JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.13117

Subtle but ubiquitous selection on body size in a natural population of collared flycatchers over 33 years

M. BJÖRKLUND & L. GUSTAFSSON

Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden

Keywords:

body size;
collared flycatcher;
multivariate selection;
reproduction;
survival;
time series.

Abstract

Understanding the magnitude and long-term patterns of selection in natural populations is of importance, for example, when analysing the evolutionary impact of climate change. We estimated univariate and multivariate directional, quadratic and correlational selection on four morphological traits (adult wing, tarsus and tail length, body mass) over a time period of 33 years (≈ 19 000 observations) in a nest-box breeding population of collared flycatchers (*Ficedula albicollis*). In general, selection was weak in both males and females over the years regardless of fitness measure (fledged young, recruits and survival) with only few cases with statistically significant selection. When data were analysed in a multivariate context and as time series, a number of patterns emerged; there was a consistent, but weak, selection for longer wings in both sexes, selection was stronger on females when the number of fledged young was used as a fitness measure, there were no indications of sexually antagonistic selection, and we found a negative correlation between selection on tarsus and wing length in both sexes but using different fitness measures. Uni- and multivariate selection gradients were correlated only for wing length and mass. Multivariate selection gradient vectors were longer than corresponding vector of univariate gradients and had more constrained direction. Correlational selection had little importance. Overall, the fitness surface was more or less flat with few cases of significant curvature, indicating that the adaptive peak with regard to body size in this species is broader than the phenotypic distribution, which has resulted in weak estimates of selection.

Introduction

The process of natural selection is one of the cornerstones of evolutionary biology. Selection acting on heritable traits will lead to changes in trait distributions between generations and thus will act as an explanation for the patterns of change over time. Consequently, a large number of studies have been devoted to understanding the impact and occurrence of selection on phenotypic traits in natural populations (see reviews by Endler, 1986; Kingsolver *et al.*, 2001, 2012; Kingsolver & Diamond, 2011; Siepielski *et al.*, 2009). These reviews show that selection can easily be found in natural populations but that it is rarely very strong. This suggests that natural populations are often more or less well adapted to their environments, but that fluctuations in environmental conditions occasionally displace them from the optimum.

Most of these studies are short-term studies with snapshots of selection in action and may or may not accurately represent the process of selection in natural populations for a variety of reasons such as publication bias due to lack of interest in 'negative' results from researchers, reviewers and editors and lack of long-term studies that can give information of whether a significant selection one year is part of a trend or a just single event. Long-term studies are very rare, but there are exceptions (reviewed by Siepielski *et al.*, 2009; Morrissey & Hadfield, 2012). In total, sufficient long-term data (longer than 10 years) were available in nine studies of

Correspondence: Mats Björklund, Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden.
Tel.: +46 18 471 2666; fax: +46 18 471 6484;
e-mail: mats.bjorklund@ebc.uu.se

eight species, resulting in a total of 28 estimates of directional selection for morphological and life-history traits using different fitness measures (from Table 1 in Morrissey & Hadfield, 2012). This should be contrasted to more than 4500 single estimates used in Kingsolver *et al.* (2012).

Even if single univariate and multivariate selection gradients can give important information on ecological and evolutionary processes, a fuller understanding of the interaction between phenotypes and their environment needs a multivariate approach to discern the complexity of nonlinear selection and selection trait combinations. The methodology has been known for decades (e.g. Lande & Arnold, 1983; Phillips & Arnold, 1989; Blows & Brooks, 2003; Blows, 2007), but the number of studies with the full multivariate scope is still surprisingly small (see reviews in Svensson & Calsbeek, 2012). By using a full multivariate approach, we can get an understanding of the fitness surface, which provides more complete information on the relationship between phenotypes and their environment. An additional set of information can be gained if we repeat this analysis over a number of years because we then can analyse whether the fitness surfaces change over time, for example, as a response to changing environmental conditions, or as a process towards adaptation (reaching an adaptive peak). We can also analyse whether the fitness surfaces differ between different groups in the population such as males and females to test the prevalence of sexually antagonistic selection (e.g. Rice, 1996).

Table 1 Percentage of significant results at the 5% level ($P < 0.05$), the test statistic for a uniform distribution of *P*-values (X^2), *P*-value for the X^2-statistic and the number of cases with a significant coefficient at a false detection rate of 0.05 (FDR). Sexes combined. s is the univariate selection gradient, g is the univariate quadratic selection coefficient, β is multivariate directional selection gradient, γ_{ii} is quadratic (stabilizing and disruptive) selection, and γ_{ij} is correlational selection.

Period	Selection	% <0.05	X^2	*P*	FDR
Fledged	s	24.2	81.5	< 0.001	12
	g	17.0	1.1	0.30	1
	β	4.2	3.5	0.55	0
	γ_{ii}	8.4	5.9	0.015	0
	γ_{ij}	3.8	1.2	0.28	0
Recruits	s	21.6	50.4	< 0.001	8
	g	13.6	1.3	0.24	0
	β	9.9	12.4	0.0004	4
	γ_{ii}	6.4	1.1	0.30	4
	γ_{ij}	6.8	2.6	0.11	4
Survival	s	12.1	8.8	0.003	2
	g	9.5	13.2	0.0003	0
	β	6.8	1.8	0.19	0
	γ_{ii}	4.9	0.003	1	7
	γ_{ij}	4.8	0.03	0.86	3

This leads to the following questions: Are there long-term trends in the patterns of selection? What does the adaptive surface look like and how does it change over time? Can we find patterns of selection in a time series that is too elusive in short-term studies? What is the relationship between univariate and multivariate selection gradients, that is the importance of phenotypic correlations among traits? How strong are the different forms of selection (directional, quadratic and correlational)? Furthermore, we can get information on the importance of sexually antagonistic selection, where sexes have different optima and hence being selected in different directions (e.g. Rice 1996), or sexually synergistic selection if the optima are the same for the two sexes. By analysing the data in a time series framework, we could get information of trends of selection, or subtle patterns, that are not available by using a single-year approach and thus get insight into how the long-term changes in climate affect body size in this species through the process of natural selection.

