Strong Reciprocity or Strong Ferocity? A Population Genetic View of the Evolution of Altruistic Punishment

Laurent Lehmann,[1,*] François Rousset,[2,†] Denis Roze,[3,‡] and Laurent Keller[4,§]

1. Morrison Institute for Population and Resource Studies, Stanford University, Stanford, California 94305;
2. Laboratoire Génétique et Environnement, Université de Montpellier II, France;
3. Evolution et Génétique des Populations Marines, Station Biologique de Roscoff, France;
4. Ecology and Evolution, Biophore, University of Lausanne, Switzerland

Submitted September 13, 2006; Accepted January 16, 2007; Electronically published May 11, 2007

Online enhancements: appendix, Mathematica notebook.

abstract: Strong reciprocity, defined as a predisposition to help others and to punish those that are not helping, has been proposed as a potent force leading to the evolution of cooperation and altruism. However, the conditions under which strong reciprocity might be favored are not clear. Here we investigate the selective pressure on strong reciprocity by letting both limited dispersal (i.e., spatial structure) and recombination between helping and punishment jointly determine the evolutionary dynamics of strong reciprocity. Our analytical model suggests that when helping and punishment are perfectly linked traits (no recombination occurring between them), strong reciprocity can spread even when the initial frequency of strong reciprocators is close to 0 in the population (i.e., a rare mutant can invade). By contrast, our results indicate that when recombination can occur between helping and punishment (i.e., both traits coevolve) and is stronger than selection, punishment is likely to invade a population of defectors only when it gives a direct fitness benefit to the actor. Overall, our results delineate the conditions under which strong reciprocity is selected for in a spatially structured population and highlight that the forces behind its evolution involves kinship (be it genetic or cultural).

Keywords: altruism, punishment, strong reciprocity, genetic kinship, cultural kinship, cultural transmission.

* E-mail: lehmann@stanford.edu.

† E-mail: rousset@isem.univ-montp2.fr.

‡ E-mail: roze@sb-roscoff.fr.

§ E-mail: laurent.keller@unil.ch.

Explaining the evolution of cooperative and altruistic behaviors in humans and other species is a fundamental problem in biology and the social sciences. Models predict that under situations devoid of reputation or repeated interactions, individuals should not help unrelated individuals (Hamilton 1964*a*; Axelrod and Hamilton 1981). However, results of experiments with humans are in apparent contrast with this prediction because they have repeatedly shown that, in one-shot interactions, individuals behave more cooperatively than predicted under one-shot interaction models (Fehr and Fischbacher 2003). It has been suggested that this paradox could be resolved by the observed tendency of cooperators to engage in costly punishment of noncooperators (Gintis 2000, 2003; Fehr and Fischbacher 2003; Gintis et al. 2003; Bowles and Gintis 2004). The general idea is that humans are strong reciprocators, defined as individuals bearing a propensity to help others and to punish those that are not helping (Gintis 2000), where both helping and punishing are assumed to be costly and result in no direct economic benefit.

The idea that punishment of defectors can enforce the evolution of helping behaviors between unrelated individuals is not new (Hirshleifer and Rasmusen 1989; Boyd and Richerson 1992). However, the conditions under which such punishment may evolve in a population consisting initially of defectors remains unclear. While several authors have proposed that strong reciprocity represents a solution to the puzzle of the evolution of helping behaviors in humans (Gintis 2000, 2003; Fehr and Fischbacher 2003; Gintis et al. 2003; Bowles and Gintis 2004), others recently suggested that punishment cannot promote the evolution of helping behaviors in any interesting situation when punishing itself is a costly behavior (Gardner et al. 2006). In order to identify the causes leading to such contrasting views, a starting point is a discussion by Boyd and Richerson (1992) of a simple model of helping and punishment. These authors studied the evolution of a so-called cooperator-punisher strategy in a randomly mixing population and showed that when there is a first round of

interaction where individuals can help each other and another round where they can punish nonhelpers, strong reciprocators can invade a population of unconditional defectors when the inequality

$$pD > C_H + (1 - p)C_P \quad (1)$$

is satisfied. The right side of this inequality is the cost C_H of expressing the act of helping plus the cost $(1 - p)C_P$ of expressing the act of punishment, where p is the frequency of strong reciprocators in the population. The left side of the inequality is the benefit of not being punished, which depends on the damage D resulting from punishment times the frequency of strong reciprocators in the population. According to the inequality, helping and punishing becomes a better option than defecting in terms of direct fitness benefits when the frequency p of strong reciprocators exceeds an initial threshold frequency determined by the costs and benefits of helping and punishment. However, when the initial frequency of strong reciprocators is close to 0 ($p \to 0$), strong reciprocators cannot invade the population because their fitness is always lower than the fitness of defectors (Sigmund et al. 2001).

Three questions emerge when considering the conditions for the evolution of strong reciprocity given by inequality (1). The first is whether limited dispersal (spatial population subdivision) and the resulting kin selection effects may allow strong reciprocity to evolve even when the initial frequency of strong reciprocators is close to 0 (i.e., whether a rare mutant can invade). The impact of population subdivision on the evolution of strong reciprocity has been studied in simulations by Bowles and Gintis (2004), who suggested that kin selection is absent from their model and that limited dispersal is the key factor promoting the evolution of altruistic punishment. However, their simulations do not allow us to disentangle the role of kin selection and direct fitness benefits. Further, their findings that strong reciprocity evolves under limited dispersal possibly stems from the fact that punishment spreads simply because punishers reduce the intensity of competition at a local scale by decreasing the fecundity of competitors, thereby increasing their own reproductive success. Such punishment would therefore qualify as being selfish (self-interest) rather than altruistic because the act of punishment results in an increase of the expected number of adult offspring of a focal punisher (i.e., fitness sensu Hamilton 1964*a*, 1970; e.g., Grafen 1985; Rousset 2004; Lehmann and Keller 2006). The role of population subdivision has also been studied analytically by Nakamaru and Iwasa (2005, 2006). However, their analyses assume overlapping generations, a feature that by itself can lead to the evolution of helping in spatially structured populations (Taylor and Irwin 2000; Irwin and Taylor 2001), thus making it difficult to determine whether it is punishment or overlapping generations that is the predominant factor influencing the evolution of helping. A more direct evaluation of the condition of invasion of strong reciprocity is required to delineate the impact of limited dispersal and kin selection on both the direct and indirect fitness benefits received by strong reciprocators.

The second question is whether strong reciprocity can actually evolve when both helping and punishment coevolve as independent traits. There is indeed no a priori reason to assume that no recombination occurs between helping and punishment so that both traits cannot segregate from each other (Gardner and West 2004; Lehmann and Keller 2006). Gardner et al. (2006) have studied an analytical model of the coevolution of punishment and helping in panmictic populations and have demonstrated that helping and punishment cannot invade the population unless Hamilton's rule is satisfied. In other words, with recombination occurring between helping and punishment, strong reciprocity does not promote the evolution of helping in any special situation in panmictic populations, and one might ask if this result holds in spatially subdivided populations as well.

The third question is whether a cultural transmission of strong reciprocity may affect the selective pressure on the trait. While several authors have suggested similarities between the selective pressures on genetically and culturally determined helping traits (Werren and Pulliam 1981; Feldman et al. 1985; Allison 1991), others (Gintis 2000; Bowles et al. 2003; Fehr and Fischbacher 2003) have stressed that standard evolutionary theory cannot explain helping traits in humans and that alternatives such as "cultural group selection" are needed. Such claims are currently not supported by thorough mathematical analyses, and the conditions under which cultural and genetic transmission can lead to different selective pressures on the evolution of strong reciprocity remain to be identified.

In this article we present a mathematical model that extends previous work on the coevolution of helping and punishment (Boyd and Richerson 1992; Sigmund et al. 2001; Bowles and Gintis 2004; Gardner and West 2004; Nakamaru and Iwasa 2005, 2006; Brandt et al. 2006; Gardner et al. 2006) by letting both recombination and spatial structure jointly determine the evolutionary dynamics of punishment and helping. While our model assumes a genetic transmission of the traits, we also discuss variants of this model that allow us to consider cultural transmission as well. The main aim of our analysis is to establish the conditions under which expressing punishment conditionally in subdivided populations can in itself promote the evolution of costly helping and punishment without any other confounding factor possibly leading to selection

on these traits (e.g., repeated interactions, overlapping generations, kin recognition, special modes of dispersal, special modes of population demography, special modes of competition between groups, enforcement by an institution or a cultural norm, specific cognitive preferences). A second aim is to determine the nature of the selective forces involved in the evolution of strong reciprocity and under what situations it qualifies as altruistic sensu Hamilton (1964*a*, 1970).

