Evolutionary Ecology Research, 2008, **10**: 1201–1216

Comparison of two methods for analysing the biological factors contributing to assortative mating or sexual isolation

Andrés Pérez-Figueroa[1], Jacobo de Uña-Alvarez[2], Paula Conde-Padín[1] and Emilio Rolán-Alvarez[1]

[1]*Departamento de Bioquímica, Genética e Immunologia, Facultad de Biología and* [2]*Departamento do Estadistica e Investigación Operativa, Facultad de Ciencias Económicas y Empresariales, Universidad de Vigo, Vigo, Spain*

ABSTRACT

Question: How can we establish the biological factors that contribute to variation in assortative mating (based on a quantitative or qualitative trait)?

Key assumptions: Assortative (or disassortative) mating for a particular trait can produce sexual isolation between ecotypes or incipient species. The individual contribution to population assortative mating for a quantitative trait can be estimated by means of the r_i statistic, which is an additive decomposition of the Pearson correlation coefficient. The mating pair contribution to population sexual isolation can be estimated by the *PSI* coefficient. These statistics can be used to quantify the variability in assortative mating/sexual isolation in a particular population.

Search method: It was recently proposed that both the r_i statistic and the *PSI* coefficient could be used as dependent variables in a multiple regression approach to determine which of a set of independent variables explains the greatest variation in the dependent variable. We describe both statistics and undertake simulations to compare the efficiency of each statistic to infer assortative mating when it is caused *a priori* by a mate choice decision based on a quantitative or a qualitative trait.

Conclusions: The r_i statistic outperforms the *PSI* coefficient when trying to infer the causes of both assortative mating and sexual isolation. The applicability of both methods to other cases is discussed.

Keywords: estimation properties, incipient speciation, mate choice, mate discrimination, mating behaviour, regression, speciation.

Correspondence: E. Rolán-Alvarez, Departamento de Bioquímica, Genética e Immunologia, Facultad de Biología, Universidad de Vigo, 36200 Vigo, Spain. e-mail: rolan@uvigo.es

INTRODUCTION

Mating is a fundamental characteristic of sexual organisms, and partners need to be carefully chosen to avoid wasted time and effort due to unsuccessful mating within or between species (Andersson, 1994). Within a particular species, the occurrence of similar or dissimilar mating pair types that exceeds random mating is called assortative or disassortative mating respectively (Lewontin *et al.*, 1968). One of the most frequent cases of assortative mating occurs for body size (when mating preferentially occurs between similar-sized individuals), and this has been documented extensively in insects, molluscs, reptiles, birds, and humans (Crespi, 1989; Staub and Ribi, 1995; Jonson, 1999; Delestrade, 2000; Forero *et al.*, 2001; Masello and Quillfeldt, 2003; Silventoinen *et al.*, 2003). It has been demonstrated that size assortative mating is involved in the reproductive isolation of two morphs or ecotypes that differ in size (Nagel and Schluter, 1998; Cruz *et al.*, 2004a; McKinnon *et al.*, 2004; Rolán-Alvarez, 2007). In fact, mating behaviour is one of the main mechanisms able to produce isolation barriers between incipient species [sexual isolation, *sensu* Coyne and Orr (2004)]. A full understanding of the biological mechanisms responsible for sexual isolation is required to explain how reproductive isolation evolves *in situ* as well as how it can be reinforced after secondary contact (Coyne and Orr, 2004). In addition, the behavioural mechanisms responsible for sexual isolation have been key parameters in some theoretical models of speciation (Turelli *et al.*, 2001; Kirkpatrick and Ravigné, 2002; Gavrilets, 2004).

The degree of assortative mating for a particular quantitative trait is quantified by the Pearson (or related) correlation coefficient (Johannesson *et al.*, 1995; Masumoto 1999; Silventoinen *et al.*, 2003). This correlation provides an estimate of assortative mating for the whole population, but not for each mating pair separately. An index for the estimation of the individual contribution (of each pair) to overall assortative mating in the population would help us to understand the causes of assortative mating and sexual isolation. This approach has been satisfactorily adopted by employing multiple regression to investigate the causes and consequences of sexual selection in the wild (Arnold and Wade, 1984a, 1984b; Cruz *et al.*, 2001; reviewed in Brodie *et al.*, 1995), especially when sexual selection estimates were available for each specimen. This methodology could also be applied to the study of assortative mating if a similar index was available.

Recently, a study of sexual isolation and parallel ecological divergence in the marine gastropod *Littorina saxatilis* (Conde-Padín *et al.*, 2008) used an additive decomposition of the Pearson correlation coefficient (called r_i), which can be estimated for each mating pair individually and, therefore, allows one to investigate the contribution of each mating pair to assortative mating for any trait. This statistic was used together with multiple regression to examine the biological variables that contribute the most to the variation in assortative mating. Among a series of shape and size morphological measurements, the square of male size was the main factor responsible for the observed variation in size assortative mating (Conde-Padín *et al.*, 2008). The relationship is quadratic because under size assortative mating, mating pairs showing the greatest deviations from mean size also have the most pronounced size assortative mating. In addition, size assortative mating was closely linked to sexual isolation in *L. saxatilis*, since this species commonly presents size assortative mating, and in the population studied there were two ecotypes showing clear size differences between them (Rolán-Alvarez, 2007). Conde-Padín and colleagues also used an estimate of sexual isolation per mating type, the *PSI* coefficient (Rolán-Alvarez and Caballero, 2000), in combination with the same multiple regression approach, to study the causes of sexual isolation. The trend observed was very similar using both statistics, although the results obtained

suggested that r_i was better than the *PSI* for predicting variation in assortative mating using a regression approach. Here, we describe the sampling properties of the r_i statistic, use a simulation study to compare the ability of the two statistics to infer the causes of size assortative mating and sexual isolation, and present an example.

DESCRIPTION OF THE r_i STATISTIC AND ITS SAMPLING PROPERTIES

The proposed estimator is based on the Pearson correlation coefficient (Pearson, 1894) but is calculated for each mating pair separately. We will refer to it as the individual correlation coefficient (r_i), and it is the product of the standardized values of a mating pair for a given variable:

$$r_i = Z_{\mathrm{m}} \times Z_{\mathrm{f}},$$

where Z_{m} and Z_{f} are the values of the variable standardized within each sex and sample. Thus, each Z value from any individual requires knowing the mean (μ) and the standard deviation (σ) for males and females independently for each sample obtained. For example, for each individual male with value x_{m}, $Z_{\mathrm{m}} = (x_{\mathrm{m}} - \mu_{\mathrm{m}})/\sigma_{\mathrm{m}}$, and analogously for females. The mean r_i value in the whole population is algebraically equal to the parametric Pearson correlation coefficient. The use of standardized variables has the advantage of allowing comparisons of the same variable in populations differing in their means and variances (Sokal and Rohlf, 1995).

