When do individuals maximize their inclusive fitness?

Laurent Lehmann* and François Rousset†

2019-10-02

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. We demonstrate that the maximization (in the best-response sense) of class-specific inclusive fitness obtains in uninvadable population states (meaning that all deviating mutant go extinct). This defines a genuine actor-centered perspective on adaptation. But this inclusive fitness is assigned to all bearers of a mutant allele in a given class and depends on distributions of demographic and genetic contexts. These distributions, in turn, usually depend on events in previous generations and are thus not under individual control. This prevents, in general, from envisioning individuals themselves as autonomous fitness-maximizers, each with its own inclusive fitness. For weak selection, however, the dependence on earlier events can be neglected. We then show that each individual in each class appears to maximize its own inclusive fitness when all other individuals exhibit fitness-maximizing behavior. This defines a genuine individual-centered perspective of adaptation and justifies formally, as a first-order approximation, the long-heralded view of individuals appearing to maximize their own inclusive fitness.

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

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. This causes organisms to become adapted to their

*Department of Ecology and Evolution, University of Lausanne, Switzerland.
†University of Montpellier, CNRS, IRD, IHPE, Institut des Sciences de l'Évolution, France.

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 changes in trait composition in the population induced by natural selection. 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. Indeed, 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 a particular environment if no alternative trait combination or behavioral schedule can be produced by mutation, which would result in further allele frequency change (Fisher, 1930; Williams, 1966; Grafen, 1988; Reeve and Sherman, 1993; Dawkins, 1996). In this long-term perspective, the maximization of the geometric growth ratio of a mutant allele when rare–referred to here as *invasion fitness*–in a large population where individuals express a resident allele, provides a condition of uninvadability of mutant traits (all deviating mutants from some feasible set of traits go extinct). Uninvadability 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). It is in terms of this notion of best-response that maximization of (invasion) fitness can actually be conceived in the long-term evolutionary perspective and this holds regardless of the underlying genetic details (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. This 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). It has even been argued that any focus on individual survival and reproduction to understand adaptation is misleading and should be abandoned altogether (Dawkins, 1978). The gene-centered perspective of adaptation (Hamilton, 1963, 1996; Dawkins, 1976, 1982; Haig, 1997b, 2012) has in fact distanced itself from ideas of maximization of individual survival and reproduction long ago, and focuses instead on the differential transmission of alleles to understand adaptation.

In spite of the logical primacy of the gene-centered perspective, trying to interpret natural selection in individuals terms appears necessary for anyone observing individuals rather than genes. 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. While this inclusive fitness is not a distinct property of each individual in a population, individual inclusive fitness maximization has nevertheless 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 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 related individuals.

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 objective function (i.e., 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 appears on our reading not to cover social interactions in any broad sense. It has also been shown that, in age-structured population without social interactions and in group-structured populations with social interactions, individuals appear to maximize a weighted average individual fitness (respectively Grafen, 2015 and Lehmann et al., 2015), which is distinct from inclusive fitness. More generally, connections between individual fitness, inclusive 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 these discussions often do not emphasize enough the distinction between the gene-centered and the individual-centered perspective of adaptation, and generally do not cover 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 inclusive fitness in an uninvadable population state. To that aim, we use the common framework of evolutionary invasion analysis and proceed as follows in four steps. We start by presenting a model of evolution in a diploid group-structured population with limited dispersal and class-structure within groups. In this model, we then consider three perspectives on adaptation, formalized by three fitness measures, all of which are maximized (in the best-response sense) in an uninvadable population state. First, we consider a "recipient-centered perspective", which is formalized as a weighted mean of the fitness of individuals who bear an allele. Second, we consider an "actor-centered perspective", which is formalized as an exact class-specific version of inclusive fitness. This maximand is assigned to all bearers of an allele in a given class. Finally, we consider a "rational actor-centered perspective". This is the only perspective that fully captures

the idea that individuals maximize *their own* inclusive fitness, and is formalized as a class-specific inclusive fitness that can be assigned to individuals without knowing their genotype.

By formalizing these three perspectives on adaptation and discussing similarities and differences thereof, we unify and extend previous results on the relationship between maximization 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 full connection between evolutionary invasion analysis, the different perspectives on inclusive fitness theory, and game-theoretic approaches. Readers who find the technical details in the following text challenging may nevertheless have a look at the expression for inclusive fitness (eq. 4) and read the Discussion. Therein, results are summarized with a number of take-home messages about the interpretation of adaptation in terms of fitness maximization. Finally, readers interested in how selection on traits can be decomposed into direct effect on actors and indirect effects on recipients, but not on whether individuals do really maximize their own inclusive fitness, should read the section "The actor-centered perspective of adaptation", but can skip the more complicated section "The rational actor-centered perspective of adaptation".

The model

Assumptions

In order to formalize and compare the different perspectives on 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 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 individuals in a group belongs to a class (e.g., queen and worker, male or female, young or old) and n_{c} denotes the number of classes, which is assumed fixed and finite. The number of individuals in each class can differ within a group (but this difference is the same for all groups).

Each individual in each group can express a class-specific trait that affects its own survival, reproduction, and dispersal and possibly those of group neighbors. Let us focus on a focal group where individuals are labelled from 1 to n and let x_i denote the trait expressed by focal individual i from that group. When this individual is of class a, its trait is taken from the set $\mathcal{X}_a$ of feasible traits available to an individual of class a (i.e., $x_i \in \mathcal{X}_a$). We collect the class-specific traits

expressed by all of the $n-1$ neighbors of i into the vector $\boldsymbol{x}_{-i}$ of group-neighbor traits. We then let $w_{ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ denote the expected number of surviving class-u offspring *per haplogenome* produced over a demographic time period by a class-a individual i with trait $x_i \in \mathcal{X}_a$ when group neighbors have trait profile $\boldsymbol{x}_{-i}$ in a population where the average individual trait profile across classes is $\bar{x} \in \mathcal{X}$ (here $\mathcal{X}$ denotes the set of feasible traits across all classes; see Appendix A, eq. A.3, for a formal definition of fitness 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 thus gives the average number of replicate gene copies produced by an individual per homologous gene and so the unit of measurement is here gene copy number. In writing individual fitness as $w_{ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ we made, for simplicity of analysis and presentation, a number of assumptions on fitness. First, individual fitness depends here on the focal's own phenotype only through the trait of the class in which the focal resides. This thus excludes interdependence among own traits throughout the lifespan of the focal when it can change class (such as in an age-structured population when traits affect body size and thus may have effects on different age classes). Second, the focal's individual fitness depends on the trait of each neighbor only through the neighbor's class-specific trait. This thus also excludes interdependence among traits throughout the lifespan of neighbors that can change class (such as in an age-structured population when traits affect body size and transfer of resources among individuals depends on body such size). Finally, 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. While making these assumptions on fitness do not allow us to cover all biological scenarios of interest (see section "Scope of our results" for a discussion of the restrictive assumptions of our model and how to relax them), our model does cover the evolution of traits of arbitrary complexity in disjoint classes (i.e. when the class of an individual is fixed at birth), like workers and queens or males and females. As such, our model applies in broad generality to the paradigmatic biological scenarios for which inclusive fitness theory has been initially developed.

Evolutionary assumptions

No assumptions so far have been made on the genetic composition of the population and each individual may express a different trait and thus be phenotypically distinct from any other individual. In order 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). Accordingly, 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 $y = (y_1, y_2, ..., y_{n_c}) \in \mathcal{X}$ of traits, one for each class of individuals, where y_a is the trait of a (homozygote) individual of class a. Let a heterozygote mutant individual have trait vector $x = (x_1, x_2, ..., x_{n_c}) \in \mathcal{X}$. 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 $z_a(x_a, y_a)$ of mutant homozygotes of class a as a function of the traits x_a of heterozygote and y_a of resident homozygotes of that class (Appendix eqs. A.13–A.14 for a formal definition). This also covers the haploid case where we simply assign trait x to mutant and trait y to residents. So, for both haploid and diploid populations, traits are fully specified by x and y.

Characterizing adaptation by way of uninvadability

With the assumptions of the last section, the fate (extinction or spread) of a mutant allele with heterozygote trait x introduced into a monomorphic resident class-structured population with homozygote trait y can be determined by the *invasion fitness* $W(x, y)$. This is the geometric growth ratio of the mutant allele when rare. Henceforth, the mutant allele cannot invade the population when

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

Suppose now that a given resident trait, say $x^* = (x_1^*, x_2^*, ..., x_{n_c}^*)$, is uninvadable. Namely, it is resistant to invasion by any alternative trait from the set of all possible traits $\mathcal{X}$ (i.e., eq. 1 holds for any mutant given the resident x^*). This trait x^* must then be a best response to itself, meaning that if we vary invasion fitness $W(x, x^*)$ by varying x, the uninvadable trait x^* must maximize invasion fitness with respect to $x \in \mathcal{X}$. This means that x^* results in the highest invasion fitness among all alternatives traits given in the set $\mathcal{X}$ of feasible traits, for the resident population at the uninvadable state. 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" with "fitness" being invasion fitness and "a specified set of variants" being $\mathcal{X}$. This definition of adaptation is useful for two reasons. First, it has a direct formal encapsulation as uninvadability. Second, it captures the apparent purposefulness of traits as an outcome of cumulative natural selection, which has forcefully been argued as being the defining feature of adaptation (Fisher, 1930; Williams, 1966; Grafen, 1988; Dawkins, 1996).

The above characterization of adaptation is gene-centered, since it is obtained in terms of the fitness of mutant alleles. In order to assess whether individuals maximize their inclusive fitness, however, we must characterize adaptation in terms of individual-centered concepts. For that purpose, we contrast three individual-centered perspectives on adaptation:

(1) The recipient-centered perspective. Here, we provide a representation of invasion fitness expressed in terms of the class-specific fitness components (the w_{ua} functions introduced above) of a typical carrier of the mutant allele. This perspective directly flows out from the evolutionary model and is recipient-centered because we consider the fitness of individuals bearing the mutant allele and how this fitness is affected by the behavior of others. Hence, behavioral effects are grouped by recipients.

(2) The actor-centered perspective. Behavioral effects, however, are the outcomes of actors expressing traits. We thus ask whether it is possible to obtain a class-specific representation of fitness that is maximized (in the best-response sense) at uninvadability and where behavioral effects are grouped by the actor. That is, can we define a (class-specific) inclusive fitness that is maximized for each class in an uninvadable population state?

(3) The rational actor-centered perspective. Here, we go one step further than the actor-centered perspective. We ask whether it is possible to obtain an actor-centered representation of fitness where the trait of each individual is different and where this fitness is still maximized at uninvadability, independently for each individual in each class. This is the (behavioral ecology) question of whether we can we look at adaptation as if each individual is an autonomous decision-maker having free choice of action and *as if* these actions are guided by a striving to maximize a measure of inclusive fitness.

Perspective (1) obtains when an allele's fitness is viewed as a (possibly weighted) mean of the fitness of individuals who bear the allele. It is therefore a straightforward translation of the gene-centered perspective in individual-centered terms; in the context of this paper, it is hardly distinguishable from the gene-centered perspective. Perspective (2) changes the focus from the own fitness of allele bearers to their accumulated fitness effects on different recipients. This is the classic inclusive-fitness interpretation introduced by Hamilton (1964). However, in that interpretation, a single value (the class-specific inclusive fitness) is still assigned to all bearers of the allele in a given class. By contrast, only perspective (3) may capture the idea of individuals maximizing *their* inclusive fitness, and thus it is perspective (3) that seems most often invoked in behavioral ecology to understand adaptation. We now develop each perspective in turn to assess their relative levels of generality. We formally describe these perspectives and all arguments subtending our analysis in an extensive Appendix, which fully details, generalizes, and discusses more in depth technical concepts. The Appendix also contains some additional material and references as our analysis connects to many different ideas and intertwined concepts (for instance inclusive fitness itself can be formalized in different ways, see Appendix B). The main text presents the key formalization and results of our analysis.

The recipient-centered perspective of adaptation

Average direct fitness

In Appendix A, we provide the relationship of invasion fitness to the individual fitness components w_{ua} by suitably averaging over the distribution of class states in which a carrier of the mutant allele can reside. We denote this distribution by the vector $\boldsymbol{\phi}(x,y)$ whose ath entry is the probability $\phi_a(x,y)$ that a random copy of the mutant allele finds itself in class a. Invasion fitness of the (heterozygote) mutant x in a resident y can then be written

$$W(x,y) = \frac{1}{V(x,y)} \sum_{a=1}^{n_c} w_{\mathrm{DF}a}(x,y)\phi_a(x,y), \tag{2}$$

where the *average direct fitness* $w_{\mathrm{DF}a}(x,y)$ of a class-a mutant (with "DF" standing for direct fitness) is a weighted average $\sum_{u=1}^{n_c} v_u(y) w_{\mathrm{DF}ua}(x,y)$, over descendants of class u, of the expected number $w_{\mathrm{DF}ua}(x,y)$ of u-type offspring produced (per haplogenome) by an individual of class a bearing a copy of the mutant allele; $v_u(y)$ is the neutral reproductive value of a gene copy in class u (i.e., the asymptotic contribution to the gene pool of a single class-u gene copy in a monomorphic resident population); and $V(x,y)$ 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 copy that is assigned the total offspring (reproductive) values of a resident gene copy. Hence, invasion fitness can be represented as the average direct fitness of a copy of the mutant allele relative to the average direct fitness this allele copy would have if it was assigned the individual fitness components of resident individuals.

We can express $w_{\mathrm{DF}ua}(x,y)$ as an average $\mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim \boldsymbol{q}_a^{\mathrm{D}}(x,y)}\Big[w_{ua}(x_i,\boldsymbol{x}_{-i},y)\Big]$ over the distribution $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ of group trait profiles $(x_i,\boldsymbol{x}_{-i})$ experienced by a mutant gene copy [in a diploid population in particular, the trait x_i of the carrier of that gene copy of class a can be either that of a homozygote or heterozygote ($x_i \in \{z_a(x_a,y_a),x_a\}$), while the trait x_j of any neighbor j of class s may takes values in $\{z_s(x_s,y_s),x_s,y_s\}$]. The distribution $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ accounts for correlated phenotypic effects within groups due to identity-by-descent experienced by a carrier of class a of the mutant allele. Hence, the $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ distribution captures kin selection effects experienced by a class-a individual, which occurs whenever the fecundity and survival of an individual is affected by the genetic trait expressed by one or more other individuals who are genetically related to the actor in a nonrandom way at the loci determining the trait (Michod, 1982, p. 4). Invasion fitness depends on the average direct fitness of each class and thus on the collection $\boldsymbol{q}^{\mathrm{D}}(x,y) = (\boldsymbol{q}_a^{\mathrm{D}}(x,y))_{a\in\mathcal{C}}$ of class-specific genotype distributions. In particular,

$$w_{\mathrm{DF}a}(x,y) = \sum_{u=1}^{n_c} v_u(y)\mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim \boldsymbol{q}_a^{\mathrm{D}}(x,y)}\Big[w_{ua}(x_i,\boldsymbol{x}_{-i},y)\Big], \tag{3}$$

which means that the average direct fitness of a class-a mutant is the neutral reproductive value-weighted average of the average, over the distribution of trait profiles $(x_i,\boldsymbol{x}_{-i})$ experienced by a mutant gene copy, of the expected number, given the group genotypes, of u-type offspring produced per haplogenome by an individual of class a bearing a random copy of the mutant allele (see Box 1 for an example of average direct fitness).

The importance of genetic contexts

The expression for invasion fitness $W(x,y)$ (eqs. 2–3) makes explicit that invasion fitness is the average individual fitnesses components w_{ua} over a distributions of demographic states (captured by $\boldsymbol{\phi}(x,y)$) and genetic states (captured by $\boldsymbol{q}^{\mathrm{D}}(x,y)$) of a copy of the mutant allele. Importantly, these two distributions depend on mutant and resident traits. 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 focus on gene replication epitomizes the gene-centered perspective of evolution. Accordingly, 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 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 (changes in w_{ua}), but also on changes in the probabilities of *contexts* (Kirkpatrick et al., 2002, p. 1728) in which the allele at a given locus can be found [as measured by $\boldsymbol{\phi}(x,y)$ and $\boldsymbol{q}^{\mathrm{D}}(x,y)$]. 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 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 survival or reproduction, 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 $\boldsymbol{\phi}(x,y)$ and $\boldsymbol{q}^{\mathrm{D}}(x,y)$ as the contextual distributions.