We demonstrate that when helping and punishment are perfectly linked traits with no recombination occurring between them, strong reciprocity can invade a population of defectors when rare (i.e., $p \to 0$). However, this is true only when the population is spatially structured and when two individuals sampled from the same deme are more likely to bear the same genes (or memes) inherited from a common ancestor than two individuals sampled from two different demes. Different modes of cultural transmission may increase, decrease, or not alter at all the selective pressure on strong reciprocity obtained under genetic transmission. But our models also demonstrate that when the recombination between helping and punishment is stronger than selection, strong reciprocity cannot invade a population of defectors when its initial frequency is low (this being true whatever the structure of the population). Then, strong reciprocity can be selected for only if an external mechanism allows it to reach a threshold initial frequency and if kin selection is operating. While our models show that strong reciprocity can evolve in groups of large size, none of the models of genetic and cultural group selection indicate that strong reciprocity is likely to evolve in a situation of anonymous and nonrepeated interactions. Hence, our results suggest that critics (Hagen and Hammerstein 2006; West et al. 2006) of the interpretations of cooperation in experimental games have to be taken very seriously.

Model

Life Cycle

Let us posit that evolution occurs in a population following Wright's infinite island model of dispersal where individuals live in demes of finite size N (see table 1 for a list of symbols). We assume that each individual in each deme interacts with its $N-1$ neighbors and that an interaction consists of two stages. In the first stage, individuals help each other. We assume that bearing a mutant helping allele H at a first locus results in a total fecundity cost C_H for the actor and a total fecundity benefit B for the recipients of the behavior. Accordingly, during each of its $N-1$ interactions, an actor provides a benefit $B/(N-1)$ to its partner at a direct fecundity cost $C_H/(N-1)$ to himself. During the second stage, individuals bearing a mutant punishment allele P at a second locus conditionally express an act of punishment on their partners who have not expressed helping during the first stage of the interaction. During a single interaction, punishment results in a fecundity cost $C_P/(N-1)$ to the actor and decreases the fecundity of the recipient by $D/(N-1)$. Since each actor interacts with $N-1$ neighbors, expressing the punishment allele results in a fecundity cost C_P for the actor and a total decrease by D of the fecundity of neighbors whenever there are no helpers among the neighbors of the actor. Those individuals that bear the resident allele at the punishment and/or helping locus do not express any phenotype at that locus. We assume a haploid life cycle (e.g., Seger 1985; Kirkpatrick et al. 2002) with events occurring in the following order: (1) individuals produce a large number of juveniles according to their fecundity determined by the round of social interactions; (2) each juvenile disperses independently from each other with probability m to another deme and adults die; (3) juveniles fuse randomly to produce diploid zygotes (syngamy), which is immediately followed by meiosis with a recombination rate r between the helping and punishment loci to produce a new generation of haploid individuals; (4) regulation occurs with the effect that only N individuals settle in each patch.

We investigate two different variants of this life cycle. First, we assume that recombination is absent ($r = 0$) and evaluate the selective pressure on strong reciprocators (individuals bearing the helping and punishment alleles) when introduced in a monomorphic population of unconditional defectors (who never cooperate and never punish). Second, we assume that recombination occurs as described for stage 3 of the life cycle. This leads to the presence of four gametes coevolving in the population, and we ask whether helping (allele H) and punishment (allele P) are selected for when introduced at low frequency in the population.

For the interpretation of the results, we recall that when interactions between individuals consist only of the helping stage (no punishment), the direction of selection on the helping allele is given by

$$-C_H > 0. \tag{2}$$

Helping is selected for only if the actor's fecundity, that is, the number of its juveniles counted before any competition stage, is increased (Taylor 1992*a*). Taylor (1992*b*), has further demonstrated that this result is true whatever the dispersal distribution in the population (e.g., island model of dispersal as described above, stepping-stone dispersal). Here, we ask whether introducing conditional punishment can in itself affect Taylor's $-C_H > 0$ rule and

Table 1: List of symbols

Symbol	Definition
N	Deme size
m	Migration rate
r	Recombination rate between gametes
C_H	Fecundity cost of expressing the helping allele
B	Fecundity benefit generated by expressing the helping allele
C_P	Fecundity cost of expressing punishment
D	Reduction in fecundity resulting from expressing punishment
w_{ij}	Fitness of individual j breeding in deme i
f_{ij}	Effects of actors on the fecundity of individual j breeding in deme i
f_i	Effects of actors on the average fecundity in deme i
f	Effect of actors on the average fecundity of demes different than i
$-c$	Net effect on its fitness of an individual bearing both the helping and the punishment allele (i.e., strong-reciprocator)
$-c_H$	Net effect on its fitness of an individual bearing the helping allele
$-c_P$	Net effect on its fitness of an individual bearing the punishment allele
p	Frequency of strong reciprocators in the population
p_H	Average frequency of the helping allele in the population
p_P	Average frequency of the punishment allele in the population
$\zeta_{A(ij)} = p_{A(ij)} - p_A$	Centered variable at locus A
R	Relatedness between actor and recipient (evaluated in a neutral model)
D_{PH}	Covariance between an allele H and an allele P sampled in the same individual. In the neutral case, $D_{PH} = 0$.
$D_{H/H}$	Covariance between two alleles H sampled in two individuals from the same deme. In the neutral case, $D_{H/H} = p_H(1 - p_H)R$.
$D_{P/P}$	Covariance between two alleles P sampled in two individuals from the same deme. In the neutral case, $D_{P/P} = p_P(1 - p_P)R$.
$D_{P/H}$	Covariance between one allele H and one allele P sampled in two individuals from the same deme. In the neutral case, $D_{P/H} = 0$.
$D_{P/H/H}$	Association between two alleles H and one allele P sampled in three different individuals from the same deme. In the neutral case, $D_{P/H/H} = 0$.
$D_{P/P/H}$	Association between two alleles P and one allele H sampled in three different individuals from the same deme. In the neutral case, $D_{P/P/H} = 0$.
$D_{HP/H}$	Association between two alleles H and one allele P where one allele H and allele P are sampled from the same individual and the second allele H is sampled from a different individual. In the neutral case, $D_{HP/H} = 0$.
$D_{HP/P}$	Association between two alleles P and one allele H where one allele P and allele H are sampled from the same individual and the second allele P is sampled from a different individual. In the neutral case, $D_{HP/P} = 0$.

whether recombination between helping and punishment may alter this effect.

Gene Dynamics

In order to investigate the coevolutionary dynamics of helping and punishment, we use the multilocus framework presented in Kirkpatrick et al. (2002) or Gardner et al. (2006) and extended by Roze and Rousset (2005) to include subdivided populations with finite deme size. The change in frequency of allele A (H or P) over one generation in the population can be written as

$$\Delta p_A = \mathrm{E}[w_{ij} p_{A(ij)}] - p_A, \tag{3}$$

where w_{ij} is the expected number of offspring that will breed in the next generation (i.e., fitness) of individual j breeding in deme i, $p_{A(ij)}$ is the frequency of allele A in that individual (0 or 1), and p_A designates the average frequency of allele A in the population. The expectation in equation (3) is taken over all individuals and all demes. Since the population is assumed to be of constant size, the mean fitness is equal to 1 ($\mathrm{E}[w_{ij}] = 1$), and one can recognize equation (3) as being the Price equation (Hamilton 1970; Price 1970). The change in allele frequency is equivalently given by

$$\Delta p_A = \mathrm{E}[w_{ij} \zeta_{A(ij)}], \tag{4}$$

where $\zeta_{A(ij)} = p_{A(ij)} - p_A$ is a centered variable.

Fitness Function

The fitness w_{ij} of individual j in deme i depends on both its expected number of offspring reaching adulthood in deme i and on those reaching adulthood in other demes after dispersing. These two components of fitness depend on the fecundity of individual j in deme i. Without loss of generality, it is sufficient to consider the fecundity of an individual relative to the fecundity in a population without helping or punishment. Relative fecundity will be written as $1 + f_{ij}$, where f_{ij} is the total effect of individuals bearing the mutant allele on the fecundity of individual j in deme i, composed of the effects of the individual himself and its $N - 1$ neighbors, which is given by

$$f_{ij} = \frac{1}{N-1}\sum_{k=1,k\neq j}^{N}[-C_H p_{H(ij)} + Bp_{H(ik)} - C_P(1 - p_{H(ik)})p_{P(ij)} - D(1 - p_{H(ij)})p_{P(ik)}]. \quad (5)$$

In this equation $p_{H(ik)}$ and $p_{P(ik)}$ designate, respectively, the frequencies (0 or 1) of the helping and punishment alleles in individual k breeding in deme i.

