PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

**Cite this article:** Fromhage L, Jennions MD. 2019 The strategic reference gene: an organismal theory of inclusive fitness. *Proc. R. Soc. B* **286**: 20190459. http://dx.doi.org/10.1098/rspb.2019.0459

Received: 23 February 2019
Accepted: 12 May 2019

**Subject Category:**
Evolution

**Subject Areas:**
behaviour, evolution, theoretical biology

**Keywords:**
social evolution, kin selection, adaptation, Hamilton's rule, causality, selfish gene

**Author for correspondence:**
Lutz Fromhage
e-mail: lutz.fromhage@jyu.fi

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4510190.

THE ROYAL SOCIETY PUBLISHING

# The strategic reference gene: an organismal theory of inclusive fitness

Lutz Fromhage[1] and Michael D. Jennions[2]

[1]Department of Biological and Environmental Science, University of Jyväskylä, PO Box 35, 40014 Jyväskylä, Finland
[2]Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, Australian Capital Territory 2601, Australia

LF, 0000-0001-5560-6673; MDJ, 0000-0001-9221-2788

How to define and use the concept of inclusive fitness is a contentious topic in evolutionary theory. Inclusive fitness can be used to calculate selection on a focal *gene*, but it is also applied to whole *organisms*. Individuals are then predicted to appear designed as if to maximize their inclusive fitness, provided that certain conditions are met (formally when interactions between individuals are 'additive'). Here we argue that applying the concept of inclusive fitness to organisms is justified under far broader conditions than previously shown, but only if it is appropriately defined. Specifically, we propose that organisms should maximize the sum of their offspring (*including* any accrued due to the behaviour/phenotype of relatives), plus any effects on their relatives' offspring production, weighted by relatedness. By contrast, most theoreticians have argued that a focal individual's inclusive fitness should exclude any offspring accrued due to the behaviour of relatives. Our approach is based on the notion that long-term evolution follows the genome's 'majority interest' of building coherent bodies that are efficient 'vehicles' for gene propagation. A gene favoured by selection that reduces the propagation of unlinked genes at other loci (e.g. meiotic segregation distorters that lower sperm production) is eventually neutralized by counter-selection throughout the rest of the genome. Most phenotypes will therefore appear as if designed to maximize the propagation of any given gene in a focal individual and its relatives.

> Perhaps we should not feel entirely confident about generalizing our principle until a more comprehensive mathematical argument, with inclusive fitness more widely defined, has been worked out. Hamilton [1, p. 18]

## 1. Introduction

What, if anything, are organisms shaped by evolution adapted to achieve [2–4]? To answer this question, consider the fact that natural selection is roughly analogous to trial-and-error learning: mutations create gene variants which affect the phenotypes of organisms expressing them; variants then spread if their causal effects on the world, mediated by how they affect the phenotype, aid their propagation [5]. Accordingly, it is a truism that any naturally selected trait can be said to have evolved because genes contributing to the trait in past generations were more successful than their alternatives at leaving copies in the present. But what kinds of phenotypes will successful genes contribute to building? Hamilton made a major breakthrough in answering this question [6,7]. He distinguished two causal pathways by which a gene, expressed in a given organism, can aid its propagation. It can enhance the organism's own reproduction (direct fitness), and it can cause the organism to enhance the reproduction of others that carry the gene's identical copies (indirect fitness). To capture this insight, he defined *inclusive fitness* ($IF_{Hamilton}$) as a combined measure of direct and indirect fitness components [6, p. 8]:

> Inclusive fitness may be imagined as the personal fitness which an individual actually expresses in its production of adult offspring as it becomes after it has been first stripped and then augmented in a certain way. It is stripped of all components which can be considered as due to the individual's social environment, leaving the fitness which he would express if not exposed to any of the harms or benefits of that environment. This quantity is then augmented by certain fractions of the quantities of harm and benefit which the individual himself causes to the fitnesses of his neighbours. The fractions in question are simply the coefficients of relationship appropriate to the neighbours whom he affects: unity for clonal individuals, one-half for sibs, one-quarter for half-sibs, one-eighth for cousins, ... and finally zero for all neighbours whose relationship can be considered negligibly small.

Hamilton showed that $IF_{Hamilton}$ works as a *genetic accounting tool* to predict when a focal gene is positively selected, which occurs when an individual expressing it enjoys increased inclusive fitness. He inferred from this that the long-term outcome of successive genes being selected in this way is that organisms shaped by natural selection should be adapted to maximize $IF_{Hamilton}$. This would make $IF_{Hamilton}$ a *phenotypic maximand* [3]. The concept of a phenotypic maximand is useful for studying adaptation because we can then envisage individual organisms as maximizing agents with a defined biological purpose [8]. It allows us to predict that an organism's (naturally selected) traits tend to be shaped to cause a higher expected value of the maximand than feasible alternative traits. Organism-centred usage of inclusive fitness requires that $IF_{Hamilton}$ is a measurable property of an individual organism. To meet this requirement, $IF_{Hamilton}$ must use a concept of relatedness that is applicable to entire organisms (i.e. that approximately measures genetic similarity across the genome), rather than being only applicable to one gene at a time.

By contrast, inclusive fitness models often focus on a single gene, predicting that it will spread if it satisfies Hamilton's rule $rb - c > 0$ (where $r$ is relatedness, $-c$ and $b$ are changes caused to the reproduction of 'self' and 'other', and the left-hand side is defined as the gene's *inclusive fitness effect* [6,9]). This approach calls for a gene-specific (genic) definition of relatedness [10] which—unlike 'pedigree relatedness' between organisms—accounts for genetic similarity between individuals for a focal gene that can arise by processes that do not apply equally to all genes (e.g. non-random assortment of organisms with the focal gene). This difference in relatedness concepts indicates that the connection between gene-level selection and organism-level adaptation is not straightforward. Indeed, some theorists have even concluded that inclusive fitness is not a meaningful property of an organism [11–13]. If true, this precludes it being a phenotypic maximand (but see [3,9]). But do we really want to abandon the use of inclusive fitness when we study adaptations, which are usually complex traits determined by the effects of many genes?