The actor-centered perspective of adaptation

Inclusive fitness

What is missing to understand how selection targets traits in the representations of invasion fitness given by average direct fitness, eqs (2)–(3), is twofold. First, it is a simple and intuitive quantification of the effect of limited dispersal (and thus kin selection), which summarizes the variation on fitness introduced by the distribution over genetic contexts (the $\boldsymbol{q}^{\mathrm{D}}(x,y)$ distribution). Second, it is a quantification of the contribution to fitness of trait expressions in the different classes of individuals who really contribute to allele transmission. For instance, in the social insect example described in Box 1, a male has positive individual fitness but its trait is not under selection, while a worker has zero individual fitness, but its trait is affected by selection. Then, can one identify the force of selection on the actor's trait in a fitness measure? In other words, we aim to find an *inclusive fitness* $w_{\mathrm{IF}a}(x,y)$ for a class-a carrier of the mutant allele in a resident y population, which, by varying the class-specific trait $x_a \in \mathcal{X}_a$ maximizes the function $w_{\mathrm{IF}a}$ with respect to that trait and for any actor class a in an uninvadable population state x^*. Hence, such maximization means that x_a^* results in the highest class-a inclusive fitness among all traits values in $\mathcal{X}_a$ in an uninvadable population at x^*.

In Appendix B, we show that

$$w_{\mathrm{IF}a}(x,y) = v_a(y) + \sum_{u=1}^{n_{\mathrm{c}}} v_u(y) \left[-c_{ua}(x,y) + \sum_{s=1}^{n_{\mathrm{c}}} r_{s|a}(x,y)\, b_{us\leftarrow a}(x,y) \right] \tag{4}$$

satisfies this maximization problem. Here, extending an established population genetics terminology, $-c_{ua}(x,y)$ is the *average effect* on the number of class-u mutant gene copies produced by a single class-a individual when expressing a copy of the mutant instead of the resident allele (originally, the average effect of an allele substitution on a quantitative trait; e.g., Fisher, 1941, Ewens, 2004, p. 63), and $b_{us\leftarrow a}(x,y)$ is the average effect on the expected number of class-u offspring produced (per haplogenome) by all class-s neighbors in a group, and stemming from a single class-a gene copy switching to expressing a copy of the mutant instead of the resident allele. These costs and benefits hold regardless of the number of group partners, and have been reached by using a multiplayer and class-specific generalization of the two-predictor regression of individual fitness used in the exact version of kin selection theory (Queller, 1992; Frank, 1997; Gardner et al., 2011; Rousset, 2015; see Appendix B for further considerations on regressions and inclusive fitness). As such, $-c_{ua}(x,y)$ and $b_{us\leftarrow a}(x,y)$ describe additive effects on fitness resulting from two distinct gene substitutions, as if each was independently brought up by mutation. Each such effect accounts for changes in individual fitness when everything else is held constant, in particular holding constant the average effects of interactions between individual traits (although those will be modified by an allelic substitution).

Finally, inclusive fitness (eq. 4) depends on

$$r_{s|a}(x,y) = \frac{r_{\mathrm{n}\,s|a}(x,y)}{r_{\mathrm{f}a}(x,y)}, \tag{5}$$

which is the relatedness between a class-a actor and a class-s recipient. Here, $r_{\mathrm{f}a}(x,y)$ is the probability that, conditional on an haplogenome in a focal individual of class a carrying the mutant allele (hence the subscript "f"), a randomly sampled homologous gene in that individual is mutant, and $r_{\mathrm{n}\,s|a}(x,y)$ 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 the subscript "n"). 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 $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ distribution) and as such summarizes the statistical effects of limited dispersal on mutant-mutant interactions.

The subunits of adaptation

Class-specific inclusive fitness $w_{\mathrm{IF}a}(x,y)$ is the reproductive value of a class-a individual augmented by the average effect of that individual switching to expressing a copy of the mutant allele on the reproductive-value weighted number of mutant gene copies produced by all recipients of its action(s). Thus all behavioral effects are grouped by actor in eq. (4) and we emphasize that this holds also within classes (see Appendix B for proofs and for details on the connection

to the neighbor-modulated formulation of inclusive fitness). Crucially, inclusive fitness is defined at the allele level and this is consistent with the original formulation of this concept (Hamilton, 1964, pp. 3-8). But our own formulation (eq. 4) extends it to class-structured populations, and further shows that it holds regardless of the complexity of the evolving trait and the strength of selection on the mutant allele.

The fundamental difference between average direct fitness $w_{\mathrm{DF}a}(x,y)$ (eq. 3) and inclusive fitness $w_{\mathrm{IF}a}(x,y)$ (eq. 4) 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 their own individual fitness and/or that of other individuals in the population (see the concrete social insect example in Box 2). As a result, in an uninvadable population state the inclusive fitness of each class appears to be maximized with respect to the class-specific x_a trait, while the average direct fitness $w_{\mathrm{DF}a}(x,y)$ does not.

Inclusive fitness also makes explicit that a fitness comparison is made between expressing or not the mutant allele, since this involves comparing successful number of gene copies gained and lost through behavioral interactions. Hence, inclusive fitness allows one to fasten attention on the pathways determining fitness 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, say sterile workers, which can thus be seen as contributing to the fitness of their gene lineage. The inclusive fitness formulation thus shifts attention from those individuals that are passive carriers of alleles to those individuals whose trait actively affects the transmission of alleles. In other words, the unit of adaptation is the gene (Dawkins, 1978, 1982; Haig, 2012), and its subunits are the replicate gene copies expressed differently in particular classes of individuals, whose traits can be regarded as the outcomes of maximizing a class-specific inclusive fitness.

The rational-actor perspective of adaptation

Fitness as-if

Class-specific inclusive fitness $w_{\mathrm{IF}a}(x,y)$ (eq. 4) seems satisfying from a population genetics point of view to understand selection on traits affecting relatives. Many behavioral ecologists, however, observe individual behavior instead of genes. What is then missing to understand how selection targets traits in the class-specific representation of inclusive fitness $w_{\mathrm{IF}a}(x,y)$ is a prediction of adaptive traits among the alternatives that an individual can potentially express, *given* the actions of social partners. The simplest and most widely used concept for the prediction of individual behavior is that of a Nash equilibrium trait profile, compared to which no individual 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). Here, an individual is envisioned as an autonomous decision-maker, "freely choosing" its action independently of each other individual and with a striving to maximize payoff. In a Nash equilibrium, the individual then makes the best decision for itself in terms of payoff, based on all others individuals making the best decisions for themselves. Our aim now is to find such a payoff for a focal individual of class a, denoted

$w_{\mathrm{I}a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$. Here we use the same notation as before for individual traits, but x_i is to be understood as any trait value that a given individual i could express instead of the (genetically determined) trait that it actually expresses. We refer to $w_{\mathrm{I}a}$ as the *fitness as-if* function of an individual of class a, because the individual acts *as if* it maximizes this function in an uninvadable state by choosing the appropriate trait value $x_i \in \mathcal{X}_a$ (see eq. C.1 and Appendix C for more details on the construction of fitness as-if).

The previously considered invasion fitness, class-specific average direct fitness, and class-specific inclusive fitness do not fulfill the role of a fitness as-if since they depend on mutant and resident traits. The fundamental difference between fitness as-if $w_{\mathrm{I}a}$ maximization and, say inclusive fitness $w_{\mathrm{IF}a}$ maximization, is that each individual can vary its own trait in fitness as-if, independently of each other individual. By contrast, inclusive fitness is maximized with respect to a mutant trait and this implies that it takes into account correlated changes in the traits expressed by co-bearers. In this respect, the previous perspectives were all gene-centered, while the rational-actor perspective is distinctly individual-centered. In reaching a definition of a fitness as-if, we will still need to take correlated trait changes into account, by adjusting the function definition rather than by allowing its argument $\boldsymbol{x}_{-i}$ to vary with x_i.

A general rational actor-centered maximand?

Appendix C shows that it is possible, in theory, to construct an inclusive fitness as-if maximand that individuals appear to maximize in an uninvadable population state (see eq. C.21). This maximand, however, requires that the contextual distributions are written in terms of the actor's trait and the average in the population. It is thus as if the actor *controlled* the genetic and demographic contexts it experiences. In reality, however, the contextual distributions cannot be under the actor's control. These distributions do not depend on the actor's behavior, but on the reproduction and survival of ancestors that determine the present genetic and demographic contexts. This precludes, in our understanding, a general rational actor-centered representation of adaptation.

If the contextual distributions, $\boldsymbol{\phi}(x, y)$ and $\boldsymbol{q}^{\mathrm{D}}(x, y)$, were to be independent of the mutant trait, then a fitness as-if could be constructed with these distributions exogenous to an individual's own behavior. There are least two ways to achieve exactly this and both hinge on *weak-selection* approximations implying that, to first-order, the distribution of genetic and class contexts will no longer be dependent on the mutant allele. Such first-order approximations are reached either by assuming that the phenotypic effects of the mutant is small ("small-mutation" weak selection, in which case the individual fitness functions w_{ua} are assumed differentiable), or that parameters determining both mutant and resident phenotypic effects on fitness are small ("small-parameter" weak selection). In both cases, the contextual distributions depend at most on the resident trait ($\boldsymbol{\phi}(x, y) \sim \boldsymbol{\phi}(y)$ and $\boldsymbol{q}^{\mathrm{D}}(x, y) \sim \boldsymbol{q}^{\mathrm{D}}(y)$, see Appendix C.5.1 for more details). The key implication is that under weak-selection a mutant allele will not affect the genealogical and/or demographic 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 to obtain

a representation of individual maximizing behavior in the absence of social interactions in age-structured populations (Grafen, 2015; Lessard and Soares, 2016) and in the presence of social interactions in group-structured populations (Lehmann et al., 2015). These two cases can be generalized by the results provided in Appendix C, and we now present a fitness as-if that takes the form of inclusive fitness.

Maximization of inclusive fitness as-if under weak selection

Assuming weak selection, we show in Appendix C that the inclusive fitness as-if function $w_{\mathrm{I}a}$ of a class a individual defined by

$$w_{\mathrm{I}a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = v_a(\bar{x}) + \sum_{u=1}^{n_{\mathrm{c}}} v_u(\bar{x}) \left[-c_{\mathrm{I}ua}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + \sum_{s=1}^{n_{\mathrm{c}}} r_{s|a}(\bar{x})\, b_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) \right] \tag{6}$$

is maximized in an uninvadable population state. In this inclusive fitness as-if, $-c_{\mathrm{I}ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is an additive marginal 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 an additive marginal effect of a single class-a individual on the number of class-u offspring produced by all class-s neighbors. These marginal effects, costs to self and benefits to others, 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 trait expression, whereby the effects of switching trait expression can be assessed. This allele is taken to have the same distribution as the mutant allele in the population-genetic model and ensures that the inclusive fitness as-if (eq. 6) of a class a individual aligns, at uninvadability, with the class-specific inclusive fitness of the population genetic model (eq. 4). The key difference between the cost $c_{\mathrm{I}ua}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ in the as-if and the cost $-c_{ua}(x, y)$ in the population-genetic model (recall eq. 4) is then that all individuals within groups have distinct traits in the rational-actor 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 allele is constant with respect to the actor's trait. Hence, it is equivalent to the standard relatedness in a monomorphic population with trait $\bar{x}$ (sometimes called pedigree relatedness) and is independent of actor genotype.

Eq. (4) provides an inclusive-fitness representation of fitness as-if that individuals from each class appear to maximize in an uninvadable population state (see Appendix C.5.2 for a proof of this result). It may be felt to be a stretch to define the inclusive fitness as-if, which is not a population genetic quantity. Such an elaborate construction may nevertheless be needed to assign a coherent meaning to the untold number of statements found in the literature that individuals maximize their own inclusive fitness (rather than the inclusive fitness, eq. (4), of an allele in the population genetic model). Of particular note here is that our construct can be assigned to individuals without knowing their genotype. On the other hand, its representation (eq. 6) still makes explicit that individuals in each class adjusting their behavior will strive to do so by maximizing their genetic contribution to the next generation and by treating others according

to the degree of genetic relatedness between actor and recipient.

Scope of our results

We have identified a rational actor-centered maximand, which individuals appear to maximize in uninvadable population states under weak selection (or if, for other reasons, the $\boldsymbol{\phi}$ and $\boldsymbol{q}^{\mathrm{D}}$ contextual distributions are independent of selection). This result, as well as the actor-centered maximand (eq. 4), 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), which describes common practices in theoretical evolutionary biology (e.g., Brown and Vincent, 1987; Taper and Case, 1992; Geritz et al., 1998; Dercole and Rinaldi, 2008; Lion, 2018), to 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 and thus result in environmental feedbacks on fitness. For weak selection, these distributions or values can then be approximated as a function of the resident traits (see Ronce et al., 2000; Rousset and Ronce, 2004 for concrete examples of fitness functions and distributions covering both demographic and environmental fluctuations, in particular examples where the number of individual in each class can fluctuate between groups, as well as examples where the total number of individuals within groups can fluctuate, respectively). 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, different number of individuals across classes in different groups), 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 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 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 (e.g., they can be of arbitrary dimension, combining both discrete and continuous trait values, be reaction norms), we assumed that there is no kin recognition or discrimination based on using genetic cues (most models in the literature do not allow for this case either). Yet, other forms of kin discrimination, such as different behaviors expressed toward

sisters, cousins, etc., are readily accounted by our results once sisters, cousins, etc., are recognized as different classes of actors. Our model also does not cover complicated conditional class-specific traits expressions, such as when an individual helps its mother as long as she is alive and upon her death starts to help its siblings. Including such biologically 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 with any finite number of genotypes coexisting in equilibrium. Here, previous results (Eshel and Feldman, 1984; Liberman, 1988; Eshel et al., 1998) suggest, again, that an extension of the number of contexts will allow to define a corresponding fitness as-if. In conclusion, all the above scenarios involve considering more complicated $\boldsymbol{\phi}$ and $\boldsymbol{q}^{\mathrm{D}}$ 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 to generalize both class-specific inclusive fitness (eq. (4)) and class-specific inclusive fitness as-if (eq. (4)), but are unlikely to alter our conclusions, since they all rely only on broadening the number of contexts (be it demographic or genetic).

Discussion

Summary and take-home messages

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 get a hold on the conditions under which individuals can be envisioned as maximizing their own inclusive fitness in an uninvadable population state (i.e., a population where all deviating mutants go extinct). In particular, we considered three individual-centered perspectives on adaptation (Fig. 1), and defined two class-specific inclusive fitness maximands. The first inclusive fitness maximand is assigned to bearers of a mutant allele, but here only identically to all bearers in a given class, rather than identically to all bearers of an allele as in Hamilton (1964). Our second maximand is assigned to each individual in the population separately, and is a generalisation of our class-specific inclusive fitness, that a behavioral ecologist can directly assign to any observed individual without knowing genotypes. Instead of depending on genotype, it has to depend on the traits expressed by all social partners. This "behavioral" inclusive fitness nevertheless reduces to the class-specific inclusive fitness of a mutant allele when the resident population is an uninvadable population state and under a weak-selection approximation. We thereby defined a rational actor-centered inclusive fitness that is maximized at an evolutionary equilibrium. Our formal analysis leads to the following main take-home messages.

(1) Actor-centered inclusive fitness maximization obtains generally. Uninvadable traits can be characterized in terms of mutant alleles attempting to maximize (in the best-response sense) their own transmission across generations. We showed that the traits

expressed by individuals in each class maximize class-specific inclusive fitness in an uninvadable population state (eq. 4). This provides a genuine actor-centered perspective of adaptation.

(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, the class-specific inclusive fitness is not the fitness of a single individual, but that of an average class-specific carrier of a mutant allele sampled from the distributions of genetic contexts it experiences.

(3) Rational actor-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, we showed that each individual in each class appears to maximize its own inclusive fitness (eq. 6) in an uninvadable population state. This provides a genuine individual-centered perspective of adaptation that can be assigned to an individual without knowing its genotype.

Point (1) follows 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. This logic can be transposed down at the class level so that adaptation in the long-term evolutionary perspective can be envisioned as the maximization of the class-specific inclusive fitness of a mutant allele that holds for any trait complexity and selection strength. This inclusive fitness consists of the class-specific reproductive value of an allele augmented by the inclusive fitness effect, which is a decomposition of the force of selection, in terms of direct and indirect effects on transmission of replica copies of this allele. The inclusive fitness effect in class-structured populations is commonly represented, to the first-order, as an average effect on the reproductive value weighted-number of offspring produced by a recipient of a given class (Taylor, 1990; Frank, 1998; Rousset, 2004). Our class-specific inclusive fitness, which collects effects by actors of a given class, better embodies the notion of class-specific inclusive fitness and holds generally. Further, since any (indirect) effect that an actor from a given class has on the survival and reproduction of a relative in another class is not an effect on the actor's own fitness, the maximization of class-specific average direct fitness is generally not compatible with uninvadability, in contrast to class-specific inclusive fitness.

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 the fitness effects under the control of an allele (the set of all copies of an allele) may include changes of class (demographic) and genetic contexts.

As such, inclusive fitness is a gene-centered fitness measure (consistent with Hamilton's original definition), whose components, like relatedness, cannot be under the control of a single individual. This precludes a general interpretation of uninvadability as individuals maximizing *their* inclusive fitness. 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. This warrants the view of individuals from different classes as autonomous decision-makers, each maximizing its own inclusive fitness.

We finally delineate two implications highlighted by our analysis.

Is it empirically expected that $-c + rb > 0$?