A fraction $1 - m$ of the offspring of individual j in deme i remain philopatric and compete with a relative number $1 + f_i$ of juveniles produced in deme i and an average relative number $1 + f$ of immigrant juveniles produced in different demes. The effect of actors in deme i on the average relative fecundity of individuals in that deme reads

$$f_i = \frac{1}{N}\sum_{k=1}^{N} f_{ik}, \quad (6)$$

while the effect of actors on the average relative fecundity of individuals in different demes is given by

$$f = \frac{1}{n_d - 1}\sum_{h=1,h\neq i}^{n_d} f_h, \quad (7)$$

which according to our infinite island model assumptions is evaluated in the limit of an infinite number n_d of demes ($n_d \to \infty$).

A complementary fraction m of the offspring of individual j in deme i disperse and enter in competition with a relative number $1 + f$ of offspring produced in different demes than i. Collecting all components of fitness gives the fitness of individual j in deme i as

$$w_{ij} = (1 - m)\left[\frac{1 + f_{ij}}{(1 - m)(1 + f_i) + m(1 + f)}\right] + m\left(\frac{1 + f_{ij}}{1 + f}\right). \quad (8)$$

Following Hamilton (1964*a*, 1970) and Grafen (1985), we categorize behaviors according to their effects on the fitness of actors and recipients. An action is altruistic (vs. spiteful) with respect to a particular recipient when it decreases the fitness of the actor and increases (vs. decreases) the fitness of the recipient. Individual j in deme i is altruistic with respect to a neighbor when he helps the neighbor because the action decreases its fitness w_{ij}. Similarly, individual j in deme i is spiteful with respect to a neighbor when he punishes the neighbor because the act of punishment decreases w_{ij}. Other authors (e.g., Gintis 2000; Bowles and Gintis 2004) define altruism from effects on fecundity. In that case, individual j in deme i is altruistic with respect to a neighbor when he increases the fecundity of the neighbor while decreasing its own fecundity f_{ij}. This procedure misses out consequences of the action of an individual on its fitness (f_{ij} is also involved in f_i in eq. [8]), a point that will be illustrated below.

Intensities of Selection

Expanding equation (8) into a Taylor series with respect to the phenotypic effects on fitness (C_H, C_P, B, and D) allows us to investigate the coevolutionary dynamic of helping and punishment under various intensities of selection. The classical "weak selection" approach of population genetics usually evaluates the change in allele frequency to the first order in phenotypic effects only (first-order Taylor expansions around $C_H = 0$, $C_P = 0$, $B = 0$, and $D = 0$), thus neglecting terms of higher order of magnitude. The change in frequency of allele A (here H or P) under such conditions is given by

$$\Delta p_A = \mathrm{E}[(f_{ij} - f - (1 - m)^2(f_i - f))\zeta_{A(ij)}] + O(\delta^2), \quad (9)$$

where δ is the largest of the four phenotypic effects on fecundity of our model. Expressing all the gene frequencies appearing in the expectation of this equation in terms of centered variables ($p_{A(ij)} = p_A + \zeta_{A(ij)}$) and taking the expectation gives a decomposition of the selective pressure in terms of coefficients of selection and associations of alleles within and between individuals (see the appendix in the online edition of the *American Naturalist*). For instance, the associations within individuals involve the genetic variance (e.g., $\mathrm{E}[\zeta^2_{H(ij)}]$) and the linkage disequilibrium ($\mathrm{E}[\zeta_{H(ij)}\zeta_{P(ij)}]$), while the associations between

individuals involve intraloci (e.g., $\mathrm{E}[\zeta_{\mathrm{H}(ij)}\zeta_{\mathrm{H}(ik)}]$) and interloci (e.g., $\mathrm{E}[\zeta_{\mathrm{H}(ij)}\zeta_{\mathrm{P}(ik)}]$) measures of relatedness between individuals (Kirkpatrick et al. 2002; Roze and Rousset 2005; Gardner et al. 2006). The resulting selective pressures on helping and punishment are presented in equations (A10) and (A12) in the appendix in the online edition of the *American Naturalist*.

In general, selection will affect genetic associations. However, we can neglect this effect (of order δ) when we compute the change in frequency of the mutant allele to the first order in δ because associations are multiplied by selection coefficients, which are of order δ. Therefore, it is sufficient to compute associations under neutrality. In other words, the effect of selection on the distribution of the mutant alleles in demes is neglected, with the result that the linkage disequilibrium will be 0 at equilibrium (Roze and Rousset 2005). However, relatedness will build up, and the selective pressure on an allele will depend on intralocus kin selection effects.

Here, we also investigate the change in allele frequency to the second order in the largest of the four phenotypic effects in absolute value (C_{H}, C_{P}, B, and D), thus allowing the expression of stronger forms of selection that may generate linkage disequilibrium. In this case, the selective pressure on an allele will depend on genetic covariances within and between individuals at the same and at different loci (i.e., this takes into account intralocus as well as interloci kin selection effects). When the change in gene frequency of allele A (here H or P) is evaluated to the second order in the phenotypic effects, we have

$$\begin{aligned}\Delta p_{\mathrm{A}} = {} & \mathrm{E}[(f_{ij} - f - (1-m)^2(f_i - f))\zeta_{\mathrm{A}(ij)}] \\ & - \mathrm{E}[(f_{ij}f - (1-m)^2\{f_j(2mf - f_{ij}) + (1-m)f_j^2 + ff_{ij}\} \\ & - m\{1 + m(1-m)\}f^2)\zeta_{\mathrm{A}(ij)}] + O(\delta^3). \qquad (10)\end{aligned}$$

The first expectation involves only the linear effect on fitness, and the associations in this expression must now be evaluated to the first order in phenotypic effects so that the change in gene frequency is ascertained by all second-order terms (quadratic in phenotypic effects) generated by selection. The associations evaluated to the first order in δ are equal to the associations evaluated in the neutral case plus a deviation due to the effect of selection (Roze and Rousset 2005). Hence, the effect of selection on the distribution of the mutant allele within and between demes is now taken into account. By contrast, the second expectation already involves a quadratic effect on fitness (e.g., B^2, CD), and it is thus sufficient to evaluate the associations in it in the neutral process only.

In all our calculations, we assume that both the migration rate and the recombination rate are stronger than selection so that the genetic associations reach their steady state before any significant change in allele frequency has occurred at the level of the population. This "quasi-equilibrium" assumption is widely used in population genetic theory (e.g., Kimura 1965; Nagylaki 1993; Bürger 2000; Kirkpatrick et al. 2002; Roze and Rousset 2005) and allows us to conveniently express the associations in terms of gene frequencies and model parameters. However, the calculation of the various associations appearing in equation (10) remains extremely tedious, and we used the recursions for associations as automated by F. Rousset and D. Roze (unpublished manuscript). This program was used to obtain the explicit form of the selective pressure (given by equation [10]) on both helping and punishment under second-order effects, and these equations are presented in the Mathematica notebook "Strong reciprocity" in the online edition of the *American Naturalist*.[1] Readers are invited to contact the authors for future updates of this Mathematica notebook.

Results

No Recombination between Helping and Punishment

To study the evolution of strong reciprocity in the absence of recombination (i.e., $r = 0$), we set $p_{\mathrm{H}(ij)}$ and $p_{\mathrm{P}(ij)}$ equal to a single value, $p_{(ij)}$, in equation (5). The change in the frequency p of strong reciprocators can then be obtained from the appendix either from equation (A10) (by substituting all subscripts H with P) or from equation (A12) (by substituting all subscripts P with H). The resulting change in frequency (eq. [A13]) is a complicated function of the parameters, which, to the first order in $1/N$, becomes

$$\begin{aligned}\Delta p = {} & p(1-p)\{-C_{\mathrm{H}} - (1-p)C_{\mathrm{P}} + pD \\ & + [B + (1-p)C_{\mathrm{P}} - pD]R \qquad (11) \\ & - (1-m)^2[B - C_{\mathrm{H}} - (1-2p)(C_{\mathrm{P}} + D)]R^{\mathrm{R}}\},\end{aligned}$$

where

$$R^{\mathrm{R}} = \frac{1}{N} + \left(\frac{N-1}{N}\right)R \qquad (12)$$

is the relatedness between a focal individual and an individual sampled with replacement from its deme (thus including the focal individual with probability $1/N$). The coefficient R measures the relatedness between two different individuals sampled in the same patch (given here

[1] Code that appears in the *American Naturlist* has not been peer-reviewed, nor does the journal provide support.

by Wright's (1951) measure of population structure $R \equiv F_{ST}$), which is related to the preceding equation by

$$R = (1 - m)^2 R^R, \tag{13}$$

because with probability $(1 - m)^2$, two individuals descend from the same deme, and with probability $1/N$, they descend from the same parent.