Here we argue that invoking $IF_{Hamilton}$ as a general phenotypic maximand is problematic, but that these problems are surmounted if we redefine inclusive fitness. We start from the observation that genes with opposing phenotypic effects can simultaneously be selected for, due to gene-specific patterns of inheritance and expression (e.g. meiotic driver genes versus those for balanced meiosis). We then invoke a broad interpretation of the principle of the 'parliament of genes' [14] to predict how such opposing forces are likely to be resolved over evolutionary time. To operationally characterize the genome's 'majority interest', we invoke an idealized 'reference gene' whose interest in which phenotype is expressed always aligns with that of most other genes in the same organism. We then propose a modified definition of inclusive fitness based on a quantity whose maximization best serves the genome's 'majority interest'. Our goal is not to paint a precise picture of population genetic processes, but rather to argue for a higher-level principle that tends to guide cumulative phenotypic evolution in a coherent direction: namely, towards optimized design of individual organisms.

We consider a wide range of potential objections to our approach, which is likely to be controversial. However, to avoid too many asides, we relegate many of these objections to a 'questions and answers' list (electronic supplementary material). We also include a video that gives a non-technical overview of our ideas.

## 2. Reference genes and the parliament of genes

Any quantity that qualifies as a phenotypic maximand should tend to be increased through phenotypic changes induced by gene frequency changes due to natural selection. But, of course, organisms are integrated units shaped by selection on thousands of loci over long time spans, so not every positively selected gene needs to be a step towards increasing the maximand. Once a focal gene has spread and propelled a population along an evolutionary trajectory in phenotypic space, genetic variation at other loci determines how the trajectory continues. The focal gene's contribution could either be retained or eliminated. When studying long-term evolution, the common guiding question 'what kind of gene will be positively selected?' should therefore be complemented by adding '... such that its phenotypic effect is not eliminated in the long run'. A similar point was made by Leigh [14, p. 249] to account for fair meiosis being overwhelmingly common, despite the huge selective advantage that segregation distorter genes can enjoy. Leigh wrote: 'It is as if we had to do with a parliament of genes: each acts in its own self-interest, but if its acts hurt others, they will combine together to suppress it.' He explained the rarity of segregation distorters by invoking the principle that genes that oppose the genome's 'majority interest' are eliminated by counter-selection at other loci. Here we combine this idea with Dawkins's [4] vision of individual organisms as *vehicles* for gene propagation. Specifically, we postulate that the genome's 'majority interest' is to build an organism with high *vehicle quality*, which we define as an organism's general capacity to propagate its genes and their identical copies. To quantify vehicle quality, we envisage a hypothetical *reference gene* (more precisely, an allele) which is: (i) present in the focal organism, (ii) rare in the population, (iii) subject to Mendelian inheritance, and (iv) rarely or never expressed (i.e. low penetrance; assuming that other alleles at the same locus are never expressed). These properties are chosen in part (i, ii) to facilitate measuring gene propagation (essentially, by counting copies), and in part (iii, iv) so that the reference gene's evolutionary interest as to what phenotype should be expressed (i.e. the ranking of possible phenotypes with respect to how well they propagate the reference gene) aligns with the common interest of the organism's other genes.

We measure vehicle quality as the number of reference gene copies that can be causally attributed to the focal organism. These are the net number of additional copies that arise because the focal organism exists. Every sexually produced offspring of the focal organism contributes $s$ copies, and every offspring produced by a relative of degree $r$ accounts for $sr$ copies. Here, $s$ is the probability of transmitting a reference gene copy to a given offspring, which is given by the focal organism's consanguinity [15] with itself; in diploid, outbreeding populations, $s = 0.5$. The pedigree relatedness $r$ is the coefficient of relatedness [15] as applied to weakly selected genes due to coancestry. The number of propagated reference gene copies then sums to $s \cdot \sum (r \cdot \Delta n_r)$, where $\Delta n_r$ is the net number of offspring[1] produced (or not produced) by relatives of degree $r$ because the focal organism exists. It includes *all* of the focal organism's own offspring, for which $r = 1$. An individual's vehicle quality is maximized by the phenotype that causes the greatest representation of the reference gene in future generations. This occurs when the individual maximizes the expected value of $\sum (r \cdot \Delta n_r)$, which is the sum of its own offspring number, plus its effects on its relatives' number of offspring, weighted by relatedness. We call this the *folk definition of inclusive fitness*, $\mathrm{IF_{folk}}$, which has been described as a 'common misdefinition of inclusive fitness' [16, p. 426]. Unlike $\mathrm{IF_{Hamilton}}$, there is no 'stripping' of the social environment for $\mathrm{IF_{folk}}$. That is, all of the focal individual's own offspring count as being caused by it, in the sense that they would not have been produced if the focal organism had not existed and exhibited a phenotype with the requisite fertility.

To summarize, we postulate that the genome's 'majority interest' is to build an organism with high vehicle quality. Here, vehicle quality is the general capacity for gene propagation, which we propose to quantify as the number of reference gene copies that can be causally attributed to the organism. Since that number is proportional to $\mathrm{IF_{folk}}$, the number of reference gene copies is maximized when $\mathrm{IF_{folk}}$ is maximized. So, if evolution mainly follows the genome's majority interest, organisms should express traits that maximize their $\mathrm{IF_{folk}}$. The reference gene's property of being rarely expressed (hence weakly selected) justifies using a pedigree-based concept of relatedness for $\mathrm{IF_{folk}}$, which is also relevant for multi-locus evolution because coancestry is the only source of genetic similarity that promotes wide agreement across the genome as to what traits best serve each constituent gene's propagation [3,10].

What do we mean when we claim that organisms should behave so as to maximize their $\mathrm{IF_{folk}}$? In general, maximization occurs when a mathematical or physical function reaches its highest achievable output value through changes, within a specified range, in the values of its input arguments. In the present case, the function of interest is $\mathrm{IF_{folk}}$, and its argument is the individual organism's phenotypic strategy (including its propensity to help or harm, but also non-social traits). Formally, we can write this as $\mathrm{IF_{folk}}[\pi] = \sum (r \cdot \Delta n_r) | \mathrm{do}(\mathrm{phenotype} = \pi)$, where the 'do' operator (adopted from Pearl's causal modelling framework [17,18]) stands for 'set phenotype to $\pi$'. This formulation conveys the idea that any given phenotype belongs to a set of feasible options that could be generated by appropriate genotypes, or by experimental intervention. We then predict that organisms tend to exhibit phenotypes that yield higher $\mathrm{IF_{folk}}$ than feasible alternatives. Crucially, while $\mathrm{IF_{folk}}$ is useful for comparing phenotypes at a given time, in a given social environment, it does not measure changes in absolute fit between organisms and their environment over evolutionary time. Hence our prediction that phenotypes yielding higher $\mathrm{IF_{folk}}$ tend to evolve should not be misinterpreted as a claim that $\mathrm{IF_{folk}}$ increases over evolutionary time, towards a maximum at equilibrium. The environment that sets the background for evaluating $\mathrm{IF_{folk}}$ changes over time due to both abiotic and biotic factors, including frequency-dependent traits.

