$. A more compact representation of this fitness as-if is

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \mathrm{E}_{\tilde{\boldsymbol{x}}_{-i} \sim \sigma(x_i, \boldsymbol{x}_{-i}, \bar{x})} \left[w(x_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x}) \right], \tag{C.9}$$

where the notation $\sim$ specifies that variable $\tilde{\boldsymbol{x}}_{-i}$ follows the distribution $\sigma(x_i, \boldsymbol{x}_{-i}, \bar{x})$ (recall eq. A.12).

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

C.2.2 Diploids with classes

We can now generalize the construction of the instrumental distribution to a diploid class-structured population. For this case, the hypothetical distributions $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$ for the realized profile of traits $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})$ in a group when individual i is of class s is defined as follows: the trait $\tilde{x}_i$ of individual i takes values in $\{z(x_i, \bar{x}), x_i\}$ and each element $\tilde{x}_j$ of $\tilde{\boldsymbol{x}}_{-i} = (\tilde{x}_1, , ..., \tilde{x}_{i-1}, \tilde{x}_{i+1}, ...x_N)$ takes values in the set of traits belonging to the class of the individual under scrutiny; that is, if individual j is of class a then $\tilde{x}_j \in \{z_a(x_{a(i)}, x_{a(j)}), x_{a(i)}, x_{a(j)}\}$. Thus any value of $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})$ is a hypothetical patch trait profile, where element $x_{a(i)}$ of the true profile may have been replaced by $z_a(x_i, \bar{x})$ and element $x_{a(j)}$ for $j \neq i$ may have been replaced by either $x_{a(i)}$ or $z_a(x_i, x_j)$. The hypothetical profile $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})$ is distributed according to $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$, a distribution that has support in the set $\mathcal{S}_{s(i)} = \{z(x_i, \bar{x}), x_i\} \times \prod_{a \in \mathcal{C}} \prod_{j \neq i}^{(n_a - \delta_{sa})} \{z_a(x_{a(i)}, x_{a(j)}), x_{a(i)}, x_{a(j)}\}$.

Given the instrumental distribution $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$, the average direct fitness as-if of an individual of class a with trait x_i in a group with neighbor trait profile $\boldsymbol{x}_{-i}$ in a population with average group trait $\bar{x}$ is defined as a reproductive value-weighted sum of expected numbers of offspring of different classes u:

$$w_{\mathrm{I},a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \sum_{u \in \mathcal{C}} v_u(\bar{x}) \mathrm{E}_{(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}) \sim \boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})} [w_{ua}(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x})] . \tag{C.10}$$

In order to illustrate the notation and better understand the expectation in eq. (C.10) for diploidy, we consider the case of $N = 2$ without class structure (hence the neigbor trait profile is the singleton $\boldsymbol{x}_{-i} = x_{-i}$). Then, we write the direct fitness as-if of an individual with trait x_i as

$$\begin{aligned} w_{\mathrm{I}}(x_i, x_{-i}, \bar{x}) = {} & w(x_i, x_{-i}, \bar{x}) \sigma_{\mathrm{S,O}}(x_i, x_{-i}, \bar{x}) + w(x_i, x_i, y) \sigma_{\mathrm{O,O}}(x_i, x_{-i}, \bar{x}) \\ & + w(x_i, z(x_i, \bar{x}), \bar{x}) \sigma_{\mathrm{F,O}}(x_i, x_{-i}, \bar{x}) + w(z(x_i, \bar{x}), x_j, \bar{x}) \sigma_{\mathrm{S,F}}(x_i, x_{-i}, \bar{x}) \\ & + w(z(x_i, \bar{x}), x_i, \bar{x}) \sigma_{\mathrm{O,F}}(x_i, x_{-i}, \bar{x}) + w(z(x_i, \bar{x}), z(x_i, \bar{x}), \bar{x}) \sigma_{\mathrm{F,F}}(x_i, x_{-i}, \bar{x}). \end{aligned} \tag{C.11}$$

Here, the second subscript $k \in \{\mathrm{O}, \mathrm{F}\}$ in $\sigma_{j,k}(x_i, x_{-i}, \bar{x})$ denotes that the instrumental substitute to individual i can be of two possible types, either it is "outbred" ($k = \mathrm{O}$), in which case its

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

(objective) fitness w depends on trait x_i, or it is "inbred" ($k = \mathrm{F}$), in which case its fitness depends on trait $z(x_i, \bar{x})$. The first subscript $j \in \{\mathrm{S}, \mathrm{O}, \mathrm{F}\}$ denotes that the instrumental substitute to the group neighbor can express three different traits: either it expresses trait x_{-i} ($j = \mathrm{S}$ for "self"), or it expresses x_i ($j = \mathrm{O}$), or $z(x_i, \bar{x})$ ($j = \mathrm{F}$). With these notations, $\sigma_{j,\mathrm{O}}(x_i, x_{-i}, \bar{x})$ is the instrumental probability that, given trait profile $(x_i, x_{-i}, \bar{x})$, individuals i is of type "outbred" and its neighbor expresses the trait of type $j \in \{\mathrm{S}, \mathrm{O}, \mathrm{F}\}$, while $\sigma_{j,\mathrm{F}}(x_i, x_{-i}, \bar{x})$ is the instrumental probability that individual i is "inbred" and its neihbor expresses trait of type j. In terms of these probabilities, we can write the instrumental distribution of profiles experienced by individual i as

$$\sigma(x_i, x_{-i}, \bar{x}) = \left(\{\sigma_{j,\mathrm{O}}(x_i, x_{-i}, \bar{x})\}_{j\in\{\mathrm{S},\mathrm{O},\mathrm{F}\}}, \{\sigma_{j,\mathrm{F}}(x_i, x_{-i}, \bar{x})\}_{j\in\{\mathrm{S},\mathrm{O},\mathrm{F}\}} \right). \tag{C.12}$$