First, in any population that is subject to density-dependent regulation and that is in an uninvadable state, the average individual fitness, and the average inclusive fitness, are equal to one. In practice, populations will not be exactly at uninvadable states, but, when they depart from such states, they are not expected to depart in any consistent way in terms of expressing social behavior (e.g., the trait level expressed by individuals could be as well below or above the uninvadable one). Hence, in contrast to the hypothesis considered by Bourke (2014), we do not necessarily expect a tendency for the inclusive fitness effect ("$rb - c$", the difference between the baseline fitness of "1" and inclusive fitness) of a "more social" act to be positive even when kin selection operates through positive indirect effects ($rb > 0$). This argument applies to the class-specific inclusive fitness as well (the difference between $w_{\mathrm{IF}a}(x, y)$ and the reproductive value in eq. 4, which will be zero at uninvadability).

Bourke (2014) 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, $rb - c$, and more generally the class-specific inclusive fitness effect, should be negative for any mutant trait in an uninvadable population state. This is so, since by definition, the inclusive fitness effect 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 tends to vanish.

By contrast to the inclusive fitness effect ($rb - c$ in the absence of classes), 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, at an equilibrium, is not informative about the importance of kin selection, and the indirect fitness effect, which for all conditions quantifies how the force of

selection on a mutant depends on relatedness. Nevertheless, the inclusive fitness of a particular class of individuals (eq. 4) is itself informative about the importance of kin selection, since it assigns fitness contributions even to individuals that do not reproduce.

So, when do individuals maximize their inclusive fitness?

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, 1979, 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 this way, they should keep in mind the underlying weak-selection assumption (points 2 and 3), the assumption that the population needs to be close to an uninvadable state, and, the definition of inclusive fitness for which it holds (eq. 6). Namely, it is a function of the traits of an individual and of its social partners all assumed distinct, yet which coincides with the allelic inclusive fitness of an allele 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 do not involve any phenotypic interactions (Grafen, 2006a)[1], whereby ruling out social interactions in any broad sense.

While Fisher (1930) and Hamilton (1996, p.27–28) have 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). 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 and N. Raihani for useful comments on previous version of the manuscript. We gratefully acknowledge extensive and extremely useful comments on the paper by D. Bolnick, A. Grafen, S. Lion and an anonymous reviewer. These comments fundamentally improved the manuscript, the results as well as the presentation.

[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.

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).
$w_{us}(x_i, \boldsymbol{x}_{-i}, \bar{x})$	Expected number of surviving class-u offspring per haplogenome produced by an individual of class-s (possibly including self) over one demographic time period (eq. A.3).
$v_s(y)$	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).
$V(x,y)$	Average of the $v_u(y)$ reproductive values (eq. A.20).
$w_{\mathrm{DF}a}(x,y)$	Average direct fitness of a class-a mutant gene copy (eqs. A.24, A.29).
$w_{\mathrm{DF}ua}(x,y)$	Average (over the $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ distribution) of the expected number of u-type offspring produced (per haplogenome) by an individual of class a bearing a copy of the mutant allele (eq. A.29).
$w_{\mathrm{IF}a}(x,y)$	Inclusive fitness of a class-a mutant gene copy (eq. B.38).
$w_{\mathrm{I}a}(x, \boldsymbol{x}_{-i}, \bar{x})$	Fitness as-if of a class-a individual (eqs. C.10, C.28, C.34).
$\phi_s(x,y)$	Probability that a mutant gene copy resides in a class s individual (eq. A.21).
$\boldsymbol{\phi}(x,y)$	Distribution for $\phi_s(x,y)$.
$\boldsymbol{q}_a^{\mathrm{D}}(x,y)$	Conditional distribution of identity for a gene-copy in class a. The $\boldsymbol{q}_a^{\mathrm{D}}(x,y)$ distribution accounts for correlated phenotypic effects within groups due to identity-by-descent and thus captures the kin selection effects experienced by a class-a individual (eqs. A.10, A.17).
$r_{s\|a}(x,y)$	Conditional relatedness, with a class-s individual, of a gene copy taken in a class-a individual (eq. B.24). 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 identity with a gene copy randomly taken in a class-s individual, of a gene copy taken in a class-a individual (eq. B.24).
$r_{\mathrm{f}a}(x,y)$	Conditional identity with a gene copy randomly taken in a class-a individual, of a gene copy taken in that same individual (eq. B.24).
$c_{ua}(x,y)$	Average effect on its own fitness through class-u offspring of a gene substitution in a class-a individual (eq. B.14).
$b_{us\leftarrow a}(x,y)$	Average effect of a gene substitution in a single class-a individual on the fitness of all class-s recipients (eq. B.31) 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.13) 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.13) through class-u offspring.

Box 1: Social insect example. Let us thus consider a seasonal population of diploid social insects (e.g., termites rather than ants) 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 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 expected number of class-u offspring produced by a class-a mutant individual per haplogenome can be written $w_{\mathrm{DF}ua}(x,y)$, where trait $x = (x_{\mathrm{f}}, x_{\mathrm{m}}, x_{\mathrm{o}})$ collects, respectively, the traits of females, males, and workers. 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 in eq. (2) with the direct fitnesses given by

$$\begin{aligned} w_{\mathrm{DFf}}(x,y) &= v_{\mathrm{f}}(y) w_{\mathrm{DFff}}(x,y) + v_{\mathrm{m}}(y) w_{\mathrm{DFmf}}(x,y) \\ w_{\mathrm{DFm}}(x,y) &= v_{\mathrm{f}}(y) w_{\mathrm{DFfm}}(x,y) + v_{\mathrm{m}}(y) w_{\mathrm{DFmm}}(x,y) \\ w_{\mathrm{DFo}}(x,y) &= 0, \end{aligned} \tag{B.1}$$

and the distribution of class states in which a carrier of the mutant allele can reside being $\boldsymbol{\phi}(x,y) = (\phi_{\mathrm{f}}(x,y), \phi_{\mathrm{m}}(x,y), \phi_{\mathrm{o}}(x,y))$. 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.32) 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 $x = (x_{\mathrm{f}}, x_{\mathrm{m}}, x_{\mathrm{o}})$.

Box 2: Individual and inclusive fitness for social insect example. As an example of the individual fitnesses in eq. (B.1), 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

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

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.2) is for the case where the worker is heterozygote and the second when it is homozygote resident. The denominators in eq. (B.2) 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

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

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

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

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,

$$\begin{aligned} w_{\mathrm{IFf}}(x,y) &= v_{\mathrm{f}}(y) \\ w_{\mathrm{IFm}}(x,y) &= v_{\mathrm{m}}(y) \\ w_{\mathrm{IFo}}(x,y) &= v_{\mathrm{f}}(y)\left(\frac{P(x_{\mathrm{o}})-P(y_{\mathrm{o}})}{1+P(y_{\mathrm{o}})}\right) + v_{\mathrm{m}}(y)\left(\frac{P(x_{\mathrm{o}})-P(y_{\mathrm{o}})}{1+P(y_{\mathrm{o}})}\right) \end{aligned} \tag{B.5}$$

(see Appendix B.3 for a proof).

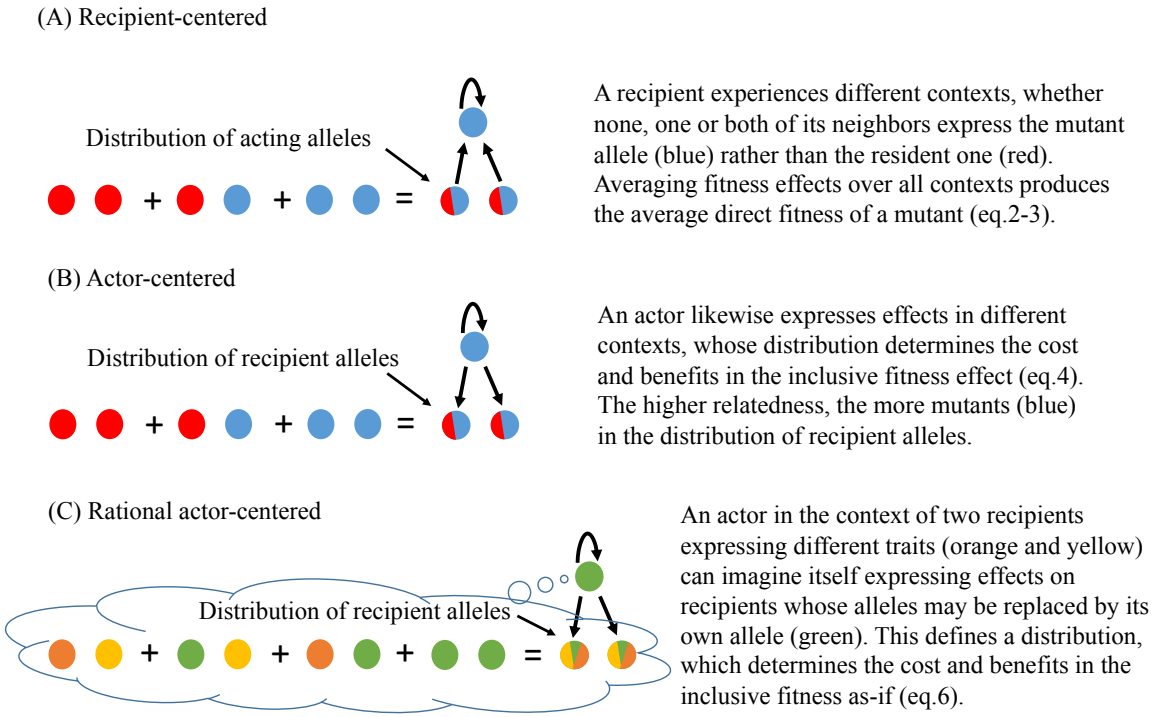

Figure 1: Three perspectives on adaptation

The three perspectives are here illustrated in groups of three haploid individuals, and ignoring any class structure. In the "recipient-centered" perspective (panel A), interactions between group members are described by arrows representing effects of group neighbors on the fitness of a focal recipient carrying a mutant allele. In the "actor-centered" perspective (panel B) and "rational-actor centered" perspective (Panel C), interactions between group members are described by arrows representing effects of a focal actor on the fitness of average group recipients.

Appendix

Contents

A Invasion fitness as average direct fitness **24**
A.1 Building blocks . . . 24
A.1.1 Individual fitness . . . 24
A.1.2 Distinct and indistinct individuals . . . 25
A.2 Average direct fitness without classes . . . 26
A.2.1 Haploids . . . 26
A.2.2 Diploids . . . 28
A.3 Average direct fitness with classes . . . 29
A.3.1 Haploids . . . 29
A.3.2 Diploids and social insects . . . 32

B Inclusive fitness **34**
B.1 Inclusive fitness for haploids without classes . . . 34
B.1.1 Regression with respect to neighbors . . . 35
B.1.2 Regression with respect to focal and neighbors . . . 35
B.1.3 Comparing single- and two-predictor regression . . . 37
B.2 Inclusive fitness for diploids with classes . . . 38
B.2.1 Multiplayer class-structured regression . . . 38
B.2.2 Average allele frequency . . . 40
B.2.3 Average inclusive fitness . . . 41
B.2.4 Grouping effects by actor . . . 42
B.2.5 Class-specific inclusive fitness maximization . . . 43
B.3 Example: inclusive fitness for social insects . . . 45
B.3.1 Regressions for female fitness components . . . 46
B.3.2 Regressions for male fitness components . . . 48
B.3.3 Inclusive fitness effects . . . 48

C Fitness as-if **49**
C.1 Rational-actor payoff maximization . . . 49
C.2 Average direct fitness as-if . . . 50
C.2.1 The instrumental distribution . . . 50
C.2.2 Haploids without classes . . . 50
C.2.3 Diploids with classes . . . 51
C.3 Inclusive fitness as-if . . . 53
C.3.1 Multiplayer regression for class-structure . . . 53
C.3.2 Actor-centered regression coefficients . . . 55
C.4 Connecting the gene-centered and the rational-actor centered perspectives . . . 58
C.5 Weak selection . . . 59
C.5.1 Weak-selection concepts . . . 59
C.5.2 Class-specific inclusive fitness as-if maximization under weak selection . . . 60

Appendix A: Invasion fitness as average direct fitness

In this Appendix, we derive the expressions for the average direct fitness given in the main text, eqs. (2)–(3). As explained therein, we limit our 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). In each group, there is a finite number of classes and we 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 $|\mathcal{C}| = n_c$), 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).

In such a population, a resident trait x^* is uninvadable if it is a best response to itself, meaning that if we can vary invasion fitness $W(x, x^*)$ by varying the mutant trait x in the set $\mathcal{X}$ of feasible traits,[2] an uninvadable trait must be a trait maximizing invasion fitness for the resident at x^* (formally invasion fitness is the function $W : \mathcal{X}^2 \to \mathbb{R}_+$ with $W(y, y) = 1$ for all $y \in \mathcal{X}$). It then follows that an uninvadable trait x^* satisfies

$$x^* \in \arg\max_{x\in\mathcal{X}} W(x, x^*), \tag{A.1}$$

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 x^* (in eq. A.1, x^* belongs to a set because an uninvadable trait is not necessarily unique).

Since (even assuming the island model) notations and concepts, to express the function W in terms of individual-centered components in the presence of class-structure and diploidy, become rapidly complicated, we will progressively introduce different cases (haploid, diploid, etc.), concepts, and notations. We start by defining formally the central building block of our analysis, which is individual fitness.

A.1 Building blocks

A.1.1 Individual fitness

In the absence of class-structure, we define the *individual fitness function* as

$$w : \mathcal{X} \times \mathcal{X}^{n-1} \times \mathcal{X} \to \mathbb{R}_+, \tag{A.2}$$

such that $w(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is the expected number of successful offspring produced (*per haplogenome*) by a focal individual $i \in \mathcal{I} = \{1, 2, .., n\}$ with trait $x_i \in \mathcal{X}$, where neighbors in group $\mathcal{I}$ have trait profile $\boldsymbol{x}_{-i} \in \mathcal{X}^{n-1}$ in a population where the average trait over all individuals is $\bar{x} \in \mathcal{X}$.

[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-valued traits, combination of thereof, etc.), and is the space for which general results concerning function maximization exists (Alipantris and Border, 2006, pp. 581-585).

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, and following the assumptions and explanations of the main text, the individual fitness functions is defined as

$$w_{us} : \mathcal{X}_s \times \prod_{a \in \mathcal{C}} \mathcal{X}_a^{n_a - \delta_{as}} \times \mathcal{X} \to \mathbb{R}_+ \quad \forall (u, s) \in \mathcal{C}^2, \tag{A.3}$$

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 individual $i \in \mathcal{I}$ when in class-s (*per haplogenome*) with trait $x_i \in \mathcal{X}_s$ 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 $n-1$ 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.2–A.3) 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 of all ordered trait profiles (i.e., all ordered group states are considered; for instance, in the absence of class-structure, this is $\mathcal{X}^n$). 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 characterization of the population, where one 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 states. Since neighbors are exchangeable, the individual fitness of an individual i with given trait x_i is identical for all permutations of neighbors' traits in $\boldsymbol{x}_{-i}$. Thus,

$$w(x_i, \boldsymbol{x}_{-i}, \bar{x}) = w(x_i, \boldsymbol{x}_k, \bar{x}) \qquad \forall \boldsymbol{x}_{-i} \in \mathcal{S}_k, \tag{A.4}$$

where $\boldsymbol{x}_k = (x, x, \dots y, y, \dots)$ is the vector of dimension $n-1$ with the first $k-1$ entries equal to x and the subsequent $n-k$ entries equal to y, and $\mathcal{S}_k$ is the set of all distinct permutations of $\boldsymbol{x}_k$. The number of distinct permutations is the binomial coefficient

$$\mathcal{B}(n, k) = \binom{n-1}{k-1}. \tag{A.5}$$

In the class-structured case, permutation-invariance is on the trait profile of neighbors belonging to the same class. Permutation-invariance is not an assumption, it is part of what allows one to determine whether different neighbors belong to the same class of actors. When permutation-invariance does not hold, individuals belong to different classes.

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

A.2 Average direct fitness without classes

A.2.1 Haploids

Indistinct individuals. In the absence of within-group class structure (homogeneous individuals), the invasion fitness can be written

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

which takes the form of average direct fitness. Indeed, here $w(x,\boldsymbol{x}_k,y)$ is the individual fitness given by eq. (A.4) 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 ($\sum_{k=1}^{n} q_k(x,y) = 1$).

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,

$$q_k(x,y) = \frac{ku_k(x,y)}{\sum_{j=1}^{n} ju_j(x,y)}, \tag{A.7}$$

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:

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

The jkth entry of $\boldsymbol{A}(x,y)$, denoted by a_{jk}, gives the expected number of groups with $j \geq 1$ mutants descending over one time step from a group with $k \geq 1$ mutant, and $u_j(x,y)$ is the

stationary probability that there are j mutants in a group, conditional on there being at least one mutant. A proof of eq. (A.6) follows by left-multiplying eq. (A.8) with a vector $\mathbf{n}$ whose entry j is equal to the number j of mutants, noting that $\sum_{j=1}^{n} j a_{jk} = w(x, \boldsymbol{x}_k, y)k$ is the expected number of successfull mutant copies produced over one time step by all mutant gene copies in a group in state k, and rearranging terms; see Lehmann et al. (2016) for a detailed proof and a more detailed characterization of the multitype branching process underlying mutant dynamics.