## 3. Rogue genes

Despite the parliament of genes, selection need not always increase vehicle quality. At least in the short term, the opposite can occur. Here we use the term *rogue genes* for genes that can generate selection for traits that reduce vehicle quality. Rogue genes include *Mendelian outlaw genes*, *greenbeard genes* and a previously undescribed type that we call *mirror effect rogue* (MER) *genes*. The existence of these kinds of genes is partly why some theoreticians are dubious about the usefulness of applying inclusive fitness to individual organisms. *Mendelian outlaw genes* spread at the expense of unlinked genes in the same organism by violating the laws of Mendelian inheritance. A meiotic drive gene that ends up in more than half of an organism's zygotes may spread, despite reducing the organism's reproductive output. However, a driver gene also selects for unlinked modifier genes that neutralize its phenotypic effect [19]. *Greenbeard genes* can spread by causing their bearer to (i) exhibit an cue (e.g. a green beard) and (ii) behave altruistically towards others bearing the cue [4,20]. Once a greenbeard gene has spread, the maintenance of its phenotypic effects relies on the genetic constraint that the cue (which enhances vehicle quality) cannot be expressed without the altruistic behaviour (which reduces vehicle quality). Eventually, this constraint should be undermined through selection for modifier genes that suppress the altruistic behaviour, but not the cue [21,22]. *MER genes* are particularly pertinent to deciding whether $\mathrm{IF_{folk}}$ qualifies as a phenotypic maximand, but we defer their definition until §5 as we must first introduce some additional concepts.

## 4. The mirror effect

There is a conceptual distinction between genes with and without a 'mirror effect'. The 'mirror effect' is a gene's tendency to be simultaneously expressed in interacting individuals that carry the gene. The term alludes to the idea that an individual expressing a gene with a mirror effect will tend to find its own phenotype 'mirrored' by relatives who share the gene. In an interaction between individuals who share a gene, the mirror effect's strength is quantified as the conditional probability that the gene is expressed in the non-focal individual, given that it is expressed in the focal individual. When this probability is zero or negligibly small, we speak of a 'gene without mirror effect'. Population genetic models (including Hamilton's [6]) often assume that a gene is always expressed, thereby implicitly assuming the mirror effect is maximally strong. There are, however, two mechanisms by which a gene can be exempt from the mirror effect. First, if the expression of a behaviour is conditional on an asymmetry between social partners (e.g. in size, residency, caste, social

dominance or any arbitrary variable), the underlying gene is exempt from the mirror effect [23]. Second, if a gene has low penetrance (i.e. probability of being expressed), it will rarely be simultaneously expressed in both the actor and the recipient during a social interaction—even if both parties carry the gene. This makes the mirror effect negligibly weak. The mirror effect presents a difficulty for quantifying the causal effects of a gene, because it is expressed both in a focal organism and in other organisms that make up its social environment (figure 1). For example, if we compare organisms that either do or do not have a helping gene (with mirror effect), $IF_{folk}$ overestimates the gene's causal effect because it counts the benefit of helping twice—both when giving and receiving help [16]. The conventional remedy for this 'double accounting' is to use $IF_{Hamilton}$, which, by 'stripping' the effect of the social environment, isolates the causal effect of a gene when it is expressed in the focal organism. However, inspired by Pearl's causal modelling framework [17,18], we suggest an alternative remedy that is analogous to measuring causality in a controlled experiment. We can measure a gene's causal effect on a focal organism's $IF_{folk}$ by comparing the observed value of $IF_{folk}$ with the counterfactual value $\widehat{IF}_{folk}$ that would arise if the individual were experimentally prevented from expressing the gene (see legend to figure 1). This heuristic recovers the correct inclusive fitness effect when interactions are additive (i.e. when the effects of an individual's actions are independent of the phenotype of others; figure 1). As importantly, it also predicts the direction of multi-locus evolution for the kinds of non-additive interactions that have stymied attempts to 'strip' the effects of the social environment on the focal individual's inclusive fitness (electronic supplementary material S5, Q15). Instead of being a mere technicality that needs accounting for, the mirror effect can sometimes affect the direction of selection by biasing the flow of social benefits towards particular genotypes in non-additive interactions (i.e. when the benefits provided to a recipient partly depend on the recipient's phenotype; figure 2).

## 5. Mirror effect rogue genes

Intriguingly, opposite phenotypes (e.g. help versus do not help) can be selected for depending on whether or not a gene has a mirror effect (figure 2). In this context, we define an *MER gene* as an allele that reduces the *vehicle quality* of the organisms expressing it, but is still selected for due to the mirror effect (i.e. because the mirror effect biases the flow of social benefits towards particular genotypes at that locus). Here, an organism's reduction in vehicle quality is measured relative to the counterfactual situation where only the focal organism, in its given social environment, expresses an alternative phenotype to that induced by the MER. This definition implies that any unlinked modifier gene will be selected for if it slightly reduces an MER gene's probability of being expressed. This follows because the modifier gene meets our definition of a reference gene in being rarely expressed (only in rare instances where its effect on the MER gene is realized), implying that more copies of it are propagated when the focal organism's vehicle quality is increased (due to the MER gene's negative effect being negated by the modifier). MER genes can occur when there are non-additive social interactions in which matching phenotypes interfere with each other (e.g. mutual help is less efficient than unilateral help; electronic supplementary material, S1). Loosely speaking, these are conditions where a rational actor would prefer to help, unless she anticipates that her actions will be 'mirrored' by relatives. In the example given in figure 2, helping increases vehicle quality when it is rare; however, an MER allele for 'not helping' can spread to fixation when the helping allele is always expressed (i.e. is subject to a mirror effect), thereby failing to generate indirect fitness benefits for its carriers due to interference. We should emphasize that MER genes do not merit discussion because there is empirical evidence for them, but rather because many theoretical models [24–27] have made assumptions under which MER genes occur. This has prompted conclusions which appear to contradict our prediction that evolution tends toward the maximization of $IF_{folk}$.