C.2.3 Fitness as-if of outbred and inbred individuals

Eq. (C.11) shows that fitness as-if is defined as an average over cases where individuals are "outbred" or "inbred", i.e., have the trait of an heterozygote or homozygote, and so varying x_i varies the trait both when the substituted individual is heterozygote (given by x_i itself) and when it is homozygote (given by $z(x_i, x_j)$). This construction, which is used to later prove sufficient conditions where eq. (C.2) holds, ultimately owes to the fact that in the original reproductive process individuals express different traits upon being heterozygote or homozygote (e.g., eq. A.27), a standard modeling assumption for diploids (e.g., Nagylaki, 1992; Gillespie, 2004; Hartl and Clark, 2007). But any actor whose fitness as-if is considered can itself be homozygote our heterozygote, and may thus express a different trait accordingly. Although the fitness as-if in either case will be distinct, their maximization will identify the same best-response. To see this, suppose that an heterozygote individual has trait x_i and fitness as-if $w_{\mathrm{I}}(x_i, x_{-i}, \bar{x})$, while an homozygote has trait $z(x_i, \bar{x})$ and fitness as-if $w_{\mathrm{I}}(z(x_i, \bar{x}), x_{-i}, \bar{x})$. A best-response satisfies $x^* \in \arg\max_{x\in\mathcal{X}} w_{\mathrm{I}}(x, \boldsymbol{x}^*_{-i}, x^*)$, where x is any "dummy" variable. Hence, owing to the convexity assumption eqs. (A.13)–(A.14), $z(x_i, \bar{x}) \in \mathcal{X}$ for all $x_i \in \mathcal{X}$ and $\bar{x} \in \mathcal{X}$, and by using $z(x_i, \bar{x}) \in \mathcal{X}$ instead of x_i as argument in $w_{\mathrm{I}}(\cdot, x_{-i}, \bar{x})$ we are just making a change of variable and not chang-

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

ing the nature of the maximization problem. Hence, fitness as-if can be defined also for inbred individuals.

C.3 Connecting the gene- and individual-centered perspectives

C.3.1 Is fitness as-if compatible with uninvadability?

In order to connect the gene- and individual-centered perspective, we note that the distribution $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$ takes values in the set $\Delta(\mathcal{S}_{s(i)})$, which is the simplex generated by the support $\mathcal{S}_{s(i)}$. The distribution $\boldsymbol{q}_s^{\mathrm{D}}(x, y)$ determining invasion fitness (recall eq. A.27) in the population takes values in the simplex $\Delta(\mathcal{S}_s)$ generated by $\mathcal{S}_s$, but these two simplices are the same $\Delta(\mathcal{S}_s) = \Delta(\mathcal{S}_{s(i)})$. Hence, we can choose $\boldsymbol{q}_s^{\mathrm{D}}$ as the instrumental distribution:

$$\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \boldsymbol{q}_s^{\mathrm{D}}(x_i, \bar{x}), \tag{C.13}$$

whereby we consider that the instrumental probabilities of events, defined by replacing elements of the trait profile by the focal individual's trait, are identical to the probabilities of ordered trait profiles in the population genetic model (or equivalently, to the probabilities of joint genetic identity in the group).

Using eq. (C.13) in the average direct fitness as-if given by eq. (C.8), average direct fitness as-if then satisfies eq. (C.1). To prove this result, it suffices to compare the right-hand side of eq. (C.8) by setting $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \boldsymbol{q}_s^{\mathrm{D}}(x_i, \bar{x})$ to the right-hand side of eq. (A.11), whereby we see that

$$W(x, x^*) = w_{\mathrm{I}}(x, \boldsymbol{x}^*, x^*), \tag{C.14}$$

which implies eq. (C.1) and where $\boldsymbol{x}^* = (x^*, ..., x^*)$ is an $N-1$-dimensional vector with each entry x^*. In other words, in an uninvadable population state individuals appear to maximize $w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \mathrm{E}_{\tilde{\boldsymbol{x}}_{-i} \sim \boldsymbol{q}^{\mathrm{D}}(x_i, \bar{x})}[w(x_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x})]$, which, when $N = 2$, reduces to $w_{\mathrm{I}}(x_i, x_{-i}, \bar{x}) = [1 - q(x_i, \bar{x})]w(x_i, x_{-i}, \bar{x}) + q(x_i, \bar{x})w(x_i, x_i, \bar{x})$, which is eq. (11) of the main text.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

C.3.2 Break-down of the individual-centered perspective

As explained in the main text section "A general individual-centered maximand?", eq. (C.9) is actually not a convincing as-if fitness as it entails that an individual controls the instrumental distribution describing the number of group neighbors expressing the same trait as self. Indeed, consider a class-structured population and suppose that in an uninvadable population state, an individual of class a maximizes either fitness as-if given by eq. (C.4) or by eq. (C.6). Then, we see that an individual of class a must also control the probability that it is of class a (since $\boldsymbol{\phi}(\tilde{x}_a, x^*)$ appears in both eq. C.4 and eq. C.6). But since the $\boldsymbol{\phi}$ contextual distribution is a population-level property that depends on the mutant trait, we cannot meaningfully view this distribution as under the control of a particular individual. For this reason, even for a panmictic population (no limited dispersal) we were unable to obtain a biologically convincing representation of uninvadability in terms of individual-centered maximization as long as the class distribution $\boldsymbol{\phi}$ depends on the mutant allele.