In this study, we analyse selection in the collared flycatchers (*Ficedula albicollis*) over a time period of 33 years using a uni- and multivariate approach using four morphological traits. We analyse the strength of directional, quadratic (stabilizing and disruptive) selection as well as correlational selection in all years for males and females separately. We are not aware of any other study of a natural population with the same scope, neither in terms of number of years nor in terms of the extent of multivariate analysis. We are particularly interested in selection in adult male and female body size, and its components. Body size has many times been shown to be under selection in a large number of species, including birds, because it relates to, for example, territory acquisition, inter- and intraspecific competition, flight manoeuvrability and sexual selection (e.g. Endler, 1986; Kingsolver *et al.*, 2001, 2012; Siepielski *et al.*, 2009; Kingsolver & Diamond, 2011).

There is debate about the extent to which selection is temporally fluctuating (Siepielski *et al.*, 2009), or is temporally consistent (Morrissey & Hadfield, 2012). Over the time of the study, environmental conditions have changed both between years and over the whole time period with an increasing spring temperature (Björklund & Gustafsson, 2013; Evans & Gustafsson, 2017). We have already shown that the opportunity for selection is correlated with mean temperature in April and that in some years we can find selection on the level of the pair (Björklund & Gustafsson, 2013). In addition, selection on the forehead patch in this species has changed in relation to climate change, and population density has decreased during the same time period (Evans & Gustafsson, 2017). An increased ambient winter temperature has been shown to correlate with an increase in body size in another passerine, the citril finch (Carduelis citrinella), and a response to selection

for a larger size (longer wings) was inferred to be the most likely explanation (Björklund *et al.*, 2015). Trends of change in size related to change in temperature have indeed been found in other birds (see review by Teplitsky & Millien, 2013), but the trends shown have very rarely been connected to a change in selection (but see Evans & Gustafsson, 2017).

Our results show that there are patterns of selection, but they are weak and most often discernable only when viewed in a long-term time series perspective. The main conclusion is that body size in this species is only weakly related to fitness, indicating a rather flat adaptive surface.

Materials and methods

Data collection

All data were collected in the same population of collared flycatchers on southern Gotland, Sweden (57°30′N, 18°33′E), during the years 1981–2013. The field procedures have been described in a large number of papers, and for details, we refer to, for example, Qvarnström *et al.* (2006). In total, the data set consists of 8080 and 10 154 observations on males and females at the fledged stage, 8886 and 10 732 observations on males and females for recruits and survival. The details broken down by year are given in Table S1. Over the years, a number of experiments have been carried out on this population, which to a very large extent has affected different fitness measures. Thus, to avoid any influence on these experiments in this analysis, all breeding records that included an experiment were excluded. Mean temperature data in June, which is the time where offspring are raised, were taken from Hoburg Meteorological Station, approx. 20 km from the main nest-box area.

We used the following traits: wing length, tarsus length, tail length and body mass. As there are systematic changes in body mass over the season in females, we controlled for date of measurement in relation to hatching time. This was performed because females are drastically reducing their weight at the time of hatching, and thus, we added a categorical variable (before/after hatching) to the model for females. This does not happen in males, and hence, this was not included in the male models. We controlled for age when we analysed wing length because second-year birds have on average longer wings than first-year birds. We checked for outliers for every year and trait, and removed data points that were larger (smaller) than three SDs from the mean. All traits were normally distributed.

We separated fitness into three components; first, we used the number of fledged offspring in a given year (fledged), which measures the impact of body size on a given reproductive event, that is, if body size is related to parental quality. Second, we used the number of offspring returning to the population during subsequent years (recruits), which is measure of the impact of body size on the contribution to the future generations, that is, a measure that is more closely related to fitness. Finally, we used adult survival from one year to the next (survival), which is a measure of the impact of body size on the survival during the non-breeding season including migration and interactions at the winter grounds. By using three different fitness measures, rather than using life-time reproductive success, we can analyse whether the strength and pattern of selection differ between the different time periods and the way of measuring fitness (Scranton *et al.*, 2016). Hence, even if these fitness measures can be correlated, they clearly describe different opportunities for selection.

Estimation of selection

We calculated the opportunity for selection, I, defined as the variance in relative fitness (e.g. Arnold, 1986).

We used generalized linear models (GLM) to estimate selection gradients using Poisson (fledged and recruits) and binomial (survival) links (O'Hara & Kotze, 2010) and standardized data with zero mean and unit variance. We estimated univariate selection using a model with two coefficients (one directional, s, and one quadratic term, g). We fitted a multivariate model with all possible selection coefficients (e.g. Lande & Arnold, 1983); four directional (β), four quadratic (stabilizing, disruptive, γ_{ii}) and six correlational selection (γ_{ij}) coefficients, in total 14 coefficients. The quadratic coefficients were estimated using the standardized values squared and then halved (Stinchcombe *et al.*, 2008; Morrissey & Goudie, 2016). The correlational selection coefficients were estimated as the different traits multiplied to each other. The coefficients obtained from a GLM are, however, not the same as in a regular Lande–Arnold least-squares regression, and thus, we used the transformations developed in Morrissey & Goudie (2016). Tests of significance were made after calculating the standard errors of the transformed estimates using Jacobians and calculating $t = s/se$ (see Morrissey & Goudie, 2016 for details) and P-values were then taken using the Student t-distribution. It should be noted that these transformations assume that quadratic coefficients are fairly low; otherwise, the estimates and the standard errors could be heavily inflated as they scale with $1/(1-g)$, where g is the quadratic estimate (Morrissey & Goudie, 2016), and this might have happened in this study in a few cases.

The directional selection gradients obtained using this approach show the change in fitness in relation to a change in standard deviations of the trait. This is the by far most used way of expressing directional selection gradients and thus allows comparison with other studies. However, mean-standardized gradients are more

easily interpreted in terms of strength of selection as the selection on fitness itself is then 1.0 (Hereford *et al.*, 2004; Hansen & Houle, 2008). Thus, we also used the absolute mean-standardized directional gradients. We adjusted for bias resulting from using absolute values on numbers that have confidence intervals overlapping zero using the approach by Hereford *et al.* (2004; see also Morrissey, 2016), where values with a bias larger than the estimate were set to zero. This approach works for directional selection only, and to get an idea of the strength of quadratic selection, we expressed the obtained estimates in relation to the maximum possible given as $\sqrt{2}$I to give a measure of effect size (Arnold, 1986).