To assess the properties of the sampling distribution of this statistic, we developed explicit formulae for the standard deviation, the asymmetry, and the kurtosis of the r_i statistic. This is interesting, since little is known about the distribution of a product of possibly correlated random variables (Nadarajah, 2006). Assume that (X, Y) is a random vector that follows a bivariate normal distribution. Let μ_X, μ_Y, σ_X, σ_Y, σ_{XY} denote the mean of X, the mean of Y, the standard deviation of X, the standard deviation of Y, and the covariance between X and Y respectively. Furthermore, let ρ_{XY} denote the Pearson correlation coefficient between X and Y, and let $r = Z(X)Z(Y) = (X - \mu_X)(Y - \mu_Y)/\sigma_X\sigma_Y$, where $Z(X)$ stands for a standardized X variable. Asymptotically, since the sample means and standard deviations converge to their population counterparts, we can approximate the distribution of r_i by that of r. We use the properties of the conditional expectation to write:

$$E(r) = E[Z(X)E(Z(Y)/X)].$$

Since (X, Y) is normally distributed, the conditioned variable Y/X is again normally distributed with mean $\mu_Y + \rho_{XY}(\sigma_Y/\sigma_X)(X - \mu_X)$ and variance $\sigma_Y^2(1 - \rho_{XY}^2)$. From this it follows that $E(Z(Y)/X) = \rho_{XY}(X - \mu_X)/\sigma_X$ and then $E(r) = \rho_{XY}$ is obtained. Now write:

$$E(r^2) = E[Z(X)^2E(Z(Y)^2/X)].$$

Taking into account the above mentioned properties of Y/X, it is easily shown that

$$E(Z(Y)^2/X) = 1 - \rho_{XY}^2 + \rho_{XY}^2(X - \mu_X)^2/\sigma_X^2.$$

This gives

$$E(r^2) = 1 - \rho_{XY}^2 + \rho_{XY}^2E(Z(X)^4) = 1 + \rho_{XY}^2,$$

where we have used the zero kurtosis property of the normal distribution $E(Z(X)^4) - 3 = 0$.

The variance of r can now be derived as $\mathrm{Var}(r) = E(r^2) - (E(r))^2 = 1 + \rho_{XY}^2$, which gives our first formula for the standard deviation:

$$s.d.(r_i) \approx \sqrt{1 + \rho_{XY}^2}. \tag{1}$$

For the asymmetry formula we need to derive first the moment of order 3 of r, $E(r^3)$. Now,

$$E(r^3) = E[Z(X)^3 E(Z(Y)^3/X)].$$

Again, assuming Y/X is normally distributed with mean $\mu_Y + \rho_{XY}(\sigma_Y/\sigma_X)(X - \mu_X)$ and variance $\sigma_Y^2(1 - \rho_{XY}^2)$, we easily obtain

$$E(Z(Y)^3/X) = (3\,\sigma_{XY}(X - \mu_X)\,\sigma_Y^2(1 - \rho_{XY}^2)/\sigma_X^2 + (\sigma_{XY}/\sigma_X^2)^3(X - \mu_X)^3)/\sigma_Y^3.$$

This equality and the properties of the standardized normal distribution,

$$E[Z(X)^3] = 3 \qquad E[Z(X)^6] = 15$$

give

$$E(r^3) = 3\,\rho_{XY}(3 + 2\rho_{XY}^2).$$

Since the asymmetry of r is given by

$$a(r) = E[((r - E(r))/\mathrm{Var}(r)^{1/2})^3] = [E(r^3) - 3E(r)E(r^2) + 2(E(r))^3]/\mathrm{Var}(r)^{3/2},$$

we obtain (after simple algebra)

$$a(r_i) \approx \frac{2\rho_{XY}(3 + \rho_{XY}^2)}{(1 + \rho_{XY}^2)^{\frac{3}{2}}}. \tag{2}$$

Finally, note that the kurtosis coefficient is written as:

$$k(r) = E[((r - E(r))/\mathrm{Var}(r)^{1/2})^4] - 3 = [E(r^4) - 4E(r)E(r^3) + 6E(r^2)(E(r))^2 - 3(E(r))^4] - 3.$$

Thus, to obtain an explication of the kurtosis we need to investigate the fourth-order moment of r. Using the normal distribution of Y/X, we obtain (after some lengthy although straightforward calculations)

$$E(Z(Y)^4/X) = 3(1 - \rho_{XY}^2)^2 + 6\,\rho_{XY}^2(1 - \rho_{XY}^2)Z(X)^2 + \rho_{XY}^4 Z(X)^4$$

and hence

$$E(r^4) = E[Z(X)^4 E(Z(Y)^4/X)] = 9 + 72\rho_{XY}^2 + 24\rho_{XY}^4,$$

and thus obtain

$$k(r_i) \approx 6 + \frac{24\rho_{XY}^2}{(1 + \rho_{XY}^2)^2}. \tag{3}$$

The three parameters (equations 1–3) are increasing functions of the correlation. For independent variables ($\rho_{XY} = 0$), we have $s.d.(r_i) \approx 1$, $a(r_i) \approx 0$, and $k(r_i) \approx 6$ (a symmetric, but not Gaussian, distribution). Under a perfect correlation ($\rho_{XY} = 1$), the parameter values are those corresponding to a chi-squared distribution with one degree of freedom.

SAMPLING PROPERTIES AT LOW SAMPLE SIZE

We simulated the sampling of mating pairs from a finite population to determine whether the above parameters are maintained at biological sample sizes. We assumed X and Y values (of the studied trait in each mating pair) obtained from different bivariate distributions, different sampling sizes, and different levels of *a priori* correlation between the values. There were three models involving alternative background distributions:

1. *Normal model*, using a normal distribution with mean equal to 0 and standard deviation equal to 1.
2. *Uniform model*, using a uniform distribution with values between 0 and $\sqrt{12}$, to obtain a standard deviation of 1, as in the previous case.
3. *Chi-squared model*, using a chi-squared distribution with 10 degrees of freedom.

To obtain correlations between X and Y values, they were calculated as follows. First, a value of X was obtained from the corresponding distribution. After that, a Y value was sampled following the linear model $Y = X + E$, where E is an error variable independent of X, with null mean and a standard deviation (σ_E) determined by the correlation level (ρ_{XY}) desired between X and Y. Thus, for the normal and uniform models, $\sigma_E = \sqrt{1 - \rho_{XY}^2}$, and for the chi-squared model, $\sigma_E = \sqrt{20} \times \sqrt{1 - \rho_{XY}^2}$. We simulated different sampling sizes (20, 50, 100, 500, and 1000 mating pairs). After the sampling of all pairs of X and Y values, these were standardized for their sample (e.g. $(X - \mu_x)/\sigma_x$, with μ_x and σ_x the sample mean and standard deviation of variable X). Finally, we calculated the r_i statistic for every pair and its distribution, defined by the mean, standard deviation, asymmetry, and kurtosis.