C.4 Individual maximands under weak selection

C.4.1 Weak selection

We now finally turn to deriving fitness as-if functions under weak selection (see section "Weak selection concepts" of the main text for an informal discussion). To take weak selection more formally into account, we let the matrix $\boldsymbol{A}(x, y)$ describing the growth of the mutant when rare in the population with classes (eq. A.8 with elements giving the expected number of groups in state $\mathbf{i}$ that descend from a group in state $\mathbf{j}$) be of the form

$$\boldsymbol{A}(x, y) = \boldsymbol{A}(y) + \epsilon \tilde{\boldsymbol{A}}(x, y) + O(\epsilon^2), \tag{C.15}$$

where matrix $\boldsymbol{A}(y)$ has leading positive eigenvalue equal to 1 and is independent of the mutant trait, $\tilde{\boldsymbol{A}}(x, y)$ is a matrix depending on both mutant and resident traits, and ϵ is a small parameter.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

The representation given in eq. (C.15) captures the two kinds of weak selection that we discussed in section "Weak selection concepts" of the main text (see also Box 3 therein). First, one can consider that the parameters determining both mutant and resident phenotypic effects are small. In this case of "small-mutation" selection, the matrix $\boldsymbol{A}(y)$ depends only the resident trait y and matrix $\tilde{\boldsymbol{A}}(x,y)$ is a first-order polynomial in mutant trait x. Second, one can consider traits affecting some material payoff (e.g., calory intake), or any other phenotypic feature, which itself affects only weakly a background reproduction and survival ("small-parameter" weak selection). For this case, matrix $\boldsymbol{A}(y) \to \boldsymbol{A}$ is actually independent of both mutant and resident traits and the perturbation matrix $\tilde{\boldsymbol{A}}(x,y)$ can take any form.

For weak selection, $\epsilon \to 0$ (e.g, Nagylaki, 1993; Lessard and Soares, 2016), the remainder $O(\epsilon^2)$ in eq. (C.15) is neglected and

$$q_{\mathbf{k}|a}(x,y) \to q_{\mathbf{k}|a}(y) \qquad \text{and} \qquad \phi_a(x,y) \to \phi_a(y), \tag{C.16}$$

where the left-hand sides depend at most on the resident traits and where $\mathbf{k}$ can describe either a haploid or diploid group state (in the latter case, $\mathbf{k}$ must account for heterozygotes and homozygotes within each class), and is independent of the evolving traits altogether under "small-parameter" weak selection. Eq. (C.16) follows from Lessard and Soares (2016, eqs. 59-67) who show that when $\epsilon \to 0$, the distribution over states of the mutant when rare in the population is described by the right unit eigenvector $\boldsymbol{u}(y)$ of $\boldsymbol{A}(y)$, and this vector subtends $q_{\mathbf{k}|a}(y)$ and $\phi_a(y)$ (e.g., $q_{\mathbf{k},a}(y) = k_a u_{\mathbf{k}}(y) / \left[\sum_{\mathbf{k}\in I}\sum_{a\in\mathcal{C}} k_a u_{\mathbf{k}}(y)\right]$, eq. A.21, eq. A.25 and explanations below eq. A.18 for the haploid case). Hence, not only the reproductive value $v_u(y)$ but also the genealogical and class structure no longer depend on mutant traits. In other words, the population-level properties may vary with the resident trait but are held constant on variation of the mutant trait. This argument also applies to the case of distinct individuals (remember Section A.1.2).

By collecting all components $q_{\mathbf{k}|a}(x,y)$ into the distribution $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ of genetic group states,

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

we have for weak selection that

$$\boldsymbol{q}_a^{\mathrm{D}}(x,y) \to \boldsymbol{q}_a^{\mathrm{D}}(y). \tag{C.17}$$

We will next apply eq. (C.17) to derive explicit expressions for fitness as-if under weak selection.

We are now ready to derive explicit as-if fitness representations. Fully endorsing weak selection, we denote from now on by $\tilde{w}_{ua}(x_i,\boldsymbol{x}_{-i},\bar{x})$ the weak-selection approximation of the class-specific fitness function $w_{ua}(x_i,\boldsymbol{x}_{-i},\bar{x})$ (as this covers both types of small-mutant and of small-parameter weak selection). We first recover two expressions for direct fitness as-if from the literature, which will allow us to point to limitations of this maximand, and finally to formalize inclusive fitness as-if.

C.4.2 Average direct fitness as-if: connection to previous results and limitations

First, recall that in the absence of class structure, we showed that individuals appear to maximize average direct fitness, eq. (C.8), and so, under weak selection, individuals will appear to maximize the weak selection version of average direct fitness:

$$w_{\mathrm{I}}(x_i,\boldsymbol{x}_{-i},\bar{x}) \;=\; \mathrm{E}_{\tilde{\boldsymbol{x}}_{-i}\sim\boldsymbol{\sigma}(x_i,\boldsymbol{x}_{-i},\bar{x})}\left[\tilde{w}(x_i,\tilde{\boldsymbol{x}}_{-i},\bar{x})\right]. \tag{C.18}$$

This was used as a fitness as-if in Lehmann et al. (2015, eqs. C-7-C.9), who further expressed individual fitness $\tilde{w}(x_i,\boldsymbol{x}_{-i},\bar{x})$ in terms of material payoff (e.g., energy intake), which was assumed to weakly affect the baseline vital rates (hence the focus was on "small parameter" weak selection). This allowed to consider more proximate components that individuals appear to be maximizing in an uninvadable population state.

Second, suppose we have a panmictic age-structured population with $\mathcal{C} = \{0,1,2,...\}$ and assume there are no effects of the traits expressed by an actor at any age on the fitness of that actor at later ages (no within-individual inter-class trait effects). Then, we can write the average

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

direct fitness as-if of an individual of age class a as

$$w_{\mathrm{I},a}(x_{a(i)}, \boldsymbol{x}_{-i}, \bar{x}) = v_0(\bar{x})\, \tilde{w}_{0a}(x_{a(i)}, \boldsymbol{x}_{-i}, \bar{x}) + v_{a+1}(\bar{x})\, \tilde{w}_{(a+1)a}(x_{a(i)}, \boldsymbol{x}_{-i}, \bar{x}), \tag{C.19}$$

where $\tilde{w}_{0a}(x_{a(i)}, \boldsymbol{x}_{-i}, \bar{x})$ is the number of newborns of an individual of age a and $\tilde{w}_{(a+1)a}(x_{a(i)}, \boldsymbol{x}_{-i}, \bar{x})$ its survival probability (thus eq. C.19 is equivalent to eq. 38 in Grafen, 2015). Since the $v_a(\bar{x})$'s do not depend on the behavior of the maximizer, this fitness as-if turns out to be a maximand that individuals appear to be maximizing in an uninvadable population state.

