Methods in Ecology and Evolution British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 538–548 doi: 10.1111/2041-210X.12498

The measurement of selection when detection is imperfect: How good are naïve methods?

John Waller* and Erik I. Svensson

Evolutionary Ecology Unit, Department of Biology, Lund University, SE-223 62 Lund, Sweden

Summary

1. The life spans of animals can be measured in natural populations by uniquely marking individuals and then releasing them into the field. Selection on survival (a component of fitness) can subsequently be quantified by regressing the life spans of these marked individuals on their trait values. However, marked individuals are not always seen on every subsequent catching occasion, and for this reason, imperfect detection is considered a problem when estimating survival selection in natural populations.

2. Capture–mark–recapture methods have been advocated as a powerful means to correct for imperfect detection. Here, we use simulated and field data sets to evaluate the effect of assuming perfect detection ('naïve methods'), when detection is really imperfect. We compared the performance of the naïve methods with methods correcting for imperfect detection (mark–recapture methods, or MR).

3. Although the effects of trait-dependent recapture probability are mitigated when recapture probability is high, mark–recapture methods still provide the safest choice when recapture probability might be trait-dependent. In our simulations, mark–recapture methods had a power advantage over naïve methods, but all methods lost statistical power at low recapture probabilities.

4. The main advantage of mark–recapture methods over naïve methods is the ability to control for hidden trait-dependent recapture probability, as it is often hard to tell *a priori* if trait dependence is an issue in a particular study. However, when trait-dependent recapture probability is weak, naïve methods and mark–recapture methods perform similarly as long as recapture rates do not become too low, and the main problem of survival selection studies is still low statistical power. We provide a R package (EasyMARK) alongside with this paper to facilitate future integration between MR methods and classical selection studies. EasyMARK provides the opportunity to convert the regression coefficients from MR-approaches in to classical standardized selection gradients.

Key-words: capture–mark–recapture, directional selection, mark, natural selection, selection gradients, simulation, stabilizing selection

Introduction

The detection and quantification of natural and sexual selection in the field has a long history in evolutionary biology. A considerable milestone in this field was the highly influential publication by Russel Lande and Steve Arnold in the early nineteen eighties (Lande & Arnold 1983), and interest was further stimulated by John Endler's classical volume about natural selection in the wild a few years later in 1986 (Endler 1986). Since these early influential contributions, thousands of selection estimates have been published from plant and animal populations, and our knowledge about the strength, mode and occurrence of sexual and natural selection has increased considerably (Kingsolver *et al.* 2001, 2012). Several more recent meta-analyses and statistical developments have since been published in this very active research field. Recent studies deal with important methodological challenges such as critical sample sizes and their possible effects on selection gradients (Kingsolver *et al.* 2001; Knapczyk & Conner 2007), whether observed magnitudes of selection can be considered weak or strong (Conner 2001; Kingsolver *et al.* 2001; Hereford, Hansen & Houle 2004), the role of temporal (Siepielski, DiBattista & Carlson 2009; Siepielski *et al.* 2011) and spatial variation in selection (Gosden & Svensson 2008; Calsbeek *et al.* 2012; Siepielski *et al.* 2013) and how measurement error of both phenotypic traits and fitness measures can effect selection estimates (Kingsolver & Diamond 2011; Morrissey & Hadfield 2012). Other important issues are the inferences that could be made from a purely correlative approach to studying selection (Mitchell-Olds & Shaw 1987) and the ecological causes of selection and the need to identify selective agents (Wade & Kalisz 1990; Svensson & Sinervo 2000; MacColl 2011).

Yet one issue in survival selection studies that has largely been ignored is imperfect detection. Clobert (1995) pointed out that evolutionary ecologists tended to ignore imperfect detection and assume any effect of unobserved individuals was negligible, using the last date seen in the field instead. Later,

*Correspondence author. E-mail: john.waller@biol.lu.se

Gimenez *et al.* (2008) showed, using a field data set, that even the mode of selection (directional or stabilizing) might be wrongly inferred when the recapture probability is less than one. Although imperfect detection has therefore been recognized as a problem, the severity of this problem has never been quantified using systematic simulations or in field studies, and it is very seldom discussed. This raises some important questions: Should past studies be reanalysed? How low does recapture probability have to be before it becomes a problem? In which direction would imperfect detection effect the estimates of selection – if any? How good are currently available statistical techniques at correcting such errors?

One statistical solution to imperfect detection is capture–mark–recapture methods (MR). To correct for imperfect detection, MR uses the extra information within individual capture histories to estimate recapture probability. For example, if an individual is seen on day 1, and then not seen on day 2, but then is seen later on day 3, (a capture history of '101') a researcher knows that individual was definitely alive on day 2, but was simply not detected. Describing MR methods in detail and their many implementations is outside the scope of this paper. Instead, we point readers to the key literature in the field (Cormack 1964; Seber 1965; Jolly 1982; Lebreton *et al.* 1992; Seber & Schwarz 2002; Williams, Nichols & Conroy 2002).

Here, our primary aim is instead to quantify the risk of flawed inference by comparing the estimates from two 'naïve' regression methods versus using the MR-approach (the 'informed method'). With the two naïve methods, we use the minimum life span of individuals in the field as the fitness measure. This is a common fitness measure used by evolutionary ecologists. We evaluate these different approaches on both real and simulated data sets. We primarily focus on stabilizing selection, as this mode of selection is of principal interest on both micro- and macro-evolutionary time-scales (Haller & Hendry 2013). Stabilizing selection is also notoriously difficult to estimate, due to low statistical power in most field studies with limited sample sizes (Kingsolver *et al.* 2001) and the fact that such stabilizing selection tends to 'erase its traces' by removing variation around the fitness optimum (Haller & Hendry 2013).

Our main finding is that all methods lose statistical power with decreasing recapture probability, and trait-dependent recapture probability becomes an issue mainly when recapture probabilities are trait-dependent and low. Note that we use 'trait dependence' in the broad sense, such as when recapture probabilities are heterogeneous is some way. Such heterogeneous recapture probabilities are not restricted to quantitative phenotypic traits, but could also include time or discrete phenotypic categories, such as sex effects. Nevertheless, it is often hard to tell *a priori* if trait dependence or low recapture probability is an issue within a data set, making the use of MR methods the safest choice even in scenarios when it may only have low impact on the conclusions. We provide R scripts and a new R package EasyMARK (Supporting: EasyMARK) along with this article to facilitate research in this area.

Methods

SIMULATING INDIVIDUAL CAPTURE HISTORIES

We developed a simple mark–recapture simulator, which produced individual capture histories under a range of different conditions. The simulator was written in the R programming language v3.0.2 (R Development Core Team 2008). The situation we model can be thought of as a short-lived organism (e.g. an insect), with no or negligible senescence. For more information, see Supporting Information: Simulating Individual Capture Histories (Fig. S1). Simulations were run in parallel using the R package foreach v1.4.1 (Revolution Analytics & Weston 2013) on a Windows 7, Intel Core i7-3930K CPU @ 3·20 GHz and 32 GB of RAM.

Two scenarios were simulated as follows: (i) linear trait-dependent recapture probability and (ii) negative quadratic ('stabilizing') selection (Table 1). These data sets will be referred to as trait-dependent recapture probability and stabilizing selection, respectively. These two scenarios represent two opposite cases:

(1) trait-dependent simulations: only trait-dependent recapture probability and no actual survival selection (Table 1).

(2) stabilizing selection simulations: only survival selection and imperfect detection, but no trait dependence on recapture probability (Table 1).