## 6. The folk definition of inclusive fitness

Based on our definition of vehicle quality and $IF_{folk}$, we advance a heuristic argument about cumulative change, and a deductive argument about evolutionary stability, to infer the most likely outcome of long-term natural selection. Consider a positively selected focal gene (of any effect size, hence subject to any strength of selection) for a trait that increases vehicle quality through an initially inefficient mechanism, as is likely for novel traits. Other genes elsewhere in the genome that enhance the trait's efficiency will then increase vehicle quality further and be selected for. In this way, traits that increase vehicle quality have the potential to evolve through complementary, cumulative contributions from unlinked genes. This potential is crucial because many genes (with various effect sizes) are usually involved in producing finely adapted and/or complex traits. It is exceedingly rare for such traits to arise in a single mutational step. Conversely, if a focal gene promotes development of a trait that reduces vehicle quality while facilitating its own propagation (i.e. a rogue gene), the trait faces counter-selection from elsewhere in the genome. The likely success of the 'parliament of genes' in countering a rogue gene is aided by the architectural principle that complex structures are more easily destroyed than built. For example, if trait development depends on a suite of genes that interact in a coherent fashion, then mutations disrupting any of these myriad interactions will tend to derail its development. These twin considerations suggest that traits that increase vehicle quality will prevail in the long run, even if selection for rogue genes temporarily reverses the trend.

We next make an argument about evolutionary stability. Consider a mutant gene whose expression in a focal individual induces a phenotypic change that increases the individual's $IF_{folk}$. If this gene meets our definition of a reference gene, it is guaranteed to be positively selected because $IF_{folk}$ is defined by a reference gene's propagation success. Hence, no phenotypic strategy is evolutionarily stable unless the organisms adopting it already maximize their $IF_{folk}$. To reach this conclusion, all we need to assume is that mutations can arise with any degree of penetrance. Even if evolutionary dynamics are largely driven by high-penetrance genes under strong selection, evolutionary stability has to be evaluated allowing for mutant genes with any degree of penetrance. To the extent that the availability of suitable alleles poses a genetic

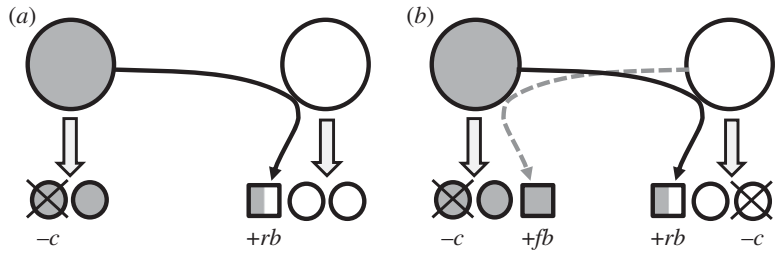


**Figure 1.** Performance of inclusive fitness measures as an accounting tool for genes without or with mirror effect. Big circles represent adults; the shaded one is the focal actor. Small circles represent offspring produced without help from the social environment. Crossed-out small circles represent offspring not produced as a result of a costly helping act. Small squares represent offspring produced through helping. The shading of small squares represents such offspring's relatedness to the focal individual, relative to its own offspring. Black arrows represent helping acts performed by the focal individual, pointing to the resultant offspring produced by the non-focal individual. Dashed arrows represent helping acts received by the focal individual from its social environment. We compare $IF_{Hamilton}$ with $IF_{folk}$, which differs from $IF_{Hamilton}$ in that none of the focal individual's offspring are stripped away. (*a*) In a population where by default each individual produces two offspring without giving or receiving help (*baseline* $= 2$), a mutant gene *without mirror effect* causes the focal individual to help a relative, yielding an indirect fitness benefit $rb$, at cost $-c$. Because the focal individual's behaviour is not mirrored by its relative, we have $IF_{Hamilton} = IF_{folk} =$ baseline $+ rb - c$, and the gene is positively selected if IF $>$ baseline (i.e. $rb - c > 0$). Thus, both $IF_{Hamilton}$ and $IF_{folk}$ work as an accounting tool for this type of gene. (*b*) Similar to (*a*), but *with mirror effect*: here the mutant gene which causes the focal individual to help is also expressed in any relatives that carry its identical copies. As a result, the focal individual produces $fb$ additional offspring, where $f = f[r, p]$ is the probability of receiving help, which is a function of relatedness $r$, the gene's frequency $p$, as well as the gene's penetrance. (Moreover, looking beyond the simplistic case where all helping in the population is due to the focal allele, the $f$ term should also account for help received due to behaviour encoded by other loci.) This situation yields $IF_{Hamilton} =$ baseline $+ rb - c$ (not including the $fb$ offspring produced due to the social environment) and $IF_{folk} =$ baseline $+ rb - c + fb$. Now $IF_{Hamilton} >$ baseline still correctly predicts selection on the focal gene (provided fitness effects are additive [24]), because it isolates the gene's causal effects from the correlational component $fb$ that would arise even if the gene in the focal organism were not expressed. By contrast, $IF_{folk} >$ baseline does *not* correctly predict selection because the term $fb$ includes a benefit (in the focal individual) whose cost (in another individual) is unaccounted for [16]. However, rather than being a shortcoming of $IF_{folk}$, this merely reflects the general difficulty of inferring a causal effect from correlational data. In a causal modelling framework [17,18], this difficulty is readily avoided by calculating $\widehat{IF}_{folk}$[do not help] = baseline $+ fb$ as the focal individual's counterfactual value of $IF_{folk}$ that would arise if the focal individual did not help. Then the focal gene's causal effect on the focal organism's $IF_{folk}$ is positive if $IF_{folk}$[help] $- \widehat{IF}_{folk}$[do not help] $= rb - c > 0$, which recovers the correct inclusive fitness effect. Thus, the focal gene is selected for if expressing it causes the focal organism's $IF_{folk}$ to increase.

constraint, even a low frequency of mutations should eventually overcome this constraint. Hammerstein [28, p. 523] made a similar point about non-social evolution: 'If genetic constraints keep a population away from a phenotypically adaptive state, there is a possibility for a new mutant allele to code for phenotypes that perform better than the population mean.' It follows that the maximization of $IF_{folk}$ is necessary for evolutionary stability under far broader conditions than have been previously reported [27], including non-additive interactions and mutations of various step sizes, both large and small.

We emphasize that the above argument neither assumes nor implies that low-penetrance genes are more important for evolutionary stability than high-penetrance genes. However, it is stability against low-penetrance mutations that implies organismal maximizing behaviour. This is because a low-penetrance gene, when expressed, induces exactly the kind of change we envision in our definition of $IF_{folk}$ being a function of phenotype: namely, there is a change in the focal organism but no correlated (mirrored) change in its social environment. The gene's causal effect on its own propagation thus corresponds exactly to its causal effect on the focal organism's $IF_{folk}$. And this correspondence ensures that only organisms that already maximize their $IF_{folk}$ cannot be modified by a low-penetrance gene to gain a propagation advantage.