In a longitudinal study over different episodes and traits, a large number of tests are made, which will increase the probability of performing a type I error. We used the method by Pike (2011) using false detection rates (FDR). An important question here is what constitutes *m,* the number of tests to adjust for. This is not the total number of tests made (> 4500) as this would be far too conservative. Thus, we have to define what constitutes the appropriate family of tests where adjustment is needed. In each year, we are testing 14 multivariate coefficients (four directional, four quadratic and six correlational coefficients) in each sex. Thus, the number of tests to control for is $2 \times 14 = 28$. For the univariate case, we tested four traits with two estimates and two sexes $= 4 \times 2 \times 2 = 16$ tests. Thus, we used the classic one-stage method, which satisfies the condition $P_i < iq/m$, where $q = 0.05$ and $m = 28$ (16). We used the same logic when testing for trends and other time series of selection coefficients, where $m = 33$ years.

An important issue is power, that is the ability to find a significant estimate given the sample size and the size of the estimate for a given level of significance. We used the approach by Hersch & Phillips (2004) used the notebook CorrPower provided by Phillips (http://pages.uoregon.edu/pphil/software.html). Power was shown to depend on the parametric correlation between trait(s) and fitness (ρ), and in the univariate case, sufficient power (0.8) is reached when $\rho = 0.28$, 0.2, 0.16 and 0.14 for $N = 100$, 200, 300 and 400, respectively. The corresponding figures for the multivariate case taking the number of tests into account are $\rho = 0.36$, 0.26, 0.22, 0.19 and 0.17. Thus, with the sample sizes used in this study, we are likely to find even fairly low levels of selection.

Multivariate analyses

To evaluate the multivariate slope of the fitness landscape, we used the estimated multivariate selection gradients (β_i) and calculated the multivariate directional selection gradient vector $\boldsymbol{\beta} = [\beta_{tail}, \beta_{tarsus,} \beta_{wing}, \beta_{mass}]$, and the intensity of selection is then the length of this vector. The longer the vector is, the stronger is the selection. We also used the univariate selection gradient vector $\mathbf{s} = [s_{tail}, s_{tarsus}, s_{wing}, s_{mass}]$ for comparisons with the multivariate equivalent.

We were interested in the relation between the **s** and $\boldsymbol{\beta}$-vectors, and the statistics of interest is the difference in length and the difference in direction (angle). One way to test the angle between any two vectors is to generate a null distribution of random vectors and compare the observed angle to the random ones. However, the null distribution was in preliminary simulations found to be far too wide to be of use; hence, the power of the test is very weak. Instead, we compared the vectors to an isometric size vector (size = [0.5, 0.5, 0.5, 0.5]). This is the vector of selection gradients if selection is uniformly acting on body size. We calculated the difference in angle to the size vector for both the **s** and $\boldsymbol{\beta}$-vector, and created new vectors taking estimation error into account by resampling the observed gradients each year from a normal distribution with a mean equal to the observed values and a standard deviation equal to the standard error for each gradient. We repeated this 1000 times and compared the angles to the size vector. If there is no difference in direction between the **s** and $\boldsymbol{\beta}$-vector, we would expect the difference to have a mean of zero. The proportion of runs where the difference is larger than zero equals the *P*-value.

The curvature of the fitness landscape is given by the $\boldsymbol{\gamma}$-matrix with the quadratic coefficients (γ_{ii}) at the diagonal and the correlational coefficients (γ_{ij}) off-diagonal. The curvature is then given by the eigenvalues of this matrix (Phillips & Arnold, 1989). A large vector of eigenvalues indicates a highly curved landscape. To test the different aspects of the curvature, we used the randomization procedure by Reynolds *et al.* (2010) by using the same data set but with the fitness values randomized among individuals. This leads to new data sets with no relation between fitness and traits keeping the distribution of fitness values constant. *P*-values were then calculated as the proportion of times the observed value was larger than the randomized value. We used 1000 randomizations in each case. We have no information on the power of these tests. One indication of low power would be whether there are more cases with $0.05 < P < 0.1$ than expected by chance (5%).

First, we tested the overall strength of nonlinear selection using the length of the eigenvalues of the $\boldsymbol{\gamma}$-matrix. Second, if there is no correlational selection (all zero off-diagonal elements), then the largest eigenvalue of the $\boldsymbol{\gamma}$-matrix will be equal to the largest quadratic coefficient (diagonal element). This means that the larger the difference between the largest eigenvalue and the largest quadratic coefficient, the more important is the correlational selection (Blows & Brooks, 2003). Thus, this can serve as a test of the overall importance of correlational selection.

Time series analysis

Analysing time series of selection gradients could reveal consistent, but subtle, selection that can be difficult to see on a single-year basis. The null hypothesis in all tests of the time series is that the estimates are random numbers with mean zero and a certain variance, that is a white noise time series. In other words, this is a null hypothesis of no selection, but with estimates differing from zero due to sampling errors. If so, we would not expect trends over time such as successively increasing or decreasing estimates, or consistently positive or negative values. Moreover, there should not be any correlation between time series either between traits within a sex or between sexes.

To test this, we used four different tests. First, we used the Kruskal–Wallis test in order to test whether the estimates were taken from a distribution with zero mean, that is testing for consistent selection in one direction. Second, to test for trends, we used the Mann–Kendall test for trends taking autocorrelations into account (Hamed & Rao, 1998). Third, in a random walk model with zero mean, the sum of all the deviations from the mean over time should equal zero with a variance given by the variance of the deviations. Thus, a simple test is the sum of the time series divided by the standard deviation, which should follow a t-distribution. In the first two tests, we tested the time series statistics by creating 5000 random time series modelled as moving average MA (0)-processes with a zero mean and a standard deviation equal to the standard deviation of the time series observed for each trait. *P*-values are then the proportion of random series with a test statistic larger than the observed value.