To prove this result, first note that the invasion fitness $W(\tilde{x}_a, x^*)$ for a mutant $\tilde{x}_a$ under the assumptions leading to eq. (C.19) can be written (by using eq. A.23) as

$$W(\tilde{x}_a, x^*) = \frac{1}{V(x^*, x^*)} \left[w_{\mathrm{T},a}(\tilde{x}_a, x^*)\phi_a(x^*) + \underbrace{\sum_{u \neq a} w_{\mathrm{T},u}(x^*, x^*)\phi_u(x^*)}_{k(x^*)} \right], \tag{C.20}$$

where $k(x^*)$ is a constant with respect to trait x_a. Second, comparing eq. (A.24) to eq. (C.19) (under the assumptions leading to eq. C.19) yields

$$w_{\mathrm{I},a}(x_a, \boldsymbol{x}^*_{-i}, x^*) = w_{\mathrm{T},a}(\tilde{x}_a, x^*), \tag{C.21}$$

and substituting this equation into eq. (C.20), shows that we can write fitness as-if as

$$w_{\mathrm{I},a}(x_a, \boldsymbol{x}^*_{-i}, x^*) = \frac{1}{\phi_a(x^*)} \left(V(x^*, x^*) W(\tilde{x}_a, x^*) - k(x^*) \right). \tag{C.22}$$

This shows that fitness as-if $w_{\mathrm{I},a}(x_a, \boldsymbol{x}^*_{-i}, x^*)$ is an affine function of invasion fitness $W(\tilde{x}_a, x^*)$, since the right-hand side of eq. (C.22) depends on x_a only through $W(\tilde{x}_a, x^*)$ and all other terms depend at most on x^*. Hence, the maximization of $w_{\mathrm{I},a}(x_a, \boldsymbol{x}^*_{-i}, x^*)$ with respect to x_a is equivalent to the maximization of $W(\tilde{x}_a, x^*)$ with respect to x_a, which itself returns the uninvadable population trait of class a (recall eq. C.3). Hence, in an uninvadable population state individuals of each class act as if they aimed to maximize their fitness as-if defined by eq. (C.19). This is an

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

individual-centered perspective of adaptation that also holds in the presence of social interactions, a case not covered in Grafen (2015).

Eq. (C.20) also points to the limitations of using average direct fitness as-if as a biologically meaningful individual-centered maximand. Indeed, in the presence of indirect fitness effects, where an actor of class a carrying the mutant allele affects the fitness of another individual carrying the mutant (say a worker affecting the reproduction of a queen), the terms for $u \neq a$ in expression (C.20) for invasion fitness will no longer be independent of the mutant trait x_a of a class-a individual. In this case, the as-if fitness of a class-a individual, also represented by eq. (C.20), also depends on these terms, and in particular on $w_{\mathrm{T},u}(\tilde{x}_a, x^*)$, even though the individual is not in any class $u \neq a$. Hence the biological interpretation of an individual of class a as an autonomous decision maker maximizing $w_{\mathrm{T},a}(\tilde{x}_a, x^*)$ at an evolutionary equilibrium, breaks down. To circumvent this problem, we now finally derive the inclusive fitness as-if of a class a individual (we note that for an age-structured panmictic population, the problem of defining a class-specific fitness as-if can also be circumvented by introducing a so-called *Hamiltonian* function as is done in life-history theory, e.g., Schaffer, 1982).

C.4.3 Inclusive fitness as-if

To construct inclusive fitness as-if for diploidy, we consider the same regression model as in the population genetic model (recall eqs. B.10–B.11) but we will evaluate its regression coefficients under a distribution of traits different from the population genetic model and therefore with different costs and benefits (minimizers of a sum of squares dependent on the distribution of traits). Namely, we focus on individual i with trait x_i in a group with neighbor trait profile $\boldsymbol{x}_{-i}$, and consider a hypothetical switch in behavior to expressing trait $\tilde{x}_i$ in a group with neighbor trait profile $\tilde{\boldsymbol{x}}_{-i}$, and write the fitness of individual i in the altered group as

$$\tilde{w}_{us}(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x}) = w_{us}(\bar{x}, (\bar{x}, ..., \bar{x}), \bar{x}) - c_{\mathrm{I},us}(x_i, \boldsymbol{x}_{-i}, \bar{x}) p_{\mathrm{I},i} + \sum_{a \in \mathcal{C}} b^{\mathrm{N}}_{\mathrm{I},us \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) p_{\mathrm{I},a(i)} + \text{residual}. \tag{C.23}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

The functional form of this equation is the same as eq. (B.10), but its interpretation differs and is as follows. We consider a group state where each of the gene copies from the original group of the focal individual i may be replaced in any individual by 2, 1, or 0 copies of an "I" allele; that is, where the new trait values $\tilde{x}_j$ of individual j of class a, distinct from the focal originally with trait value x_i, is within the set $\{z_a(x_{a(i)}, x_{a(j)}), x_{a(i)}, x_{a(j)}\}$, which stems, respectively, from individual j expressing 2, 1, or 0 copies of the "I" allele, with the total frequency of allele "I" in class a being $p_{\mathrm{I},a(i)}$. For the focal individual itself, the new trait value $\tilde{x}_i$ is within the set $\{z(x_i, \bar{x}), x_i, \bar{x}\}$, when it expresses 2, 1, or 0 copies of the "I" allele with $p_{\mathrm{I},i}$ being the frequency of allele "I" in individual i. Thus any value of $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})$ is a hypothetical group trait profile, resulting from a switch of allele expression and where, by construction, eq. (C.23) must hold for all $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}) \in \{z(x_i, \bar{x}), x_i, \bar{x}\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa})} \{z_a(x_{a(i)}, x_{a(j)}), x_{a(i)}, x_{a(j)}\}$. Hence, the main structural difference between eq. (C.23) and its population genetic counterpart, eq. (B.10), is that individuals are distinct in eq. (C.23).