Although necessary, maximization of $IF_{folk}$ is not sufficient for evolutionary stability. Even when $IF_{folk}$ is maximized and it cannot be increased by changing a focal organism's phenotype in its current environment, a large-effect mutation with mirror effect might perturb the social environment so as to render a new phenotype optimal. For example, if there are synergistic benefits of mutual cooperation, cooperator genes with mirror effect can invade (and then increase $IF_{folk}$ in the new local environment they create) even when unilateral switching to cooperation would decrease $IF_{folk}$ (electronic supplementary material, S1).

Earlier work that rejected the principle of $IF_{folk}$ maximization made the restrictive assumption that genes with incomplete penetrance and/or conditional expression do not exist [25,26]. Consequently, mutant genes could not change the phenotype of the individual they were expressed in without immediately facing a correlated change in relatives carrying the same gene. Given interference between matching phenotypes, which is when MER genes can arise, this prevented organisms from evolving the optimal phenotype for their social environment (figure 2). Here we show that equilibria established by MER genes (at which $IF_{folk}$ is not maximized) are unstable against invasion by mutant genes without mirror effect, whereas the corresponding equilibria at which $IF_{folk}$ is maximized are stable against mutant genes both with and without mirror effect (electronic supplementary material, S1). We then use simulations to show that the principle of $IF_{folk}$ maximization is realized ever more closely when the genetic system is more flexible (electronic supplementary material, S2). This flexibility can arise due to either a one-locus multi-allele system (electronic supplementary material, figure S1) or a multi-locus system (electronic supplementary material, figures S2–S4). Our results suggest that, barring permanent genetic constraints that seem biologically implausible, interference between matching phenotypes (that allows for MER genes) poses no unsurmountable impediment to organisms evolving the optimal phenotype for their environment in the long term.

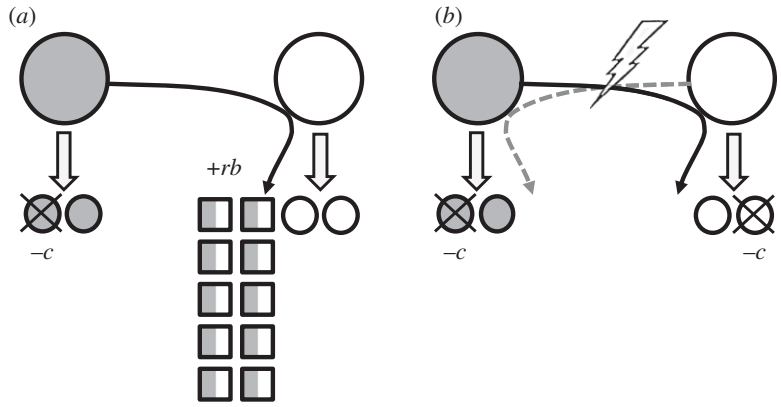


**Figure 2.** Example of how the mirror effect, in combination with non-additive interactions between individuals, can generate selection for a trait that reduces vehicle quality. Consider a population where siblings interact (i.e. pedigree relatedness $r = 0.5$), and where unilateral help (*a*) is highly efficient (e.g. $b = 10$, $c = 1$), whereas mutual help (*b*) is completely inefficient due to strong interference between matching phenotypes (symbolized by lightning bolt; $d = -10$ in the notation of electronic supplementary material, S1). In this situation, helping cannot evolve based on a gene with full penetrance, because benefits accrue exclusively to individuals who lack the helping gene. Thus, when a full-penetrance helping gene (which is subject to the mirror effect) is introduced at low frequency into the population, its alternative allele (which can be considered a full-penetrance non-helping gene) will quickly spread back to fixation. This occurs even though at the phenotypic level, individuals could increase their vehicle quality by switching to unilateral helping, thus reaping the indirect benefits shown in *a*. Even though defection to non-helping reduces vehicle quality, it spreads to fixation based on an *MER gene*—leading to an equilibrium where helping does not occur. In other words, organisms end up making no use of the huge indirect fitness benefit that would accrue from unilateral helping, which contradicts the idea that individuals are selected to maximize their $IF_{folk}$. Crucially, however, this equilibrium without helping is only stable under the restrictive assumption that mutations without mirror effect cannot arise (electronic supplementary material, S1). If such mutants arise (e.g. a low-penetrance gene; or a gene for helping your younger sibling, conditional on being the older one), they generate selection for helping due to the indirect benefits shown in (*a*).

## 7. Hamilton's inclusive fitness

Does maximizing $IF_{folk}$ instead of $IF_{Hamilton}$ actually make a difference? Do we really need to abandon $IF_{Hamilton}$? To be a quantity which an individual could meaningfully be said to be maximizing, $IF_{Hamilton}$, like $IF_{folk}$, must be a function of an individual organism's phenotype. This raises the question of how to interpret the 'stripping procedure' in Hamilton's definition. Hamilton stated that $IF_{Hamilton}$ is 'stripped of all components which can be considered as due to [i.e. that are causal effects of] the individual's social environment, leaving the fitness which he would express if not exposed to any of the harms or benefits of that environment' [6, p. 8]. We take this to mean that, if a non-focal individual performs a social act that causes the focal organism's reproduction to change (compared to the counterfactual situation where it is not performed), then the magnitude of that change must be stripped from the focal individual's IF. This worked in Hamilton's original set-up because the assumption of additive interactions ensures that every consequence is attributable to a single act and actor. Additivity ensures that the components to be stripped are unaffected by the focal organism. By contrast, non-additivity introduces the difficulty that causal effects of non-focal individuals' behaviour depend on a focal individual's phenotype. There are at least three approaches to dealing with this challenge:

(i) One approach to $IF_{Hamilton}$, which we used, is to apply the original stripping procedure. That is, if a non-focal individual performs an act that causes the focal organism's reproduction to change (compared to if the act did not occur), then we calculate $IF_{Hamilton}$ as if this act did not occur (i.e. 'stripping'). This leads to the conclusion that $IF_{Hamilton}$ fails as a phenotypic maximand, because it unduly neglects a component of reproductive success that the focal individual *can* influence. Creel's paradox [12] neatly exemplifies the problem this creates when trying to account for obviously adaptive traits: $IF_{Hamilton}$ implies that it is better to be a helper than a breeder in a cooperative breeding system (figure 3; electronic supplementary material, S3).

(ii) Alternatively, anticipating the inadequacy of approach (i) to capture all of a focal organism's causal effects, one might conclude that $IF_{Hamilton}$ simply cannot be applied in non-additive situations. This might be called the 'Grafen–Nowak approach', after [9] ('the question of how to define inclusive fitness in the absence of additivity has not been settled, and so fundamental theory on the non-additive case can hardly yet begin') and [29] ('since non-linear, synergistic phenomena cannot be attributed to individual actors, there is in general no meaningful way to define an individual's inclusive fitness').