The distribution of the r_i statistic under different levels of *a priori* correlation is presented in Fig. 1. The sample size is very large (1000 pairs) and the scenario is the normal model. With no correlation ($\rho_{XY} = 0$), the distribution is almost symmetric. In that case, averaging 100 replicates, we obtain $\mu(r_i) = 0.00 \pm 0.00$, $\sigma(r_i) = 1.00 \pm 0.00$, $a(r_i) = -0.04 \pm 0.05$, and $k(r_i) = 5.97 \pm 0.32$. As the correlation increases, the distribution becomes more asymmetrical and leptokurtic, up to the point with perfect correlation ($\rho_{XY} = 1$), where the distribution fits to a chi-squared one with one degree of freedom (averaging 100 replicates: $\mu(r_i) = 1.00 \pm 0.00$, $\sigma(r_i) = 1.41 \pm 0.01$, $a(r_i) = 2.77 \pm 0.04$, $k(r_i) = 11.05 \pm 0.41$).

Table 1 shows the values of parameters (averaging 100 replicates) defining the r_i distribution with different sample sizes under the normal model. The shape of the distribution remains mostly constant up to sample sizes of 20 mating pairs. Kurtosis is the only parameter affected by the small sample size. Thus, as sample size decreases, the distribution becomes less leptokurtic, especially at greater degrees of correlation. Despite this, the r_i statistic seems appropriate for use with biological sample sizes (at least 20 mating pairs) because it is not biased by the sampling and maintains its distribution as a function of the degree of correlation.

A comparison of three models of initial distribution for the mating pairs values is presented in Fig. 2. There is little difference across the normal, uniform, and chi-squared models, except for the kurtosis in the latter. Thus, the r_i statistic could be used under conditions where the normality of the trait in the population cannot be assumed.

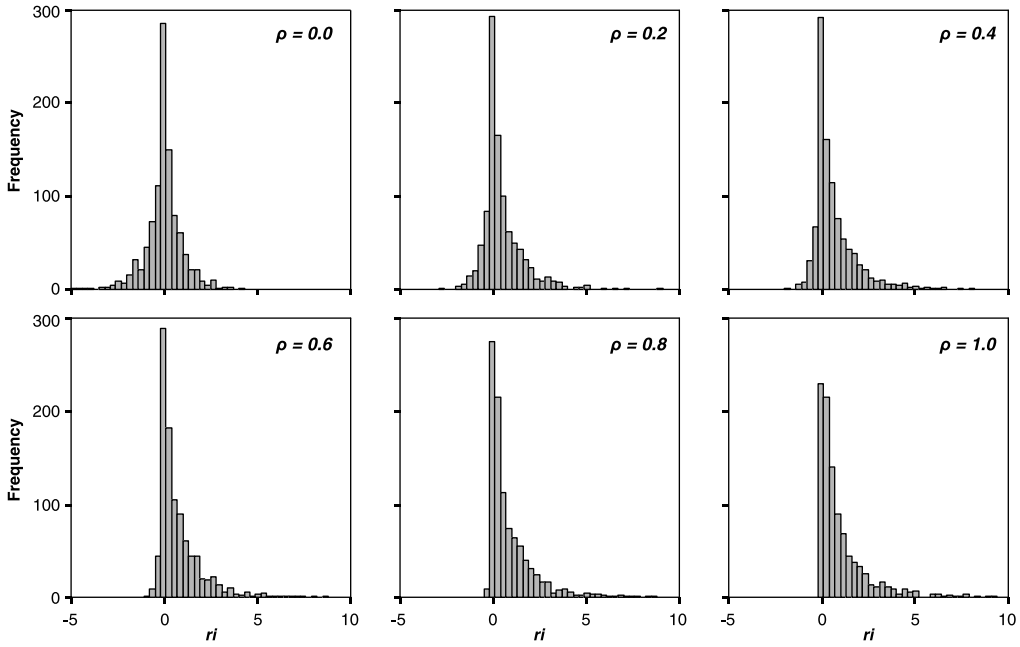

Fig. 1. Distribution of r_i under the normal model with different degrees of correlation (ρ_{XY}) between mating pairs. Large sample size (1000 pairs).

Table 1. Mean (μ), standard deviation (σ), asymmetry coefficient (a), and kurtosis (k) of the distribution of r_i values, averaged for 100 replicates, with different sample sizes, under the normal model and correlation (ρ_{XY}) between pairs (numbers in parentheses represent standard error)

	$\rho_{XY} = 0$				$\rho_{XY} = 1$			
Sample size	$\mu(r_i)$	$\sigma(r_i)$	$a(r_i)$	$k(r_i)$	$\mu(r_i)$	$\sigma(r_i)$	$a(r_i)$	$k(r_i)$
∞	0.00	1.00	0.00	6.00	1.00	1.41	2.82	12.00
500	0.01	1.00	0.00	5.58	1.00	1.41	2.75	10.93
	(0.00)	(0.00)	(0.06)	(0.28)	(0.00)	(0.01)	(0.05)	(0.65)
100	–0.02	0.99	–0.11	5.06	0.99	1.36	2.39	7.31
	(0.01)	(0.01)	(0.11)	(0.38)	(0.00)	(0.01)	(0.06)	(0.53)
50	0.00	0.96	0.10	3.77	0.98	1.38	2.43	7.41
	(0.02)	(0.01)	(0.11)	(0.34)	(0.00)	(0.02)	(0.08)	(0.59)
20	0.01	0.94	0.10	3.05	0.95	1.31	2.04	4.90
	(0.02)	(0.02)	(0.13)	(0.31)	(0.00)	(0.03)	(0.08)	(0.44)

BRIEF DESCRIPTION OF THE *PSI* STATISTIC

The only alternative approach for estimation of the individual contribution (of each pair) to overall assortative mating in the population is the *PSI* (pair sexual isolation) coefficient (Rolán-Alvarez and Caballero, 2000). The *PSI* coefficient is defined for every pair combination (for a categorical trait, i.e. ecotype) as the number of observed pair types divided by