The regression coefficients $c_{\mathrm{I},us}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ and $b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$, are now obtained by following the same line of argument as in the population genetic model. First, we note that eq. (C.23) says that we predict fitness with the same linear regression as in the population genetic model (recall eq. B.11):

$$\hat{w}_{us}\left(c_{\mathrm{I},us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I},us}, p_{\mathrm{I},i}, p_{\mathrm{I},a(i)}, \bar{x}\right) = w_{us}(\bar{x}, (\bar{x}, ..., \bar{x}), \bar{x}) - c_{\mathrm{I},us} p_i + \sum_{a\in\mathcal{C}} b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a} p_{a(i)}, \tag{C.24}$$

where $\boldsymbol{b}^{\mathrm{N}}_{\mathrm{I},us} = (b^{\mathrm{N}}_{\mathrm{I},us\leftarrow 1}, ..., b^{\mathrm{N}}_{\mathrm{I},us\leftarrow|\mathcal{C}|})$. Second, we denote by $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ and $\tilde{\boldsymbol{\sigma}}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ the distribution of group states in a population where the vector of frequencies of allele "I" across classes is $\boldsymbol{p}$ (the $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ and $\tilde{\boldsymbol{\sigma}}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ distributions have, respectively, support in $\{z(x_i, \bar{x}), x_i\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa})} \{z_a(x_{a(i)}, x_{a(j)}), x_{a(i)}, x_{a(j)}\}$ and $\{x_i, \bar{x}\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa}))} \{z_a(x_{a(i)}, x_{a(j)}), x_{a(i)}, x_{a(j)}\}$). This, in turn, allows us to define the quadratic expression

$$\begin{aligned} Q_{\mathrm{I},us}\left(c_{\mathrm{I},us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I},us}, x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}\right) &= \mathrm{E}_{(\tilde{x}_i,\tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_s(x_i,\boldsymbol{x}_{-i},\boldsymbol{p})}\left[\left(\hat{w}_{us}\left(c_{\mathrm{I},us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I},us}, p_{\mathrm{I},i}, p_{\mathrm{I},a(i)}, \bar{x}\right) - \tilde{w}_{us}(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}, \boldsymbol{p})\right)^2\right] p_s \\ &+ \mathrm{E}_{(\tilde{x}_i,\tilde{\boldsymbol{x}}_{-i})\sim\tilde{\boldsymbol{\sigma}}_s(x_i,\boldsymbol{x}_{-i},\boldsymbol{p})}\left[\left(\hat{w}_{us}\left(c_{\mathrm{I},us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I},us}, p_{\mathrm{I},i}, p_{\mathrm{I},a(i)}, \bar{x}\right) - \tilde{w}_{us}(\tilde{x}, \tilde{\boldsymbol{x}}_{-i}, \boldsymbol{p})\right)^2\right] (1 - p_s), \end{aligned} \tag{C.25}$$

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

and we solve $\partial Q_{\mathrm{I},us}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})/\partial c_{\mathrm{I},us} = 0$ and $\partial Q_{\mathrm{I},us}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})/\partial b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a} = 0$ for all $a \in \mathcal{C}$ for the corresponding regression coefficients.

Next, we assume that

$$\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) = \boldsymbol{q}^{\mathrm{D}}_s(\bar{x}, \boldsymbol{p}) \quad \text{and} \quad \tilde{\boldsymbol{\sigma}}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) = \tilde{\boldsymbol{q}}^{\mathrm{D}}_s(\bar{x}, \boldsymbol{p}), \tag{C.26}$$

so that the distribution of allele "I" across classes is the distribution of the mutant allele frequency under the true reproductive process. Recalling eq. (B.15), we obtain the regression coefficients of eq. (C.23) as

$$c_{\mathrm{I},us}(x, y) = \lim_{p\to 0} c_{\mathrm{I},us}(x, y, \boldsymbol{p}) \qquad b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a}(x, y) = \lim_{p\to 0} b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a}(x, y, \boldsymbol{p}). \tag{C.27}$$

We now further note that when $\boldsymbol{q}^{\mathrm{D}}_s(\bar{x}, \boldsymbol{p}) \to \boldsymbol{q}^{\mathrm{D}}_s(\bar{x})$ and $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) \to \boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$ as $p \to 0$, we have

$$\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \boldsymbol{q}^{\mathrm{D}}_s(\bar{x}), \tag{C.28}$$

and

$$\mathrm{E}_{(\tilde{x}, \tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})}[p_{\mathrm{I},i}] = \mathrm{E}_{(x_i, \boldsymbol{x}_{-i})\sim\boldsymbol{q}^{\mathrm{D}}_s(\bar{x})}[p_{\mathrm{I},i}] = r_{\mathrm{f},s}(\bar{x}), \tag{C.29}$$

where the first equality follows from eq. (C.28) and the definition of $p_{\mathrm{I},i}$, which takes exactly the same frequency as the mutant allele within an individual under assumption (C.28); while the second equality follows from the definition of the within-individual identity in state under neutrality (recall eq. B.16). Likewise, we have

$$\mathrm{E}_{(\tilde{x}, \tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})}\left[p_{a(i)}\right] = \mathrm{E}_{(x_i, \boldsymbol{x}_{-i})\sim\boldsymbol{q}^{\mathrm{D}}_s(\bar{x})}\left[p_{a(i)}\right] = r_{\mathrm{n},a|s}(\bar{x}), \tag{C.30}$$

where the first equality follows from eq. (C.28) and the definition of $p_{a(i)}$, while the second equality follows from the definition of the between-individual identity in state under neutrality

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

and the fact that the distribution of allele "I" is the same as that of the mutant allele (recall eq. B.17). In terms of the $r_{\mathrm{f},s}(\bar{x})$ and $r_{\mathrm{n},a|s}(\bar{x})$ probabilities, we have that

$$r_{s|a}(\bar{x}) = \frac{r_{\mathrm{n},s|a}(\bar{x})}{r_{\mathrm{f},a}(\bar{x})}, \tag{C.31}$$

which is the (neutral) relatedness coefficient between a class-s actor and a class-a recipient in a population monomorphic for trait value $\bar{x}$.