Our trait-dependent simulations had the following form:

$$\mathrm{logit}(p) = I_p + b_p z \qquad \text{eqn 1}$$

Here, p is survival probability, I_p is an intercept term, z is a normally distributed vector of trait values (mean = 0, var = 1), and b_p is the linear trait dependence coefficient. The linear trait dependence coefficient (b_p) was set in our simulator at 0·05, 0·15 and 0·25, which we refer to as 'low', 'medium' and 'high' trait dependence, respectively (but see Figs S6–S7, where we also tested with higher values). Unfortunately, it is generally unknown what a typical or reasonable value in practice for this coefficient would be. The intercept term (I_p) controlled the mean recapture probability. Mean recapture probability was varied using the intercept from 0·1 to 0·9 at a 0·1 interval, resulting in 9 mean recapture probability values (i.e. using values $I_p = -2{\cdot}2, -1{\cdot}39, -0{\cdot}85, -0{\cdot}41, 0, 0{\cdot}41, 0{\cdot}85, 1{\cdot}39, 2{\cdot}2$). The normally distributed continuous trait (z) had mean zero and variance of one. Survival probability (Φ) was set constant at 0·70 and was not dependent on the trait (z). These recapture probabilities (p) and survival probabilities (Φ) for each individual were then used to create individual capture histories.

Our stabilizing selection simulations had the following form:

$$\mathrm{logit}(\Phi) = I_\Phi + b_\Phi z + g_\Phi z^2 \qquad \text{eqn 2}$$

Here, Φ is the survival probability, I_Φ is an intercept term, z is a normally distributed vector of standardized phenotypic trait values (mean = 0, var = 1), b_Φ is the linear coefficient, and g_Φ is the quadratic coefficient. When simulating stabilizing selection, the linear coefficient (b_Φ) was set to zero, which is equivalent to assuming that the popula-

Table 1. Summary of the simulations and assumptions of the two scenarios

Simulation	Selection of Survival	Recapture Probability
Stabilizing selection	Yes	Not dependent on trait
Trait-dependent recap. prob	No	Dependent on trait

tion sits on its adaptive peak and is no longer subject to directional selection. The quadratic coefficient (g_Φ) was set at values of −0·03, −0·07 and −0·11. This corresponds to transformed values of gamma (γ) of −0·06, −0·12, −0·18 in terms of classical variance-standardized quadratic selection gradients (Lande & Arnold 1983; Janzen & Stern 1998; Stinchcombe *et al.* 2008; Morrissey & Sakrejda 2013, M. Morrissey, pers. comm.). See Supporting Information: Comparing Estimates.

Our simulated selection strengths are similar in magnitude to those that have been estimated previously in natural populations and which were included in meta-analyses. These estimates should therefore presumably reflect the true average strength of selection in natural populations (Hoekstra *et al.* 2001; Kingsolver *et al.* 2001). An intercept (I_Φ) of one was used in survival simulations, resulting in a mean life span of about 3 days (occasions), maximum life span of around 20 days, and an average daily survival probability of approximately 0·70 (Fig. S1). The recapture probability (p) was varied from 0·1 to 0·9 at a 0·1 interval, resulting in nine recapture probability values, which were constant and not dependent on the trait (z).

Three sample sizes were used for both data sets: 100, 250, and 500 individuals. A 1000 runs were made for each combination of recapture probability, sample size, and selection or trait dependence strength.

Minimum life span was calculated as the time elapsed from marking to the last day detected. For example, individuals with capture histories of '1010100', '1111100' and '1000100' would all have minimum life spans of 5 days. The full output and associated R scripts will be uploaded to DRYAD (http://datadryad.org/). For more information, see Supporting: Simulating Individual Capture Histories. Our simulated life spans resulted in a non-normal and highly skewed survival distribution (Fig. S1), that was very similar to the distribution of estimated life spans from the field data set. An implementation of this simulator is also available in our R package EasyMARK. See Supporting: EasyMARK.

FITTING THE MODELS TO THE SIMULATED DATA

All analyses were conducted in R. We fitted two general linear model types from the R base and MASS packages with different assumptions about the underlying distribution of errors (normal and negative binomial) and a mark–recapture model from the R package mra v2·13 (McDonald 2012; Table 2). Hereafter, these three approaches will be referred to as LA ('Lande-Arnold'), NB (negative binomial), and MR (mark–recapture). Note that we use the term 'Lande-Arnold method' here to mean a regression model with a Gaussian error structure. The LA- and NB-approaches can both be seen as naïve linear methods which do not take recapture probability into account and which only use minimum life span as the fitness measure. In contrast, the MR-method is a more informed method, which corrects for recapture probability, and the fact that individuals that are still alive might not be observed every time they are looked for by the investigator in the field. A general goal in this study is to compare these two naïve methods with the MR-method, and to understand how robust the former are with respect to deviations from the assumptions of perfect detection, as past survival selection studies have often assumed perfect detection.

There are very close relationships between per-interval logistic-scale regression coefficients of survival, and log-scale regression coefficients, such as those generated by our NB analysis, and selection gradients acting through longevity (Morrissey & Sakrejda 2013; M.B. Morrissey & I.B.J. Goudie, unpubl. data; M. Morrissey, pers. comm.; See Supporting information for our implementation). We applied the formulae given in Morrissey and Goudie equations 7 and 8 to the logistic and log-scale regression coefficients from our MR and NB analysis to render them directly comparable to the selection gradients estimated by LA.

Table 2. Summary of methods and their abbreviations in this study

Method	Full Name	Fitness Measure	Recapture Probability
LA	Lande-Arnold	Minimum Life span	Ignores
NB	Negative Binomial	Minimum Life span	Ignores
MR	Mark–Recapture	Survival Probability	Corrects

For LA, but not for NB or MR, the fitness response variable of each individual was divided by the sample mean fitness (Lande & Arnold 1983). In our case, this meant that the minimum life span of each individual was divided by the mean life span of the sample. Additionally, for LA, but not for NB or MR, we multiplied the resulting quadratic coefficient by two (Stinchcombe *et al.* 2008). Furthermore, for NB and MR, but not for LA, we needed to transform the estimates numerically to the data scale (Morrissey & Sakrejda 2013; M.B. Morrissey & I.B.J. Goudie, unpublished manuscript; M. Morrissey, pers. comm.). We will use beta (β) and gamma (γ) to denote estimates that represent selection gradients interpretable within the classical quantitative genetics framework. We address these transformation issues more in Supporting information: Comparing Estimates.

Mark–recapture was implemented with the R package mra (McDonald 2012). In addition to the assumptions of standard MR models (Cormack 1964; Seber 1965; Jolly 1982; Lebreton *et al.* 1992), we did not assume any time effects in our mark–recapture model. We direct readers to mra's documentation (McDonald 2012) and the source literature (Cormack 1964; Seber 1965; Jolly 1982; Lebreton *et al.* 1992) for more details. We chose the negative binomial over a Poisson distribution because we observed zero-inflation and over-dispersion when attempting to a fit a Poisson distribution (Fig. S1; Hoef & Boveng 2007).

ANALYSIS OF THE SIMULATED DATA

With the simulated data sets, we were interested in which method (LA, NB, MR) would perform better in our two scenarios (i) inferring correctly that there is no survival selection in the case of the trait-dependent recapture data sets and (ii) inferring correctly stabilizing selection in the case where there was no trait dependence on recapture probability, but only stabilizing selection on survival probability.