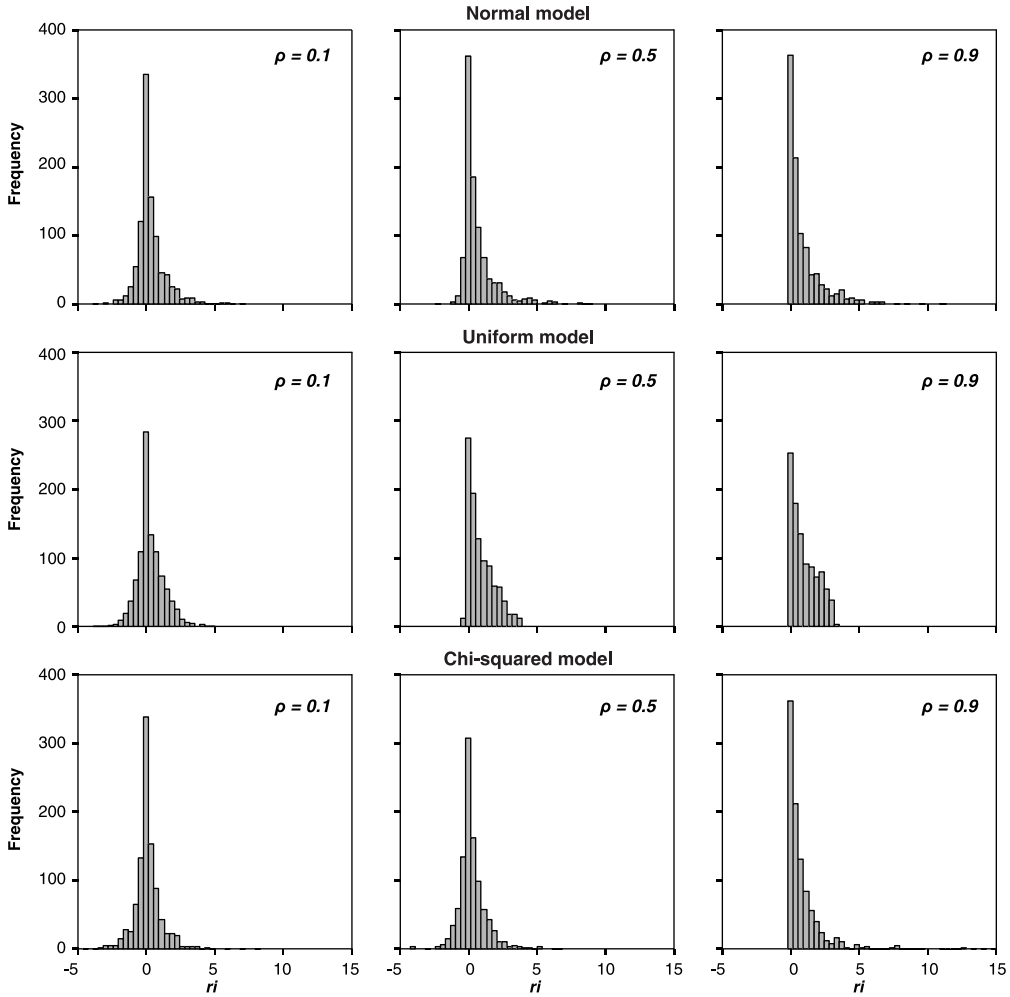

Fig. 2. Distribution of r_i under the different models studied with different degrees of correlation (ρ_{XY}) between mating pairs. Large sample size (1000 pairs).

the expected pair types from mates: let A and B be the two types of males (with A' and B' the corresponding frequencies of females) in a particular natural or laboratory population. After a particular period of time, the observed number of mates for each male and female type is aa, ab, ba, and bb, respectively, for a total number t. Then, the PSI for mates of male A and female B is $PSI_{ab} = \dfrac{ab \times t}{(aa + ab) \times (ab + bb)}$. In the same way, we can compute the PSI coefficient for the other mate types.

The PSI coefficient has been shown to be useful for estimating sexual isolation by applying the joint isolation index on the PSI:

$$I_{PSI} = \frac{(PSI_{aa} + PSI_{bb}) - (PSI_{ab} + PSI_{ba})}{(PSI_{aa} + PSI_{ab} + PSI_{ba} + PSI_{bb})}$$

(Rolán-Alvarez and Caballero, 2000). This estimator, the I_{PSI}, is presently considered one of the most accurate and unbiased sexual isolation estimators, especially at biological sample sizes (Pérez-Figueroa *et al.*, 2005).

Although the *PSI* is defined for qualitative traits, it could be used to analyse assortative mating by a quantitative trait by dividing it into a few discrete classes. In this way, the *PSI* could be used as the dependent variable in a multiple regression to investigate the biological variables that contribute the most to the variation in assortative mating (Conde-Padín *et al.*, 2008).

STATISTICAL COMPARISON BETWEEN r_i AND *PSI* REGRESSION APPROACHES

Computer simulations were developed to compare the ability of the different statistics to estimate different levels of assortative mating caused by different traits. We simulated the formation of mating pairs in a population assuming different *a priori* levels of assortative mating for a quantitative or a qualitative trait. Thus, we obtained a random sample of mates from such a population and, *a posteriori*, we could estimate the assortative mating (r_i and I_{PSI}). One of the objectives is to check *a posteriori* which estimator is the best predictor of the simulated (*a priori*) assortative mating. In addition, we could also check which of the estimators (r_i or *PSI*) is the most useful under the exploratory tool developed to infer the causes of assortative mating following Conde-Padín *et al.* (2008). This can be done by linear regression, using the r_i and *PSI* statistics as dependent variables and the square of the trait (qualitative or quantitative) as an independent variable. Note that the square of the trait is used because the relationship between any trait and the assortative mating should be quadratic. For example, in the case of size assortative mating, both low and high values of size will contribute to a pattern of size assortative mating. The coefficient of determination (r^2) can be used as an estimate of the efficiency of a particular regression equation for predicting the variation in the dependent variable (Sokal and Rohlf, 1995). This coefficient can be interpreted (when multiplied by 100) as the percentage of variation explained by the regression model.

The simulation procedure was as follows. First, we obtained a base population of 100,000 individuals. Each individual consisted of a value for a quantitative trait, taken from a normal distribution with mean equal to 0 and standard deviation equal to 1, and a state for a related discriminant trait, 0 if the value for the quantitative trait was negative and 1 otherwise. Note that these traits are causally correlated (empirical correlation in the simulated population was 0.88). There were two possible scenarios depending on the *a priori* mechanism of assortative mating. First, if assortative mating was caused by the quantitative trait, then two randomly chosen individuals mate with a probability equal to $1 - (| x - y | \times \rho)$, where x and y are the values for the quantitative trait in each pair of mates, and ρ is the desired *a priori* level of assortative mating (range between 0 and 1). This produced a better chance of mating for those pairs with a small trait difference or a small assortative mating trend (or both). For example, under random mating the mating probability is 1. Alternatively, under high *a priori* assortative mating, only rather small differences in the trait between mates will result in a reasonable probability of mating.