(iii) One can abandon the task of calculating $IF_{Hamilton}$ as a property of an organism, and instead calculate the *inclusive fitness effect* of a focal gene or trait. This can be done with methods such as neighbour-modulated fitness (electronic supplementary material, S4) that automatically 'strip' appropriate components of only the effects of a particular gene or trait. One version of this approach, called the Taylor–Frank method [30,31], is very useful for constructing models, albeit without directly engaging with the phenotypic maximand concept. Another version, called the 'general form of Hamilton's rule' [32–34], defines a focal gene's inclusive fitness effect so as to make it positive by definition for any positively selected gene—even if it is a rogue gene that lowers vehicle quality. Although this formulation creates the impression of selection having a coherent direction, it does not resolve the question of how the opposing phenotypic effects of rogue genes and other genes play out in evolutionary time.

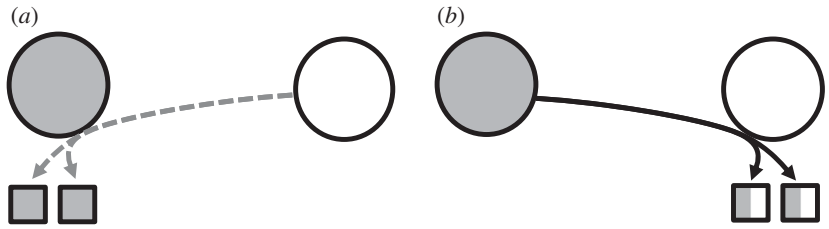


**Figure 3.** Creel's paradox, modified after Queller [12]: in an obligate cooperative breeding system where reproduction requires exactly one breeder and one helper, the focal individual has a choice between becoming the breeder (*a*) or the helper (*b*), while the non-focal individual (based on some asymmetry) must take the remaining role. Since offspring produced due to the social environment are excluded from $IF_{Hamilton}$, the focal individual has lower $IF_{Hamilton}$ in (*a*) than (*b*) (0 versus 2*r*), despite transmitting more genes as a breeder. Invoking $IF_{Hamilton}$ as a phenotypic maximand predicts wrongly that the focal individual should prefer to become the helper (electronic supplementary material, S3). By contrast, the focal individual's $IF_{folk}$ is higher in (*a*) than (*b*) (2 versus 2*r*), predicting correctly that the focal individual should prefer to become the breeder. This matches Queller's [12] prediction, which he obtained (without invoking inclusive fitness as a property of an organism) by applying Hamilton's rule separately to two genes, each expressed conditionally in one of the two roles.

Approaches (i) and (ii) both support our conclusion that $IF_{Hamilton}$ is not a general phenotypic maximand; and approach (i) makes it explicit why $IF_{Hamilton}$ fails. Approach (iii) is silent on what, if any, property of an organism qualifies as a phenotypic maximand, as it is unconcerned with calculating IF as a property of an organism. Unfortunately, this limitation is frequently obscured by the practice of equating the *inclusive fitness effect* (applicable to a gene or trait) with inclusive fitness itself.

For example, consider a focal organism that produces $X$ offspring, and causes its relatives of relatedness $r$ to produce another $Y$ offspring, by expressing several different traits. $IF_{folk}$ is readily defined as $X + rY$. But what is the focal organism's $IF_{Hamilton}$? According to approach (i), we can answer this question by measuring the component to be stripped, as the change in the focal organism's reproduction that would ensue from preventing all social acts of non-focal individuals. According to approach (ii), the question is meaningless unless all fitness interactions are additive, because the focal individual's $IF_{Hamilton}$ is not defined in the general case. And according to approach (iii), we cannot answer the question as the components to be stripped will differ from trait to trait, yielding no overall measure of $IF_{Hamilton}$ as a property of an individual.

Although $IF_{Hamilton}$ is the orthodox way to define inclusive fitness, we conclude that it is only a phenotypic maximand when interactions are additive. It only applies when the number of offspring which the social environment causes an individual to produce is unaffected by any aspect of the focal individual's phenotype that could be selected for [9]. In that special case, it makes no difference whether we think of $IF_{Hamilton}$ or $IF_{folk}$ as being maximized: they are both maximized by the same strategy, a point which has been made in a more general form by Okasha & Martens [26].

## 8. Discussion

The most profound achievement of evolutionary theory is to explain the origin of complex organismal design that was once attributed to supernatural creation. According to the theory of natural selection, complex design arises gradually because changes in numerous phenotypic dimensions, induced by many genes, are predominantly guided in a coherent direction. The guiding principle that gives directionality to this process was identified by Darwin [2, p. 84] as 'the improvement of each organic being in relation to its organic and inorganic conditions of life', and refined by Hamilton [1,6] as the improvement of inclusive fitness. Here we have tried to emphasize and strengthen these core ideas by modifying some of the theory's details.

One of these modifications bears on the fiery debate between critics and defenders of inclusive fitness ignited by Nowak *et al*. [35]. As we see it, both sides of the controversy make some valid claims. The critics are correct that inclusive fitness, when defined as $IF_{Hamilton}$, is a meaningful property of individual organisms (and hence a candidate phenotypic maximand) only under narrow conditions. But the defenders of inclusive fitness are equally correct to counter that organismal design can be understood, under very general conditions, in terms of inclusive fitness maximization [36]. We suggest that the discrepancy between these statements is resolved by replacing $IF_{Hamilton}$ with $IF_{folk}$, which, we have argued, is a more general maximand.

We advocate the idea that long-term phenotypic evolution tends to follow the genome's 'majority interest'. Our rationale is that, although only genes that actually affect a given trait matter for its evolution, the genes that matter can change over time [28,37]. Even if a trait is currently affected by only one or a few loci, in the long term the whole genome is a target for mutations whose effects can modify those of these few loci. This makes it relevant to ask what modifier genes would be selected for. Are they those that strengthen or those that undermine a given phenotypic effect? The phrase 'the trait serves/opposes the genome's majority interest' is shorthand for: the trait selects for unlinked modifiers improving/undermining it. Accordingly, the genome's 'majority interest' (formally encapsulated in a reference gene's interest) should manifest over evolutionary time because traits that align with it tend to be improved through complementary, cumulative contributions from unlinked genes, whereas traits opposed to it will tend to be eliminated.