With the trait-dependent simulations, for LA and NB we fitted two model types: the linear model ($w = I_w + b_w z$) and intercept only model ($w = I_w$), the latter assuming no fitness dependence on the trait (i.e. trait neutrality). Here relative longevity (w) is minimum life span divided by the mean life span of the sampled population. We counted a run as failed for LA and NB if the linear model ($w = I_w + b_w z$) was the top performing model by ΔAIC value >2. For MR, we fitted four model types.

Two of these models included b_Φ terms:

$$\mathrm{logit}(p) = I_p; \mathrm{logit}(\Phi) = I_\Phi + b_\Phi z \qquad \text{eqn 3}$$

$$\mathrm{logit}(p) = I_p + b_p z; \mathrm{logit}(\Phi) = I_\Phi + b_\Phi z \qquad \text{eqn 4}$$

Two models did not include the b_Φ terms:

$$\mathrm{logit}(p) = I_p + b_p z; \mathrm{logit}(\Phi) = I_\Phi \qquad \text{eqn 5}$$

$$\text{logit}(p) = I_p; \text{logit}(\Phi) = I_\Phi \qquad \text{eqn 6}$$

For MR, we counted a trial as failed if a model with a linear survival term (b_Φ) was the top performing model by at least two ΔAIC greater than a model without that term.

For the stabilizing selection simulations, for LA and NB we compared nested models in ordered complexity: Intercept only ($w = I_w$), intercept plus linear term ($w = I_w + b_w z$), and intercept and linear plus quadratic term ($w = I_w + b_w z + g_w z^2$). We considered the run successful, for LA and NB, if the model with the g_w term was the best model by at least two ΔAIC. For MR, we compared three models:

$$\text{logit}(p) = I_p; \text{logit}(\Phi) = I_\Phi + b_\Phi z + g_\Phi z^2 \qquad \text{eqn 7}$$

$$\text{logit}(p) = I_p; \text{logit}(\Phi) = I_\Phi + b_\Phi z \qquad \text{eqn 8}$$

$$\text{logit}(p) = I_p; \text{logit}(\Phi) = I_\Phi \qquad \text{eqn 9}$$

Similarly, we considered a run successful, for MR, if the model with the g_Φ term was the best model by at least two ΔAIC than the next best model.

Transformed estimate values from each method (LA, NB, MR) were all compared at each recapture probability and trait dependence or selection strength. For the trait-dependent recapture probability simulations, we looked at estimates of linear selection (β), and for the stabilizing selection simulations, we looked at estimates of gamma (γ).

QUANTIFYING SURVIVAL SELECTION ON *CALOPTERYX* DAMSELFLIES IN THE FIELD

We also evaluated the three approaches on a field data set of marked individuals of damselflies (Insecta: Odonata). This field data set comes from an individually marked population of the banded demoiselle (*Calopteryx splendens*) in the southern part of the Swedish province Skåne at Sövdemölla gård near Sövdesjön. This study population is located within the same river system (Klingavälsån) as previous work that has been performed on *C. splendens* further upstream (Svensson, Eroukhmanoff & Friberg 2006; Svensson & Friberg 2007; Wellenreuther, Vercken & Svensson 2010; Wellenreuther, Larson & Svensson 2012). All data in this study comes from the field season of 2012 (June and July). From this data set, we only analysed male *C. splendens* individuals, which at this site had a recapture probability of only around 10%, and a daily mean survival probability of around 80% (see Supporting information: Field Data set, Fig. S2).

SELECTION ANALYSIS OF THE FIELD DATA

The *C. splendens* field data set was analysed in a similar manner as the simulated data. We primarily focus on two traits: total body length (tbl) and front wing length (fwl), although we also followed up this analysis by estimating selection for a total of 11 phenotypic traits in these damselflies. We fitted the methods described above (LA, NB, MR) to the individual recapture histories, while also accounting for trait-dependent recapture probability on tbl and fwl (Table 3). See Supporting information: Analysis of Field Data set.

Results

SIMULATED DATA SETS

With the stabilizing selection data set, all methods (LA, NB, MR) lost statistical power to infer the correct form of selection with decreasing recapture probability (Fig. 1). Under conditions of weak selection and low sample sizes, no method performed well in detecting stabilizing selection, even when it was present (Fig. 1). In contrast, under strong selection and at high recapture probabilities, all methods performed similarly (Fig. 1), although MR generally performed better at inferring stabilizing selection across most recapture probabilities (Fig. 1).

Estimates of stabilizing selection between the methods (LA, NB, MR) were largely in agreement (Fig. 2). However, the estimates from NB and LA became stronger (more negative) with decreasing recapture probability (Fig. 2). This is likely reflects that minimum observed longevity divided by true longevity varies with true longevity (M. Morrissey, pers. comm.). In addition, MR estimates of stabilizing selection tended to have lower variance than the naïve methods.

In the simulations where we incorporated trait-dependent recapture probabilities, all methods (LA, NB, MR) performed similarly at high recapture probabilities, but LA and NB began to fail at low recapture probabilities and high trait dependence (Figs 3, S6–S7). Although the estimates of β, from all three methods were similar at high recapture probabilities, LA and NB were biased at lower recapture probabilities and when trait dependence was high (Figs 4, S7).

Table 3. Results from the field dataset. Survival selection for total body length (tbl) and front wing length (fwl). Note that all MR models also included a term controlling for trait-dependent recapture probability logit(p) = I_p + tbl + fwl. The delta column shows the ΔAIC from the top ranking model. Selection estimates were taken from the full model for each method: NB and LA: $w = I_w$ + tbl + fwl + tbl^2 + fwl^2. MR: logit (p) = I_p + tbl + fwl; logit(Φ) = I_Φ + tbl + fwl + tbl^2 + fwl^2. See Supporting: Analysis of the Field Dataset for more details

MR	MR. delta	LA	LA. delta	NB	NB. delta	Term	MR.est	LA.est	NB.est
I + fwl	0	I + fwl	0	I + fwl	0	tbl	−0·012	0·184	0·134
I + tbl + fwl	1·9	I + tbl	0·2	I + tbl	0·7	fwl	0·34	0·181	0·217
I + tbl + fwl + fwl^2	2·3	I + tbl + fwl	0·9	I + tbl + fwl	1·4	tbl^2	−0·028	−0·118	−0·208
I + tbl + fwl + tbl^2	3·2	I + tbl + fwl + tbl^2	2·5	I + tbl + fwl + tbl^2	2·1	fwl^2	−0·018	−0·004	0·043
I + tbl + fwl + tbl^2 + fwl^2	4·2	I + tbl + fwl + fwl^2	2·9	I	2·8				
I + tbl	7·8	I	3	I + tbl + fwl + fwl^2	3·1				
I	10·5	I + tbl + fwl + tbl^2 + fwl^2	4·5	I + tbl + fwl + tbl^2 + fwl^2	4·1				