When assortative mating was caused by differences in a qualitative trait (scaled with values 0 and 1), we used a different approach. In this case, four different mating pair combinations are possible: matings between 1/1 (male 1 with female 1), 1/0 (male 1 with female 0), 0/1 (male 0 with female 1), and 0/0 (male 0 with female 0). Under an assortative

mating situation, we would expect the following mating preferences for the above corresponding mating pair combinations: 1, $1 - c$, $1 - c$, and 1. Then, by using the I_{PSI} algorithm, it is possible to obtain the best estimate of the isolation (ρ):

$$\rho = \frac{(1 + 1) - ((1 - c) + (1 - c))}{2 + 2c}.$$

From the above equation it is possible to estimate c:

$$c = \frac{2\rho - 2}{-2 - 2\rho}.$$

The simplest way to simulate assortative mating is to allow matings if they have the same state (1/1 or 0/0), but allowing them to mate only with a probability equal to c if they have different states (1/0 or 0/1).

We obtained 100 mating pairs by each of the two *a priori* mechanisms. Four values ($\rho = 0$, 0.3, 0.6, and 0.9) of *a priori* levels of assortative mating were run. In those simulated pairs, we calculated *a posteriori* the r_i and *PSI* coefficients for every mating pair (note that the *PSI* coefficient for a mating pair stands for the one corresponding to its mate pair combination: 1/1, 1/0, 0/1 or 0/0). We expect that r_i and I_{PSI} will be close to ρ if they work properly. In addition, given that assortative mating is only caused by one trait in this model, the percentage of explained variation in assortative mating (r_i or *PSI*) by the squared trait (quantitative or qualitative) gives us the possibility of quantifying their contribution *a posteriori*, and so indirectly to evaluate which of the variables causes it. If this approach is useful we would expect a lower percentage of explained variation under low levels of *a priori* assortative mating and a higher percentage of explained variation in the opposite situation. Each sample of mating pairs was re-sampled from the simulated populations 10,000 times.

Table 2 shows the estimated *a posteriori* overall assortative mating level (using I_{PSI} and r_i coefficients) and the variance explained (r^2 under a regression approach) by the squared quantitative/qualitative variables on the estimated assortative mating level (r_i or *PSI*). When the *a priori* assortative mating (given by ρ) is caused by a preference on a quantitative trait, r_i is a better predictor of ρ than I_{PSI}, and vice versa when it is based on a qualitative difference. The latter is expected from the estimation properties of I_{PSI} (described in Pérez-Figueroa *et al.*, 2005) given that this index is unbiased in a wide range of situations. However, when these estimators were used under the regression approach to infer the causes of assortative mating, the r_i statistic outperformed the *PSI* coefficient in all circumstances – that is, the former typically showed a better fit regression with the putatively causal variables of the assortative mating (Table 2). This happened even when the assortative mating was causally produced by the qualitative trait and so the I_{PSI} is the best predictor of the assortative mating. The reason for such an apparent contradiction may be that the *PSI* coefficient is not in fact a random variable in the sample, but rather a qualitative index for each mate type combination and so it may produce very poor fit with regression of individual pairs. In addition, the degree of variance explained by the regression model increases linearly with the level of *a priori* assortative mating simulated; the larger the ρ values, the larger the percentage of variance explained by the association. For example, the correlation between ρ and r^2_{qt} was 0.98 when assortative mating was caused by the quantitative trait and 0.93 when

Table 2. Results of comparative analyses for detecting traits contributing to the assortative mating

	A priori quantitative						*A priori* qualitative					
	r_i			*PSI*			r_i			*PSI*		
ρ	$\overline{r}_i$	r^2_{qt}	r^2_{ql}	I_{PSI}	r^2_{qt}	r^2_{ql}	$\overline{r}_i$	r^2_{qt}	r^2_{ql}	I_{PSI}	r^2_{qt}	r^2_{ql}
0.0	0.00	0.037	0.004	0.00	0.010	0.00	0.00	0.037	0.004	−0.01	0.010	0.001
	(0.00)	(0.05)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.05)	(0.01)	(0.00)	(0.01)	(0.01)
0.3	0.32	0.194	0.008	0.21	0.029	0.003	0.19	0.056	0.005	0.30	0.011	0.005
	(0.00)	(0.12)	(0.01)	(0.00)	(0.03)	(0.01)	(0.00)	(0.07)	(0.01)	(0.00)	(0.01)	(0.01)
0.6	0.68	0.657	0.036	0.46	0.100	0.019	0.38	0.122	0.011	0.60	0.011	0.025
	(0.00)	(0.10)	(0.03)	(0.00)	(0.04)	(0.03)	(0.00)	(0.01)	(0.02)	(0.00)	(0.01)	(0.04)
0.9	0.82	0.826	0.060	0.59	0.128	0.052	0.57	0.274	0.025	0.90	0.015	0.156
	(0.00)	(0.08)	(0.04)	(0.00)	(0.05)	(0.06)	(0.00)	(0.13)	(0.03)	(0.00)	(0.02)	(0.18)

Note: The regression approach is applied on two different estimators of assortative mating (r_i and *PSI*), when it is *a priori* caused by quantitative or qualitative traits across different degrees of *a priori* assortative mating (ρ). For each scenario, the table shows the assortative mating estimate in the population (mean r_i and I_{PSI}), and the variance explained by the regression of traits on assortative mating. In the latter case, the ability to infer the causes of assortative mating is estimated by the squared correlation of r_i (or *PSI*) with the squared trait (quantitative = r^2_{qt}; qualitative = r^2_{ql}). Such an *a posteriori* squared correlation (r^2) should be positively related with the true *a priori* degree of assortative mating (ρ). Numbers in parentheses are standard deviations.

it was caused by the qualitative trait. Somewhat weaker correlations were also observed between ρ and r^2_{ql} (0.97 and 0.80, respectively). This indicates that r_i (and to a lesser extend also the *PSI*) can be useful to indirectly infer the factors causing assortative mating, although the former will always show a higher prediction capability (higher r^2).

EXAMPLE OF APPLICATION

In addition, we used published data from Conde-Padín *et al.* (2008) to apply the r_i and *PSI* statistics as exploratory tools to detect which variables contribute the most to the individual variability in assortative mating using multiple regression analysis. We will briefly introduce a real example to illustrate a practical application of the statistic.