Distinct individuals. We now make the link to characterizing invasion fitness by considering all ordered groups states. To obtain this representation, we note that from eq. (A.4), 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 with $k-1$ neighbors bearing the mutant allele. 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 $\boldsymbol{x}_{-i} \in \mathcal{S} = \{x, y\}^n$ contains exactly $k-1$ individuals also carrying the mutant (hence $\sum_{k=1}^{n} \sum_{\boldsymbol{x}_{-i} \in \mathcal{S}_k} q_k^{\mathrm{D}}(x, y) = 1$ and the superscript D stands for a reminder that the distribution is over profiles of traits for distinct individuals). On substituting eqs. (A.9)-(A.10) into eq. (A.6), we can write the invasion fitness of a 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.

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 sample space in $\mathcal{S} = \{x, y\}^n$. With this, we can write invasion fitness as

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

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. Let $z \in \mathcal{X}$ denote 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

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

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

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

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 $\boldsymbol{q}^{\mathrm{D}}(x, y)$ of ordered group profiles of traits, determined by the distribution of contexts of copies of the mutant allele. The sample space of the distribution of ordered group profiles of strategies is $\mathcal{S} = \{z(x, y), x\} \times \{z(x, y), x, y\}^{n-1}$. Eq. (A.15) shows that invasion fitness can, as in the haploid case and regardless of the mating system, be expressed as an average of the direct 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. Indeed, for the diploid case, the matrix $\boldsymbol{A}(x, y)$ defining the closed "dynamically sufficient" system determining the fate of the mutant when rare (recall eq. A.8) has entries $a_{\mathbf{jk}}$, which gives the expected number of groups in state $\mathbf{j}$ descending over one time step from a group in state $\mathbf{k}$, and where a state specifies the number of homozygote and heterozygote individuals in a group. More specifically, $\mathbf{j} = (j_y, j_x, j_z)$, where j_y denotes the number of homozygote residents, j_x the number of heterozygotes, and j_z the number of homozygote mutants in a group. The state space of the process is $\mathcal{G} = \{\mathbf{j} : j_y + j_x + j_z = n \text{ and } j_x + j_z > 0\}$ and the stationary distribution satisfying eq. (A.8) is $\boldsymbol{u}(x, y) = (u_{\mathbf{j}}(x, y))_{\mathbf{j} \in \mathcal{G}}$.

A proof of eq. (A.15) then follows by left-multiplying eq. (A.8) with a vector $\mathbf{n}$ whose entry $\mathbf{j}$ is equal to the number $j_x + 2j_z$ of mutant gene copies in that state and rearranging terms by noting that $\sum_{\mathbf{j} \in \mathcal{G}} (j_x + 2j_z)\, a_{\mathbf{jk}} = w(x, \boldsymbol{x}_{\mathbf{k}}, y) k_x + w(z(x, y), \boldsymbol{x}_{\mathbf{k}}, y) 2k_z$ is the expected number of success-

ful mutant copies produced over one demographic time step by all mutant gene copies in a group in state $\mathbf{k}$, and where $\boldsymbol{x}_{\mathbf{k}}$ denotes the vector of neighbor group profiles when a group is in state $\mathbf{k}$. With this, $W(x,y) = \sum_{\mathbf{k}\in\mathcal{G}} \left[w(x,\boldsymbol{x}_{\mathbf{k}},y)k_x + w(z(x,y),\boldsymbol{x}_{\mathbf{k}},y)2k_z\right] u_{\mathbf{k}}(x,y) / \left[\sum_{\mathbf{k}\in\mathcal{G}} (k_x + 2k_z)\, u_{\mathbf{k}}(x,y)\right]$ regardless of the assumptions on the mating system. Expanding therein the individual fitnesses in terms of ordered group profiles (like in eq. A.9) one obtains eq. (A.15). The distribution $\boldsymbol{q}^{\mathrm{D}}(x,y)$ of ordered group states experienced by a typical copy of the mutant allele will thus be expressed in terms of $\boldsymbol{u}(x,y)$, of the numbers k_x and $2k_z$, and combinatorial terms. We skip the explicit expression of $\boldsymbol{q}^{\mathrm{D}}(x,y)$ since it requires to define notations to take into account that a mutant allele copy can be in a homozygote or an heterozygote individual (see eq. A.16 for an example) but the logic to obtain $\boldsymbol{q}^{\mathrm{D}}(x,y)$ is as in the haploid case (see eq. A.7 and eq. A.10).

As an example of eq. (A.15), let us consider the case $n = 2$, then

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

where $q_{j,\mathrm{he}}(x,y)$ is the probability that, given a gene copy is mutant, its carrier is heterozygote and its group neighbor has j copies of the mutant ($j = 0,1,2$, then indicate the cases, respectively, for the neighbor to be homozygote resident, heterozygote, and homozygote mutant), while $q_{j,\mathrm{ho}}(x,y)$ is the probability that, given a gene copy is mutant, its carrier is homozygote and its group neighbor has j copies of the mutant. For this case, the distribution over group contexts 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).

A.3 Average direct 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 $x = (x_1, x_2, ..., x_{n_c}) \in \mathcal{X}$, where x_a is the trait of a mutant individual when of class a. Likewise, we have $y = (y_1, y_2, ..., y_{n_c}) \in \mathcal{X}$. Using eq. (A.3), we let $w_{us}(x_s, \boldsymbol{x}_{\mathbf{k}}, y)$ be the expected number of class-u offspring produced by a class-s mutant when in a group in state $\mathbf{k} = (k_1, ..., k_{n_c})$, 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), where $\mathbf{0}$ indicates the value of $\mathbf{k}$ with no mutants at all. 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:

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

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

Eq. (A.19) is in terms of unordered neighbor profiles charaterized by $\mathbf{k}$. In this formalism, invasion fitness $W(x, y)$ still satisfies eq. (A.8), if the matrix $\boldsymbol{A}(x, y)$, describing the growth of the mutant lineage when rare in the population, now has entries $a_{\mathbf{jk}}$ giving the expected number of mutant copies in context $\mathbf{j}$ that descend from a mutant copy in context $\mathbf{k}$, and the $q_{\mathbf{k},s}(x, y)$ distribution is then expressed in terms of the leading right eigenvector $\boldsymbol{u}(x, y) = (u_{\mathbf{k}}(x, y))_{\mathbf{k} \in I}$ of this matrix; namely,

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

A detailed proof of eq. (A.19) can be found in Lehmann et al., 2016, Appendix F. It follows by left-multiplying eq. (A.8) with a vector $\mathbf{n}$ whose entry $\mathbf{j}$ is equal to the reproductive value-weighted number $\sum_{u \in \mathcal{C}} j_u v_u(y)$ of mutant gene copies in context $\mathbf{j}$. This yields $\mathbf{n} \cdot \boldsymbol{A}(x, y) \boldsymbol{u}(x, y) =$

$\sum_{\mathbf{k}\in I}\sum_{\mathbf{j}\in I}\sum_{u\in\mathcal{C}} j_u v_u(y) a_{\mathbf{jk}} u_{\mathbf{k}} = W(x,y)\,\mathbf{n}\cdot\boldsymbol{u}$, wherein the central expression can be simplified (and the whole expression rearranged) by noting that $\sum_{\mathbf{j}\in I} j_u a_{\mathbf{jk}} = \sum_{s\in\mathcal{C}} w_{us}(x,\boldsymbol{x}_{\mathbf{k}},y)k_s$ is the expected number of successful mutant copies in class u produced over one demographic time step by all mutant gene copies in a group in state $\mathbf{k}$.

We now write eq. (A.19) as

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

where

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

is the expected reproductive value-weighted fitness of a class s mutant gene copy 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 carrying the mutant allele 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 $w_{\mathrm{DF}s}(x,y)$ 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), we 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, 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 rare to predict the direction of average allele frequency change by a scalar fitness measure at all allele frequencies under weak selection (e.g., 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.3.2 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 $x = (x_1, x_2, ..., x_{n_c}) \in \mathcal{X}$ and the trait profile of a homozygote resident individual is $y = (y_1, y_2, ..., y_{n_c}) \in \mathcal{X}$ (following eqs. A.13–A.14, we assume that for each a, $z_a(x_a, y_a)$ is obtained by assuming that heterozygote traits are a convex combination of the homozygotes' traits). With this, let $z(x, y) = (z_1, z_2, ..., z_{n_c}) \in \mathcal{X}$ denote 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

$$W(x, y) = \frac{1}{V(x, y)} \sum_{s \in \mathcal{C}} w_{\mathrm{DF}s}(x, y)\phi_s(x, y), \tag{A.27}$$

where

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

and

$$w_{\mathrm{DF}us}(x, y) = \mathrm{E}_{(x_i, \boldsymbol{x}_{-i}) \sim \boldsymbol{q}_s^{\mathrm{D}}(x, y)}\Big[w_{us}(x_i, \boldsymbol{x}_{-i}, y)\Big], \tag{A.29}$$

which is the expected reproductive value-weighted fitness of a class s mutant gene copy. Here, $x_i \in \{z_s(x_s, y_s), x_s\}$ if individual i is of class s 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.29), the couple $(x_i, \boldsymbol{x}_{-i})$ follows the distribution $\boldsymbol{q}_s^{\mathrm{D}}(x, y)$ of ordered focal group trait profiles determined by the distribution of contexts of copies of the mutant allele in

individuals of class s. The distribution of trait profiles has sample space

$$\mathcal{S}_s = \{z_s(x_s, y_s), x_s\} \times \prod_{a \in \mathcal{C}} \{z_a(x_a, y_a), x_a, y_a\}^{(n_a - \delta_{sa})}, \tag{A.30}$$

since among the neighbors of an individual of class s we have n_a individuals of class $a \neq s$ and $n_s - 1$ class-s individuals. An algebraic proof of eqs. (A.27)–(A.29) follows by combining the line of arguments used to derive the invasion fitness for diploids without classes (eq. A.15) and that for haploids with classes (eq. A.19).

A special case of eq. (A.29) is when there is only a single adult 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). As a concrete example, we work out the model of a seasonal population of social insects presented in Box 1, where $x = (x_{\mathsf{f}}, x_{\mathsf{m}}, x_{\mathsf{o}})$, which collects, respectively, the traits of females, males, and workers so that the set of classes is $\mathcal{C} = \{\mathsf{f}, \mathsf{m}, \mathsf{o}\}$. 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

$$\mathbf{A}(x, y) = \begin{bmatrix} w_{\mathrm{DFff}}(x, y) & w_{\mathrm{DFfm}}(x, y) & 0 \\ w_{\mathrm{DFmf}}(x, y) & w_{\mathrm{DFmm}}(x, y) & 0 \\ w_{\mathrm{DFof}}(x, y) & w_{\mathrm{DFom}}(x, y) & 0 \end{bmatrix}, \tag{A.31}$$

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

$$\phi_{\mathsf{f}}(x, y) = \frac{w_{\mathrm{DFff}}\left(w_{\mathrm{DFff}} + w_{\mathrm{DFmm}}\right)}{X}, \ \phi_{\mathsf{f}}(x, y) = \frac{w_{\mathrm{DFmm}}\left(w_{\mathrm{DFff}} + w_{\mathrm{DFmm}}\right)}{X} \ \text{ and } \ \phi_{\mathsf{o}}(x, y) = \frac{w_{\mathrm{DFff}} w_{\mathrm{DFof}}}{X}, \tag{A.32}$$

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

$$W(x, y) = \frac{1}{V(x, y)} \left[w_{\mathrm{DFf}}(x, y) \phi_{\mathsf{f}}(x, y) + w_{\mathrm{DFm}}(x, y) \phi_{\mathsf{m}}(x, y) \right]. \tag{A.33}$$

The two direct fitnesses appearing in this equation are given by eq. (B.1) 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 $\phi_{\mathsf{f}}(y, y) = \phi_{\mathsf{m}}(y, y) = \phi_{\mathsf{o}}(y, y) = 1/3$ and the reproductive values, normalized so as to satisfy eq. (A.26), are

$$v_{\mathsf{o}}(y) = 0, \quad v_{\mathsf{f}}(y) = \frac{3}{2}, \quad v_{\mathsf{m}}(y) = \frac{3}{2}. \tag{A.34}$$

Appendix B: Inclusive fitness

The main aim of this Appendix is to derive from invasion fitness the expression for the inclusive fitness of a class-a individual given in the main text (eq. 4) and then to show that it is maximized for the class-specific trait x_a in an uninvadable state x^*. As in the previous section, we do so by progressively introducing the different concepts. Before delving into the calculations, it is worth recalling that "inclusive fitness" can be regarded as achieving two related decompositions of the force of selection on a mutant allele. First, it is a partition of selection into direct and indirect fitness effects (the "cost" and "benefit" of Hamilton's rule, e.g., Hamilton, 1964; Frank, 1998; Rousset, 2004), where the indirect effects are weighted by relatedness coefficient(s). This allows to usefully classify behaviors in different categories ("selfishness", "altruism", "spite", etc.) and is most easily achieved in an evolutionary model by a neighbor-modulated representation where fitness effects are grouped by recipient of actions (e.g., Hamilton, 1970, Frank, 1998, Rousset, 2004, Fig. 7.1). Second, and as further explained in the main text, one may seek to go one step further and group the indirect fitness effects by actor. We will refer to this as the actor-centered approach to inclusive fitness.

For a model with arbitrary strength of selection on a mutant allele without class structure, a general expression for the decomposition into direct and indirect effects 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 and on the frequency of the mutant in its neighbors (Queller, 1992; Frank, 1997; Gardner et al., 2011), thus reaching a neighbor-modulated representation of "inclusive fitness". Importantly, it has been shown that such a *two-predictor* regression automatically yields an actor-centered representation of such effects (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 decomposition of the force of selection into direct and indirect effects for haploid class-structured populations. The single-predictor regression, however, only leads to a neighbor-modulated representation of direct and indirect effects.

In order to obtain a general actor-centered representation of fitness effect for diploid class structured populations and avoid confusions between approaches, we first delineate in the case of haploids the differences between the partitions of fitness by single and two-predictor regressions, and by neighbor-modulated and inclusive fitness effects. In a second time, we turn to the general class-structured populations analysis.

B.1 Inclusive fitness for haploids without classes

We start by deriving a decomposition into direct and indirect effects from invasion fitness (eq. A.6) for the haploid case and without class structure. 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 $r(x,y) = q_2(x,y)$).

B.1.1 Regression with respect to neighbors

For a one-predictor regression, 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 on a focal's fitness of allele frequency in neighbors, and $(k-1)/(n-1)$ is the frequency of the mutant allele among neighbors of a 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, 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 by p this frequency, and by a slight abuse of notation, we denote by $w(x,\boldsymbol{x}_k,p)$ 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 frequency p in a resident y population is:

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

where c and b^{N} are regression coefficients (the superscript N will from now on stand as a reminder that the regression is, by construction, neighbor-modulated, and we also regress on the number rather than the frequency of mutant neighbors, as this will be useful later), $q_k(x, y, p)$ 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 (this probability can be obtained from the full dynamical system for a non-rare mutant, but since we do not need to compute this probability explicitly, we do not specify this dynamical system). 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 (again this probability can be obtained from the full dynamical system for a non-rare mutant). Minimizing the quadratic form $Q(c, b^{\mathrm{N}}, x, y, p)$ (by solving $\partial Q(c, b^{\mathrm{N}}, x, y, p)/\partial c = 0$ and $\partial Q(c, b^{\mathrm{N}}, x, y, p)/\partial b = 0$) we then obtain the regression coefficients $c = c(x, y, p)$ and $b^{\mathrm{N}} = b^{\mathrm{N}}(x, y, p)$, which depend on the population allele frequency.

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

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

where the residual and the cost and benefit will depend on mutant trait, resident trait, and are limits as $p \to 0$ of frequency-dependent terms, i.e., $c(x, y) = \lim_{p \to 0} c(x, y, p)$ and $b^{\mathrm{N}}(x, y) = \lim_{p \to 0} b^{\mathrm{N}}(x, y, p)$. 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 groups is described by the mean matrix $\boldsymbol{A}$, which is also the matrix of the linearized dynamical system around $p = 0$, and so $q_k(x, y, p) = q_k(x, y) + O(p)$ for all k. The residuals are orthogonal to the regressors when regression coefficients minimize the quadratic form (Cox and Wermuth, 1996, section 3.3.2). Here the regressors include both an intercept and the focal allele frequency, and then the expectation of the residuals is zero whether an individual carries the mutant or the resident allele. Thus, the residuals disappear from the average of expression (B.6) over the conditional distribution $q_k(x, y, p)$ of k given an individual carries the mutant allele, and we can then write invasion fitness as:

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

for the relatedness coefficient defined in eq. (B.1). We now give an interpretation of this result. The interpretation of $-c(x,y)$ is the additive marginal effect on the number of successful gene copies produced by an individual when it expresses the mutant instead of the resident allele, while $b^{\mathrm{N}}(x,y)$ can be interpreted in the two following ways.