In terms of the so-obtained regression (eq. C.27) and relatedness (eq. C.31) coefficients, we can now define

$$b_{\mathrm{I},us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \frac{b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) n_s}{n_a} \tag{C.32}$$

and the inclusive fitness as-if of an individual of class a as

$$w_{\mathrm{I},a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \sum_{u\in\mathcal{C}} v_u(\bar{x}) \left[-c_{\mathrm{I},ua}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + \sum_{s\in\mathcal{C}} r_{s|a}(\bar{x})\, b_{\mathrm{I},us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) \right]. \tag{C.33}$$

With fitness as-if defined in this way, eq. (C.2) holds for all $\bar{x}$. In words: in an uninvadable population state an individual in any class will appear to be maximizing their inclusive fitness as-if, for all other individuals in the population doing the same.

This is the key result of our analysis and a proof follows from first noting that under weak selection, the inclusive fitness as-if defined by eq. (C.33) equals the inclusive fitness effect defined in eq. (B.29):

$$w_{\mathrm{I},a}((x_a, x^*_{-a(i)}), \boldsymbol{x}^*_i, x^*) = \Delta w_{\mathrm{IF},a}\left(\tilde{x}_a, x^*\right). \tag{C.34}$$

This equation holds since under weak selection $\phi_a(x,y) \to \phi_a(y)$, $r_{\mathrm{f},s}(x,y) \to r_{\mathrm{f},s}(y)$, and $r_{s|a}(x,y) \to r_{s|a}(y)$ (owing to eq. C.16), and further we have $c_{\mathrm{I},ua}(x, \boldsymbol{x}^*, x^*) = c_{ua}(x, x^*)$ and $b^{\mathrm{N}}_{\mathrm{I},us\leftarrow a}(x, \boldsymbol{x}^*, x^*) = b^{\mathrm{N}}_{us\leftarrow a}(x, x^*)$, since for indistinct neighbors with trait x^* of a focal individual with trait x_i the quadratic form for the individual-centered model (eq. C.25) under the assump-

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

tion eq. (C.26) is equivalent to the one of the gene-centered model (eq. B.12):

$$Q_{\mathrm{I},us}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x_i, x^*, \boldsymbol{p}) = Q_{us}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x_i, x^*, \boldsymbol{p}), \tag{C.35}$$

as all quantities are evaluated under the same population genetic distributions depending only on x^*. Finally, note that for weak selection, $r_{\mathrm{f},a}(x^*) = r_{\mathrm{f},a}(\tilde{x}_a, x^*)$ and $\phi_a(\tilde{x}_a, x^*) = \phi_a(x^*)$ in eq. (C.6), and so $w_{\mathrm{I},a}((x_a, x^*_{-a(i)}), \boldsymbol{x}^*_i, x^*)$ defined by eq. (C.33) and satisfying eq. (C.34) is an affine function of $W_{\mathrm{IF}}(\tilde{x}_a, x^*)$ and thus eq. (C.5) is satisfied.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

References

Akçay, E. and J. Van Cleve. 2016. There is no fitness but fitness and the lineage is its bearer. *Philosophical Transactions of the Royal Society B* 371:20150085.

Alcock, J. 2005. Animal Behavior: An Evolutionary Approach. Sinauer Associates, Massachusetts.

Alexander, R. D. 1990. Epigenetic rules and Darwinian algorithms: the adaptive study of learning and development. *Ethology and Sociobiology* 11:241–303.

Alipantris, C. D. and K. C. Border. 2006. Infinite Dimensional Analysis: a Hitchiker's Guide. Springer, Berlin, 3th edn.

Altenberg, L. and M. W. Feldman. 1987. Selection, generalized transmission and the evolution of modifier genes. I. The reduction principle. *Genetics* 117:559–572.

Binmore, K. 2007. Playing for Real: a Text on Game Theory. Oxford University Press, Oxford.

Birch, J. 2017. The inclusive fitness controversy: finding a way forward. *Royal Society Open Science* pp. 1–12.

Bourke, A. 2011. Principles of Social Evolution. Oxford University Press, Oxford.

Bourke, A. 2014. Hamilton's rule and the causes of social evolution. *Proceedings of the Royal Society B-Biological Sciences* 369:20130362.

Bürger, R. 2000. The Mathematical Theory of Selection, Recombination, and Mutation. John Wiley and Sons, New York.

Buss, D. M. 2005. The Handbook of Evolutionary Psychology. Wiley, New Jersey.

Caswell, H. 2000. Matrix Population Models. Sinauer Associates, Massachusetts.

Charlesworth, B. 1994. Evolution in Age-Structured Populations. Cambridge University Press, Cambridge, 2th edn.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Cox, D. R. and N. Wermuth. 1996. Multivariate Dependencies: Models, Analysis and Interpretation. Chapman and Hall, London.

Crespi, B. J. 2014. The insectan apes. *Human Nature* 25:6–27.

Damour, T. 2016. Gravitational waves and binary black holes. In *Ondes Gravitationnelles, Séminaire Poincaré XXII*, pp. 1–51.

Darwin, C. R. 1859. On the Origin of Species by Means of Natural Selection: or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

Davies, N. B., J. R. Krebs, and S. A. West. 2012. An Introduction to Behavioural Ecology. Wiley-Blackwell, New Jersey, 5th edn.

Dawkins, R. 1976. The Selfish Gene. Oxford University Press, Oxford.

Dawkins, R. 1978. Replicator selection and the extended phenotype. *Zeitschrift für Tierpsychologie* 47:61–76.