According to equation (11), the change in frequency p of strong reciprocators in the population is frequency dependent and is a function of three components. The first is $-C_H - (1 - p)C_P + pD$, which is equivalent to the selective pressure on strong reciprocity obtained in a panmictic population (see eq. [1]). The second is $[B + (1 - p)C_P - pD]R$, which is the change in the relative fecundity of a focal strong reciprocator resulting from its neighbors being strong reciprocators. It depends on the benefit BR conferred by strong reciprocators to the focal individual, on a benefit $(1 - p)C_P R$ resulting from the focal individual not expressing punishment to the same extent as in a panmictic population because its neighbors are likely to be strong reciprocators, and on the cost pDR because the advantage for a strong reciprocator relative to defectors of not being punished decreases when its neighbors are also strong reciprocators. Finally, the third is a term (third line) that is the change in competition in the focal deme resulting from all strong reciprocators in the focal deme expressing helping and punishment. This term is weighted by $(1 - m)^2$, which represents the probability that two offspring produced in the same deme compete against each other.

Substituting equation (13) into equation (11) informs us that the fecundity benefit B conferred by a strong reciprocator to its neighbor, is canceled out by the concomitant increase in kin competition in exactly the same way as occurs in Taylor's (1992*a*) model. By contrast, the effect on fecundity D does not cancel out because punishment is expressed only conditionally, and an individual expressing punishment will never be punished because he also expresses helping. Consequently, punishment can only benefit relatives indirectly (through the reduction of competition), and it cannot directly harm them. Limited dispersal increases the selective pressure on strong reciprocators because common genealogy results in the association of individuals bearing identical genotypes. This means that strong reciprocators are likely to benefit from a reduced local competition stemming from other strong reciprocators expressing harming toward defectors. This raises the question of whether this kin selective pressure may favor the spread of strong reciprocators when their initial frequency is close to 0. By letting $p \to 0$ in the term in brackets in equation (11), we find that this occurs when

$$R(D + 2C_P + C_H) > C_H + C_P. \tag{14}$$

Whether a focal strong reciprocator is altruistic when strong reciprocity is favored by selection when rare in the population depends on the effect $-c$ of its behavior on its own fitness. This effect on fitness is obtained from equation (A13) by setting $R = 0$ and $p = 0$ in the braces. Thus,

$$-c = -C_H - C_P + \frac{(1 - m)^2(C_P + C_H + D - B)}{N}, \tag{15}$$

which depends on both the direct costs from expressing strong reciprocity and the indirect benefits resulting from the decrease in competition in the focal deme. The change in competition induced by a strong reciprocator depends on the additional number $(C_P + C_H + D - B)(1 - m)$ of offspring coming in competition in the focal deme resulting from the focal strong reciprocator expressing the behavior and on the probability $(1 - m)/N$ that this increment in competition will displace its own offspring.

Comparing equations (14) and (15) reveals that there is a range of parameter values under which $-c < 0$ and where strong reciprocity can spread in the population (i.e., altruistic strong reciprocity can invade and is stable). In figure 1, we compare the selective pressure on strong reciprocity (the term in braces in eq. [11]) and the effect of an actor on its fitness (eq. [15]) as a function of the migration rate m and deme size N. Increased rates of dispersal or greater deme size generally lead to a shift from strong reciprocity being selfish to being altruistic because the act of punishment provides less benefits to an actor in terms of reduced competition in its deme. But when the dispersal rate or patch size become large, the relatedness between patch members tends toward 0. In that case, strong reciprocity is counterselected since it provides no indirect benefits to relatives. Finally, from equation (15) we also see that strong reciprocity is more likely to be altruistic when the benefit of helping B is high and when the damage of punishment D is small.

Recombination between Helping and Punishment

Weak Selection (First-Order Effects). With recombination, we must follow the change in frequency of both the helping (p_H) and punishment (p_P) alleles. From equation (A10), the change in frequency of the helping allele is given to the first-order phenotypic effects on fitness (weak selection) by

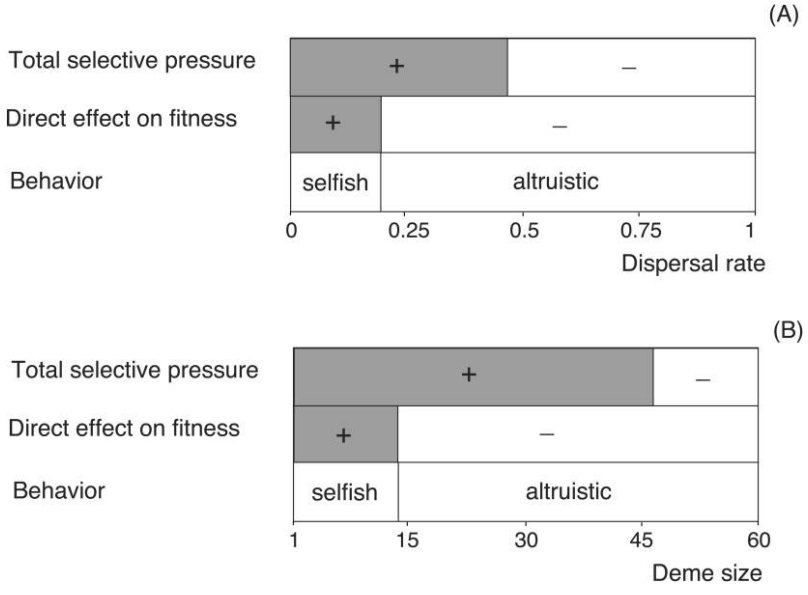

Figure 1: Signs of both the total selective pressure on strong reciprocity (eq. [14]) and the direct effect $-c$ of an actor on its fitness (eq. [15]) when allele p is rare ($p \to 0$). Gray shading indicates positive signs, white shading indicates negative signs. *A*, Signs as a function of the dispersal rate m. Parameter values are $N = 10$, $C_H = 0.01$, $C_P = 0.01$, $B = 0.2$, and $D = 0.5$. *B*, Signs as a function of deme size N. Same parameter values as panel *A* except $m = 0.2$.

$$\begin{aligned}\Delta p_H = p_H(1 - p_H)\{&-C_H + BR \\ &- (1 - m)^2(B - C_H)R^R \\ &+ p_P[D + C_P R - (1 - m)^2(D + C_P)R^R]\}.\end{aligned} \tag{16}$$

The first term (C_H) in the braces of this equation is the direct cost of helping, the second term (BR) is the kin-selected benefits of helping, and the third term ($(1 - m)^2(B - C_H)R^R$)is the increase in kin competition resulting from all actors in the focal deme expressing helping. The remaining terms in equation (16) are the additional selective pressures on helping stemming from actors punishing individuals that do not express helping. This term depends of the frequency p_P of punishers in the population and consists of three components. The first is the direct benefit of expressing helping and thus not being punished. The second is an indirect benefit, which results from a punisher that also bears the helping allele and who does not express the cost of punishment because the focal actors express helping. Finally, the last term is an indirect cost resulting from the increase in competition in the focal deme associated with punishers bearing the helping allele and not expressing any punishment in the focal patch as a result of the focal individual expressing helping.

Substituting equation (13) into equation (16) and simplifying informs us that helping is selected for when

$$Dp_P - C_H > 0. \tag{17}$$

Helping is a better strategy than the alternate option of defecting if the frequency of punishers in the population times the decrease in fecundity resulting from punishment exceeds the cost of helping. As was the case for the previous model, the fecundity benefit B that an actor of helping confers to its neighbors is canceled out by the concomitant increase in kin competition generated by the act of helping. Helping spreads only if it increases the fecundity of the focal individual, a result that can be interpreted as being an application of Taylor's model (1992*a*; 1992*b*). Therefore, the condition $Dp_P - C_H > 0$ is likely to hold, whatever the dispersal distribution (e.g., island model of dispersal, stepping-stone model of dispersal).

Helping can be altruistic if it results in a negative effect on the fitness of the actor. The effect of an act of helping on the fitness of the actor is obtained from equation (16) by setting the relatedness between different individuals in the braces of equation (16) to 0 ($R = 0$, $R^R = 1/N$), which gives

$$-c_H = Dp_P - C_H - \frac{(1 - m)^2(B - C_H) - p_P(D + C_P)}{N}, \tag{18}$$

where the term weighted by $1/N$ is the change in the intensity of competition in the focal deme resulting from an actor expressing helping. From this equation we see that there is a range of parameter values where $Dp_P - C_H > 0$ is consistent with a negative effect of helping on the fitness of the actor.