In Galician exposed rocky shores, a striking polymorphism of the marine snail *Littorina saxatilis* is found associated with different shore levels and habitats (Johannesson *et al.*, 1993; Quesada *et al.*, 2007; reviewed by Rolán-Alvarez, 2007). On the upper shore, the RB (ridged and banded) ecotype is associated with the barnacle belt, while the SU (smooth and unbanded) ecotype is associated with the mussel belt on the lower shore. Mussels and barnacles overlap at the mid-shore, creating a patchy micro-habitat, where these two pure ecotypes meet and occasionally mate, and a variable percentage of intermediate fertile forms (called hybrids; HY) are observed. These two ecotypes differ for many morphological, behavioural, and even life-history characteristics, mostly due to the existence of disruptive selection acting across the vertical environmental gradient (Rolán-Alvarez *et al.*, 1999; Cruz *et al.*, 2001, 2004b, 2004c; Conde-Padín *et al.*, 2007). In spite of the hybridization at the mid-shore, some partial sexual isolation (70% of the maximum possible on average) contributes to the maintenance of the ecotype differences across the environmental gradient (Johannesson *et al.*, 1995; Rolán-Alvarez *et al.*, 1999, 2004; Cruz *et al.*, 2004a; Quesada *et al.*, 2007). This ecotype assortative mating was indirectly caused by the existing size assortative mating and the mean size difference between ecotypes (see Cruz *et al.*, 2004a; Rolán-Alvarez *et al.*, 2004; Rolán-Alvarez, 2007; Conde-Padín *et al.*, 2007). Additionally, it has been proposed that approximately half of the ecotype assortative mating can be achieved by snail micro-aggregation, perhaps caused by active search for refuges (Kostylev *et al.*, 1997; Erlandsson *et al.*, 1999), or by the existence of different preferences in RB and SU ecotypes for mussel and barnacle micro-patches at the mid-shore (Otero-Schmitt *et al.*, 1997; Carballo *et al.*, 2005). In summary, it could be advanced that size assortative mating in this population should be influenced to some extent by the particular micro-habitat conditions.

The mating pairs were captured during May and June 2006 in Silleiro and Centinela in the *L. saxatilis* hybrid zone of Galicia (NW of Spain), during low tide directly on the rocky shore. The specimens from mating pairs were classified as belonging to a particular ecotype (RB, SU or HY) and sex. Shell size was estimated as the distance between the shell apex and the shell base (for further details, see Conde-Padín *et al.*, 2008). With these data we could obtain the individual contribution of each pair to the assortative mating by the r_i and the *PSI* statistics (see above) to be used as dependent variables in the exploratory multiple regression analysis.

A plastic circle (20 cm in diameter) placed over each mating pair was used to obtain some environmental/demographic variables inside those micro-areas. We took a digital photograph of the micro-area and in the laboratory we divided the photograph into 16 large quadrates, each one divided into 16 small quadrates (256 small squares in total). Thus we obtained from such images the *relative abundance of mussels* (number of small quadrates covered by mussels), and the *relative abundance of barnacles* within the micro-area. The *aggregation of mussels* and the *aggregation of barnacles* were obtained by dividing the mean

number of small quadrates by their variance across large quadrates (following Taylor, 1984; Margalef, 1991). We also obtained two linear profiles (*horizontal* and *vertical profiles* in relation to the shoreline) of the surface of the micro-area following Kostylev *et al.* (1997). We used the length of these profiles, and the mean, as estimates of the surface rugosity (Conde-Padín *et al.*, 2008). A more detailed analysis of morphological variables is presented elsewhere (Conde-Padín *et al.*, 2008). Some of those environmental variables might partially contribute to the existing size assortative mating in these populations, and so they were used as independent variables using a multiple regression approach.

EXPLORATORY REGRESSION ANALYSIS IN *L. SAXATILIS*

Mating pairs and corresponding environmental variables were used to explore the environmental factors contributing to or affecting assortative mating in the wild. The r_i and *PSI* values of pairs were used as dependent variables in a multiple regression analysis, using the square of the associated environmental variables (*relative abundance of mussels* and *barnacles*, *aggregation of mussels* and *barnacles*, *horizontal*, *vertical*, and *mean profiles*) as independent ones. Thus we used seven environmental variables under a step-wise multiple regression approach, with forward ($P < 0.05$) and backward ($P < 0.1$) criteria to determine the significant variables contributing to the size assortative mating. All these calculations were done with the SPSS/PC software version 14.0.

When using r_i under multiple regression (Table 3), one environmental variable explained part of the variation in assortative mating in Silleiro (*horizontal profile*^{2}) and Centinela (*mean profile*^{2}). Interestingly, the regression was not significant when using the *PSI* coefficient (Table 3). The relationship was positive in Centinela (suggesting that both low and high rugosity contribute to assortative mating) and negative in Silleiro (suggesting that intermediate rugosity contributes most to assortative mating). In summary, one environmental factor (substrate rugosity) could also contribute to the pattern of assortative mating. However, the fact that this relationship is different in sign between localities perhaps suggests that it has no causative role in explaining the pattern of assortative mating in the wild.

Table 3. Results of step-wise linear regression of the environmental variables explaining the individual contribution to the assortative mating (r_i) in Silleiro and Centinela

Dependent variable	Locality	Independent environmental variables	r^2	Coefficient of partial regression
r_i	Silleiro	*horizontal profile*^{2}	0.124	−0.353*
	Centinela	*mean profile*^{2}	0.196	0.442***
PSI	Silleiro	—	—	—
	Centinela	—	—	—

Note: The variance explained by the only variable introduced (r^2) in the model and its partial regression coefficients are shown. The regression of environmental variables on *PSI* coefficients was not significant.

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

DISCUSSION

A modified version of the Pearson correlation coefficient is described here to study the causes of assortative (or disassortative) mating in any quantitative trait. The statistic shows relatively robust sampling properties under low sample size or for non-normal variables, and the development of its sampling standard deviation will allow it to be used under parametric hypothesis testing or for generating sampling confidence intervals efficiently. In addition, this index allows exploration of the causes of assortative mating directly in the wild for a particular species using a regression approach. The r_i statistic was compared with another statistic [*PSI* (from Rolán-Alvarez and Caballero, 2000)] to allow this kind of exploration.

The r_i method outperformed the *PSI* in all situations when used to infer the causes of assortative mating under a regression approach, which suggests that the *PSI* statistic should be avoided for such exploratory analysis unless there is no alternative available. The r_i showed a high predictive ability to detect the causal relationship of variables with assortative mating. Although the simulation conditions were simple, they show that the percentage of variance explained by the regression is directly related to the *a priori* degree of assortative mating, allowing this methodology to be used in more complex situations. For example, this method has also been employed to determine the morphological traits (from male or females) that are most relevant to mate choice in the marine gastropod *Littorina saxatilis* (Conde-Padín *et al.*, 2008), which represents a case of sympatric ecological speciation (Rolán-Alvarez, 2007; Quesada *et al.*, 2007). In this case, the dependent variable was a canonical discriminant score used to distinguish between these ecotypes, while the independent variables were shell size and shape traits. The results suggest that male (not female) size is the main determinant of the degree of ecotype assortative mating in this model system. These results were independently corroborated by a laboratory experiment in which males could choose among different types of females [with different ecotypes and sizes (Conde-Padín *et al.*, 2008; see also Cruz *et al.*, 2004a; Rolán-Alvarez *et al.*, 2004)]. The results from our simulations corroborate the previous use of this approach.