(1) neighbor-modulated interpretation. Here, $b^{\mathrm{N}}(x,y)$ gives the average effect, on the expected number of offspring (per haplogenome) produced by a focal individual, and stemming from a randomly sampled neighbor expressing a copy of the mutant instead of the resident allele.

(2) Actor-modulated interpretation. Here, $b^{\mathrm{N}}(x,y)$ gives the average effect, on the expected number of offspring (per haplogenome) produced by a randomly sampled group neighbor, of an individual expressing a copy of the mutant instead of the resident allele.

This dual interpretation follows from the fact that the regression (e.g. eq. B.5) averages over all contexts, mutant and resident neighbors affecting the fitness of a focal recipient, which itself can be mutant or resident ("two-predictor regression"), and so, on average, recipient and actor individuals can be interchanged (see Rousset, 2015 for more details on this dual perspective). As such, eq. (B.7) provides a genuine actor-centered representation of inclusive fitness and in order to obtain a more compact expression, we let

$$b(x,y) = (n-1)\, b^{\mathrm{N}}(x,y) \tag{B.8}$$

denote the additive effect, on the expected number of offspring (per haplogenome) produced *by all group neighbors*, of an individual expressing a copy of the mutant instead of the resident allele. Thereby, selection favors the mutant ($W(x,y) > 1$) when Hamilton's rule is satisfied:

$$r(x,y)b(x,y) - c(x,y) > 0. \tag{B.9}$$

B.1.3 Comparing single- and two-predictor regression

The key 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 (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

$$\begin{aligned}\gamma(x,y) &= w(y,y,y) - w(x,y,y)\\ \beta(x,y) &= w(x,x,y) - w(x,y,y).\end{aligned} \tag{B.10}$$

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

$$\begin{aligned}-c(x,y) &= \frac{1}{1+r(x,y)}\left(w(x,y,y) - w(y,y,y)\right) + \frac{r(x,y)}{1+r(x,y)}\left(w(x,x,y) - w(y,x,y)\right)\\ b(x,y) &= \frac{1}{1+r(x,y)}\left(w(y,x,y) - w(y,y,y)\right) + \frac{r(x,y)}{1+r(x,y)}\left(w(x,x,y) - w(x,y,y)\right).\end{aligned} \tag{B.11}$$

In the actor-modulated interpretation, $1/[1 + r(x,y)]$ and $r(x,y)/[1 + r(x,y)]$ weigh in both $-c$ and b the case where the neighbor of a focal individual is either resident or mutant, respectively. To understand where these weights come from, let P denote the probability of an (x,x) focal-neighbour pair in a group and Q the probability of an (x,y) focal-neighbour pair [these probabilities reducing respectively to $pr(x,y)$ and to $p(1 - r(x,y))$ for vanishing p]. Then the weights are proportional to $P + Q$ versus P, rather than Q versus P, for the following reason. The derivative of the sum of squares with respect to c is proportional to

$$\begin{aligned}&Q(w(x,y,y) + c - w(y,y,y)) + P(w(x,x,y) + c - b - w(y,y,y))\\ &\qquad = (Q+P)(w(x,y,y) + c - w(y,y,y)) + P(w(x,x,y) - b - w(x,y,y)).\end{aligned} \tag{B.12}$$

The effect of least-square regression is to predict $w(x,y,y)$ and $w(y,x,y)$ with identical prediction residuals: $w(y,x,y) - w(y,y,y) - b = w(x,y,y) - w(y,y,y) + c$, and thus the derivative is proportional to

$$(Q+P)(w(x,y,y) + c - w(y,y,y)) + P(w(x,x,y) + c - w(y,x,y)), \tag{B.13}$$

meaning that we have represented the original term $w(x,x,y) - w(y,y,y)$ as the effect of two allelic substitutions, each with effect $-c$. This recovers the solution for $-c$. The same logic holds for the weights in the expression for b and this argument holds at all allele frequencies p.

B.2 Inclusive fitness for diploids with classes

B.2.1 Multiplayer class-structured regression

We now turn to deriving an expression for inclusive fitness for diploids with class structure 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 u of descendants and each class s of parent, the fitness (per haplogenome) of a focal individual i with trait x_i in a group with trait profile $\boldsymbol{x}_{-i}$ as

$$w_{us}(x_i, \boldsymbol{x}_{-i}, y) = w_{us}(y, \boldsymbol{x}_{\mathbf{0}}, y) - c_{us}(x, y)p_i + \sum_{j \neq i} b^{\mathrm{N}}_{us \leftarrow k(j)}(x, y)p_j + \text{residual}. \tag{B.14}$$

Here, p_i denotes the frequency of the mutant allele in individual $i \in \mathcal{I}$ [zero, one-half, or one, so that the individual expresses, respectively, trait y, x or $z(x, y)$]. The function $k : \mathcal{I} \to \mathcal{C}$ assigns to each individual in group $\mathcal{I}$ its class in $\mathcal{C}$, such that $k(j) = a$ when individual j is in class a. The mutant allele frequencies p_i in each individual are thus predictor variables of fitness and $c_{us}(x, y)$ and $b^{\mathrm{N}}_{us \leftarrow k(j)}(x, y)$ are regression coefficients depending on mutant and resident trait values. Indeed, eq. (B.14) says that we seek to obtain the predictor

$$\hat{w}_{us}\left(\mathbf{c}_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, \mathbf{p}_i, y\right) = w_{us}(y, \boldsymbol{x}_{\mathbf{0}}, y) - c_{us}(x, y)p_i + \sum_{j \neq i} b^{\mathrm{N}}_{us \leftarrow k(j)}(x, y)p_j. \tag{B.15}$$

of class-u fitness of a class-s gene copy as a linear regression on the mutant allele frequency carried by all actors on that fitness [here the vector $\boldsymbol{b}^{\mathrm{N}}_{us}$ collects all the $b_{us \leftarrow k(j)}$ regressors and the vector $\mathbf{p}_i$ collects mutant frequencies p_i in all individuals].

Eq. (B.14) must hold for all trait profiles $(x_i, \boldsymbol{x}_{-i}) \in \{z_s(x_s, y_s), x_s, y_s\} \times \prod_{a \in \mathcal{C}} \{z_a(x_a, y_a), x_a, y_a\}^{(n_a - \delta_{sa})}$ and the regression coefficients are determined by the following argument, which generalizes the one developed in the absence of class structure (section B.1). Let then $w_{us}(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ denote 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 $\boldsymbol{p} = (p_1, ..., p_{n_c})$, where p_s is the average mutant allele frequency in class s in the population. Further, let $\boldsymbol{q}^{\mathrm{D}}_s(x, y, \boldsymbol{p})$ denote the ordered distribution of group traits, determined by the distribution of contexts of copies of the mutant allele in class s. This $\boldsymbol{q}^{\mathrm{D}}_s(x, y, \boldsymbol{p})$ distribution has the same sample space as $\boldsymbol{q}^{\mathrm{D}}_s(x, y)$ (recall eq. A.30) and generalizes it to arbitrary allele frequency. Likewise, $\tilde{\boldsymbol{q}}_s(x, y, \boldsymbol{p})$ denotes the ordered distribution of group traits for non-rare mutant frequency, determined by the distribution of contexts of copies of the resident allele in class s [$\tilde{\boldsymbol{q}}_s(x, y, \boldsymbol{p})$ has sample space $\{x_s, y_s\} \times \prod_{a \in \mathcal{C}} \{z_a(x_a, y_a), x_a, y_a\}^{(n_a - \delta_{sa})}$]. With these notations, the expected sum of squares to be minimized by the regression coefficients can be written

$$\begin{aligned} Q_{us}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x, y, \boldsymbol{p}) = {} & \mathrm{E}_{(x_i, \boldsymbol{x}_{-i}) \sim \boldsymbol{q}^{\mathrm{D}}_s(x, y, \boldsymbol{p})} \left[\left(\hat{w}_{us}\left(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, \mathbf{p}_i, y\right) - w_{us}(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) \right)^2 \right] p_s \\ & + \mathrm{E}_{(x_i, \boldsymbol{x}_{-i}) \sim \tilde{\boldsymbol{q}}^{\mathrm{D}}_s(x, y, \boldsymbol{p})} \left[\left(\hat{w}_{us}\left(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, \mathbf{p}_i, y\right) - w_{us}(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) \right)^2 \right] (1 - p_s). \end{aligned} \tag{B.16}$$

By solving $\partial Q_{us}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x, y, \boldsymbol{p}) / \partial c_{us \leftarrow a} = 0$ and $\partial Q_{u,s}(c_{us}, \boldsymbol{b}^{\mathrm{N}}_{us}, x, y, \boldsymbol{p}) / \partial b^{\mathrm{N}}_{us \leftarrow k(j)} = 0$ for all $j \neq i$, we obtain the regression coefficients $c_{us \leftarrow a}(x, y, \boldsymbol{p})$ and $b^{\mathrm{N}}_{us \leftarrow k(j)}(x, y, \boldsymbol{p})$. These depend on

the population state and we note that that

$$b^{\mathrm{N}}_{us\leftarrow k(j)}(x,y,\boldsymbol{p}) = b^{\mathrm{N}}_{us\leftarrow a}(x,y,\boldsymbol{p}) \qquad \text{for all } j \text{ when } k(j)=a, \tag{B.17}$$

since individuals from the same class carrying similar traits have the same effect on the recipients of their actions. Computing the regression coefficients explicitly involves solving linear systems of equations with complicated terms. But all in all, this is not more involved that computing invasion fitness to begin with, since this requires computing eigenvectors (recall eq. A.8). Hence, no computational complexity is added if the regression coefficients are to be computed explicitly. That said, the interpretative nature of inclusive fitness does require the explicit computation of the regression coefficients.

B.2.2 Average allele frequency

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

$$p_a = L_1 u_a(x,y) \tag{B.18}$$

where $u_a(x,y)$ is (up to a constant factor) the frequency of the mutant allele in class a under the quasi-stationary distribution $\boldsymbol{u}(x,y)$. 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

$$p_a = p\frac{u_a(x,y)}{\sum_{a\in\mathcal{C}} \alpha_a(y)u_a(x,y)} = p\frac{u_a(x,y)}{\sum_{a\in\mathcal{C}} u_a(x,y)}\frac{\sum_{a\in\mathcal{C}} u_a(x,y)}{\sum_{a\in\mathcal{C}} \alpha_a(y)u_a(x,y)}. \tag{B.19}$$

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): $\phi_a(x,y) = u_a(x,y)/\sum_{a\in\mathcal{C}} u_a(x,y)$. The last fraction in eq. (B.19) is then the inverse of the fraction $\sum_{a\in\mathcal{C}} \alpha_a(y)\left[u_a(x,y)/\sum_{a\in\mathcal{C}} u_a(x,y)\right] = \sum_{a\in\mathcal{C}} \alpha_a(y)\phi_a(x,y)$, and

$$p_a = p\frac{\phi_a(x,y)}{\sum_{a\in\mathcal{C}} \alpha_a(y)\phi_a(x,y)}. \tag{B.20}$$

Substituting eq. (B.20) into $c_{us}(x,y,\boldsymbol{p})$ and $b^{\mathrm{N}}_{us\leftarrow a}(x,y,\boldsymbol{p})$, and recalling eq. (B.17), we compute the regression coefficients of eq. (B.14) as

$$c_{us\leftarrow a}(x,y) = \lim_{p\to 0} c_{us\leftarrow a}(x,y,\boldsymbol{p}) \qquad \text{and} \qquad b^{\mathrm{N}}_{us\leftarrow a}(x,y) = \lim_{p\to 0} b^{\mathrm{N}}_{us\leftarrow a}(x,y,\boldsymbol{p}) \tag{B.21}$$

(see section B.3 for a concrete application and explicit computation of such coefficients). 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

$$r_{\mathrm{f}s}(x,y) = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim \boldsymbol{q}^{\mathrm{D}}_s(x,y)}[p_i]\,, \tag{B.22}$$

which is the probability that, conditional on a random gene copy of class s carrying the mutant allele, a randomly sampled homologous gene in that individual is a mutant. For $k(j) = a$, we have

$$r_{\mathrm{n}\,a|s}(x,y) = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim \boldsymbol{q}^{\mathrm{D}}_s(x,y)}\left[p_j\right] \tag{B.23}$$

which is the probability that, conditional on a random gene copy of class s carrying the mutant allele, a 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

$$r_{a|s}(x,y) = \frac{r_{\mathrm{n}\,a|s}(x,y)}{r_{\mathrm{f}s}(x,y)}. \tag{B.24}$$

B.2.3 Average inclusive fitness

Now substitute eq. (B.14) into direct fitness (eq. A.29) and then into invasion fitness (eq. A.27). Then, by dint of the reproductive values recursion (eq. A.18), the reproductive values normalizer (eq. A.20), the relatedness coefficients (eqs. B.22–B.24), the relationship $\sum_{j:k(j)=a} b^{\mathrm{N}}_{us\leftarrow k(j)}(x,y) = b^{\mathrm{N}}_{us\leftarrow a}(x,y)(n_a - \delta_{sa})$ (eq. B.17), and recalling that the residual term in eq. (B.14) cancels when averaged over the $\boldsymbol{q}^{\mathrm{D}}_s(x,y)$ distribution (since they are uncorrelated with regressors), the invasion fitness of a mutant allele introduced as a single copy in a resident population can be put under the form

$$W(x,y) = 1 + \frac{1}{V(x,y)}\left[W_{\mathrm{IF}}(x,y) - 1\right], \tag{B.25}$$

where

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{u\in\mathcal{C}}\sum_{s\in\mathcal{C}} v_u(y)\left[-c_{us}(x,y) + \sum_{a\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y)(n_a - \delta_{sa})r_{a|s}(x,y)\right] r_{\mathrm{f},s}(x,y)\phi_s(x,y). \tag{B.26}$$

is the inclusive fitness of the mutant 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 $b^{\mathrm{N}}_{us\leftarrow a}$ 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.16) and otherwise follow the same line of argument.

Since $V(x,y) > 0$, we have from eq. (B.25) that

$$W(x,y) \leq 1 \iff W_{\mathrm{IF}}(x,y) \leq 1. \tag{B.27}$$

Hence, trait x^* is uninvadable if it is a best-reply to itself in terms of inclusive fitness: $x^* \in \arg\max_{x\in\mathcal{X}} W_{\mathrm{IF}}(x,x^*)$. Finally, we note that if one chooses to normalize the (neutral) reproductive such that $V(x,y) = 1$, then one would have $W(x,y) = W_{\mathrm{IF}}(x,y)$.

B.2.4 Grouping effects by actor

In eq. (B.26), the fitness effects of social interactions among individuals in the population are grouped by recipients each class s. We now first rearrange eq. (B.26) in order to obtain a grouping of fitness effects by actor in each class, so that $W_{\mathrm{IF}}(x,y)$ reads as an average over class-specific inclusive fitnesses (which justifies the subscript "Inclusive Fitness"). In a second, time we then show that class-specific inclusive fitness is maximized in an uninvadable population state.

The key steps to reach the actor-centered perspective, is to note, first, that, as was the case under the haploid model, $b^{\mathrm{N}}_{us\leftarrow a}(x,y)$ can be interpreted in the two following ways.

(1) Neighbor-modulated interpretation. Here, $b^{\mathrm{N}}_{us\leftarrow a}(x,y)$ gives the average effect, on the expected number of class-u offspring (per haplogenome) produced by a focal individual in class a, and stemming from a randomly sampled neighbor expressing a copy of the mutant instead of the resident allele.

(2) Actor-modulated interpretation. Here, $b^{\mathrm{N}}_{us\leftarrow a}(x,y)$ gives the average effect, on the expected number of class-u offspring (per haplogenome) produced by a randomly sampled group neighbor of class a, of an individual expressing a copy of the mutant instead of the resident allele.

We will thus from now on use the second interpretation and further note that the following equality holds

$$r_{\mathrm{n}\,a|s}(x,y)\phi_s(x,y) = r_{\mathrm{n}\,s|a}(x,y)\phi_a(x,y)\frac{n_s}{n_a}. \tag{B.28}$$

To check this result, we highlight that each side of the equation involves two ways of sampling gene copies. First, 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), and then 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). 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 then obtained as $\phi_s(x,y)2n_a r_{\mathrm{n}\,a|s}(x,y)$ when the first sampled gene copy is in a class-s individual, and as $\phi_a(x,y)2n_s r_{\mathrm{n}\,s|a}(x,y)$ when the first copy is sampled in a class-a individual, from which the above result follows.