From equation (A12), the change in frequency of the punishment allele is given by

$$\Delta p_P = p_P(1 - p_P)(1 - p_H) \times [-C_P - DR + (1 - m)^2(D + C_P)R^R], \quad (19)$$

which is formally equivalent to the selective pressure on helping (eq. [16]) in the absence of punishment (substitute D with $-B$ in eq. [19] and compare with the first and second line of eq. [16]). Substituting the equilibrium value of R into equation (19) and simplifying, we find that punishment spreads when

$$-C_P > 0. \quad (20)$$

Punishment can spread only if the act of punishment results in a direct fecundity benefit for the actor. The effect on fecundity D has canceled out because, contrary to the model where punishment and helping are coded by the same gene (eq. 11), an individual expressing punishment is also likely to be punished. Hence, punishment indirectly benefits relatives (through the reduction of competition) but directly costs them when they do not bear the helping allele. This model can again be interpreted as being a special application of Taylor's (1992*a*) model, and it also illustrates that the results established by Gardner et al. (2006) for panmictic populations extent to subdivided populations as well.

Punishment results in a change in the direct fitness of the actor by magnitude

$$-c_P = (1 - p_H)\left[-C_P + \frac{(1 - m)^2(D + C_P)}{N}\right]. \quad (21)$$

The second term in the brackets of this equation is positive because punishment reduces the intensity of competition in the focal deme, which increases the likelihood that an offspring of an actor will reach adulthood. Contrary to helping, punishment cannot be altruistic when the condition for its evolution is satisfied.

Stronger Selection (Second-Order Effects). In order to assess whether the results established in the previous section also hold under stronger forms of selection, we evaluated the changes in frequency of helping and punishment to the second-order phenotypic effects on fitness. This introduces two complications. First, there are additional components in the selective pressures, quadratic in B, C_H, D, and C_P, because the second expectation of equation (10) has to be taken into account. Second, the covariances in the first expectation of equation (10), which appear when the fecundities (f_{ij}, f_i, and f) are expressed in terms of centered variables ($p_{A(ij)} = p_A + \zeta_{A(ij)}$), have to be evaluated to the first order in phenotypic effects. These genetic associations are now affected by an interaction between selection and common genealogy. In order to follow the coevolutionary dynamic of helping and punishment, we thus have to track the change of both the gene frequencies and the genetic associations within and between individuals. The explicit expressions of the associations are given in equations (A18)–(A23), and the total selective pressures on punishment and helping are very cumbersome and are presented in the Mathematica notebook "Strong reciprocity" in the online edition of the *American Naturalist*. Under our quasi-equilibrium analysis, we distinguish the situation where both the helping and punishment are initially rare from the situation where the initial frequency of one allele is not vanishingly small, because they differ qualitatively.

In the situation where both the helping allele H and the punishment allele P are initially rare ($p_H \to 0$ and $p_P \to 0$), the evolutionary development of one allele is independent of the other allele because all frequency-dependent terms cancel, a consequence of the fact that allele frequency is vanishingly small. In that case, only relatedness between individuals at homologous loci matters, and we find that the second-order condition of invasion of punishment is less stringent than the first-order condition given by equation [20]. Allele P can now be selected for in the presence of a fecundity cost ($C_P > 0$). Indeed, under second-order effects, the effect of selection on relatedness is taken into account, and punishment results in a decrease in relatedness between patch members (see eq. [A22]). A focal punisher is then less likely to be punished from defecting in the focal patch, a situation decreasing the selective pressure against punishment. Actually, punishment here can simply be envisioned as an indiscriminate "harming" behavior that is likely to spread because a focal individual reduces competition for its own offspring by harming all other individuals in its patch, an action resulting in a benefit to self. From the analysis of the total selective pressure, we found that allele P will spread only if the behavior results in a positive effect on the fitness ($-c_P$) of a focal individual, which, to the first order in $1/N$, is given by

$$-c_P = -C_P + \frac{(1 - m)^2(1 - C_P)(D + C_P)}{N}. \quad (22)$$

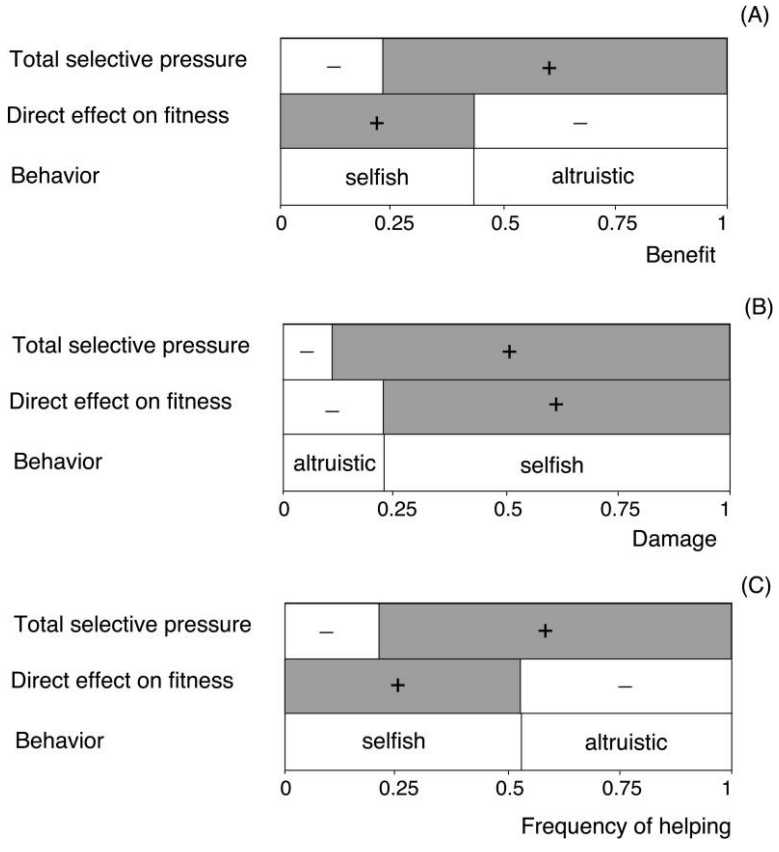

Figure 2: Signs of both the total selective pressure on punishment and the direct effect $-c_P$ of an actor on its fitness when allele p_P is rare ($p_P \to 0$). Gray shading indicates positive effect, white shading indicates negative effect. *A*, Signs as a function of the benefit B. Parameter values are $p_H = 0.5$, $N = 10$, $m = 0.1$, $r = 0.5$, $C_H = 0.01$, $C_P = 0.01$, and $D = 0.2$. *B*, Signs as a function of the damage D. Same parameter values as panel *A* except $B = 0.5$. *C*, Signs as a function of the frequency p_H of the helping allele in the population. Same parameter values as panel *A* except that $B = 0.5$ and $D = 0.25$.

Although a focal individual can bear a fecundity cost ($C_P > 0$), the allele spreads only if it results in a net fitness benefit for that individual ($-c_P > 0$). By contrast to punishment, the condition of invasion of helping with second-order phenotypic effects is more stringent than the condition with only first-order effects (eq. [17]). An explanation for this is that relatedness between two patch members is reduced when the effect of helping on relatedness is taken into account (see eq. [A20]), which further reduces the kin-selected benefits. Costly helping ($C_H > 0$) can therefore not invade in the absence of punishment. Notice that the results discussed in this paragraph corroborate the results obtained with the approximations developed in the section "Weak Selection."

By contrast, when the initial frequency of the helping and/or punishment allele(s) is not vanishingly small ($p_H > 0$ and/or $p_P > 0$), associations between the helping and the punishment alleles influence the selective pressure. The analysis of the selective pressures indicates that selection on punishment becomes positive frequency dependent with respect to the helping allele. That is, there is a threshold frequency $p_{H,T}(C_H, B, C_P, D, m, N, r)$ of the helping allele where punishment is selected for when rare ($p_P \to 0$) and which is determined by the costs, the benefits, and the demographic parameters of the population. Whether such punishment qualifies as altruistic or selfish also depends on all these parameters. Inspecting equation (A25) suggests that increasing the benefit (B) may result in altruistic punishment because by helping, punishers enhance the competition for their own offspring, thus increasing the net fitness cost of helping. By contrast, high values of the damage (D) may result in selfish strategies because the act of punishment reduces competition for the offspring of the actor. In figure 2, we compare the effect of an actor on its fitness ($-c_P$; eq. [A25]) and the selective pressure on punishment as a function of B, D, and p_H. From figure 2, it can be seen that when the frequency of helping is not vanishingly small and B is high

and/or D is low, punishment can simultaneously be altruistic and selected for. Increasing dispersal or deme size also generally leads to punishment becoming an altruistic rather than a selfish strategy. Similar to the first-order approximation (eq. [17]), selection on helping becomes positive frequency dependent with respect to the punishment allele when the frequency of punishment is not vanishingly small. There is a threshold frequency $p_{P,T}(C_H, B, C_P, D, m, N, r)$ of the punishment allele where helping is selected for when rare ($p_H \to 0$). This threshold frequency is given by C/D when phenotypic effects are of first order (eq. [17]), and it is slightly lower when phenotypic effects are of second order ($p_{P,T} < C/D$), with the result that selection on helping is slightly increased with a decrease in the dispersal rate and deme size. When the threshold frequencies are satisfied ($p_P > p_{P,T}$, $p_H > p_{H,T}$), the alleles can invade the population and go to fixation so that all individuals will ultimately behave as strong reciprocators. We finally mention that we observed some instances of stable polymorphism at the punishment locus in the situation where helping does not result in any benefit ($B = 0$).