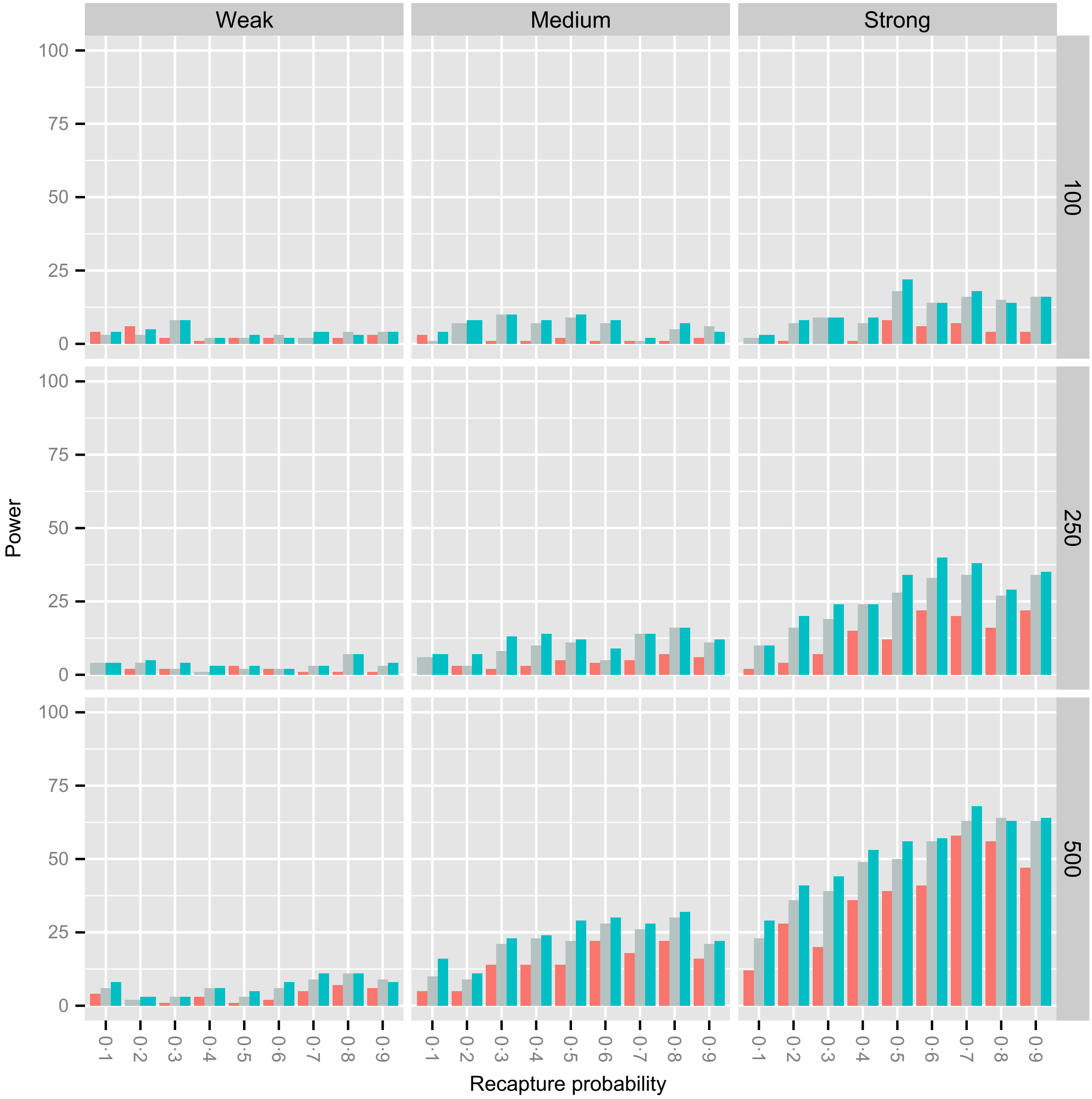

Fig. 1. Power to detect stabilizing selection in the simulated data set using three different methods: Lande–Arnold (LA; red), negative binomial (NB; gray) and mark–recapture (MR; blue). Each bar shows the percentage of times stabilizing selection was detected in our simulations under different conditions of recapture probability and selection strength. The right panel shows the sample sizes used (100, 250, 500 individuals). The upper panel shows the true underlying stabilizing selection gradient (γ: weak = −0·06, medium = −0·12, strong = −0·18). The x-axes show how we varied recapture probability from 0·1 (10% daily recapture probability) to 0·9 (nearly perfect detection). The y-axes show the percentage of runs when stabilizing selection was detected using the AIC-criterion (ΔAIC >2). Each bar represents 1000 runs. b_{Φ} values were set to 0 for all simulations in this figure, which is equivalent to assume that directional selection is absent and the population is sitting on its adaptive peak.

FIELD DATA SET

Based on AIC, all methods (LA, NB, MR) ranked the ‘model: *I* + fwl’ highest, that is, linear selection only on front wing length (fwl; Table 3). The naïve methods (LA, NB) tended to favour a models including total body length (tbl) more than MR (Table 3). Additionally, transformed estimates of linear and quadratic selection on total body length and front wing length tended to differ between the naïve methods (LA, NB) and MR, suggesting that there might be some trait-dependent effects on recapture probability in this data set (Tables 3, S1). Including an interaction between fwl and tbl did not improve model fit, so it was excluded for simplicity. Comparing selection estimates