Substituting eq. (B.28) into eq. (B.26), and rearranging we obtain

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{u\in\mathcal{C}}\sum_{a\in\mathcal{C}} v_u(y)\left[-c_{ua}(x,y) + \sum_{s\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y)(n_s-\delta_{sa})\, r_{s|a}(x,y)\right] r_{\mathrm{f}a}(x,y)\phi_a(x,y), \tag{B.29}$$

where

$$-c_{ua}(x,y) + \sum_{s\in\mathcal{C}} b^{\mathrm{N}}_{us\leftarrow a}(x,y)(n_s-\delta_{sa})\, r_{s|a}(x,y) \tag{B.30}$$

is the average effect, on the number of class-u offspring (per haplogenome) produced by all group members of class a, and stemming from an individual of class a expressing a copy of the mutant instead of the resident allele. Eq. (B.30) 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.29), we let

$$b_{us\leftarrow a}(x,y) = (n_s-\delta_{sa})b^{\mathrm{N}}_{us\leftarrow a}(x,y) \tag{B.31}$$

denote the average additive effect, on the number of class-u offspring produced (per haplogenome) by all class-s neighbors in a group, and stemming from a single class-a individual switching to expressing a copy of the mutant instead of the resident allele. Substituting eq. (B.31) into eq. (B.29), we can obtain:

$$W_{\mathrm{IF}}(x,y) = 1 + \sum_{a\in\mathcal{C}} \Delta w_{\mathrm{IF}a}(x,y) r_{\mathrm{f}a}(x,y)\phi_a(x,y), \tag{B.32}$$

where

$$\Delta w_{\mathrm{IF}a}(x,y) = \sum_{u\in\mathcal{C}} v_u(y)\left[-c_{ua}(x,y) + \sum_{s\in\mathcal{C}} b_{us\leftarrow a}(x,y) r_{s|a}(x,y)\right] \tag{B.33}$$

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

B.2.5 Class-specific inclusive fitness maximization

We here prove that the inclusive fitness effect (eq. B.33) is maximized with respect to x_a at the uninvadable state x^*, which will allow us to define a class-specific inclusive fitness that is equivalently maximized. In general, inclusive-fitness (eq. B.32) maximization does not implies maximization of the summand therein for each class with respect to all mutants $x \in \mathcal{X}$, but what

we consider here are class-specific mutants. To that end, let $\tilde{x}_a = (x_1^*, x_2^*, ..., x_{a-1}^*, x_a, x_{a+1}^*, ..., x_{n_c}^*)$, denote the trait profile of a mutation with all traits held 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 in state x^*, we have

$$\Delta w_{\mathrm{IF}v}(\tilde{x}_a, x^*) = 0 \;\; \forall v \neq a. \tag{B.34}$$

Eq. (B.34) 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 $v \neq a$ is nil. This 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.

Substituting eq. (B.34) into eq. (B.32), the inclusive fitness of a mutant inducing an unilateral deviation in class a in population state x^* is given by

$$W_{\mathrm{IF}}(\tilde{x}_a, x^*) = 1 + \Delta w_{\mathrm{IF}a}(\tilde{x}_a, x^*) r_{\mathrm{f}a}(\tilde{x}_a, x^*) \phi_a(\tilde{x}_a, x^*). \tag{B.35}$$

Since $r_{\mathrm{f}a}(\tilde{x}_a, x^*)\phi_a(\tilde{x}_a, x^*) > 0$ for all classes and traits, we have from eq. (B.35) that

$$\Delta w_{\mathrm{IF}a}(\tilde{x}_a, x^*) \leq 0 \iff W_{\mathrm{IF}}(\tilde{x}_a, x^*) \leq 1 \quad \text{for all } x_a \in \mathcal{X}_a. \tag{B.36}$$

Hence, trait x_a^* preempts invasion by any mutant inducing an unilateral deviation in class a if it satisfies $x_a^* \in \arg\max_{x_a \in \mathcal{X}_a} \Delta w_{\mathrm{IF}a}(\tilde{x}_a, x^*)$. Since this holds for each class in an uninvadable population x^*, we have

$$\Delta w_{\mathrm{IF}a}(\tilde{x}_a, x^*) \leq 0 \iff W_{\mathrm{IF}}(\tilde{x}_a, x^*) \leq 1 \;\; \forall x_a \in \mathcal{X}_a \text{ and } a \in \mathcal{C} \Leftarrow W(x, x^*) \leq 1 \quad \text{for all } x \in \mathcal{X}.$$

In other words, the inclusive fitness effect in each class is maximized in an uninvadable population state:

$$x_a^* \in \arg\max_{x_a \in \mathcal{X}_a} \Delta w_{\mathrm{IF}a}(\tilde{x}_a, x^*) \;\; \forall a \in \mathcal{C} \quad \Leftarrow \quad x^* \in \arg\max_{x \in \mathcal{X}} W(x, x^*). \tag{B.37}$$

The left-hand side can now also be written in terms of the inclusive fitness

$$w_{\mathrm{IF}a}(x, y) = v_a(y) + \Delta w_{\mathrm{IF}a}(x, y) \tag{B.38}$$

of a class-a individual (eq. 4 of the main text), which is thus also maximized in an uninvadable

population state (since $v_a(y)$ does not depend on the mutant trait):

$$x_a^* \in \arg\max_{x_a \in \mathcal{X}_a} w_{\mathrm{IF}a}(\tilde{x}_a, x^*) \;\; \forall a \in \mathcal{C} \quad \Leftarrow \quad x^* \in \arg\max_{x \in \mathcal{X}} W(x, x^*). \tag{B.39}$$

We finally mention that if we were to replace class-specific inclusive $w_{\mathrm{IF}a}(\tilde{x}_a, x^*)$ by class-specific average direct fitness $w_{\mathrm{DF}a}(\tilde{x}_a, x^*)$ in the left-hand side of eq. (B.39), then eq. (B.39) is not satisfied, unless additional assumptions are made on the form of the individual fitness functions. This negative results points to the limitations of using average direct fitness as an 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 direct fitnesses $w_{\mathrm{DF}u}(\tilde{x}_a, x^*)$ for $u \neq a$ in expression (A.27) for invasion fitness will not be independent of the mutant trait x_a of a class-a individual. In this case, the invasion fitness of trait x_a depends on these $w_{\mathrm{DF}u}(\tilde{x}_a, x^*)$ fitnesses for $u \neq a$, even though the individual expressing trait x_a is not in any class $u \neq a$. Hence the biological interpretation of an individual of class a as maximizing $w_{\mathrm{DF}a}(\tilde{x}_a, x^*)$ at an evolutionary equilibrium generally breaks down.

B.3 Example: inclusive fitness for social insects

We here derive the inclusive fitness effects for the social insect model (eq. B.5 in Box 2) from the fitness functions defined in the main text (see eqs. B.2–B.4 in Box 1) and assuming the population has reached the uninvadable sex ratio of $1/2$ for this model. Because we consider only diploidy, our "social insects" are akin to termites rather than ants. From eqs. (B.2) of Box 2, we can write the individual fitness of a female i whose worker offspring has trait $x_{\mathrm{o}(i)} \in \{z_{\mathrm{o}}, x_{\mathrm{o}}, y_{\mathrm{o}}\}$ as

$$\begin{aligned} w_{\mathrm{ff}}\left(x_{\mathrm{o}(i)}, y_{\mathrm{o}}\right) &= \frac{\left(1 + P(x_{\mathrm{o}(i)})\right)}{2\left(1 + P(y_{\mathrm{o}})\right)} \\ w_{\mathrm{mf}}\left(x_{\mathrm{o}(i)}, y_{\mathrm{o}}\right) &= \frac{\left(1 + P(x_{\mathrm{o}(i)})\right)}{2\left(1 + P(y_{\mathrm{o}})\right)}, \end{aligned} \tag{B.40}$$

which, once averaged over the cases where the offspring is heterozygote (with phenotype $x_{\mathrm{o}(i)} = x_{\mathrm{o}}$), or homozygote resident (with phenotype $x_{\mathrm{o}(i)} = y_{\mathrm{o}}$), produces eqs. (B.2) of Box 2 (and where for simplicity of presentation we only denote the traits whose variation affect fitness). Likewise, from eq. (B.4) of Box 2, the individual fitness of a male i whose worker offspring has trait $x_{\mathrm{o}(i)} \in \{y_{\mathrm{o}}, x_{\mathrm{o}}, z_{\mathrm{o}}\}$ can be written

$$\begin{aligned} w_{\mathrm{fm}}\left(x_{\mathrm{o}(i)}, y_{\mathrm{o}}\right) &= w_{\mathrm{ff}}\left(x_{\mathrm{o}(i)}, y_{\mathrm{o}}\right) \\ w_{\mathrm{mm}}\left(x_{\mathrm{o}(i)}, y_{\mathrm{o}}\right) &= w_{\mathrm{mf}}\left(x_{\mathrm{o}(i)}, y_{\mathrm{o}}\right). \end{aligned} \tag{B.41}$$

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.41), 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:

$$p_{\mathrm{ho,m}}=p_{\mathrm{ho,f}}=p^2 \qquad \text{and} \qquad p_{\mathrm{he,f}}=p_{\mathrm{he,m}}=2p(1-p). \tag{B.42}$$

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

$$Q_{j\mathrm{f}}(c_{j\mathrm{f}},b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}},b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}})=Q_{j\mathrm{f}|\mathrm{ho}(x)}p_{\mathrm{ho,f}}+Q_{j\mathrm{f}|\mathrm{he}}p_{\mathrm{he,f}}+Q_{j\mathrm{f}|\mathrm{ho}(y)}(1-p_{\mathrm{ho,f}}-p_{\mathrm{he,f}}), \tag{B.43}$$

where $Q_{j\mathrm{f}|\mathrm{ho}(x)}$, $Q_{j\mathrm{f}|\mathrm{he}}$, and $Q_{j\mathrm{f}|\mathrm{ho}(y)}$ are, respectively, the sum of squares when the female is homozygote mutant, heterozygote, and homozygote resident. Application of eqs. (B.15)–(B.16) shows that when the female is homozygote

$$\begin{aligned}
Q_{j\mathrm{f}|\mathrm{ho}(x)} &= \left(\frac{1}{2}-c_{j\mathrm{f}}+b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}+b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}}-w_{j\mathrm{f}}(z_{\mathrm{o}},y_{\mathrm{o}})\right)^2 p_{\mathrm{ho,m}}\\
&+\left[\left(\frac{1}{2}-c_{j\mathrm{f}}+b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}+\frac{b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}}}{2}-w_{j\mathrm{f}}(z_{\mathrm{o}},y_{\mathrm{o}})\right)^2\frac{1}{2}\right.\\
&\left.+\left(\frac{1}{2}-c_{j\mathrm{f}}\frac{b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}}{2}+\frac{b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{m}}}{2}-w_{j\mathrm{f}}(x_{\mathrm{o}},y_{\mathrm{o}})\right)^2\frac{1}{2}\right]p_{\mathrm{he,m}}\\
&+\left(\frac{1}{2}-c_{j\mathrm{f}}+\frac{b^{\mathrm{N}}_{j\mathrm{f}\leftarrow\mathrm{o}}}{2}-w_{j\mathrm{f}}(x_{\mathrm{o}},y_{\mathrm{o}})\right)^2(1-p_{\mathrm{ho,m}}-p_{\mathrm{he,m}})
\end{aligned} \tag{B.44}$$

where in the present example $w_{j\mathrm{f}}$ is given by eq. (B.40). 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 $Q_{j\mathrm{f}|\mathrm{he}}=(1/2)Q_{j\mathrm{f}|\mathrm{he},1}+(1/2)Q_{j\mathrm{f}|\mathrm{he},0}$, where $Q_{j\mathrm{f}|\mathrm{he},1}$ represents the case where the worker inherits the mutant allele from its mother and $Q_{j\mathrm{f}|\mathrm{he},0}$ for the case the worker does not inherit the mutant from its mother. We

find that

$$Q_{jf|he,1} = \left(\frac{1}{2} - \frac{c_{jf}}{2} + b^{N}_{jf\leftarrow o} + b^{N}_{jf\leftarrow m} - w_{jf}(z_o, y_o)\right)^2 p_{ho,m}$$
$$+ \left[\left(\frac{1}{2} - \frac{c_{jf}}{2} + b^{N}_{jf\leftarrow o} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}(z_o, y_o)\right)^2 \frac{1}{2}\right.$$
$$\left. + \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}(x_o, y_o)\right)^2 \frac{1}{2}\right] p_{he,m}$$
$$+ \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} - w_{jf}(x_o, y_o)\right)^2 (1 - p_{ho,m} - p_{he,m}) \quad \text{(B.45)}$$

and

$$Q_{jf|he,0} = \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + b^{N}_{jf\leftarrow m} - w_{jf}(x_o, y_o)\right)^2 p_{ho,m}$$
$$+ \left[\left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}(x_o, y_o)\right)^2 \frac{1}{2} + \left(\frac{1}{2} - \frac{c_{jf}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}(y_o, y_o)\right)^2 \frac{1}{2}\right] p_{he,m}$$
$$+ \left(\frac{1}{2} - \frac{c_{jf}}{2} - w_{jf}(y_o, y_o)\right)^2 (1 - p_{ho,m} - p_{he,m}). \quad \text{(B.46)}$$

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

$$Q_{jf|ho(y)} = \left(\frac{1}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + b^{N}_{jf\leftarrow m} - w_{jf}(x_o, y_o)\right)^2 p_{ho,m}$$
$$\left[\left(\frac{1}{2} + \frac{b^{N}_{jf\leftarrow o}}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}(x_o, y_o)\right)^2 \frac{1}{2} + \left(\frac{1}{2} + \frac{b^{N}_{jf\leftarrow m}}{2} - w_{jf}(x_o, y_o)\right)^2 \frac{1}{2}\right] p_{he,m}, \quad \text{(B.47)}$$

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

$$\frac{\partial Q_{jf}(c_{jf}, b^{N}_{jf\leftarrow o}, b^{N}_{jf\leftarrow m})}{\partial c_{jf}} = 0\ ,\ \frac{\partial Q_{jf}(c_{jf}, b^{N}_{jf\leftarrow o}, b^{N}_{jf\leftarrow m})}{\partial b^{N}_{jf\leftarrow o}} = 0 \text{ and } \frac{\partial Q_{jf}(c_{jf}, b^{N}_{jf\leftarrow o}, b^{N}_{jf\leftarrow m})}{\partial b^{N}_{jf\leftarrow o}} = 0 \quad \text{(B.48)}$$

for c_{jf}, $b^{N}_{jf\leftarrow o}$ and $b^{N}_{jf\leftarrow m}$. Substituting eq. (B.42) 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

$$\begin{aligned} b^{\mathrm{N}}_{\mathrm{ff}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})} \\ b^{\mathrm{N}}_{\mathrm{mf}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})}, \end{aligned} \tag{B.49}$$

where these regressions coefficients are the offspring production of the mutant worker minus that of the resident worker relative to the offspring production of the mutant worker.

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

$$\begin{aligned} b^{\mathrm{N}}_{\mathrm{fm}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})} \\ b^{\mathrm{N}}_{\mathrm{mm}\leftarrow\mathrm{o}} &= \frac{(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}}))}{1 + P(y_{\mathrm{o}})}. \end{aligned} \tag{B.50}$$

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

$$\begin{aligned} \Delta w_{\mathrm{IFf}}(x, y) &= 0 \\ \Delta w_{\mathrm{IFm}}(x, y) &= 0. \end{aligned} \tag{B.51}$$

To obtain the inclusive fitness effect for a worker, we note that from eq. B.31,

$$b_{us\leftarrow\mathrm{o}}(x, y) = b^{\mathrm{N}}_{us\leftarrow\mathrm{o}}(x, y) \tag{B.52}$$

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.33), and eqs. (B.49)–(B.50), we obtain

$$\Delta w_{\mathrm{IFo}}(x, y) = v_{\mathrm{f}}(y) \left(\frac{P(x_{\mathrm{o}}) - P(y_{\mathrm{o}})}{1 + P(y_{\mathrm{o}})} \right) + v_{\mathrm{m}}(y) \left(\frac{P(x_{\mathrm{o}}) - P(y_{\mathrm{o}})}{1 + P(y_{\mathrm{o}})} \right). \tag{B.53}$$

Adding the reproductive values to the inclusive fitness effects shown in eq. (B.51) and eq. (B.53), we obtain the inclusive fitnesses displayed in Box 2.

Appendix C: Fitness as-if

C.1 Rational-actor payoff maximization

As explained in the section "Fitness as-if" of the main text, a standard concept for the prediction of individual behavior is that of a Nash equilibrium trait profile, compared to which no individual 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). Let us now introduce such a payoff function for a class a individual, denoted $w_{\mathrm{I}a}$, and defined on the domain

$$w_{\mathrm{I}a} : \mathcal{X}_a \times \prod_{s\in\mathcal{C}} \mathcal{X}_s^{n_s-\delta_{sa}} \times \mathcal{X} \to \mathbb{R}_+ \quad \forall a \in \mathcal{C}, \tag{C.1}$$

such that $w_{\mathrm{I}a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ is the payoff to individual i, $x_i \in \mathcal{X}_a$ being now understood as any trait value that a given individual i could express instead of the trait that it actually expresses, when group neighbors express trait profile $\boldsymbol{x}_{-i}$, and in a population where the average trait expression is $\bar{x} \in \mathcal{X}$. This payoff function has exactly the same domain as the individual fitness function (eq. A.3).