In order to check the validity of our equations, we also performed simulations in which we evaluated the equilibrium allele frequencies maintained by a balance between selection and a regime of recurrent mutations. Figure 3 compares the steady state frequencies of the punishment allele (p_H) and helping allele (p_P) obtained from the second-order analytical model by incorporating mutations with those obtained from simulations. Since the simulations are only for checking purposes, no attempt was made to model mutation in a realistic manner. The simulations generally confirm the analytic results obtained by taking into account second-order phenotypic effects on fitness. Under stronger forms of selection (crudely, when phenotypic effects exceed 0.2), the selective pressures on punishment and helping obtained from the analytical model by taking second-order effects into account overestimate the selective pressure observed in the simulations. This suggests that the exact selective pressures lie in between the first- and second-order predictions obtained with the analytical models.

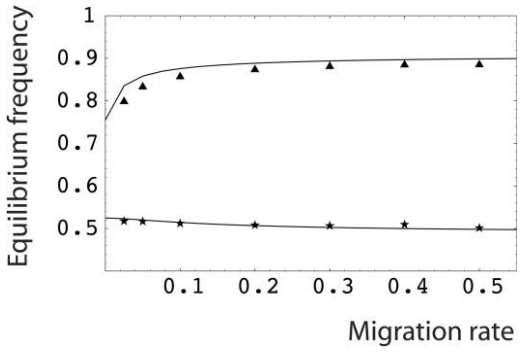

Figure 3: Equilibrium frequencies of the helping allele ($\hat{p}_H$) and punishment allele ($\hat{p}_P$) under both second-order phenotypic effects on fitness and a model of recurrent mutations (here with probability $\mu = 0.01$ an allele at a locus mutates into its alternative right before regulation) graphed as a function of migration. The upper line is the equilibrium frequency $\hat{p}_H$ of helping obtained from the analytical model under strong selection, and the triangles represent the mean frequency of allele H in the population obtained from simulations. The lower line is the equilibrium frequency $\hat{p}_P$ of punishment obtained from the analytical model, and the stars represent the mean frequencies of allele P obtained from simulations. Parameter values are $N = 10$, $r = 0.5$, $C_H = 0.005$, $C_P = 0.005$, $B = 0.2$, and $D = 0.2$. In simulations, exact expressions for fecundity were used, and regulation occurred by multinomial sampling of gametes. Simulation results are means over 100,000 generations of evolution of a population of 100 demes starting with initial conditions $\hat{p}_H = 0.5$ and $\hat{p}_P = 0.5$. Mutations were incorporated in the analytical model by using equations (27) and (28) of Kirkpatrick et al. (2002) with symmetric mutation rate μ.

Discussion

Strong reciprocity has been proposed as a potent mechanism promoting altruism and cooperation in humans, but there has been no thorough analysis of the conditions conducive to its emergence and spread within populations. Here we investigated the selective pressure on strong reciprocity and the conditions for its evolution when both recombination and spatial structure jointly determine the coevolutionary dynamics of punishment and helping. The results of our population genetic derivations suggest that in the absence of recombination between helping and punishment, strong reciprocity can invade a spatially structured population at all allele frequencies. In particular, strong reciprocity can be selected for when its initial frequency is small (i.e., rare mutants can invade). In that case, strong reciprocity qualifies as altruistic or selfish depending on the value of the parameters of the model. Further, the selective force operating is kin selection. By contrast, when recombination occurs between helping and punishment, strong reciprocity cannot invade a population of defectors when its initial frequency is low unless punishing results in a direct fecundity benefit. However, strong reciprocity can be selected for in subdivided populations if its initial frequency is greater than a given threshold that is determined by the demographic parameters of the population (table 2 summarizes the situations in which rare mutants invading the population are altruistic).

No Recombination between Helping and Punishment

Our analyses corroborate the findings that limited dispersal facilitates the spread of strong reciprocity under the

Table 2: Conditions under which rare mutants invading the population can be altruistic

Models	To be or not to be altruistic
Helping and punishment linked	Strong reciprocity can be altruistic
Helping and punishment not linked (weak selection)	Helping can be altruistic while punishment can only be selfish
Helping and punishment not linked (stronger selection):	
(a) Both mutants are rare	Helping can be altruistic while punishment can only be selfish
(b) Helping is rare, punishment is not rare	Helping can be altruistic
(c) Punishment is rare, helping is not rare	Punishment can be altruistic

assumption that both punishment and helping are completely linked traits (i.e., strong reciprocity is a single Mendelian trait; Bowles and Gintis 2004; Nakamaru and Iwasa 2005, 2006). Strong reciprocity can invade when rare and go to fixation in the population, but our model shows that the condition for the invasion may be independent of the benefits of helping. This result has to our knowledge never been discussed in the literature and stems from the fact that the fecundity benefit B that an actor confers to its neighbors is exactly canceled out by the concomitant increase in kin competition generated by the act of helping (Taylor 1992*a*, 1992*b*). Strong reciprocators are selected for not because they increase the productivity (by helping) of other individuals in the deme but because they decrease (by harming) the fitness of individuals that do not bear the trait and thus reduce locally the intensity of competition. Limited dispersal and small deme size both increase the likelihood that strong reciprocators will benefit from the reduction of competition resulting from relatives harming defectors by punishment. Helping is thus essentially used as a tag that allows one to identify individuals that do not bear the punishment allele and to exclude them from the population. This is in line with the view that punishment can be interpreted as a "spiteful" strategy (Nakamaru and Iwasa 2006). In fact, any tag associated with punishment would help detecting those individuals that do not bear the punishment allele and then reduce their fitness by harming. Accordingly, strong reciprocators are functioning like "green beards" (Hamilton 1964*b*; Dawkins 1982; Lehmann and Keller 2006). This is exactly the situation that has been reported in the fire ant *Solenopsis invicta*, where workers bearing one copy of a selfish gene kill all queens in their colony not bearing a copy of that gene (Keller and Ross 1998). By so doing, workers increase the fecundity of the other queens in their colony that do bear a copy of the green beard gene.

The condition of invasion of strong reciprocity (eq. [14]) depends on the coefficient of relatedness R between two interacting individuals. Relatedness can take substantial values for a deme of large size when migration is low (e.g., $R = 0.09$ for $m = 0.1$ and $N = 50$), but it can be further increased under different life-cycle assumptions. For instance, overlapping generations promote helping under the life cycle investigated here even without punishment (Taylor and Irwin 2000), so when generations are overlapping, the condition for the invasion of strong reciprocity will be relaxed. In particular, when reproduction follows a Moran model (only one individual dies per unit time; e.g., Ewens 2004), relatedness takes its maximal value, which for large deme size becomes $R = (1/m - 1)/N$ and takes the value of $R = 0.09$ when $m = 0.1$ and $N = 100$. Importantly, in the presence of overlapping generations, the fecundity benefit generated by helping does not cancel out in equation (11) (i.e., $R > [1 - m]^2 R^R$), and selection on strong reciprocity increases with an increase in the benefit of helping.

Recombination between Helping and Punishment

The dynamics of the coevolution of helping and punishment in the presence of recombination depends on the association of genes within and between individuals at the same and at different loci. While recombination tends to break down the association of helping and punishment within (i.e., linkage disequilibrium) and between individuals, limited migration, finite deme size, and selection work in the opposite direction (see eqq. [A18], [A19]). Limited migration and finite deme size also increase the relatedness between individuals at both the helping and punishment loci, while the effect of selection is frequency dependent (see eqq. [A20], [A21]).