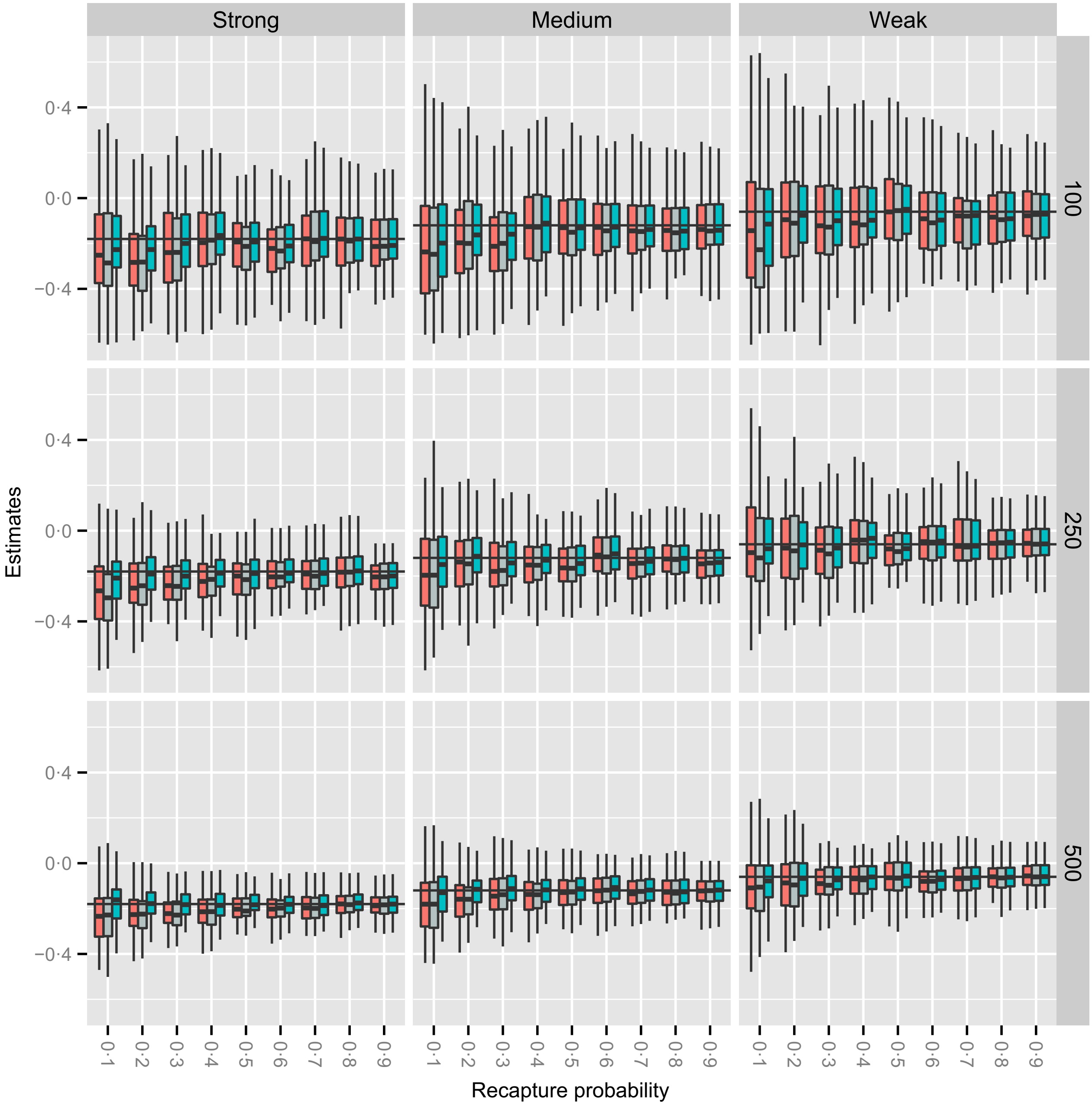

Fig. 2. Estimates of stabilizing selection in the simulated data set using three different methods: Lande–Arnold (LA; red), negative binomial (NB; grey) and mark–recapture (MR; blue). The upper panel shows the three true underlying stabilizing selection gradient categories (γ: weak = −0·06, medium = −0·12, strong = −0·18). Each box is 1000 simulated runs. The solid line indicates the true underlying selection gradient. The upper and lower limits of the boxplot give the 25th and 75th percentiles, respectively.

obtained from MR with those from the LA-approach for 11 traits in *C. splendens* revealed some concordance between the two (Fig. 5).

Discussion

Here, we have shown that the main advantage of MR methods is their ability to control for trait-dependent recapture probability (Figs 3, S6–S7). Although trait-dependent effects are mitigated at high recapture probabilities, MR methods are likely the safest choice, since it is often unclear *a priori* how strongly recapture probability depends on a given trait (Figs S6–S7). There are also certainly power advantages of MR over naïve methods at most recapture probabilities. However, we note that even MR loses significant power at low recapture probabilities (Figs 1, S3–S4).

Estimates of stabilizing selection by our naïve methods (LA, NB) were somewhat biased (becoming stronger and more negative) with decreasing recapture probability (Fig. 2), while MR estimates were largely unaffected and unbiased. This negative bias produced by naïve methods became particularly pronounced when stabilizing selection became stronger and

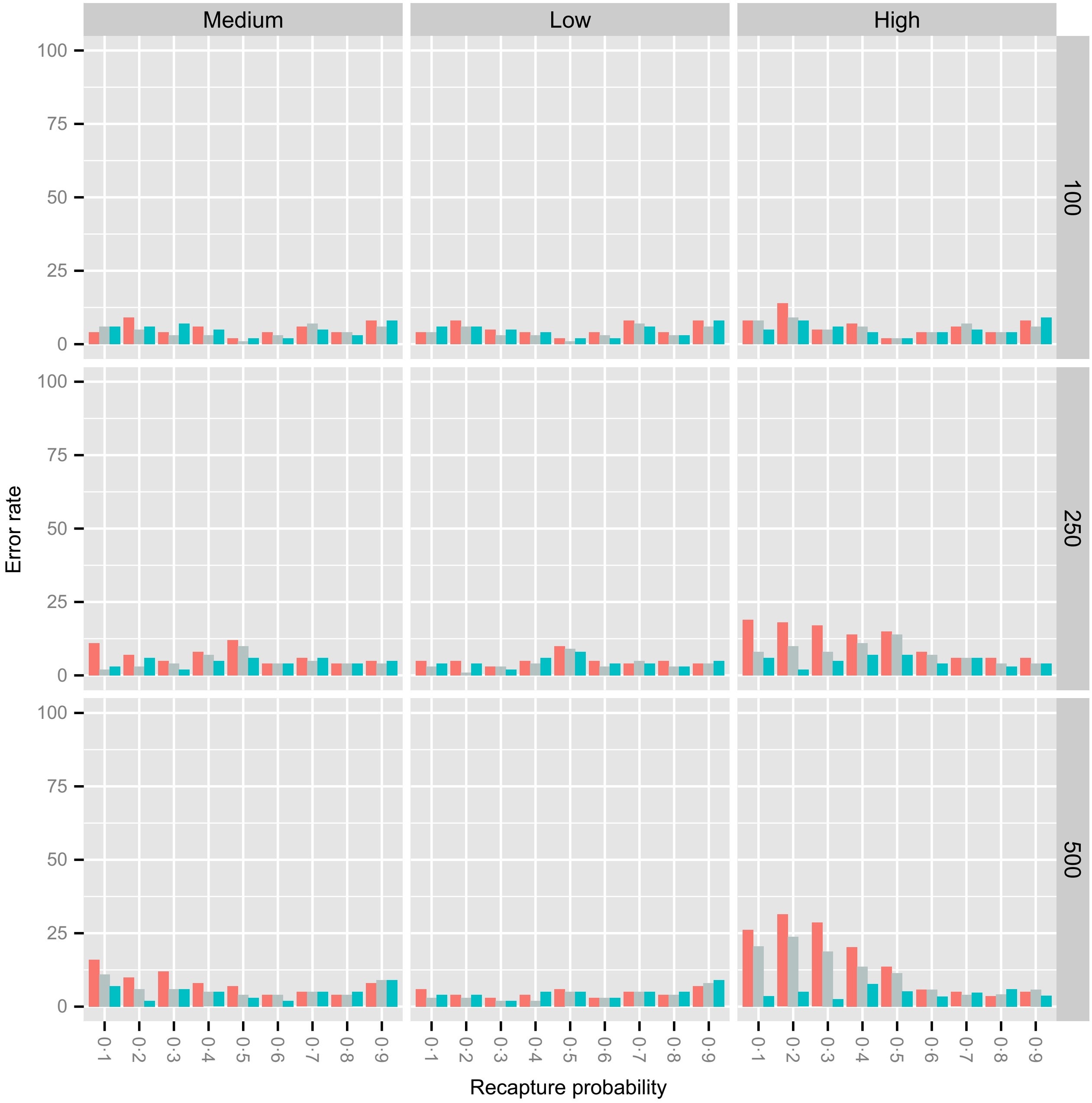

Fig. 3. Error rate from the trait dependence data set, using three different methods: Lande–Arnold (LA; red), negative binomial (NB; grey) and mark–recapture (MR; blue). Each bar represents the percentage of times a method failed our performance test. For a method to fail our performance test, it had to significantly prefer (ΔAIC >2) a model with a linear survival term (b_Φ). The trait dependence term (b_p) 0·05, 0·15 and 0·25, which we refer to as 'Low', 'Medium' and 'High' dependence respectively (See Fig. S7 for higher values). We see here the naïve methods (NB and LA) fail at high trait dependence and low recapture probabilities. The x-axes show how we varied recapture probability from 0·1 (10% daily recapture probability) to 0·9 (nearly perfect detection). Each bar represents 1000 runs.

when recapture probabilities lower (see Fig. S4), although the length of the study period also affected bias in stabilizing selection estimates in complex ways (Fig. S9). We note that such bias could potentially have affected the results and influenced the conclusions from past meta-analyses of the strength of stabilizing selection (e.g. Kingsolver *et al.* 2001). However, without empirical field data from natural populations about the magnitude recapture probabilities it is impossible to judge how severe this potential issue might be. Although not presented here, it is very likely that naïve methods might also become biased when estimating other modes of selection (directional and disruptive).