Finally, we estimated the cross-correlation between two time series after a prewhitening procedure as it is well known that temporal autocorrelation inflates the correlation coefficient. We followed the recommendations in Chatfield (2004: 158), by transforming the x-variable using

$$x_t^* = (x_t - \bar{x}) - \hat{\alpha}(x_{t-1} - \bar{x}),$$

Where $\hat{\alpha}$ is the estimated autocorrelation at lag 1. The same filter was applied to the y-variable. The standard deviation (SE) of the cross-correlation is $\sqrt{1}/N$, and the significance was estimated as $t = r/se$ and compared to a Student *t*-distribution with N-2 degrees of freedom.

All tests and simulations were carried out in Mathematica 10.4 (Wolfram 2016).

Results

Opportunity for selection

Opportunity for selection (I) was highest for recruits (mean = 2.86, S.E = 0.23), followed by survival (males mean = 1.52, SE = 0.065, females mean = 1.47, SE = 0.054) and with fledged having the lowest I (mean = 0.24, SE = 0.030). The differences in opportunity for selection between the three estimates of fitness are highly significant ($P < 0.001$, Wilcoxon test).

Directional selection

There were more significant univariate selection gradients at fledged than at recruits and survival, whereas fewer multivariate selection gradients were significant (Table 1). In total, we found 22 significant univariate gradients using the false detection rate (FDR) and only four significant multivariate gradients of 792 possible over the periods, years and sexes. Univariate selection gradients were evenly distributed around zero except at fledged and recruits (sexes combined) where the mean was significantly different from zero, with a predominance of positive values ($P = 0.00026$, and $P = 0.0013$, respectively; Fig. 1). The absolute strength of directional univariate selection (combined over years and sexes, and bias-corrected) was related to the opportunity of selection with least strength at fledged (s: mean = 0.031, SD = 0.040), more at survival (s: mean = 0.093, SD = 0.12) and at recruits (s: mean = 0.09, SD = 0.098). The difference between recruits and survival was not significant ($P = 0.29$), but both were highly significantly different from fledged ($P << 0.001$). The same pattern was found using the multivariate gradients (β: fledged mean = 0.096, SD = 0.13; recruits mean = 0.12, SD = 0.14; survival mean = 0.19, SD = 0.40; differences not significant, $P = 0.10$).

The distribution of absolute mean-standardized uni- and multivariate gradients is shown in Fig. 2. Overall, the mean univariate gradients for fledged (sexes combined) were 0.042, for recruits 0.12 and for survival 0.16. The corresponding figures for the multivariate gradients were 0.024, 0.031 and 0.063 for fledged, recruits and survival, respectively. Thus, overall, the strength of directional selection was around 4–16% of that in fitness for the univariate estimates and 2–6% for multivariate estimates. The distribution of sign changes summarized over multivariate selection gradients (linear and nonlinear) is given in Fig. 3. If the probability of changing sign of the selection coefficient between years is 50% (as it would be if changes were random with a zero mean), then we would expect 16 changes over a period of 33 years, and the observed distribution is indeed centred around 16. The number of significant changes is lower, but very close to what we would expect if 50% of the changes are significant. The same figures were found for univariate selection gradients and quadratic selection (not shown).

Viewed from a time series perspective, patterns emerged with regard to univariate selection gradients. First, in most years, we found positive univariate

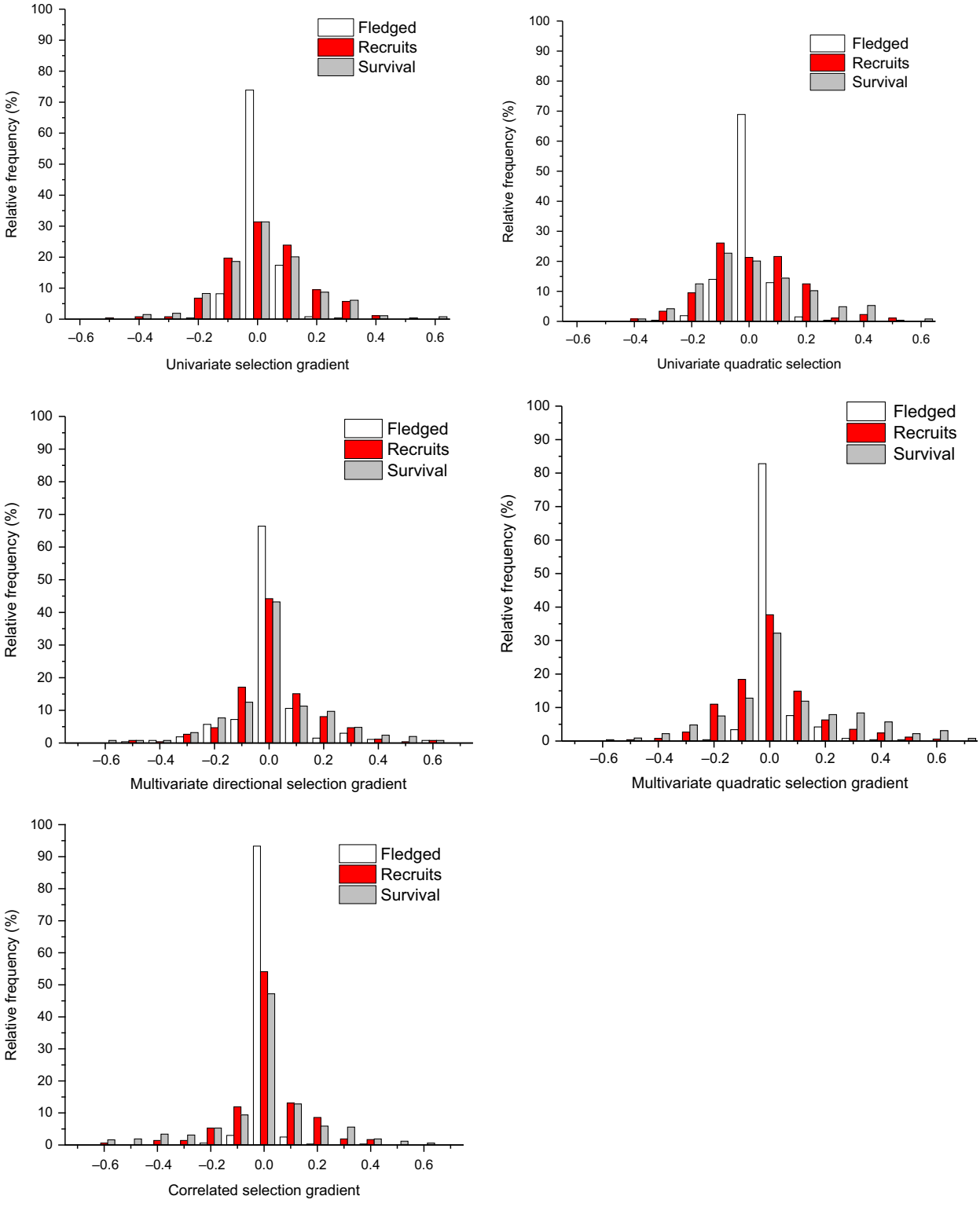

Fig. 1 Unsigned univariate and multivariate variance-standardized selection coefficients.