Under weak selection (i.e., first-order phenotypic effects on fitness), the associations of helping and punishment within and between individuals cannot build up because the intensity of selection favoring it is too weak relative to the erosion brought by recombination. In that case, only relatedness may matter, and our analysis demonstrates that punishment evolves only if it results in a direct fecundity benefit to the actor ($-C_P > 0$). By contrast, the conditions for helping to be selected for is frequency dependent, and helping becomes a better strategy than the alternate option of defecting when the frequency of pun-

ishers in the population times the damage incurred by punishment exceeds the cost of helping ($p_P D - C_H > 0$). Selection on helping is thus enforced by the threat of punishment, a feature that has been reported in several interspecific mutualisms in the wild (Pellmyr and Huth 1994; Kiers et al. 2003; Bshary and Grutter 2005). Our analysis suggest that the condition $p_P D - C_H > 0$ holds for any migration rate and patch size, but it is also likely to hold for lattice models with explicit space because this condition of invasion can be interpreted as an application of Taylor's (1992*a*; 1992*b*) result, which holds independently of the spatial structure of the population (see discussion of eq. [2]).

Introducing second-order phenotypic effects on fitness (i.e., larger phenotypic effects) generates associations between helping and punishment within and between individuals as a result of an interaction between selection and identity by descent. Under the condition that helping has a nonvanishingly small frequency in the population (i.e., $p_H > 0$), an individual bearing the punishment allele is also likely to express helping and receive helping from its neighbors. Altruistic punishment can evolve when the value of the benefit B of helping is high, the damage of punishment D is low, and when the dispersal rate and patch size are high (see fig. 2); otherwise, the trait qualifies as selfish. Importantly, however, when the initial frequency of helpers is close to 0, altruistic punishment is always counterselected (see fig. 2). Similarly, when the initial frequency of punishers tends toward 0, helping cannot evolve if it results in a fecundity cost for the actor ($C_H > 0$). Therefore, costly helping cannot evolve without punishment and costly punishment cannot evolve without helping. However, there are many different ways by which helping strategies may invade spatially subdivided populations in the first place. For instance, overlapping generations, different kin-discrimination mechanisms, various modes of dispersal, and different effects of helping on patch demography and ecology can tip the balance in favor of helping behaviors (van Baalen and Rand 1998; Taylor and Irwin 2000; Perrin and Lehmann 2001; Boyd et al. 2003; Le Galliard et al. 2003; Axelrod et al. 2004; Gardner and West 2006; Lehmann 2006; Lehmann et al. 2006). Once helping has reached a threshold frequency through these mechanisms, punishment can invade the population, and because selection on helping increases with the frequency of punishers (see eq. [17]), the selective pressure on helping can subsequently be enhanced. This feedback, in which helping raises selection on punishment and punishment then reinforces selection on helping, might result in altruism evolving in groups of larger size, as would occur in the absence of punishment. A similar process might work in panmictic populations as well, where costly helping can first evolve through interactions occurring among members of a family. However, these feedback processes are probably not irreversible once kin selection is completely suppressed, and it remains to investigate the extent to which they can explain helping in groups of large size or among unrelated individuals.

The results discussed in this section hold under the assumption that recombination is stronger than selection. Indeed, the accuracy of our quasi-equilibrium approximation (see "Model") breaks down with very low recombination rates, a situation where the selective pressure on strong reciprocity is probably underestimated in our models. Since our model with perfect linkage indicates that strong reciprocity can evolve when rare, it is plausible that such evolution will also take place when recombination is weak and selection is strong.

Genetic versus Cultural Transmission of Strong Reciprocity

We have assumed a genetic transmission of strong reciprocity. A crucial difference with cultural transmission is that genetic transmission occurs only vertically, whereas cultural transmission can also be oblique and horizontal (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985). However, a crucial similarity between these two modes of transmission is that the concept of relatedness applies to both of them. Under genetic transmission, relatedness measures the extent to which two individuals sampled from the same group are more likely to bear the same genes inherited from a common ancestor than two individuals sampled from two different groups. Similarly, under cultural transmission, relatedness measures the extent to which two individuals from the same group are more likely to bear the same cultural variant (meme) inherited from a common cultural ancestor than two individuals sampled from two different groups (Werren and Pulliam 1981; Feldman et al. 1985; Allison 1991).

Our models apply to cultural transmission under specific conditions. When helping and punishment are perfectly linked traits, strong reciprocity can be envisioned as a cultural variant affecting Darwinian fitness (e.g., Cavalli-Sforza and Feldman 1981; Feldman et al. 1985), which is transmitted from the parental to the offspring generation between the reproduction and the dispersal stage of our life cycle. With these assumptions, different modes of vertical and/or oblique transmission of the trait will affect the change in frequency of strong reciprocity (eq. [11]) only to the extent that they result in a different coefficient of kinship (see "Cultural Transmission of Strong Reciprocity" in the appendix in the online edition of the *American Naturalist*). For instance, when transmission occurs primarily by parents (i.e., each offspring inherits the cultural variant from its parent with probability τ and, with complementary probability $1 - \tau$, adopts the cultural variant

of one of the $N-1$ other parents in a deme [Feldman et al. 1985, chap. 3.11]), the cultural kinship is precisely given by equation (13) and is independent of the value of the parameter τ (see eq. [A29]). Hence, this mode of cultural transmission does not affect the condition of invasion of strong reciprocity established for genetic transmission. By contrast, when transmission occurs primarily by teachers (i.e., each offspring inherits the cultural variant from the same teacher with probability τ, and, with complementary probability $1-\tau$, adopts the cultural variant of one of the $N-1$ other adult individuals in a deme [Feldman et al. 1985, chap. 3.11]), cultural kinship may increase markedly compared with genetic transmission (see eq. [A32]). Such a "one to many" transmission scheme can then relax the condition of invasion of strong reciprocity. Our models where recombination occurs between helping and punishment can also be applied to cultural oblique transmission when cultural variants affect Darwinian fitness. In this case, the recombination rate of our models represents the probability that a given individual of the offspring generation adopts the punishment and the helping traits from two different individuals of the parental generation.

Finally, by introducing slight modifications, our models apply to the dynamics of horizontal transmission schemes where individuals imitate actions that perform better, with a probability proportional to the expected payoff obtained during the social interaction stage of our life cycle (see "Cultural Transmission by Imitation of Strong Reciprocity" in the appendix in the online edition of the *American Naturalist*). This imitation rule leads to the same dynamics as a vertically transmitted trait in panmictic populations (Hofbauer and Sigmund 1998, p. 87). However, whether this imitation rule leads to the same dynamics as genetic transmission depends on additional assumptions in spatially subdivided populations (see *D*). For instance, Boyd et al. (2003) investigated through simulations a cultural model of strong reciprocity with imitation in the presence of group extinction. In the absence of group extinction, this is very similar to our model (introducing group extinction leads in itself to a high selective pressure on helping; Lehmann et al. 2006, see their eq. [9]). Boyd et al. (2003) assume that with probability $1-m$, a focal individual encounters an individual from the focal group and with probability m, an individual from a different group. In each case, imitation occurs proportionally to payoff. Surprisingly, under these assumptions the selective pressure on strong reciprocity (eq. [A41]) is much lower than under genetic transmission (eq. [11]). This is because under transmission, through imitation of strong reciprocity, the intensity of kin competition is increased compared with a genetic transmission of the trait (cf. eq. [8] with eqq. [A39]–[A41]) with the result that greater benefit B of helping translates into greater selection against strong reciprocity.

In conclusion, our models suggest that cultural transmission may increase as well as decrease the selective pressure on helping relative to the case of genetic transmission. While genetic associations have been evaluated under myriads of life cycles, the study of the extent to which modes of cultural transmission affect cultural kinship or the fitness function in subdivided populations has fallen into oblivion. Future research should focus on how specific modes of transmission may by themselves explain the evolution of helping in groups of large size (even in the absence of punishment) and link more directly inclusive fitness theory with "cultural group selection" theory (e.g., Henrich and Boyd 2001; Richerson and Boyd 2005). This link might also be of practical relevance for constructing explicit models. Indeed, where economists strive to evaluate the conditions of invasion and stabilities of mutant strategies by computing the complete distributions of the number of copies of the strategies within and among groups (e.g., Ellison 1993; Kandori et al. 1993), adopting an inclusive fitness approach reduces the problem to the much simpler task of computing the probabilities that pairs of strategies sampled within and among groups are identical (Roze and Rousset 2003; Rousset and Ronce 2004; Rousset 2006).

Concluding Remarks

Our models suggest that kinship between interacting individuals is necessary for the evolution of strong reciprocity with both genetic and cultural transmission. In other words, strong reciprocity is not an alternative evolutionary mechanism as is sometimes implied (Gintis 2000, 2003; Fehr and Fischbacher 2003; Gintis et al. 2003; Bowles and Gintis 2004) but merely constitutes a specific proximate mechanism that generates a positive selective pressure on helping and that results in inclusive fitness benefits and/or direct benefits (self interest). Since kinship can still be substantial in groups of large size (i.e., $N \sim 50$), strong reciprocity can lead to the evolution of helping in groups of such size, especially if it is coupled with one or several of the various demographic, ecological, and life-history factors already selecting for helping in subdivided populations.