Suppose now each individual in the population is envisioned as an autonomous decision-maker, "choosing" the trait it expresses independently of each other individual and with a striving to maximize its payoff function $w_{\mathrm{I}a}$ (hence x_i can vary and is not genetically determined). Then, a Nash equilibrium $x^* = (x_1^*, x_2^*, ..., x_{n_c}^*)$ (symmetric in each class) satisfies

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

where $\boldsymbol{x}_{-i}^*$ is the trait profile of all neighbors, where entry j of $\boldsymbol{x}_{-i}^*$ is equal to x_s^* if neighbor $j \neq i$ is of class s. In such a symmetric Nash equilibrium x^*, individuals in each class make the best decision for themselves in terms of payoff, i.e., they maximize their payoff based on all others doing the same.

Our aim is to elicit a representation of the payoff function $w_{\mathrm{I}a}$ that individuals appear to maximize in an uninvadable population state. We call such a payoff a fitness as-if. More formally, a fitness as-if function $w_{\mathrm{I}a}$ satisfies

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

where the invasion fitness in the right-hand side is given by eqs. (A.27)–(A.29). Eq. (C.3) says that if x^* is uninvadable, then this equilibrium can be envisioned as a Nash equilibrium, where each individual appears to maximize its fitness as-if, when each other individual in each class exhibits fitness-maximizing behavior. In other words, in an uninvadable population state, it is as if each individual maximizes its fitness as-if.

In this Appendix, we not only prove that eq. (6) of the main text satisfies eq. (C.3), but more generally explain how to construct expressions for fitness as-if that take the form of both average

direct fitness and inclusive fitness. Thereby, this Appendix fully connects together traditional game theory, evolutionary invasion analysis and inclusive fitness theory.

C.2 Average direct fitness as-if

C.2.1 The instrumental distribution

We start by presenting a way to construct an average direct fitness as-if as this will pave the way to construct inclusive fitness as-if. Since we aim that the trait of each individual and thus of neighbors can be distinct from each other, we need to depart 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_{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 sample space 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 of the profile of traits that neighbors play, 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 sample space, beginning with a haploid population without class structure.

C.2.2 Haploids without classes

For haploids without classes, where $w_{\mathrm{I}}(x_i,\boldsymbol{x}_{-i},\bar{x})$ 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 sample space defined from the 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 set $\mathcal{P}_k(\boldsymbol{x}_{-i})$ of hypothetical neighbor trait profiles $\tilde{\boldsymbol{x}}_{-i}$ such that exactly $k-1$ components of the 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 $\boldsymbol{\sigma}(x_i,\boldsymbol{x}_{-i},\bar{x})$ on the space $\mathcal{S}_i$ 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

$$\sum_{k=1}^{n}\sum_{\tilde{\boldsymbol{x}}_{-i}\in\mathcal{P}_k(\boldsymbol{x}_{-i})}\sigma_k(\tilde{\boldsymbol{x}}_{-i};x_i,\boldsymbol{x}_{-i},\bar{x})=1. \tag{C.4}$$

The instrumental distribution $\boldsymbol{\sigma}(x_i,\boldsymbol{x}_{-i},\bar{x})$ is as yet undefined beyond its sample space. In

particular, this distribution has yet no imposed relation to the probabilities of events specified by the $\boldsymbol{q}^{\mathrm{D}}(x,y)$ distribution that occur in the actual reproductive process in the population under consideration, but it retains the ability to describe within-group correlated trait expression. In order to connect these two distributions, we note that $\boldsymbol{q}^{\mathrm{D}}(x,y)$ takes values in the simplex $\Delta(\mathcal{S})$ or ordered phenotypic groups states, which is the same as the simplex $\Delta(\mathcal{S}_{(i)})$. Thus, $\Delta(\mathcal{S}) = \Delta(\mathcal{S}_{(i)})$, and we can choose the instrumental distribution such that

$$\boldsymbol{\sigma}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \boldsymbol{q}^{\mathrm{D}}(x_i, \bar{x}). \tag{C.5}$$

So 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).

Given the instrumental distribution defined by eq. (C.5), we can define the average direct fitness as-if of an individual with trait x_i as

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) \;=\; \sum_{k=1}^{n} \sum_{\tilde{\boldsymbol{x}}_{-i} \in \mathcal{P}_k(\boldsymbol{x}_{-i})} w(x_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x}) \sigma_k(\tilde{\boldsymbol{x}}_{-i}; x_i, \boldsymbol{x}_{-i}, \bar{x}) \,, \tag{C.6}$$

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

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \mathrm{E}_{\tilde{\boldsymbol{x}}_{-i} \sim \boldsymbol{\sigma}(x_i, \boldsymbol{x}_{-i}, \bar{x})}[w(x_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x})] \,, \tag{C.7}$$

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

C.2.3 Diploids with classes

We can now generalize the construction of the instrumental distribution to a diploid class-structured population. For this case, it is useful to denote explicitly by $x_{i,a}$ the realized trait of individual $i \in \mathcal{I}$ when of class a. Then, the instrumental 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 a is defined as follows. First, the trait $\tilde{x}_i$ of individual i when of class a is assumed to take values in $\{z_a(x_{i,a}, \bar{x}_a), x_{i,a}\}$ and this parallels the case where in the genetic process an individual with the mutant allele can be heterozygote or homozygote, but our assignment is here a defining feature and is not intended to reflect any genetic reality. Second, each element $\tilde{x}_j$ of $\tilde{\boldsymbol{x}}_{-i} = (\tilde{x}_1, , ..., \tilde{x}_{i-1}, \tilde{x}_{i+1}, ...x_n)$ is defined to take values in the set of traits belonging to the class of the individual under scrutiny; that is, if individual j is of class s then, by construction $\tilde{x}_j \in \{z_s(x_{i,s}, x_{j,s}), x_{i,s}, x_{j,s}\}$. The hypothetical profile $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})$ is then defined to be distributed according to $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$, a distribution that

has sample space

$$\mathcal{S}_{s(i)} = \{z_s(x_{i,s}, \bar{x}_s), x_{i,s}\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa})} \{z_a(x_{i,a}, x_{j,a}), x_{i,a}, x_{j,a}\}. \tag{C.8}$$

Hence, the distribution $\boldsymbol{\sigma}_s$ takes values in the set $\Delta(\mathcal{S}_{s(i)})$, which is the simplex generated by the support $\mathcal{S}_{s(i)}$. Since the distribution $\boldsymbol{q}_s^{\mathrm{D}}(x, y)$ determining invasion fitness takes values in the same simplex $\Delta(\mathcal{S}_s) = \Delta(\mathcal{S}_{s(i)})$ (recall eq. A.30), we can, as in the haploid case, set the probabilities of events in the instrumental distribution identical to the probabilities of ordered trait profiles in the population genetic model:

$$\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \boldsymbol{q}_s^{\mathrm{D}}(x_i, \bar{x}). \tag{C.9}$$

Thus, we again define the instrumental probabilities, of events defined by replacing elements of the trait profile by the focal individual's trait, to be identical to the probabilities of ordered trait profiles in the population genetic model.

Given the instrumental distribution $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})$ so defined (eq. C.9), 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 neighbor 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

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

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 (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: it expresses either trait x_{-i} ($j = \mathrm{S}$ for "self"), or 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 neighbor expresses trait of type j. In terms of these probabilities,

we can write the instrumental distribution of profiles experienced by individual i as

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

Note that 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 ultimately owes to the fact that in the original reproductive process individuals express different traits upon being heterozygote or homozygote (e.g., eq. A.29), a standard modeling assumption for diploids (e.g., Nagylaki, 1992; Gillespie, 2004; Hartl and Clark, 2007).

C.3 Inclusive fitness as-if

C.3.1 Multiplayer regression for class-structure

We now present the regression analysis underlying the construction of inclusive fitness as-if and do so directly for diploids with class structure. To that end, we consider the same regression model as in the population genetic model (recall eqs. B.14–B.15) but will evaluate its coefficients under the instrumental distribution instead of the genetic contextual distribution. 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 $\tilde{\boldsymbol{x}}_{-i}$. We then write the number u of descendants of an individual i of class s in the altered group as

$$\begin{aligned} w_{us}(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}, \bar{x}) = w_{us}(\bar{x}_s, \bar{\mathbf{x}}, \bar{x}) - c_{\mathrm{I}us}(x_i, \boldsymbol{x}_{-i}, \bar{x}) p_{\mathrm{I},i} \\ + \sum_{j\neq i} b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i, \boldsymbol{x}_{-i}, \bar{x}) p_{\mathrm{I},j} + \text{residual}, \end{aligned} \tag{C.13}$$

where $\bar{\mathbf{x}}$ denotes the vector of neighbor trait profile (dimension $n-1$), all of which are evaluated at the mean (of the corresponding class) in the population and $c_{\mathrm{I}us}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ and $b^{\mathrm{N}}_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ are regression coefficients. The functional form of this equation is the same as that under the population genetic model (eq. B.14), but its interpretation slightly 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, which plays the same role as the mutant allele in being a determinant of the traits of a focal individual i and its neighbors. For the focal individual itself, the new trait value $\tilde{x}_i$ is within the set $\{z_s(x_{i,s}, \bar{x}_s), x_{i,s}, \bar{x}_s\}$, when it bears 2, 1, or 0 copies of the I allele, and $p_{\mathrm{I},i} \in \{0, 1/2, 1\}$ denotes the frequency of allele I in individual i. The new trait $\tilde{x}_j$ of a neighbor individual $j \neq i$ when of realized class $k(j) = a$ is within the set $\{z_a(x_{i,a}, x_{j,a}), x_{i,a}, x_{j,a}\}$, which stems, respectively, from individual $j \neq i$ bearing 2, 1, or 0 copies of the I allele. 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.13) must hold for all profiles $(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}) \in \{z_s(x_{i,s}, \bar{x}_s), x_{i,s}, \bar{x}_s\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa})} \{z_a(x_{i,a}, x_{j,a}), x_{i,a}, x_{j,a}\}$. Hence, the

main structural difference between eq. (C.13) and its population genetic counterpart, eq. (B.14), is that all neighbors of individual i have all distinct traits in eq. (C.13) when they do not switch to expressing allele I.

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 exactly the same line of argument as in the population genetic model, but using the instrumental distribution instead of the contextual genetic state distribution. Accordingly, we first note that eq. (C.13) says that we predict fitness with the same linear regression

$$\hat{w}_{us}\left(c_{\mathrm{I}us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}, \mathbf{p}_{\mathrm{I}}, \bar{x}\right) = w_{us}(\bar{x}_s, \bar{\boldsymbol{x}}, \bar{x}) - c_{\mathrm{I}us}(x_i, \boldsymbol{x}_{-i}, \bar{x})p_{\mathrm{I},i} + \sum_{j\neq i} b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i, \boldsymbol{x}_{-i}, \bar{x})p_{\mathrm{I},j}, \quad \text{(C.14)}$$

where vector $\boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}$ collects the $b_{\mathrm{I}us\leftarrow k(j)}$ regression coefficient and vector $\mathbf{p}_{\mathrm{I}}$ collects all the $p_{\mathrm{I},i}$ frequencies. Second, we let $\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ and $\tilde{\boldsymbol{\sigma}}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})$ denote the distribution of group states in a population where the vector of frequencies of allele I across classes is equated to the vector $\boldsymbol{p}$ of mutant allele frequencies across classes in the population genetic model (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, sample space $\{z_s(x_{i,s}, \bar{x}_s), x_{i,s}\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa})} \{z_a(x_{i,a}, x_{j,a}), x_{i,a}, x_{j,a}\}$ and $\{x_{i,s}, \bar{x}_s\} \times \prod_{a\in\mathcal{C}} \prod_{j\neq i}^{(n_a-\delta_{sa}))} \{z_a(x_{i,a}, x_{j,a}), x_{i,a}, x_{j,a}\}$). These sample spaces induce the sample simplexes as the $\tilde{\boldsymbol{q}}^{\mathrm{D}}_s$ and $\boldsymbol{q}^{\mathrm{D}}_s$ distributions (used in eq. B.16) and so we set

$$\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) = \boldsymbol{q}^{\mathrm{D}}_s(x_i, \bar{x}, \boldsymbol{p}) \quad \text{and} \quad \tilde{\boldsymbol{\sigma}}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}) = \tilde{\boldsymbol{q}}^{\mathrm{D}}_s(x_i, \bar{x}, \boldsymbol{p}). \quad \text{(C.15)}$$

We can now define the quadratic expression

$$\begin{aligned} Q_{\mathrm{I}us}\left(c_{\mathrm{I}us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}, x_i, \boldsymbol{x}_{-i}, \boldsymbol{p}\right) &= \mathrm{E}_{(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})}\left[\left(\hat{w}_{us}\left(c_{\mathrm{I}us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}, \mathbf{p}_{\mathrm{I}}, \bar{x}\right) - w_{us}(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}, \boldsymbol{p})\right)^2\right] p_s \\ &+ \mathrm{E}_{(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i})\sim\tilde{\boldsymbol{\sigma}}_s(x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})}\left[\left(\hat{w}_{us}\left(c_{\mathrm{I}us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}, \mathbf{p}_{\mathrm{I}}, \bar{x}\right) - w_{us}(\tilde{x}_i, \tilde{\boldsymbol{x}}_{-i}, \boldsymbol{p})\right)^2\right] (1 - p_s), \quad \text{(C.16)} \end{aligned}$$

and we solve $\partial Q_{\mathrm{I}us}(c_{\mathrm{I}us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}, x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})/\partial c_{\mathrm{I}us} = 0$ and $\partial Q_{\mathrm{I}us}(c_{\mathrm{I}us}, \boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}us}, x_i, \boldsymbol{x}_{-i}, \boldsymbol{p})/\partial b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)} = 0$ for all $j \neq i$. Recalling eq. (B.20), we obtain the regression coefficients of eq. (C.13) as

$$c_{\mathrm{I}us}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \lim_{p\to 0} c_{\mathrm{I}us}(x_i, \boldsymbol{x}_{-i}, \bar{x}, \boldsymbol{p}) \qquad b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \lim_{p\to 0} b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i, \boldsymbol{x}_{-i}, \bar{x}, \boldsymbol{p}). \quad \text{(C.17)}$$

We now further note that when $\boldsymbol{q}^{\mathrm{D}}_s(x_i, \bar{x}, \boldsymbol{p}) \to \boldsymbol{q}^{\mathrm{D}}_s(x_i, \bar{x})$ 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

$$\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x}) = \boldsymbol{q}^{\mathrm{D}}_s(x_i, \bar{x}), \quad \text{(C.18)}$$

and

$$\mathrm{E}_{(\tilde{x}, \tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_s(x_i, \boldsymbol{x}_{-i}, \bar{x})}[p_{\mathrm{I},i}] = \mathrm{E}_{(x_i, \boldsymbol{x}_{-i})\sim\boldsymbol{q}^{\mathrm{D}}_s(x_i, \bar{x})}[p_{\mathrm{I},i}] = r_{\mathrm{f}s}(x_i, \bar{x}), \quad \text{(C.19)}$$

where the first equality follows from eq. (C.18) and the definition of $p_{\mathrm{I},i}$, which takes exactly the same frequency as the mutant allele within an individual under assumption (C.18); while the second equality follows from the definition of the within-individual identity in state under neutrality (recall eq. B.22). Likewise, when individual $j \neq i$ is of class a, we have

$$\mathrm{E}_{(\tilde{x},\tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_{\mathrm{s}}(x_i,\boldsymbol{x}_{-i},\bar{x})}\left[p_{\mathrm{I},j}\right] = \mathrm{E}_{(x_i,\boldsymbol{x}_{-i})\sim\boldsymbol{q}_s^{\mathrm{D}}(x_i,\bar{x})}\left[p_{\mathrm{I},j}\right] = r_{\mathrm{n}\,a|s}(x_i,\bar{x}). \tag{C.20}$$

The first equality follows from eq. (C.18) and the definition of $p_{\mathrm{I},j}$, while the second equality follows from the definition of the between-individual identity in state under neutrality and the fact that the distribution of allele I is the same as that of the mutant allele (recall eq. B.23). In terms of the $r_{\mathrm{f}s}(x_i,\bar{x})$ and $r_{\mathrm{n}\,a|s}(x_i,\bar{x})$ probabilities, we have that

$$r_{s|a}(x_i,\bar{x}) = \frac{r_{\mathrm{n}\,s|a}(x_i,\bar{x})}{r_{\mathrm{f}a}(x_i,\bar{x})}, \tag{C.21}$$

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

By contrast to the regression coefficients in the population genetic model, the coefficient $b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i,\boldsymbol{x}_{-i},\bar{x})$ (eq. C.17) takes only the interpretation of a neighbor-modulated regression coefficient, since it gives the effect, on the fitness of i, of an average social partner switching to expressing trait x_i instead of its own trait x_j. This in general is not the effect of individual i on the fitness of its average social partner when switching its own trait value from $\bar{x}$ to x_i (since in general $x_j \neq \bar{x}$). We next show how to obtain an actor-centered regression coefficient out of $b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i,\boldsymbol{x}_{-i},\bar{x})$.