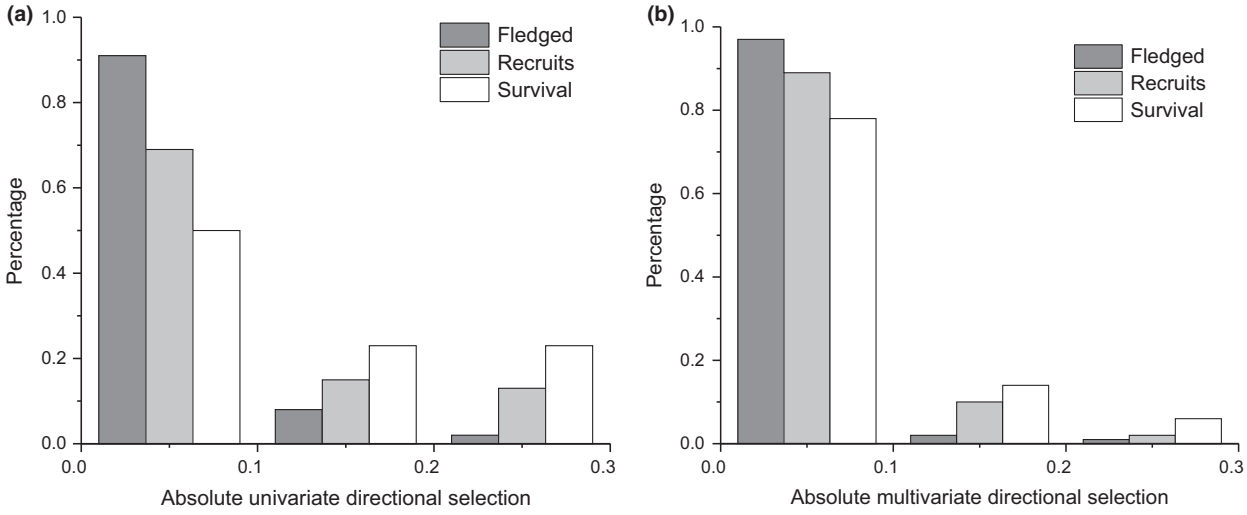

Fig. 2 Absolute values of mean-standardized directional selection as a measure of strength of selection. (a) Univariate selection gradients, (b) multivariate selection gradients.

selection gradients for wing length in males at fledged (mean = 0.034, $P = 0.000055$; Fig. 4a), as well as in females (mean = 0.023, $P = 0.0025$; Fig. 4b) with a possible trend for smaller values over time in females (b = −0.0013, $P = 0.031$, MK trend test). There was also a significant selection for larger female tarsus length at survival (mean = 0.054, $P = 0.0095$). We also found a strong negative correlation between selection on wing and tarsus length in females at fledged and recruits ($r = -0.64$ and -0.63, $P = 0.00047$ and $P = 0.00057$, respectively; Fig. 5a), as well as in males ($r = -0.46$ and $r = -0.43$, $P = 0.0073$ and $P = 0.01$, respectively, Fig. 5b)

We found clear patterns of correlated selection in the time series (to be distinguished from selection on correlations to be discussed below). However, rather than viewing them in isolation, we analysed them in meta-analytic framework combining periods and sexes. We found a positive correlation between univariate selection gradients for tail and wing length (Table 2), although there was a heterogeneity of correlations over periods and sexes. In addition, there were weaker positive correlations in three other trait combinations (Table 3). This was not carried over to the multivariate selection gradients. The correlation between tail and wing length disappeared and we now found a negative, rather than positive, correlation between tarsus and wing length (Table 3). If we compare the time series of univariate and multivariate selection gradients combined, there were strong correlations over periods and sexes for wing length and mass (Table 3). There was a consistent correlation of uni- and multivariate gradients between the sexes at fledged ($r = 0.25$, $P = 0.000029$, Q = 2.45, $P = 0.93$) and weakly so in recruits ($r = 0.14$, $P = 0.012$, Q = 2.85, $P = 0.92$), but not in survival ($r = -0.021$, $P = 0.37$, Q = 5.25, $P = 0.63$).

The mean temperature in June increased over the time analysed (b = 0.071 °C year^{-1}, SE = 0.0025, $P = 0.005$). However, we did not find any significant correlations between the selection time series and the mean temperature in June (mean correlation = −0.027, range: −0.34, 0.32, $N = 132$ correlations).

Quadratic and correlational selection

There were almost no significant univariate quadratic selection coefficients. The distribution of selection coefficients combined over traits was centred round zero and did not differ between the sexes (Fig. 1b). In a few cases (five) did we find indications of deviations from random walk but none of the deviations were not significant at a table-wise α-level. The effect sizes were on average low and were slightly larger in females at fledged (mean = 7.1 and 8.4% for males and females, respectively; $P = 0.044$) and slightly larger in males at survival (mean = 11.7% and 10.6%, for males and females, respectively; $P = 0.013$).