Dawkins, R. 1979. Twelve misunderstandings of kin selection. *Tierpsychogie* 51.

Dawkins, R. 1982. The Extended Phenotype. Oxford University Press, Oxford.

Dawkins, R. 1996. Climbing Mount Improbable. W. W. Norton, New York.

Day, T. and P. D. Taylor. 1996. Evolutionarily stable versus fitness maximizing life histories under frequency-dependent selections. *Proceedings of the Royal Society B: Biological Sciences* 263:333–338.

Day, T. and P. D. Taylor. 1998. Unifying genetic and game theoretic models of kin selection for continuous traits. *Journal of Theoretical Biology* 194:391–407.

Dobson, F. S., V. A. Viblanc, C. M. Arnaud, and J. O. Muries. 2012. Kin selection in Columbian ground squirrels: direct and indirect fitness benefits. *Molecular Ecology* 21:524–531.

Eshel, I. 1983. Evolutionary and continuous stability. *Journal ot Theoretical Biology* 103:99–111.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Eshel, I. 1991. Game theory and population dynamics in complex genetical systems: the role of sex in short term and in long term evolution. In Selten, R. (ed.), *Game equilibrium models I*, pp. 6–28. Springer.

Eshel, I. 1996. On the changing concept of evolutionary population stability as a reflection of a changing point of view in the quantitative theory of evolution. *Journal of Mathematical Biology* 34:485–510.

Eshel, I. 2019. Mutual altruism and long-term optimization of the inclusive fitness in multilocus genetic systems. *Theoretical Population Biology* In press.

Eshel, I., M. Feldman, and A. Bergman. 1998. Long-term evolution, short-term evolution, and population genetic theory. *Journal of Theoretical Biology* 191:391–396.

Eshel, I. and M. W. Feldman. 1984. Initial increase of new mutants and some continuity properties of ESS in two-locus systems. *The American Naturalist* 124:631–640.

Ewens, W. J. 2004. Mathematical Population Genetics. Springer-Verlag, New York.

Ewens, W. J. 2011. What is the gene trying to do? *The British Journal for the Philosophy of Science* 62:155–176.

Ferrière, R. and M. Gatto. 1995. Lyapunov exponents and the mathematics of invasion in oscillatory or chaotic populations. *Theoretical Population Biology* 48:126–171.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Frank, S. A. 1997. The Price equation, Fisher's fundamental theorem, kin selection, and causal analysis. *Evolution* 51:1712–1729.

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

Fudenberg, D. and J. Tirole. 1991. Game Theory. MIT Press, Massachusetts.

Gardner, A., S. A. West, and G. Wild. 2011. The genetical theory of kin selection. *Journal of Evolutionary Biology* 24:1020–1043.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Gillespie, J. H. 2004. Population Genetics: a Concise Guide. Johns Hopkins University Press, Baltimore, Maryland.

Grafen, A. 1984. Natural selection, kin selection and group selection. In Krebs, J. R. and N. Davies (eds.), *Behavioural Ecology: An Evolutionary Approach*, pp. 62–84. Blackwell Scientific Publications.

Grafen, A. 1985. A geometric view of relatedness. In Dawkins, R. and M. Ridley (eds.), *Oxford Surveys in Evolutionary Biology*, pp. 28–90. Oxford University Press, Oxford.

Grafen, A. 1988. On the uses of data on lifetime reproductive sucess. In Clutton-Brock, T. H. (ed.), *Reproductive Sucess*, pp. 454–471. The University of Chicago Press, Chicago.

Grafen, A. 2006a. Optimization of inclusive fitness. *Journal of Theoretical Biology* 238:541–563.

Grafen, A. 2006b. A theory of Fisher's reproductive value. *Journal of Mathematical Biology* 53:15–60.

Grafen, A. 2007. The formal Darwinism project: a mid-term report. *Journal of Evolutionary Biology* 20:1243–1254.

Grafen, A. 2008. The simplest formal argument for fitness optimization. *Journal of Genetics* 87:421–33.

Grafen, A. 2015. Biological fitness and the fundamental theorem of natural selection. *American Naturalist* 186:1–14.

Haig, D. 1997a. Parental antagonism, relatedness asymmetries, and genomic imprinting. *Proceedings of the Royal Society of London Series B-Biological Sciences* 264:1657–1662.

Haig, D. 1997b. The social gene. In Krebs, J. R. and N. Davies (eds.), *Behavioural Ecology: an Evolutionary Approach*, chap. 12, pp. 284–304. Blackwell, Oxford, 4th edn.

Haig, D. 2012. The strategic gene. *Biology and Philosophy* 27:461–479.

Hamilton, W. D. 1963. The evolution of altruistic behavior. *American Naturalist* pp. 354–356.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Hamilton, W. D. 1964. The genetical evolution of social behaviour, 1. *Journal of Theoretical Biology* 7:1–16.

Hamilton, W. D. 1970. Selfish and spiteful behavior in an evolutionary model. *Nature* 228:1218–1220.

Hamilton, W. D. 1996. Narrow Roads of Gene Land: Evolution of Social Behavior. W. H. Freeman and Company, New York.

Hamilton, W. D. and R. M. May. 1977. Dispersal in stable habitats. *Nature* 269:578–581.

Hammerstein, P. 1996. Darwinian adaptation, population genetics and the streetcar theory of evolution. *Journal of Mathematical Biology* 34:511–532.

Hartl, D. and A. G. Clark. 2007. Principles of Population Genetics. Sinauer, Massachusetts, 4th edn.

Hines, W. G. S. and J. Maynard Smith. 1978. Games between relatives. *Journal of Theoretical Biology* 79:19–30.

Kawecki, T. J. 1993. Age and size at maturity in a patchy environment: fitness maximization versus evolutionary stability. *Oikos* 66:309–317.

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

Krakauer, A. H. 2005. Kin selection and cooperative courtship in wild turkeys. *Nature* 434:69–72.