When recapture probability was not dependent on any phenotypic traits, all methods (LA, NB, MR) performed somewhat similarly, especially at high recapture probability (Fig. 3). However, no study can usually be sure *a priori* that there is not significant trait-dependent recapture probability present when recapture probability is less than one (Figs 3, S6–S7). Therefore, the safest choice would be to utilize MR methods even in scenarios where it might only have limited

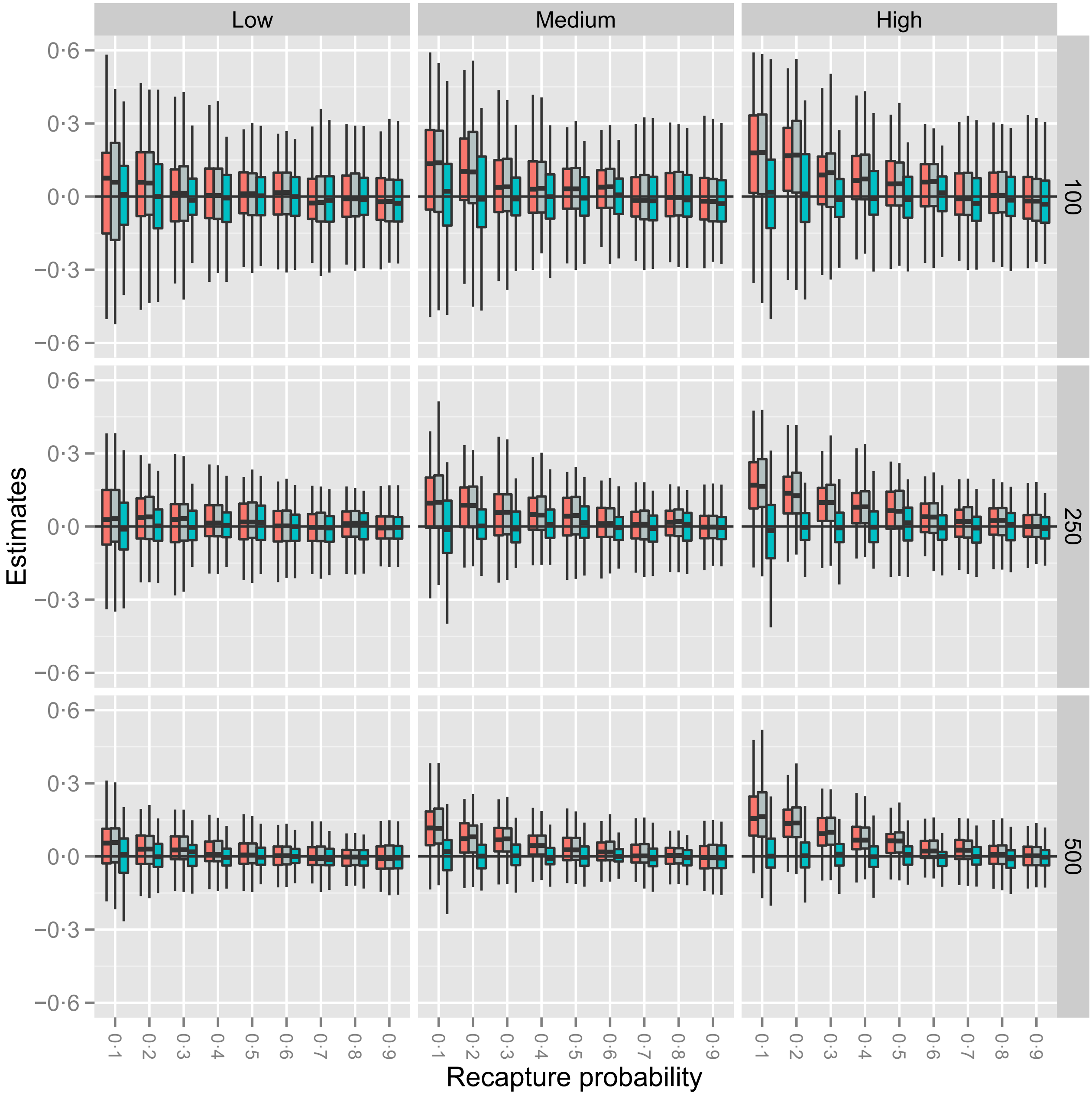

Fig. 4. Estimates of linear selection from the simulations incorporating trait-dependent recapture probability, using three different methods: Lande–Arnold (LA; red), negative binomial (NB; grey) and mark–recapture (MR; blue). Since our test with the mark–recapture method had two models with a b_Φ term, we present only estimates from the model that accounts for trait-dependent recapture probability (logit(p) = $I_p + b_p z$; logit (Φ) = $I_\Phi + b_\Phi z$). The true underlying selection ($\beta = 0$) is indicated with a solid line. b_p: Weak = 0·05, medium = 0·15, strong = 0·25. Each box is 1000 simulated estimates. The upper and lower limits of the boxplot give the 25th and 75th percentiles, respectively.

impact on the conclusions of the study, such as with high recapture probability and low sample sizes.

An alternative procedure for assessing whether MR methods are necessary would be to fit initially a naïve method, and if survival significantly depends of the observed traits, one could verify using MR methods that the observed effect is not simply an effect of trait-dependent recapture probability. However, even this procedure would not control for alternative scenarios in which recapture probability and survival probability are dependent on the observed trait but in opposite directions. Also in these types of scenarios, MR methods are still more powerful than naïve methods at most recapture probabilities (Fig. 1).

Mean survival probability also affected statistical power to detect selection. The simulations we have presented in this study were generated with a mean daily survival probability of 70%, but when this value was lowered to 30%, all methods (LA, NB, MR) lost power across all recapture probabilities (see Figs S3–S4). Similarly ending the study period early, before all individuals have died, reduces power in all methods,