C.3.2 Actor-centered regression coefficients

To introduce our argument, we first consider haploids without class structure. In that case, we set $b^{\mathrm{N}}_{\mathrm{I}us\leftarrow k(j)}(x_i,\boldsymbol{x}_{-i},\bar{x}) = b^{\mathrm{N}}_{\mathrm{I},j}(x_i,\boldsymbol{x}_{-i},\bar{x})$ since there are no class effects. It will also turn out useful to re-write the quadratic expression (eq. C.16) for the haploid case in more compact form. To that end, let $\boldsymbol{\sigma}_{\mathrm{I}}(x_i,\boldsymbol{x}_{-i},p)$ denote the full unconditional instrumental distribution (which concatenates the unconditional instrumental distributions $\boldsymbol{\sigma}(x_i,\boldsymbol{x}_{-i},p)$ and $\tilde{\boldsymbol{\sigma}}(x_i,\boldsymbol{x}_{-i},p)$ weighted by the respective allele I frequencies), and that has sample space

$$\Omega_{\mathrm{I}} = \{x_i,\bar{x}\} \times \prod_{j\neq i}\{x_i,x_j\}. \tag{C.22}$$

With this, we can then write the quadratic expression for haploids as

$$\mathrm{Q}_{\mathrm{I}}\left(c_{\mathrm{I}},\boldsymbol{b}^{\mathrm{N}}_{\mathrm{I}},x_i,\boldsymbol{x}_{-i},p\right) = \mathrm{E}_{(\tilde{x}_i,\tilde{\boldsymbol{x}}_{-i})\sim\boldsymbol{\sigma}_{\mathrm{I}}(x_i,\boldsymbol{x}_{-i},\boldsymbol{p})}\left[\left(1 - c_{\mathrm{I}}p_{\mathrm{I},i} + \sum_{j\neq i} b^{\mathrm{N}}_{\mathrm{I},j}p_{\mathrm{I},j} - w(\tilde{x}_i,\tilde{\boldsymbol{x}}_{-i},p)\right)^2\right]. \tag{C.23}$$

Eq. (C.23) illustrates that, given that we assign a distribution to the traits expressed by different individuals (the $\boldsymbol{\sigma}_{\mathrm{I}}$ distribution), the quadratic expression to be minimized can be represented as an average over a distribution of the traits. The resulting regression coefficients are contrasts of the fitness values (i.e., weighted "averages" except that the sum of the weights is zero), which can no longer be interpreted as averages over the $\boldsymbol{\sigma}_{\mathrm{I}}$ distribution, but whose contrast weights are still fully determined by this distribution (as implied by the so-called "normal equation" of linear regression). For instance, by performing from eq. (C.23) the calculations detailed in the last section, we obtain in the case $n = 2$ (where the sample space is $\{x_i, \bar{x}\} \times \{x_i, x_j\}$),

$$b^{\mathrm{N}}_{\mathrm{I},j}(x_i, x_j, \bar{x}) = \frac{1}{1 + r(x_i, \bar{x})} \left(w(\bar{x}, x_i, \bar{x}) - w(\bar{x}, x_j, \bar{x})\right) + \frac{r(x_i, \bar{x})}{1 + r(x_i, \bar{x})} \left(w(x_i, x_i, \bar{x}) - w(x_i, x_j, \bar{x})\right) \tag{C.24}$$

in the limit $p \to 0$, which can be thought of as the extension of $b(x, y)$ in eq. (B.11) to the case where the trait of each group member is distinguished.

When individuals are exchangeable (meaning that the distribution of their traits are exchangeable, and that the same fitness function holds for all individuals), then this effect is also an average $b_{\mathrm{I},j}$ of effects of the focal individual on the jth neighbor's fitness (and thus $b_{\mathrm{I},j} = b^{\mathrm{N}}_{\mathrm{I},j}$). This case holds when one considers mutant-resident dynamics (whether the resident's state is an uninvadable state, or not) and underlies the dual interpretation of the regression coefficients discussed below eq. (B.7). Otherwise, even if the instrumental distribution $\boldsymbol{\sigma}_{\mathrm{I}}$ is exchangeable between individuals, the traits (sample space) expressed by individuals that do not bear the I allele are not generally exchangeable. For instance, in the above haploid case, the trait sample space of individual i is $\{x_i, \bar{x}\}$ and $\{x_i, x_j\}$ and that of its neighbor j is $\{x_i, x_j\}$. So these are not exchangeable and we do not have $b_{\mathrm{I},j} = b^{\mathrm{N}}_{\mathrm{I},j}$.

Rather, in order to construct an additive effect $b_{\mathrm{I},j}$ of the focal individual on the jth neighbor's fitness, we let $b_{\mathrm{I},j}$ be the regression coefficient to $p_{\mathrm{I},j}$ in the quadratic expression eq. (C.23), still under the instrumental distribution defined from the context of individual i, but when the traits expressed by the focal and its neighbor j are exchanged (in general, this is also different from the regression coefficient to $p_{\mathrm{I},j}$ under the instrumental distribution defined from the context of individual j). For example, consider that x_j is equal to x_i, except that we still denote it x_j. Then, under the instrumental distribution, $p_{\mathrm{I},j}$ has no effect on the focal individual's fitness, since it has no effect on expressed trait, $x_i = x_j$. Yet, the focal individual has a distinct effect on its j-neighbor depending on its own $p_{\mathrm{I},j}$ and this must be reflected by a non-zero $b_{\mathrm{I},j}$ in the inclusive fitness as-if of individual i. This effect is recovered by switching the supports of the focal's trait and of the j-neighbor's trait, so that the focal trait now has sample space $\{x_i, x_j\}$, but more importantly the neighbor's trait now has sample space $\{x_i, \bar{x}\}$ so that we can compute a non-zero $b_{\mathrm{I},j}$ as the regression coefficient of focal fitness on neighbor's $p_{\mathrm{I},j}$.

Formally, this means that once we have an expression, for $b^{\mathrm{N}}_{\mathrm{I},j}$ from the regression of i's fitness under the instrumental distribution, as a contrast of individual fitnesses in different contexts, we

obtain $b_{\mathrm{I},j}$ by keeping the contrast weights constant, but modifying the fitness values by switching the sample spaces of the traits expressed by the I allele in individuals i and j. In the exemplar case $n = 2$ given above (eq. C.24), the switch of the sample spaces $\{x_i, \bar{x}\}$ and $\{x_i, x_j\}$ then leads to simply interchanging $\bar{x}$ and x_j in eq. (C.24). This produces

$$b_{\mathrm{I},j}(x_i, x_j, \bar{x}) = \frac{1}{1 + r(x_i, \bar{x})} \left(w(x_j, x_i, \bar{x}) - w(x_j, \bar{x}, \bar{x}) \right) + \frac{r(x_i, \bar{x})}{1 + r(x_i, \bar{x})} \left(w(x_i, x_i, \bar{x}) - w(x_i, \bar{x}, \bar{x}) \right), \tag{C.25}$$

which is an effect of the focal individual on the jth neighbor's fitness. In terms of this coefficient, we can then define the inclusive fitness as-if of individual i as

$$w_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = 1 - c_{\mathrm{I}}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + r(x_i, \bar{x}) \sum_{j \neq i} b_{\mathrm{I},j}(x_i, x_j, \bar{x}). \tag{C.26}$$

More generally, for diploidy and class-structure the same argument applies. Here, once we have the neighbor-modulated coefficient $b^{\mathrm{N}}_{\mathrm{I}us \leftarrow k(j)}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ for $k(j) = a$ from the regression of i's fitness under the instrumental distribution, as a contrast of individual fitnesses in different contexts, we can obtain $b_{\mathrm{I}uk(j) \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ as the effect of an individual i of class a on receptor j of class $k(j) = s$ by keeping the contrast weights constant, but modifying the fitness values by switching the sample spaces of the traits expressed by the focal in class s and its neighbor j when of class a; that is, by switching the supports of the focal's trait and of the j-neighbor's trait, so that the focal's trait (now of class s) has support $\{z_s(x_{s,i}, \bar{x}_s), x_{s,i}, \bar{x}_s\}$ and the neighbor's trait (now of class a) has support $\{z_a(x_{i,a}, x_{j,a}), x_{i,a}, x_{j,a}\}$. This then leads a non-zero $b_{\mathrm{I}uk(j) \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x})$ additive effect of a focal of class a on its neighbor j of class $k(j) = s$, whose sum over all recipients of class s, denoted

$$b_{\mathrm{I}us \leftarrow a} = \sum_{j \neq i, k(j) = s} b_{\mathrm{I}uk(j) \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}), \tag{C.27}$$

is an individual-centered version of eq. (B.31). In terms of the so-obtained actor-centered indirect coefficient (and the $c_{\mathrm{I}ua}$ and $r_{s|a}(x_i, \bar{x})$ coefficients computed in the previous section, recall eq. C.17 and eq. C.21), we can define the inclusive fitness as-if of an individual i of class a as

$$w_{\mathrm{I}a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = v_a(\bar{x}) + \sum_{u \in \mathcal{C}} v_u(\bar{x}) \left[-c_{\mathrm{I}ua}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + \sum_{s \in \mathcal{C}} r_{s|a}(x_i, \bar{x})\, b_{\mathrm{I}us \leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) \right]. \tag{C.28}$$

C.4 Connecting the gene-centered and the rational-actor centered perspectives

Suppose now that at an uninvadable state x^*, we have

$$\begin{aligned} c_{\mathrm{I}ua}(x_i, \boldsymbol{x}^*, x^*) &= c_{ua}(x_i, x^*) \qquad \forall x_i \in \mathcal{X}_a \\ b_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}^*, x^*) &= b_{us\leftarrow a}(x_i, x^*) \qquad \forall x_i \in \mathcal{X}_a. \end{aligned} \tag{C.29}$$

Then, the inclusive fitness as-if at x^* is equivalent to the inclusive fitness at x^*:

$$w_{\mathrm{I}a}(x_i, \boldsymbol{x}^*, x^*) = w_{\mathrm{IF}a}(x_i, x^*) \qquad \forall x_i \in \mathcal{X}_a. \tag{C.30}$$

Since we know the right-hand side is maximized at x^*, i.e., eq. (B.39) is satisfied, it follows from eq. (C.30) that the inclusive fitness as-if (eq. C.28) satisfies eq. (C.3). Thus, in an uninvadable population state, it is as if each individual aimed to maximize its inclusive fitness as-if, when all others exhibit fitness-maximizing behavior.

We now show that eq. (C.29) indeed holds. First, we have both $c_{\mathrm{I}ua}(x_i, \boldsymbol{x}^*, x^*) = c_{ua}(x_i, x^*)$ and $b^{\mathrm{N}}_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}^*, x^*) = b^{\mathrm{N}}_{us\leftarrow a}(x_i, x^*)$ because the regression model eq. (C.13) is the same as that in the population genetic model (eq. B.14), the only difference is that its regression coefficients are evaluated under a distribution where group neighbor traits are all distinct, everything else remaining the same. Hence, in a population where all neighbors of the same class of an individual have the same trait value, the regression coefficients computed under the instrumental and contextual distributions will be the same. It now remains to show that $b_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}^*, x^*) = b_{us\leftarrow a}(x_i, x^*)$. This follows by noting that in the computation leading to $b_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}^*, x^*)$, the only thing that is changed, is an exchange of supports in the evaluation of fitnesses, but the position of the argument x_i remains unchanged in all fitness functions. The nature of the maximization problem remains henceforth unchanged, because it is only the trait of neighbors that are interchanged. This has no effect on the symmetric Nash equilibrium (eq. C.3) and so when all neighbors play the same trait, the regression coefficients will be the same. Hence, in an uninvadable population state, it is as-if individuals in each class maximize (in the best-response sense) their own inclusive fitness defined by eq. (C.28).

As explained in the main text, eq. (C.28) is, however, not an operationally 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. But since the contextual distribution determining the instrumental distribution (by way of eq. C.18) 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.

C.5 Weak selection

C.5.1 Weak-selection concepts

We now finally turn to deriving an inclusive 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

$$\boldsymbol{A}(x,y) = \boldsymbol{A}(y) + \epsilon\tilde{\boldsymbol{A}}(x,y) + O(\epsilon^2), \tag{C.31}$$

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.

The representation given in eq. (C.31) captures the two kinds of weak selection that we discussed in section "Weak selection concepts" of the main text[3]. 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, in which case one can use the approximation $\boldsymbol{q}(x,y) \sim \boldsymbol{q}(y)$. 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, in which case one can use the approximation $\boldsymbol{q}(x,y) \sim \boldsymbol{q}$ and $\boldsymbol{\phi}(x,y) \sim \boldsymbol{\phi}$, 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.31) is neglected and

$$q_{\mathbf{k}|a}(x,y) \to q_{\mathbf{k}|a}(y) \qquad \text{and} \qquad \phi_a(x,y) \to \phi_a(y), \tag{C.32}$$

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-

[3] 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 1. Then, we can first Taylor-expand the fitness components, say the number of daughters produced by queens (eq. B.2), in mutant trait around the resident trait and neglect higher-order terms to obtain

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

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 $w_{\mathrm{ff}}(x,y) \sim \tilde{w}_{\mathrm{ff}}(x,y) = \frac{x_{\mathrm{f}}}{2y_{\mathrm{f}}} + \frac{x_{\mathrm{f}}\left(P(x_{\mathrm{o}}) - P(y_{\mathrm{o}})\right)}{4y_{\mathrm{f}}}$ where the right-hand side represents small-parameter approximation to fitness.

parameter" weak selection. Eq. (C.32) 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., $q_{\mathbf{k},a}(y) = k_a u_{\mathbf{k}}(y) / [\sum_{\mathbf{k}\in I}\sum_{a\in\mathcal{C}} k_a u_{\mathbf{k}}(y)]$, 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, we have for weak selection that

$$\boldsymbol{q}_a^{\mathrm{D}}(x,y) \to \boldsymbol{q}_a^{\mathrm{D}}(y). \tag{C.33}$$

We will next apply eq. (C.33) 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.5.2 Class-specific inclusive fitness as-if maximization under weak selection

First, recall that, regardless for any strength of selection, we showed that each individual of each class will appear to maximize its inclusive fitness defined by eq. (C.28) in an uninvadable population state (e.g., eq. C.30, which implies that eq. (C.3) is satisfied) regardless of the the strength of selection. Hence, eq. (C.28) will also be maximized under weak selection (eq. (C.3) is satisfied), with the only difference that we can let the population genetic structure be independent of the actor's traits. Hence, for weak selection we write the inclusive fitness as-if of individual i of class a as

$$w_{\mathrm{I}a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) = v_a(\bar{x}) + \sum_{u\in\mathcal{C}} v_u(\bar{x}) \left[-c_{\mathrm{I}ua}(x_i, \boldsymbol{x}_{-i}, \bar{x}) + \sum_{s\in\mathcal{C}} r_{s|a}(\bar{x})\, b_{\mathrm{I}us\leftarrow a}(x_i, \boldsymbol{x}_{-i}, \bar{x}) \right]. \tag{C.34}$$

This is eq. (6) of the main text and the key difference with eq. (C.28) is that relatedness is now independent of the actor's trait and the regression coefficients are evaluated under weak selection using the fitness function $\tilde{w}_{ua}$ (instead of the fitness function w_{ua}) and using the weak selection approximation of the instrumental distribution, everything else remains the same.

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. 1979. Darwinism and Human Affairs. University of Washington Press, Seattle.

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.

Brown, J. S. and T. L. Vincent. 1987. Coevolution as an evolutionary game. *Evolution* 41:66–79.

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.

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.

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.

Dercole, F. and S. Rinaldi. 2008. Analysis of Evolutionary Processes: The Adaptive Dynamics Approach and Its Applications. Princeton University Press, Princeton, NJ.

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.

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.

Fisher, R. A. 1941. Average excess and average effect of a gene substitution. *Annals of Human Genetics* 11:53–63.

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.

Geritz, S. A. H., E. Kisdi, G. Meszéna, and J. A. J. Metz. 1998. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology* 12:35–57.

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.

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.

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.

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.

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.

Lion, S. 2018. Theoretical approaches in evolutionary ecology: environmental feedback as a unifying perspective. *American Naturalist* 191:21–44.

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., 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.

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.

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

Ronce, O., S. Gandon, and F. Rousset. 2000. Kin selection and natal dispersal in an age-structured population. *Theoretical Population Biology* 58:143–159.

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.

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.

Taper, M. L. and T. J. Case. 1992. Models of character displacement adn the theoretical robustness of taxon cycles. *Evolution* 46:317–333.

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.

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.