Fortunately, the invaluable Taylor–Frank method [30] to construct kin selection models is fully compatible with our theory. This method finds evolutionarily stable values of a continuous trait such that no mutant gene can invade that slightly changes the resident trait value. This includes stability against small-effect, low-penetrance genes that meet our definition of a reference gene. Since only a population whose members already maximize $IF_{folk}$ leaves no scope for the invasion of a reference gene (§6), this implies—perhaps surprisingly—that the Taylor–Frank method finds strategies that (locally) maximize $IF_{folk}$ rather than $IF_{Hamilton}$. How

could this important implication have been overlooked? We see two likely reasons. First, $IF_{Hamilton}$ works as an accounting tool for genes with small (hence approximately additive) effects, which are the type of genes considered by the Taylor–Frank method. Some might therefore be tempted to conclude that $IF_{Hamilton}$ will also work as a phenotypic maximand. This conclusion is unjustified, however, because (approximate) additivity at the gene level does not imply additivity at the organism level. And without additivity at the organism level, $IF_{Hamilton}$ does not fully capture an organism's causal effects on gene propagation (§7). Second, ambiguity arises from the widespread use of verbal definitions that purport to describe $IF_{Hamilton}$, but, in fact, obfuscate $IF_{Hamilton}$ and $IF_{folk}$. For example, inclusive fitness has been called 'the property of an individual organism which will appear to be maximized when what is really being maximized is gene survival' [38] or 'the component of reproductive success an organism can influence' [39]. While the latter definition maps to $IF_{Hamilton}$ when applied to models with additive interactions, it maps to $IF_{folk}$ when applied to nature. In reality, no component of an organism's reproduction is *a priori* beyond its influence, in the sense that it is unaffected by any evolvable aspect of the focal organism's phenotype. For example, for an organism to convert help from others into offspring, it must allocate resources to its gonads. In nature, the 'component of reproductive success an organism can influence' therefore includes all its offspring. Similarly, when applied to nature, Hamilton's [1] 'generalized unrigorous statement of the main principle' (which does not mention any 'stripping') can arguably be read as an implicit endorsement of $IF_{folk}$: 'The social behaviour of a species evolves in such a way that in each distinct behaviour-evoking situation the individual will seem to value his neighbours' fitness against his own according to the coefficients of relationship appropriate to that situation' [1, p. 19].

Our approach is inspired by the 'gene's eye view' of adaptation made popular by Dawkins's *The Selfish Gene* [4]. According to this view, adaptive phenotypes can be identified by metaphorically adopting a gene's first-person perspective to ask: 'what phenotype should I induce to make more copies of myself?' However, in the words of Hammerstein [28, p. 530], 'a naive interpretation of the idea of the "selfish gene" can easily direct our attention to an inappropriate level of biological organization (genes instead of phenotypes). This is so because [in the multi-locus case] the genetic scene can only be described as an "incredible mess" although very clear economic principles hold—in the long run—at the phenotypic level' (quoting Dawkins [4]). Our reference gene concept is an attempt to tidy up the 'gene's eye view', by envisaging a gene that embodies the guiding principle of multi-locus evolution. This approach reflects the view of many biologists that it is usually more interesting to ask 'what phenotypes are adaptive?' than to ask 'what hypothetical gene could be selected for?' For example, undue focus on the latter question might lead us to predict fathers who kill their daughters to feed their sons (if caused by a gene on the father's Y-chromosome—a hypothetical variant of a *Mendelian outlaw gene* [20]); or to predict indiscriminate altruism between all members of a species (caused by a *greenbeard gene* gone to fixation). Such outcomes involving rogue genes are unlikely to be observed in nature because—being incompatible with the genome's majority interest—they can neither evolve through cumulative contributions of unlinked genes, nor be stable in the long term. We are left with two equivalent metaphors for long-term phenotypic evolution. We can think either of reference genes strategically 'trying' to maximize their propagation, or of organisms evolving to maximize their vehicle quality (or $IF_{folk}$). Both metaphors capture the view that organisms are integrated systems shaped over generations by the contributions of numerous genes, and, as such, are unlikely to perpetually retain traits under counter-selection from the majority of the genome.

To conclude, our present theory might confirm what many readers intuitively think—that organisms appear to be designed to maximize the weighted offspring count that defines $IF_{folk}$. The prevalence of this intuition is seen in the persistent tendency to define inclusive fitness as $IF_{folk}$ in teaching materials and other non-mathematical texts [16,40–42]. This view has, however, never been explicitly justified, and it stands in contradiction to the prevailing orthodoxy among theoreticians. Our line of argument, if valid, would create the unusual situation that orthodoxy should change to match the textbooks, rather than the other way around.

Data accessibility. Matlab code for simulations is available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.h1c0b52 [43].

Authors' contributions. L.F. had the idea and wrote the first draft. M.D.J. contributed through discussion of ideas and writing.

Competing interests. We declare no competing interests.

Funding. Academy of Finland (L.F.; grant no. 283486) and the Australian Research Council (M.D.J.).

Acknowledgements. We thank David Queller (who went way beyond the call of duty as a referee), Jono Henshaw, Jussi Lehtonen, Piret Avila and Jaakko Toivonen for discussions and comments on the manuscript; Tom Wenseleers, Zoltan Barta, Jutta Schneider, Mikael Puurtinen, Jannis Liedtke and Sara Calhim for comments on the manuscript; and Erol Akcay and Jeremy van Cleve for helpful criticism.

## Endnote

[1]To be exact, to quantify each reference gene copy's projected contribution to the future gene pool, each offspring should be weighted by $V/l$, where reproductive value $V$ is the offspring's projected contribution to the future gene pool, and ploidy level $l$ accounts for the fact that a diploid offspring's contribution is shared between two haploid genomes.