With respect to these conclusions, it is important to recall that strong reciprocity has been raised as an "explanation" of cooperation in anonymous nonrepeated interactions (Gintis 2000, 2003; Fehr and Fischbacher 2003; Bowles and Gintis 2004). Gintis et al. (2003, p. 168) specify that subjects in experimental games are unlikely to behave in a maladaptive way because they do not confuse the experimental environment (i.e., anonymous nonrepeated

interactions) with a more evolutionarily familiar situation such as a nonanonymous repeated game in which cooperation is the best option. According to these comments, neither "genetic group selection" nor "cultural group selection" of strong reciprocity can be invoked as an explanation for cooperation in anonymous nonrepeated interactions because it would imply that subjects in experimental environments misinterpret the context and behave in a way that is adaptive in an environment of nonrandom interactions (be it through genetic or cultural kinship). Our analysis of the evolution of strong reciprocity thus suggest that the solution to the puzzle of why humans engage in costly helping in anonymous and nonrepeated interactions will not be solved by inventing new models that promote strong reciprocity but rather by understanding the interpretative frames players are using in experimental games (Hagen and Hammerstein 2006).

Acknowledgments

We thank S. Bowles, H. Gintis, M. Reuter, and S. West for useful discussions. We also thank A. Gardner for comments that improved this article. L.K. and L.L. were both supported by several grants from the Swiss National Science Foundation. D.R. was supported by an European Molecular Biology Organization long-term fellowship (ALTF 280-2004). This is Institut des Sciences de l'Entreprise de Montpellier publication 07-044.

Literature Cited

Allison, P. D. 1991. Cultural relatedness under oblique and horizontal transmission. Ethology and Sociobiology 13:153–169.

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

Axelrod, R., R. A. Hammond, and A. Grafen. 2004. Altruism via kin-selection strategies that rely on arbitrary tags with which they coevolve. Evolution 58:1833–1838.

Bowles, S., and H. Gintis. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. Theoretical Population Biology 65:17–28.

Bowles, S., J. K. Choi, and A. Hopfensitz. 2003. The co-evolution of individual behaviors and social institutions. Journal of Theoretical Biology 223:135–147.

Boyd, R., and P. J. Richerson. 1985. Culture and the evolutionary process. University of Chicago Press, Chicago.

———. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13: 171–195.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson. 2003. The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the USA 100:3531–3535.

Brandt, H., C. Hauert, and K. Sigmund. 2006. Punishing and abstaining for public goods. Proceedings of the National Academy of Sciences of the USA 103:495–497.

Bshary, R., and A. S. Grutter. 2005. Punishment and partner switching causes cooperative behavior in a cleaning mutualism. Biology Letters 1:396–399.

Bürger, R. 2000. The mathematical theory of selection, recombination, and mutation. Wiley, New York.

Cavalli-Sforza, L., and M. W. Feldman. 1981. Cultural transmission and evolution. Princeton University Press, Princeton, NJ.

Dawkins, R. 1982. The extended phenotype. Oxford University Press, Oxford.

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

Ewens, W. J. 2004. Mathematical population genetics. Springer, New York.

Fehr, E., and U. Fischbacher. 2003. The nature of human altruism. Nature 425:785–791.

Feldman, M. W., L. Cavalli-Sforza, and J. L. Peck. 1985. Gene-culture coevolution: models for the evolution of altruism with cultural transmission. Proceedings of the National Academy of Sciences of the USA 82:5814–5818.

Gardner, A., and S. A. West. 2004. Cooperation and punishment, especially in humans. American Naturalist 164:753–764.

———. 2006. Demography, altruism, and the benefits of budding. Journal of Evolutionary Biology 19:1707–1716.

Gardner, A., S. A. West, and N. Barton. 2006. The relation between multilocus population genetics and social evolution. American Naturalist 2:207–228.

Gintis, H. 2000. Strong reciprocity and human sociality. Journal of Theoretical Biology 206:169–179.

———. 2003. Solving the puzzle of prosociality. Rationality and Society 15:155–187.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2003. Explaining altruistic behavior in humans. Evolution and Human Behavior 24:153–172.

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

Hagen, E. H., and P. Hammerstein. 2006. Game theory and human evolution: a critique of some recent interpretations of experimental games. Theoretical Population Biology 69:339–348.

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

———. 1964*b*. The genetical evolution of social behaviour. II. Journal of Theoretical Biology 7:17–52.

———. 1970. Selfish and spiteful behavior in an evolutionary model. Nature 228:1218–1220.

Henrich, J., and R. Boyd. 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of Theoretical Biology 208:79–89.

Hirshleifer, D., and E. Rasmusen. 1989. Cooperation in a repeated Prisoner's Dilemma with ostracism. Journal of Economic Behavior and Organization 12:87–106.

Hofbauer, J., and K. Sigmund. 1998. Evolutionary games and population dynamics. Cambridge University Press, Cambridge.

Irwin, A. J., and P. D. Taylor. 2001. Evolution of altruism in stepping-stone populations with overlapping generations. Theoretical Population Biology 60:315–325.

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

Keller, L., and G. Ross. 1998. Selfish genes: a green beard in the red fire ant. Nature 394:573–575.

Kiers, E., R. Rousseau, and S. West. 2003. Host sanctions and the legume-rhizobium mutualism. Nature 425:78–81.

Kimura, M. 1965. Attainment of quasi-linkage equilibrium when gene frequencies are changing by natural selection. Genetics 52: 875–890.

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

Le Galliard, J., R. Ferrière, and U. Dieckmann. 2003. The adaptive dynamics of altruism in spatially heterogeneous populations. Evolution 57:1–17.

Lehmann, L. 2006. The evolution of trans-generational altruism: kin selection meets niche construction. Journal of Evolutionary Biology 20:181–189.

Lehmann, L., and L. Keller. 2006. The evolution of cooperation and altruism: a general framework and a classification of models. Journal of Evolutionary Biology 19:1365–1376.

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

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

Nakamaru, M., and Y. Iwasa. 2005. The evolution of altruism by costly punishment in lattice-structured populations: score-dependent viability versus score-dependent fertility. Evolutionary Ecology Research 7:853–870.

———. 2006. The coevolution of altruism and punishment: role of the selfish punisher. Journal of Theoretical Biology 240:475–488.

Pellmyr, O., and C. J. Huth. 1994. Evolutionary stability of mutualism between yuccas and yucca moths. Nature 372:257–260.

Perrin, N., and L. Lehmann. 2001. Is sociality driven by the costs of dispersal or the benefits of philopatry? a role for kin discrimination mechanisms. American Naturalist 158:471–483.

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

Richerson, P. J., and R. Boyd. 2005. Not by genes alone. University of Chicago Press, Chicago.

Rousset, F. 2004. Genetic structure and selection in subdivided populations. Princeton University Press, Princeton, NJ.

———. 2006. Separation of time scales, fixation probabilities and convergence to evolutionarily stable states under isolation by distance. Theoretical Population Biology 69:165–179.

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

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

———. 2005. Inbreeding depression and the evolution of dispersal rates: a multilocus model. American Naturalist 166:708–721.

Seger, J. 1985. Unifying genetic models for the evolution of female choice. Evolution 39:1185–1193.

Sigmund, K., C. Hauert, and M. A. Nowak. 2001. Reward and punishment. Proceedings of the National Academy of Sciences of the USA 98:10757–10762.

Taylor, P. D. 1992*a*. Altruism in viscous populations: an inclusive fitness model. Evolutionary Ecology 6:352–356.

———. 1992*b*. Inclusive fitness in a homogeneous environment. Proceedings of the Royal Society B: Biological Sciences 240:299–302.

Taylor, P. D., and A. J. Irwin. 2000. Overlapping generations can promote altruistic behavior. Evolution 54:1135–1141.

van Baalen, M., and A. Rand. 1998. The unit of selection in viscous populations and the evolution of altruism. Journal of Theoretical Biology 193:631–648.

Werren, J., and R. Pulliam. 1981. An intergenerational transmission model for the cultural evolution of helping behavior. Human Ecology 9:466–493.

West, S. A., A. S. Griffin, and A. Gardner. 2006. Social semantics: altruism, cooperation, mutualism and strong reciprocity. Journal of Evolutionary Biology 20:415–432.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics 15:323–354.

Associate Editor: Troy Day
Editor: Michael C. Whitlock