The same patterns were found also in the multivariate selection gradient analyses. The effect sizes (strength of quadratic selection in relation to maximum possible) were low but significantly different between the sexes at fledged (mean = 7.3 and 9.8%, for males and females, respectively; $P < 0.001$) and slightly so at recruits (mean = 8.4 and 7.8%, for males and females, respectively; $P = 0.001$), but not at survival

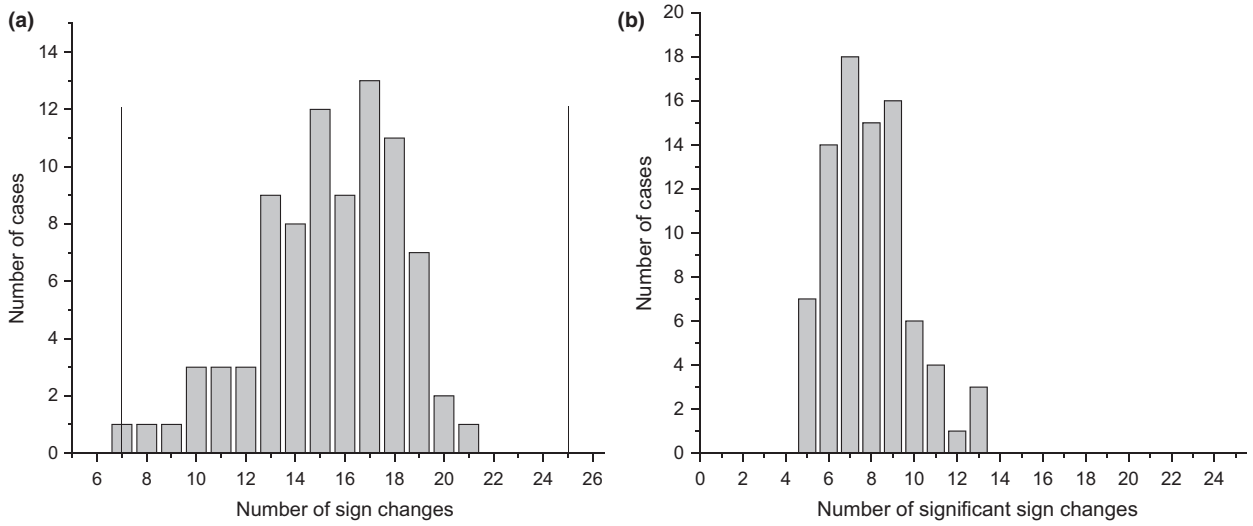

Fig. 3 Frequency distribution of changes of sign of the selection in a time series, all types of selection combined. (a) All changes. The vertical lines refer to the upper and lower 2.5% limits from a mean of 16. (b) Only significant changes.

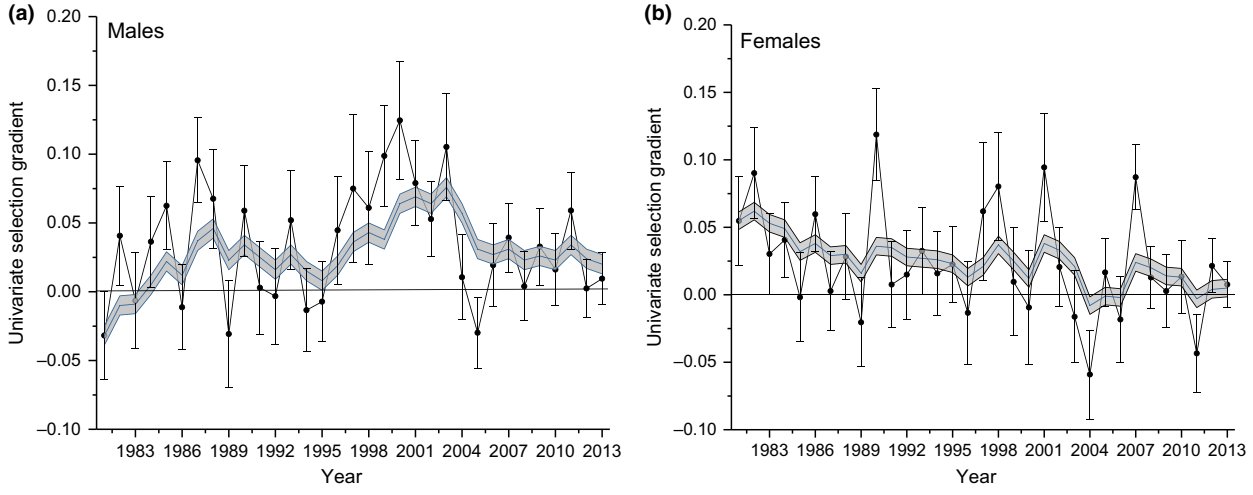

Fig. 4 Selection on wing length in (a) males and (b) females at fledge. Shown is mean ± SD, and the shaded area is the smoothed time series ± SD.

(mean = 19.8 and 14.2%, for males and females, respectively; $P = 0.094$). There was a strong positive correlation between the univariate and multivariate estimates for wing length (combined $r = 0.40$, $P < 0.001$, Q = 8.08, $P = 0.15$) and mass (combined $r = 0.64$, $P < 0.0001$, Q = 8.53, $P = 0.13$), but not for tail and tarsus length (combined $r = -0.024$ and -0.017, respectively). There were very few significant correlational selection gradients, and they were centred round zero with no sex differences (Fig. 1c). No other patterns were found at any period.

Multivariate selection

In section, we compare the vectors of directional univariate (**s**) and multivariate ($\boldsymbol{\beta}$) selection gradients and analyse the curvature of the fitness landscape based on the quadratic and correlational selection gradients. The first analysis was of the length of the vector of univariate selection gradients. We found a trend of decreasing length at survival for males (b = −0.003, $P = 0.018$).