Lehmann, L., I. Alger, and J. W. Weibull. 2015. Does evolution lead to maximizing behavior? *Evolution* 69:1858–1873.

Lehmann, L., C. Mullon, E. Akçay, and J. Van Cleve. 2016. Invasion fitness, inclusive fitness, and reproductive numbers in heterogeneous populations. *Evolution* 70:1689–1702.

Lehmann, L. and F. Rousset. 2014a. Fitness, inclusive fitness, and optimization. *Biology and Philosophy* 29:181–195.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Lehmann, L. and F. Rousset. 2014b. The genetical theory of social behaviour. *Philosophical Transactions of the Royal Society B* 369:20130357.

Lessard, S. and C. Soares. 2016. Definitions of fitness in age-structured populations: Comparison in the haploid case. *Journal of Theoretical Biology* 391:65–73.

Liberman, U. 1988. External stability and ESS: criteria for initial increase of a new mutant allele. *Journal of Mathematical Biology* 26:477–485.

Luce, R. D. and H. Raiffa. 1957. Games and Decisions. John Wiley and Sons, New York.

Macke, E., S. Magalhães, F. Bach, and I. Olivieri. 2011. Experimental evolution of reduced sex ratio adjustment under local mate competition. *Science* 334:1127–1129.

Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. Microeconomic Theory. Oxford Unversity Press, Oxford.

Maynard Smith, J. 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.

McNamara, J., A. I. Houston, and E. J. Collins. 2001. Optimality models in behavioral ecology. *SIAM Review* 43:413–466.

McPeek, M. A. 2017. The ecological dynamics of natural selection: traits and the coevolution of community structure. *American Naturalist* 189:E91–E117.

Mesterton-Gibbons, M. 1996. On the war of attrition and other games among kin. *Journal of Mathematical Biology* 34:253–270.

Metz, J. A. J. 2011. Thoughts on the geometry of meso-evolution: collecting mathematical elements for a post-modern synthesis. In Chalub, F. A. C. C. and J. Rodrigues (eds.), *The mathematics of Darwin's legacy*, Mathematics and biosciences in interaction. Birkhäuser, Basel.

Metz, J. A. J., S. D. Mylius, and O. Diekmann. 2008a. Even in the odd cases when evolution optimizes, unrelated population dynamical details may shine through in the ESS. *Evolutionary Ecology Research* 10:655–666.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Metz, J. A. J., S. D. Mylius, and O. Diekmann. 2008b. When does evolution optimize? *Evolutionary Ecology Research* 10:629–654.

Metz, J. A. J., R. M. Nisbet, and S. A. H. Geritz. 1992. How should we define fitness for general ecological scenarios? *Trends in Ecology and Evolution* 7:198–202.

Michod, R. E. 1982. The theory of kin selection. *Annual Review of Ecology and Systematics* 13:23–55.

Moran, P. A. P. 1964. On the nonexistence of adaptive topographies. *Annals of Human Genetics* 27:383–393.

Mullon, C., L. Keller, and L. Lehmann. 2016. Evolutionary stability of jointly evolving traits in subdivided populations. *American Naturalist* 188:175–195.

Nagylaki, T. 1992. Introduction to population genetics. Springer-Verlag, Heidelberg.

Nagylaki, T. 1993. The evolution of multilocus systems under weak selection. *Genetics* 134:627–647.

Okasha, S. and J. Martens. 2016. Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *Journal of Evolutionary Biology* 29:473–482.

Okun, L. 2012. ABC of Physics: a Very Brief Guide. World Scientific, London.

Otto, S. P. and T. Day. 2007. A biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press, Princeton, NJ.

Parker, G. A. and J. Maynard Smith. 1990. Optimality theory in evolutionary biology. *Science* 349:27–33.

Price, G. R. 1970. Selection and covariance. *Nature* 227:520–521.

Queller, D. C. 1992. A general model for kin selection. *Evolution* 46:376–380.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Reeve, H. K. and P. W. Sherman. 1993. Adaptation and the goal of evolutionary research. *The Quarterly Review of Biology* 68:1–32.

Rousset, F. 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton, NJ.

Rousset, F. 2015. Regression, least squares, and the general version of inclusive fitness. *Evolution* 69:2963–2970.

Rousset, F. and S. Billiard. 2000. A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. *Journal of Evolutionary Biology* 13:814–825.

Rousset, F. and O. Ronce. 2004. Inclusive fitness for traits affecting metapopulation demography. *Theoretical Population Biology* 65:127–141.

Roze, D. 2009. Diploidy, population structure and the evolution of recombination. *American Naturalist* S1:79–94.

Savage, L. J. 1954. The Foundations of Statistics. John Wiley and Sons, New York.

Schaffer, W. M. 1982. The application of optimal control theory to the general life history problem. *American Naturalist* 121:418–431.

Szulkin, M., K. V. Stopher, J. Pemberton, and J. M. Reid. 2013. Inbreeding avoidance, tolerance, or preference in animals? *Trends in Ecology & Evolution* 28:205–11.

Taylor, P. 1990. Allele-frequency change in a class-structured population. *American Naturalist* 135:95–106.

Trivers, R. L. and H. Hare. 1976. Haplodiploidy and the evolution of the social insects. *Science* 191:249–263.

Tuljapurkar, S. 1989. An uncertain life: demography in random environments. *Theoretical Population Biology* 35:227–94.

bioRxiv preprint first posted online May. 2, 2019; doi: http://dx.doi.org/10.1101/624775. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Wakker, K. F. 2015. Fundamentals of Astrodynamics. TU Delft Repository, Delft.

Weissing, F. J. 1996. Genetic versus phenotypic models of selection: can genetics be neglected in a long-term perspective? *Journal of Mathematical Biology* 34:533–555.

West, S. A. and A. Gardner. 2013. Adaptation and inclusive fitness. *Current Biology* 23:577–584.

Williams. 1966. Adaptation and Natural Selection. Princeton University Press, Princeton, NJ.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.