In addition, in a re-analysis of the above data we confirmed that one environmental variable could explain part of the variation in assortative mating (only when using r_i). This corroborates the results of simulations: the r_i statistic was able to detect a significant relationship in the example, suggesting again that it has greater statistical power than the *PSI* statistic under the regression approach. There are many cases in which the use of this methodology could facilitate the understanding of mating behaviour in the wild (e.g. Johannesson *et al.*, 1995; Erlandsson and Rolán-Alvarez, 1998; Masumoto, 1999; Shine *et al.*, 2001; Silventoinen *et al.*, 2003; Hollander *et al.*, 2005).

The applicability of the method rests on the assumption that among a group of biological traits, the one that shows the strongest correlation with the studied variable (sexual isolation or size assortative mating) is most likely causally related to the studied variable. It has been argued, however, that any correlation between two variables does not guarantee any causal relationship between them (Sokal and Rohlf, 1995), as a third variable, not included in the study, could be responsible. Such problems with correlation or regression analysis are well-known, although they have not limited the applicability of the regression model to biology, ecology or evolution (Brodie *et al.*, 1995). Some caution, however, is needed. For example, the particular variables included in the model with minor contributions ($r^2 \times 100 < 5$–10%) are not necessarily safely selected by the regression algorithm.

General progress has been made in the last few years studying the genetic causes of post-zygotic isolation (reviewed in Coyne and Orr, 2004). However, little progress has been made in understanding the genetic effects of pre-zygotic isolation, as with sexual isolation (Coyne and Orr, 2004; but see Coyne, 1996; Pugh and Ritchie, 1996; Carracedo *et al.*, 1998). This could be caused in part by the difficulty of understanding the biological mechanisms contributing to sexual isolation. The use of the r_i statistic could facilitate future research on this mechanism of sexual isolation, and it has the advantage of being applicable in any case in which two or more species are being studied (see, for example, Coyne *et al.*, 2005; Giokas *et al.*, 2006). In addition, this statistic can be also used with many biological variables – anatomical, physiological or even behavioural ones – used to explore linear and non-linear relationships, the interactions between independent variables, or measurements previously corrected for spurious effects caused by environmental variables (see Lande and Arnold, 1983; Rausher, 1992; van Tienderen and De Jong, 1994; Brodie *et al.*, 1995).

ACKNOWLEDGEMENTS

We thank A. Caballero, J.J. Pasantes, and M. Santos for useful suggestions on preliminary versions of this manuscript, as well as the following institutions for general funding: European Union (code EVK3-CT-2001-00048), Ministerio de Educación y Ciencia (code CGL2008-00135/BOS), Xunta de Galicia (code PGIDT02PXIC30101PM; PGIDT06PXIB310247PR), and University of Vigo. P. C.-P. thanks the Ministerio de Educación y Ciencia from Spain for a research grant. A.P-F. is currently funded by an Ángeles Alvariño research fellowship from Xunta de Galicia (Spain).

REFERENCES

Andersson, M. 1994. *Sexual Selection*. Princeton, NJ: Princeton University Press.

Arnold, S.J. and Wade, M.J. 1984a. On the measurement of natural and sexual selection: theory. *Evolution*, **38**: 709–719.

Arnold, S.J. and Wade, M.J. 1984b. On the measurement of natural and sexual selection: applications. *Evolution*, **38**: 709–719.

Brodie, E.D., Moore, A.J. and Janzen, F.J. 1995. Visualizing and quantifying natural selection. *Trends Ecol. Evol.*, **10**: 313–318.

Carballo, M., Caballero, A. and Rolán-Alvarez, E. 2005. Habitat-dependent ecotype micro-distribution at the mid shore in natural populations of *Littorina saxatilis*. *Hydrobiologia*, **548**: 307–311.

Carracedo, M.C., Suarez, B., Asenjo, A. and Casares, P. 1998. Genetics of hybridization between *Drosophila simulans* females and *D. melanogaster* males. *Heredity*, **80**: 17–24.

Conde-Padín, P., Carvajal-Rodríguez, A., Carballo, M., Caballero, A. and Rolán-Alvarez, E. 2007. Genetic variation for shell traits in a direct-developing marine snail involved in a putative sympatric ecological speciation process. *Evol. Ecol.*, **21**: 635–650.

Conde-Padín, P., Cruz, R., Hollander, J. and Rolán-Alvarez, E. 2008. Revealing the mechanisms of sexual isolation in a case of sympatric and parallel ecological divergence. *Biol. J. Linn. Soc.*, **94**: 513–526.

Coyne, J.A. 1996. Genetics of sexual isolation in male hybrids of *Drosophila simulans* and *D. mauritiana*. *Genet. Res.*, **68**: 211–220.

Coyne, J.A. and Orr, H.A. 2004. *Speciation*. Sunderland, MA: Sinauer Associates.

Coyne, J.A., Elwyn, S. and Rolán-Alvarez, E. 2005. Impact of experimental design on *Drosophila* sexual isolation studies: direct effects and comparison to field hybridization data. *Evolution*, **59**: 2588–2601.

Crespi, B.J. 1989. Causes of assortative mating in arthropods. *Anim. Behav.*, **38**: 980–1000.

Cruz, R., Rolán-Alvarez, E. and García, C. 2001. Sexual selection on phenotypic traits in a hybrid zone of *Littorina saxatilis* (Olivi). *J. Evol. Biol.*, **14**: 773–785.

Cruz, R., Carballo, M., Conde-Padín, P. and Rolán-Alvarez, E. 2004a. Testing alternative models for sexual isolation in natural populations of *Littorina saxatilis*: indirect support for by-product ecological speciation? *J. Evol. Biol.*, **17**: 288–293.

Cruz, R., Vilas, C., Mosquera, J. and García, C. 2004b. Relative contribution of dispersal and natural selection to the maintenance of a hybrid zone in *Littorina. Evolution*, **58**: 2734–2746.

Cruz, R., Vilas, C., Mosquera, J. and García, C. 2004c. The close relationship between estimated divergent selection and observed differentiation supports the selective origin of a marine snail hybrid zone. *J. Evol. Biol.*, **17**: 1221–1229.

Delestrade, A. 2000. Sexual size dimorphism and positive assortative mating in Alpine choughs (*Pyrrhocorax graculus*). *Auk*, **118**: 553–556.

Erlandsson, J. and Rolán-Alvarez, E. 1998. Sexual selection and assortative mating by size and their roles in the maintenance of a polymorphism in Swedish *Littorina saxatilis* populations. *Hydrobiologia*, **378**: 59–69.