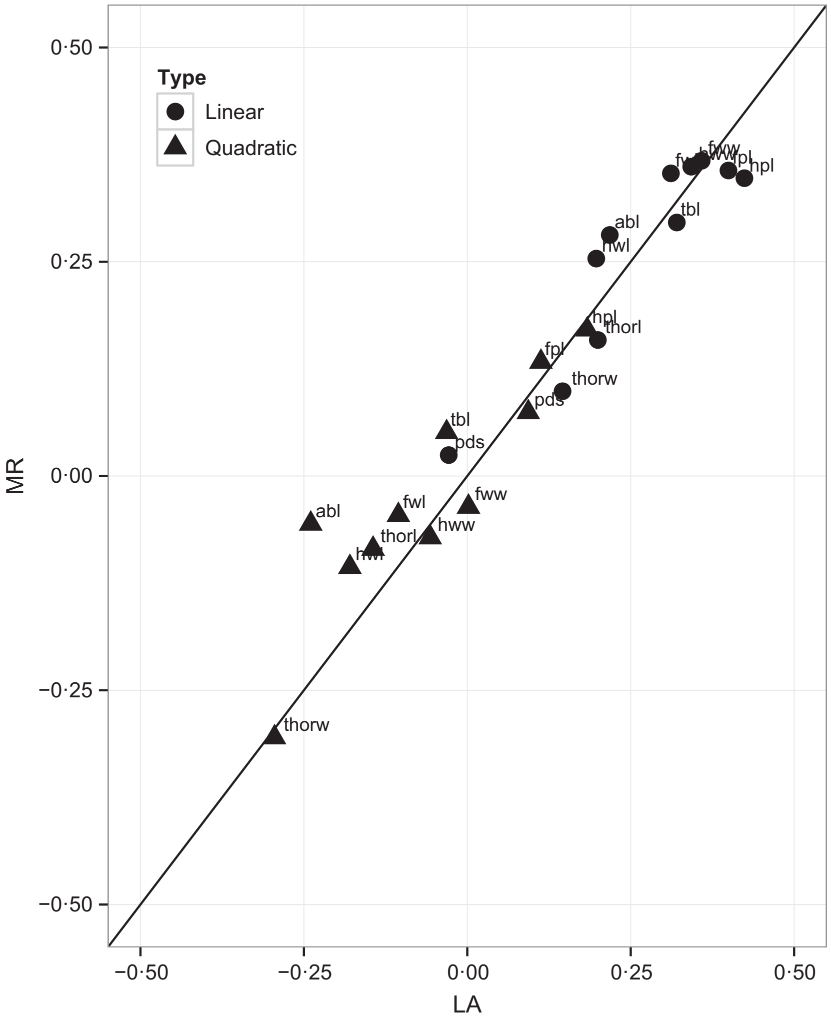

Fig. 5. Estimates of selection on eleven phenotypic traits from an empirical field data set (males in the banded demoiselle *Calopteryx splendens*). Shown is a comparison of the estimates from the classical LA approach (assuming perfect detection) and MR (taking imperfect detection into account). Each point represents an estimate of selection on an individual trait using either approach. The line represents a hypothetical perfect concordance between the approaches (slope = 1, intercept = 0). Estimates were produced from univariate models including only the focal trait (e.g. for LA: $w = I + \text{trait} + \text{trait}^2$; and for MR: $\text{logit}(p) = I + \text{trait}$, $\text{logit}(\Phi) = I + \text{trait} + \text{trait}^2$) and no other terms, so that a total of 22 separate models were estimated.

but affects naïve estimates of selection more severely (see Figs S8–S9).

The vast majority of past survival selection studies have assumed perfect detection (Conner 2001; Kingsolver *et al.* 2001). Given the weak statistical power of most past selection studies and the limited samples sizes in these (Fig. 1), the general picture of the strength of survival selection is in nature (Kingsolver *et al.* 2001) is unlikely to change with the implementation of MR methods. This is of course, assuming that trait-dependent recapture probability is not a widespread and strong force in these historical data sets, a topic that deserves further empirical attention. Of course, recapture probabilities need not only to be *trait*-dependent. Other heterogeneous aspects of recapture probability can cause problems when they are ignored. Future individual studies will nevertheless benefit from utilizing MR methods.

Comparing the selection estimates between the three methods is challenging. The coefficients typically estimated using statistical methods such as MR and NB, will not agree and are not strictly comparable with selection gradients used in quantitative genetic theory (e.g. LA; Morrissey & Sakrejda 2013; M.B. Morrissey & I.B.J. Goudie, unpubl. data; M. Morrissey, pers. comm.). All estimates used in this study have therefore been transformed to become comparable between the different methods on different scales and equivalent to selection gradients in the classical quantitative genetic framework (M.B. Morrissey & I.B.J. Goudie, unpubl. data).

Our analyses based on the field data set of *C. splendens* males were broadly in agreement with the results and conclusions from our simulations (Table 1; Fig. 5). Since our field data set had low recapture probability, at around 10%, and low to intermediate sample size ($N = 276$), our results are consistent with the situation in our simulations with low sample size and low recapture probabilities. One concern is that initial guesses about recapture probability are by no means intuitive and easily understood outside the MR-framework. With our field data set, for example, our initial guess was that we would have a daily recapture probability of about 50%, given a relatively high fraction of marked individuals that were seen at least once again in the field after the marking occasion. However, after analysing the individual field histories using MR, we found that the daily recapture probability turned out to be only 10%. We made our guess after several years of study and first-hand experience in the field. We found that recapture probability was trait-dependent for total body length (tbl) and front wing length (fwl) but with opposite sign ($b_{p.\text{tbl}} = 0{\cdot}281$, $b_{p.\text{fwl}} = -0{\cdot}482$; Table S1). This caused MR to prefer 'model: I + fwl' more strongly than our naïve methods, which also ranked 'model: I + fwl' highest, but with lower ΔAIC. While all methods ranked the 'model: I + fwl' highest, the estimates of γ and β, along with ΔAIC values diverged enough from the MR results for some concern.

Conclusions

All statistical methods (MR, LA, NB) lost power with decreasing recapture probability. Generally, no statistical method can replace the value of choosing a good study system with high recapture probability and sufficiently large sample sizes, in combination with careful study designs. Here we find that MR methods, as expected, are probably the best statistical choice in situations where recapture probability is low or when recapture probability is trait-dependent. The two naïve methods (NB and LA) that assumed perfect detection and used minimum life span performed well at high recapture probabilities. However, we stress that one can never be certain that there is not high trait dependence present within their data without the use of MR methods (Figs S6–S7). Although at present mark–recapture methods are generally more difficult to implement than our naïve methods (LA, NB), their ease of implementation are likely to improve in the future (Gimenez *et al.* 2006; Gimenez *et al.* 2007; Gimenez, Grégoire & Lenormand 2009; King 2012). We offer our R package EasyMARK as a step in this direction.

Acknowledgements

We are grateful to Thomas Gosden, Lesley Lancaster, Maren Wellenreuther, and Beatriz Willink for constructive criticisms on the first draft of this manuscript. We also thank Johanna Eklund and Anna Nordén for help in collecting the field data. We appreciate the comments and suggestions of the two reviewers of this manuscript, and would especially like to thank M. Morrissey whose comments and suggestions greatly improved the current study. Funding for this study was provided by grants from the 'Carl Tryggers Foundation' (CTS; contract number: 11:458) and The Swedish Research Council (VR; contract number: 621-2012-3768) to E.I.S.

Data accessibility

The simulation and field data sets along with associated R scripts can be accessed via DRYAD (http://datadryad.org/). Additional figures and supporting Results and discussion can be found in the online supporting information doi:10.5061/dryad.bc0d8.

References

Calsbeek, R., Gosden, T.P., Kuchta, S.R. & Svensson, E.I. (2012) Fluctuating selection and dynamic adaptive landscapes. *The Adaptive Landscape in Evolutionary Biology* (eds E.I. Svensson & R. Calsbeek), pp. 89–110. Oxford University Press, Oxford, UK.

Clobert, J. (1995) Capture-recapture and evolutionary ecology: a difficult wedding? *Journal of Applied Statistics*, **22**, 989–1008.

Conner, J.K. (2001) How strong is natural selection? *Trends in Ecology and Evolution*, **16**, 215–217.

Cormack, R.M. (1964) Estimates of survival from sighting of marked animals. *Biometrika*, **51**, 429–430.

Endler, J.A. (1986) *Natural Selection in The Wild*. Princeton University Press, Princeton, NJ, USA.

Gimenez, O., Grégoire, A. & Lenormand, T. (2009) Estimating and visualizing fitness surfaces using mark–recapture data. *Evolution*, **63**, 3097–3105.

Gimenez, O., Covas, R., Brown, C.R., Anderson, M.D., Brown, M.B. & Lenormand, T. (2006) Nonparametric estimation of natural selection on a quantitative trait using mark-recapture data. *Evolution*, **60**, 460–466.

Gimenez, O., Rossi, V., Choquet, R., Dehais, C., Doris, B., Varella, H. & Pradel, R. (2007) State-space modelling of data on marked individuals. *Ecological Modelling*, **206**, 431–438.