The length of $\boldsymbol{\beta}$ was rarely larger than expected by chance with one important exception, females at fledge,

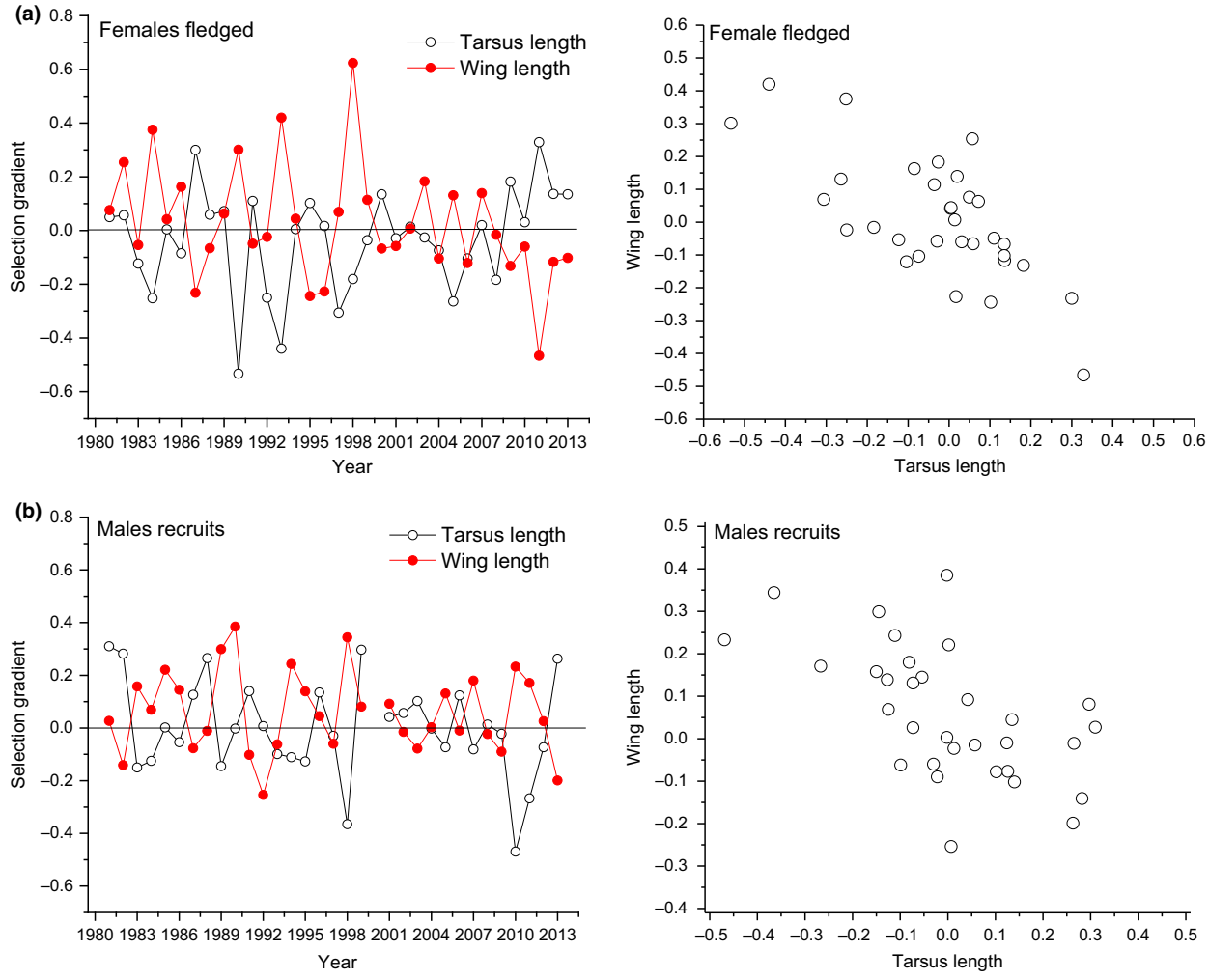

Fig. 5 Patterns of selection. (a) Time series of selection gradients for female tarsus length and wing length at fledge. (b) Time series of selection gradients for male tarsus length and wing length at recruits.

Table 2 Vector angles. Mean and variance of the angle to an arbitrary size vector for multivariate (***β***) and univariate selection gradients (**s**).

	Male			Female		
	Fledge	Recruit	Survival	Fledge	Recruit	Survival
Mean ***β***	84.1	87.5	88.4	96.2	82.5	76.2
Mean **s**	63.0	78.9	90.5	71.2	66.8	65.3
P	0.022	0.46	0.99	0.00043	0.073	0.50
Var(***β***)	491.2	574.4	511.0	531.7	490.6	849.9
Var(**s**)	1513.9	1476.8	1248.3	1366.7	1352.5	1534.8
P	0.049	0.0092	0.014	0.0092	0.017	0.035
β - **s**	0.010	0.075	0.054	0.24	0.085	0.045
P	0.17	0.00083	0.0017	< 0.0001	0.000014	0.045

where ***β*** was significant in almost all years using FDR (Table 4). We tested the direction of ***β*** in all years against ***β*** the first year to assess whether the direction of selection has been consistent over years. The direction of selection was significantly larger than expected by sampling error in nine of 33 years using FDR

Table 3 Summary of the time series correlation (r) of uni- and multivariate selection gradients (**s** vs. $\boldsymbol{\beta}$), and between traits, sexes and periods combined. Q is the test statistic for heterogeneity of the estimates.

	Trait(s)	r	P	Q	P
s vs. $\boldsymbol{\beta}$	Tail	−0.031	0.66	12.9	0.024
	Tarsus	−0.062	0.8	5.07	0.41
	Wing	0.65	< 0.0001	6.78	0.24
	Mass	0.68	< 0.0001	290.2	0.000023
s	Tail–Tarsus	−0.042	0.72	7.17	0.21
	Tail–Wing	0.49	< 0.0001	13.95	0.016
	Tail–Mass	0.075	0.16	7.2	0.21
	Tarsus–Wing	0.25	0.00034	1.09	0.95
	Tarsus–Mass	0.20	0.0034	4.19	0.52
	Wing–Mass	0.17	0.011	6.16	0.29
$\boldsymbol{\beta}$	Tail–Tarsus	−0.18	0.0068	10.31	0.067
	Tail–Wing	0.061	0.21	2.58	0.77
	Tail–Mass	−0.45	< 0.0001	0.47	0.98
	Tarsus–Wing	−0.43	< 0.0001	11.61	0.041
	Tarsus–Mass	−0.033	0.33	10.31	0.067
	Wing–Mass	−0.11	0.07	6.64	0.25

(15 years with $P < 0.05$). Overall, about 14% of all years and periods combined had a significant $\boldsymbol{\beta}$-vector in males, compared to 46% in females.