Erlandsson, J., Kostylev, V. and Rolán-Alvarez, E. 1999. Mate search and aggregation behaviour in the Galician hybrid zone of *Littorina saxatilis. J. Evol. Biol.*, **12**: 891–896.

Forero, M.G., Tella, J.L., Donazar, J.A., Blanco, G., Bertellotti, M. and Ceballos, O. 2001. Phenotypic assortative mating and within-pair sexual dimorphism and its influence on breeding success and offspring quality in Magellanic penguins. *Can. J. Zool.*, **79**: 1414–1422.

Gavrilets, S. 2004. *Fitness Landscapes and the Origin of Species*. Princeton, NJ: Princeton University Press.

Giokas, S., Mylonas, M. and Rolán-Alvarez, E. 2006. Disassociation between weak sexual isolation and genetic divergence in a hermaphroditic land snail and implications about chirality. *J. Evol. Biol.*, **19**: 1631–1640.

Hollander J., Lindegarth M. and Johannesson, K. 2005. Local adaptation but not geographic separation promotes assortative mating in a snail – support for ecological speciation. *Anim. Behav.*, **5**: 1209–1219.

Johannesson, K., Johannesson, B. and Rolán-Alvarez, E. 1993. Morphological differentiation and genetic cohesiveness over a microenvironmental gradient in the marine snail *Littorina saxatilis. Evolution*, **47**: 1770–1787.

Johannesson, K., Rolán-Alvarez, E. and Ekendahl, A. 1995. Incipient reproductive isolation between two sympatric morphs of the intertidal snail *Littorina saxatilis. Evolution*, **49**: 1180–1190.

Jonson, L.J. 1999. Size assortative mating in the marine snail *Littorina neglecta. J. Mar. Biol. Assoc. UK*, **79**: 1131–1132.

Kirkpatrick, M. and Ravigné, V. 2002. Speciation by natural selection and sexual selection: models and experiments. *Am. Nat.*, **139**: 22–35.

Kostylev, V., Erlandsson, J. and Johannesson, K. 1997. Microdistribution of the polymorphic snail *Littorina saxatilis* (Olivi) in a patchy rocky shore habitat. *Ophelia*, **47**: 1–12.

Lande, R. and Arnold, S.J. 1983. The measurement of selection on correlated characters. *Evolution*, **37**: 1210–1226.

Lewontin, R., Kirk, D. and Crow, J. 1968. Selective mating, assortative mating, and inbreeding: definitions and implications. *Eugenics Quart.*, **15**: 141–143.

Margalef, R. 1991. *Ecología*. Barcelona: Ediciones Omega.

Masello, J.F. and Quillfeldt, P. 2003. Body size, body condition and ornamental feathers of Burrowing Parrots: variation between years and sexes, assortative mating and influences on breeding success. *Emu*, **103**: 149–161.

Masumoto, T. 1999. Size assortative mating and reproductive success of the funnel-web spider, *Agelena limbata* (Aracneae; Agelenidae). *J. Insect Behav.*, **12**: 353–361.

McKinnon, J.S., Mori, S., Blackman, B.K., David, L., Kingsley, D.M., Jamieson, L. *et al.* 2004. Evidence for ecology's role in speciation. *Nature*, **429**: 294–298.

Nadarajah, S. 2006. Exact and approximate distributions for the product of inverted Dirichlet components. *Statistical Papers*, **47**: 551–568.

Nagel, L. and Schluter, D. 1998. Body size, natural selection, and speciation in sticklebacks. *Evolution*, **52**: 209–218.

Otero-Schmitt, J., Cruz, R., García, C. and Rolán-Alvarez, E. 1997. Feeding strategy and habitat choice in *Littorina saxatilis* (Gastropoda: Prosobranchia) and their role in the origin and maintenance of a sympatric polymorphism. *Ophelia*, **46**: 205–216.

Pearson, K. 1894. Contributions to the mathematical theory of evolution. *Phil. Trans. R. Soc. Lond. A*, **185**: 71–110.

Pérez-Figueroa, A., Caballero, A. and Rolán-Alvarez, E. 2005. Comparing the estimation properties of different statistics for measuring sexual isolation from mating frequencies. *Biol. J. Linn. Soc.*, **85**: 307–318.

Pugh, A.R.G. and Ritchie, M.G. 1996. Polygenic control of a mating signal in *Drosophila. Heredity*, **77**: 378–382.

Quesada, H., Posada, D., Morán, P., Caballero, A. and Rolán-Alvarez, E. 2007. Phylogenetic evidence for multiple sympatric ecological diversification in a marine snail. *Evolution*, **61**: 1600–1612.

Rausher, M.D. 1992. The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. *Evolution*, **46**: 616–626.

Rolan-Alvarez, E. 2007. Sympatric speciation as a by-product of ecological adaptation in the Galician *Littorina saxatilis* hybrid zone. *J. Mollusc. Stud.*, **73**: 1–10.

Rolan-Alvarez, E. and Caballero, A. 2000. Estimating sexual selection and sexual isolation effects from mating frequencies. *Evolution*, **54**: 30–36.

Rolán-Alvarez, E., Erlandsson, J., Johannesson, K. and Cruz, R. 1999. Mechanisms of incomplete prezygotic reproductive isolation in an intertidal snail: testing behavioural models in wild populations. *J. Evol. Biol.*, **12**: 879–890.

Rolán-Alvarez, E., Carballo, M., Galindo, J., Morán, P., Fernández, B., Caballero, A. *et al.* 2004. Non-allopatric origin of local reproductive barriers between two snail ecotypes. *Molec. Ecol.*, **13**: 3415–3424.

Shine, R., O'Connor, D., Lemaster, M.P. and Mason, R.T. 2001. Pick on someone your size: ontogenetic shifts in mate choice by male garter snakes result in size assortative mating. *Anim. Behav.*, **61**: 1133–1141.

Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R.J. and Rose, R.J. 2003. Assortative mating by body height and BMI: Finnish twins and their spouses. *Am. J. Human Biol.*, **15**: 620–627.

Sokal, R.R. and Rohlf, F.J. 1995. *Biometry*, 3rd edn. New York: Freeman.

Staub, R. and Ribi, G. 1995. Size-assortative mating in a natural population of *Viviparus ater* (Gastropoda: Prosobranchia) in Lake Zürich, Switzerland. *J. Mollusc. Stud.*, **61**: 237–247.

Taylor, L.R. 1984. Assessing and interpreting the spatial distributions of insect populations. *Annu. Rev. Entomol.*, **29**: 321–357.

Turelli, M., Barton, N.H. and Coyne, J.A. 2001. Theory and speciation. *Trends Ecol. Evol.*, **15**: 330–343.

Van Tienderen, P.H. and de Jong, G. 1994. A general model of the relation between phenotypic selection and genetic response. *J. Evol. Biol.*, **7**: 1–12.