Gimenez, O., Viallefont, A., Charmantier, A., Pradel, R., Cam, E., Brown, C.R. *et al.* (2008) The risk of flawed inference in evolutionary studies when detectability is less than one. *American Naturalist*, **172**, 441–448.

Gosden, T.P. & Svensson, E.I. (2008) Spatial and temporal dynamics in a sexual selection mosaic. *Evolution*, **62**, 845–856.

Haller, B.C. & Hendry, A.P. (2013) Solving the paradox of stasis: squashed stabilizing selection and the limits of detection. *Evolution*, **68**, 483–500.

Hereford, J., Hansen, T.F. & Houle, D. (2004) Comparing strengths of directional selection: how strong is strong? *Evolution*, **58**, 2133–2143.

Hoef, J.M.V. & Boveng, P.L. (2007) Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, **88**, 2766–2772.

Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hoang, A., Hill, C.E., Beerli, P. & Kingsolver, J.G. (2001) Strength and tempo of directional selection in the wild. *Proceedings of the National Academy of Sciences*, **98**, 9157–9160.

Janzen, F.J. & Stern, H.S. (1998) Logistic regression for empirical studies of multivariate selection. *Evolution*, **52**, 1564–1571.

Jolly, G.M. (1982) Mark recapture models with parameters constant in time. *Biometrics*, **38**, 301–321.

King, R. (2012) A review of Bayesian state-space modelling of capture–recapture–recovery data. *Interface Focus*, **2**, 190–204.

Kingsolver, J.G. & Diamond, S.E. (2011) Phenotypic selection in natural populations: what limits directional selection? *American Naturalist*, **177**, 346–357.

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gibert, P. & Beerli, P. (2001) The strength of phenotypic selection in natural populations. *American Naturalist*, **157**, 245–261.

Kingsolver, J.G., Diamond, S.E., Siepielski, A.M. & Carlson, S.M. (2012) Synthetic analyses of phenotypic selection in natural populations: lessons, limitations and future directions. *Evolutionary Ecology*, **26**, 1101–1118.

Knapczyk, F.N. & Conner, J.K. (2007) Estimates of the average strength of natural selection are not inflated by sampling error or publication bias. *American Naturalist*, **170**, 501–508.

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

Lebreton, J.D., Burnham, K.P., Clobert, J. & Anderson, D.R. (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, **62**, 67–118.

MacColl, A.D. (2011) The ecological causes of evolution. *Trends in Ecology and Evolution*, **26**, 514–522.

McDonald, T. (2012) *mra: Analysis of Mark-Recapture data*. R Foundation for Statistical Computing,Vienna, Austria. URL http://CRAN.R-project.org/package = mra [accessed 20 May 2014]

Mitchell-Olds, T. & Shaw, R.G. (1987) Regression analysis of natural selection: statistical inference and biological interpretation. *Evolution*, **41**, 1149–1161.

Morrissey, M.B. & Hadfield, J.D. (2012) Directional selection in temporally replicated studies is remarkably consistent. *Evolution*, **66**, 435–442.

Morrissey, M.B. & Sakrejda, K. (2013) Unification of regression-based methods for the analysis of natural selection. *Evolution*, **67**, 2094–2100.

R Development Core Team (2008) *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org [accessed 20 May 2014]

Revolution Analytics & Weston, S. (2013) *foreach: Foreach Looping Construct for R*. R Foundation for Statistical Computing, Vienna, Austria. URL http://CRAN.R-project.org/package = foreach [accessed 20 May 2014]

Seber, G.A.F. (1965) A note on multiple-recapture census. *Biometrika*, **52**, 249–250.

Seber, G.A.F. & Schwarz, C.J. (2002) Capture-recapture: before and after EURING 2000. *Journal of Applied Statistics*, **29**, 5–18.

Siepielski, A.M., DiBattista, J.D. & Carlson, S.M. (2009) It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecology Letters*, **12**, 1261–1276.

Siepielski, A.M., DiBattista, J.D., Evans, J.A. & Carlson, S.M. (2011) Differences in the temporal dynamics of phenotypic selection among fitness components in the wild. *Proceedings of the Royal Society B-Biological Sciences*, **278**, 1572–1580.

Siepielski, A.M., Gotanda, K.M., Morrissey, M.B., Diamond, S.E., DiBattista, J.D. & Carlson, S.M. (2013) The spatial patterns of directional phenotypic selection. *Ecology Letters*, **16**, 1382–1392.

Stinchcombe, J.R., Agrawal, A.F., Hohenlohe, P.A., Arnold, S.J. & Blows, M.W. (2008) Estimating nonlinear selection gradients using quadratic regression coefficients: double or nothing? *Evolution*, **62**, 2435–2440.

Svensson, E.I., Eroukhmanoff, F. & Friberg, M. (2006) Effects of natural and sexual selection on adaptive population divergence and premating isolation in a damselfly. *Evolution*, **60**, 1242–1253.

Svensson, E.I. & Friberg, M. (2007) Selective predation on wing morphology in sympatric damselflies. *American Naturalist*, **170**, 101–112.

Svensson, E. & Sinervo, B. (2000) Experimental excursions on adaptive landscapes: density-dependent selection on egg size. *Evolution*, **54**, 1396–1403.

Wade, M.J. & Kalisz, S.M. (1990) The causes of natural selection. *Evolution*, **44**, 1947–1955.

Wellenreuther, M., Larson, K.W. & Svensson, E.I. (2012) Climatic niche divergence or conservatism? Environmental niches and range limits in ecologically similar damselflies. *Ecology*, **93**, 1353–1366.

Wellenreuther, M., Vercken, E. & Svensson, E.I. (2010) A role for ecology in male mate discrimination of immigrant females in Calopteryx damselflies? *Biological Journal of the Linnean Society*, **100**, 506–518.

Williams, B.K., Nichols, J.D. & Conroy, M.J. (2002) *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA, USA.

Received 18 September 2015; accepted 16 October 2015
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Fig. S1. Histograms comparing minimum lifespan values generated from a random simulation and from the field dataset of damselflies (from male banded demoiselles, *Calopteryx splendens*).

Fig. S2. Quantifying morphological phenotypic variation in the banded demoiselle (*Calopteryx splendens*) from photographs.

Fig. S3. Power to detect stabilizing selection with low mean survival probability (0·3) using three different methods: Lande-Arnold (LA; red), negative binomial (NB; gray), and mark-recapture (MR; blue).

Fig. S4. Estimates of stabilizing selection with low mean survival probability (0·3) using three different methods: Lande-Arnold (LA), negative binomial (NB), and mark-recapture (MR).

Fig. S5. False positive rates (Type I error) of our three methods (LA, NB, MR).

Fig. S6. Failure rates (Type II error) from additional trait-dependent simulations, using three different methods: Lande-Arnold (LA; red), negative binomial (NB; gray), and mark-recapture (MR; blue).

Fig. S7. Estimates of β from the trait dependent simulations using three different methods: Lande-Arnold (LA), negative binomial (NB), and mark-recapture (MR).

Fig. S8. Power to detect stabilizing selection when the study period is ended early: Lande-Arnold (LA), negative binomial (NB), and mark-recapture (MR). LA (red), NB (gray), MR (blue).

Fig. S9. Estimates of stabilizing selection when the study period is ended early: Lande-Arnold (LA), negative binomial (NB), and mark-recapture (MR).

Table S1. Output from our mark-recapture analysis on the field dataset.

Data S1. Simulation and field datasets.

Data S2. R scripts used in Study.