## References

1. Hamilton WD. 1964 Genetical evolution of social behaviour II. *J. Theor. Biol.* **7**, 17–52. (doi:10.1016/0022-5193(64)90039-6)
2. Darwin C. 1859 *On the origin of species by means of natural selection*. London, UK: John Murray.
3. West SA, Gardner A. 2013 Adaptation and inclusive fitness. *Curr. Biol.* **23**, R577–R584. (doi:10.1016/j.cub.2013.05.031)

4. Dawkins R. 1976 *The selfish gene*. Oxford, UK: Oxford University Press.
5. Frank SA, Fox GA. 2017 The inductive theory of natural selection. In *The theory of evolution* (eds SM Scheiner, DP Mindell), pp. 231–261. Chicago, IL: University of Chicago Press.
6. Hamilton WD. 1964 Genetical evolution of social behaviour I. *J. Theor. Biol.* **7**, 1–16. (doi:10.1016/0022-5193(64)90038-4)
7. Frank SA. 2013 Natural selection. VII. History and interpretation of kin selection theory. *J. Evol. Biol.* **26**, 1151–1184. (doi:10.1111/jeb.12131)
8. Grafen A. 1999 Formal Darwinism, the individual-as-maximizing-agent analogy and bet-hedging. *Proc. R. Soc. Lond. B* **266**, 799–803. (doi:10.1098/rspb.1999.0708)
9. Grafen A. 2006 Optimization of inclusive fitness. *J. Theor. Biol.* **238**, 541–563. (doi:10.1016/j.jtbi.2005.06.009)
10. Grafen A. 1985 A geometric view of relatedness. *Oxford Surv. Evol. Biol.* **2**, 28–89.
11. Akcay E, Van Cleve J. 2016 There is no fitness but fitness, and the lineage is its bearer. *Phil. Trans. R. Soc. B* **371**, 20150085. (doi:10.1098/rstb.2015.0085)
12. Queller DC. 1996 The measurement and meaning of inclusive fitness. *Anim. Behav.* **51**, 229–232. (doi:10.1006/anbe.1996.0020)
13. Grafen A. 1984 Natural selection, kin selection and group selection. In *Behavioural ecology: an evolutionary approach* (eds NB Krebs, JR Davies), pp. 62–84. Oxford, UK: Blackwell Scientific Publications.
14. Leigh EG. 1971 Adaptation and diversity. San Francisco, CA: Freeman, Cooper & Co.
15. Bulmer M. 1994 *Theoretical evolutionary ecology*. Sunderland, MA: Sinauer Associates.
16. Grafen A. 1982 How not to measure inclusive fitness. *Nature* **298**, 425–426. (doi:10.1038/298425a0)
17. Pearl J, Mackenzie D. 2018 *The book of why: the new science of cause and effect*. New York, NY: Basic Books.
18. Pearl J. 2009 *Causality: models, reasoning, and inference*, 2nd edn. New York, NY: Cambridge University Press.
19. Alexander RD, Borgia G. 1978 Group selection, altruism, and the levels of organization of life. *Annu. Rev. Ecol. Syst.* **9**, 449–474. (doi:10.1146/annurev.es.09.110178.002313)
20. Dawkins R. 1982 *The extended phenotype*. Oxford, UK: Oxford University Press.
21. Ridley M, Grafen A. 1981 Are green beard genes outlaws? *Anim. Behav.* **29**, 954–955. (doi:10.1016/S0003-3472(81)80034-6)
22. Gardner A, West SA. 2010 Greenbeards. *Evolution (NY)* **64**, 25–38. (doi:10.1111/j.1558-5646.2009.00842.x)
23. Parker GA. 1989 Hamilton's rule and conditionality. *Ethol. Ecol. Evol.* **1**, 195–211. (doi:10.1080/08927014.1989.9525523)
24. van Veelen M, Allen B, Hoffman M, Simon B, Veller C. 2017 Hamilton's rule. *J. Theor. Biol.* **414**, 176–230. (doi:10.1016/j.jtbi.2016.08.019)
25. Grafen A. 1979 The hawk–dove game played between relatives. *Anim. Behav.* **27**, 905–907. (doi:10.1016/0003-3472(79)90028-9)
26. Okasha S, Martens J. 2016 Hamilton's rule, inclusive fitness maximization, and the goal of individual behaviour in symmetric two-player games. *J. Evol. Biol.* **29**, 473–482. (doi:10.1111/jeb.12808)
27. Lehmann L, Alger I, Weibull J. 2015 Does evolution lead to maximizing behavior? *Evolution (NY)* **69**, 1858–1873. (doi:10.1111/evo.12701)
28. Hammerstein P. 1996 Darwinian adaptation, population genetics and the streetcar theory of evolution. *J. Math. Biol.* **34**, 511–532. (doi:10.1126/science.273.5278.1029e)
29. Allen B, Nowak MA. 2016 There is no inclusive fitness at the level of the individual. *Curr. Opin. Behav. Sci.* **12**, 122–128. (doi:10.1016/j.cobeha.2016.10.002)
30. Taylor PD, Frank SA. 1996 How to make a kin selection model. *J. Theor. Biol.* **180**, 27–37. (doi:10.1006/jtbi.1996.0075)
31. Taylor PD, Wild G, Gardner A. 2007 Direct fitness or inclusive fitness: how shall we model kin selection? *J. Evol. Biol.* **20**, 301–309. (doi:10.1111/j.1420-9101.2006.01196.x)
32. Queller DC. 1992 A general model for kin selection. *Evolution (NY)* **46**, 376–380.
33. Gardner A, West SA, Wild G. 2011 The genetical theory of kin selection. *J. Evol. Biol.* **24**, 1020–1043. (doi:10.1111/j.1420-9101.2011.02236.x)
34. Gardner A. 2017 The purpose of adaptation. *Interface Focus* **7**, 20170005. (doi:doi.org/10.1098/rsfs.2017.0005)
35. Nowak MA, Tarnita CE, Wilson EO. 2010 The evolution of eusociality. *Nature* **466**, 1057–1062. (doi:10.1038/nature09205)
36. Abbot P *et al.* 2011 Inclusive fitness theory and eusociality. *Nature* **471**, E1–E4. (doi:10.1038/nature09831)
37. Marrow P, Johnstone RA, Hurst LD. 1996 Riding the evolutionary streetcar: where population genetics and game theory meet. *Trends Ecol. Evol.* **11**, 445–446. (doi:10.1016/0169-5347(96)30036-0)
38. Dawkins R. 1978 Replicator selection and the extended phenotype. *Z. Tierpsychol.* **47**, 61–76. (doi:10.1111/j.1439-0310.1978.tb01823.x)
39. West SA, El Mouden C, Gardner A. 2011 Sixteen common misconceptions about the evolution of cooperation in humans. *Evol. Hum. Behav.* **32**, 231–262. (doi:10.1016/j.evolhumbehav.2010.08.001)
40. Wilson EO. 1975 *Sociobiology*. Cambridge, MA: Harvard University Press.
41. Alcock J. 2005 *Animal behavior: an evolutionary approach*. Cambridge, MA: Sinauer Associates.
42. Zimmer C, Emlen D. 2016 *Evolution - making sense of life*, 2nd edn. New York, NY: WH Freeman and Co.
43. Fromhage L, Jennions MD. 2019 Data from: The strategic reference gene: an organismal theory of inclusive fitness. Dryad Digital Repository. (doi:10.5061/dryad.h1c0b52)