The length of the $\boldsymbol{\beta}$-vector was significantly larger than that of the **s**-vector for females (fledged: difference = 0.024, $P < 0.001$, recruits difference = 0.085, $P = 0.000014$, survival difference = 0.045, $P = 0.045$) and for males at recruits (difference = 0.075, $P = 0.00083$) and survival (difference = 0.054, $P = 0.0017$), but not for fledged (difference = 0.01, $P = 0.17$). There was a significant difference between the $\boldsymbol{\beta}$-vector and the **s**-vector in the angle to the arbitrary size vector for females at fledged ($x\boldsymbol{\beta} = 96.2$, $x\mathbf{s} = 71.2$, $P = 0.00043$) and perhaps also in males at fledged ($x\boldsymbol{\beta} = 84.1$, $x\mathbf{s} = 63.0$, $P = 0.022$). For all periods and sexes was the variance in angles larger for the **s**-vector than for the $\boldsymbol{\beta}$-vector (Table 4).

Table 4 Number of tests for each sex and period at different significance classes for multivariate selection gradients ($\boldsymbol{\beta}$) and eigenvalues of the γ-matrix.

		$\boldsymbol{\beta}$			γ		
Sex	Period	FDR	< 0.05	< 0.1	FDR	< 0.05	< 0.1
Males	Fledged	0	5	10	0	1	3
	Recruits	1	3	8	0	1	7
	Survival	1	6	8	0	3	6
Sum		2	14	26	0	5	16
Females	Fledged	32	32	33	7	9	12
	Recruits	3	8	13	1	6	8
	Survival	0	6	9	0	6	8
Sum		35	46	54	8	21	28

The test of the importance of correlational selection shows that the largest eigenvalue was significantly larger than the largest quadratic coefficient in only a few cases (sexes and periods combined). The length of the eigenvalues of the $\boldsymbol{\lambda}$-matrix was not significant in males in any period using FDR. In females, we found 21% of the years having a significant curvature at fledged (FDR).

Discussion

The standard tests of trait- and year-wise estimates of selection showed only a few significant results, indicating that selection is either too weak to be detected, or simply absent. This would mean that body size is basically a neutral trait in the collared flycatcher, in contrast to what was found for the forehead patch (Evans & Gustafsson, 2017). The absolute strength of directional selection was found to be in the order of <10% of selection on fitness itself (using mean-standardized data), hence weak. The variance-standardized data show that the fitness difference over the range of variation in body size in the population is small, about 0.03 units difference in fitness for every standard deviation at fledged, which corresponds to a difference of about 0.1 fitness units between individuals that differ by four standard deviations in a trait. As fitness at this period is measured in the number of offspring, it is clear that selection is very weak. The corresponding figures for recruits and survival were higher (around 0.1 units/SD) corresponding to about 0.45 fitness unit difference between extreme phenotypes. Thus, the selection load of the population with regard to body size is low (Hereford *et al.*, 2004). This was strengthened when analysing the fitness surface, which in a vast majority of cases was effectively flat. We found no indications of temporal consistency of selection as the probability of sign changes was as we would expect by chance alone. This is consistent with a conclusion that the selection coefficients obtained here are a result of random sampling, rather than describing important eco-evolutionary processes.

However, when we combined the results over years and analysed selection in a time series framework, a number of patterns emerged. For example, there was a selection for longer wings in males and females at fledged consistently over years, however significant only in a few years. Estimates of selection on tail and wing length were positively correlated, but negative between tarsus and wing length over time, showing that even if the analyses of each trait in isolation did not show anything, viewed in combination patterns arise. Furthermore, even if the selection on individual traits was nonsignificant in females, the multivariate selection gradient was significant in almost all years at fledged. There was also more nonlinear selection in females than in males as evident from the analysis of

the γ-matrix. This can be due to larger quadratic coefficients and/or large amounts of correlational selection. However, the analysis shows very little evidence for an importance of correlational selection. Generally, there was more selection acting on females, but there were no indications of sexually antagonistic selection; on the contrary, selection in males and females was positively correlated. This indicates that the factors acting on reproductive success and survival affect the sexes in a similar way when it comes to body size, but more so in females at the actual time of breeding.

This shows two things; first, that there is selection acting in various ways over traits, sexes and periods separately and in combination. Thus, the estimates of selection are real and reveal something about the interaction between the phenotype and fitness. Secondly, the selection that is acting is weak. Even if we can find patterns of selection using this extensive data set, the overall strength of selection was clearly not strong. It can be argued that the reason we could not find significant selection is due to low power of the tests as a result of low sample sizes. This might well be true in some years, but in many years sample size counts in several hundreds, yet the significant results were not to be found. The answer to this is that the effect sizes were in most cases very low, down to a few percentage of the selection acting on fitness, in which case several thousands of data points are needed. The only way to find patterns of selection was to combine a large number of years and using a time series approach, which shows that selection is ubiquitous but subtle. This in turn prompts the question of the biological relevance of this low level of selection.

One way of addressing this point is to compare the evolutionary effects of the strength of selection we have found here to other random processes such as genetic drift and pure sampling between years. As there was a consistent selection on male wing length at fledged, we used that as an example and calculated the magnitude of expected response between years to selection using data on the additive genetic variance from Björklund *et al.* (2013). This was compared to the magnitude of between-year change due to drift alone and due to sampling. For details, see Appendix 1. As can be seen in Fig. 6, there is considerable overlap between the three distributions (drift vs. sampling 67%, drift vs. selection 55.8% and sampling vs. selection 73.8%), and the main difference is the predicted shift to larger wings due to selection. Thus, the change imposed by selection is only slightly larger than that by drift and sampling. Even if the figures differ between traits, the differences are not large enough to dispute a conclusion that there are certainly patterns of selection, but with questionable biological relevance.

It should be stressed that we are only analysing adults. It is easy to imagine that selection might occur at different stages before reaching adulthood. For example, highly deviant offspring can be selected against at a very early stage and hence at the time of fledging they are already removed from the populations. Likewise, the time between leaving the nest and returning as a recruit at the next breeding period is riddled with factors that select against the most extreme individuals. Hence, overall fitness landscape can be seen as a birthday-cake being essentially flat over a certain area but with steep boundaries. When individuals at the sides are removed by selection, only individuals at the flat part are left, showing no (or only little) relation between body size and fitness. However, it should be remembered that truly deviant individuals are rare in any population and thus selection is weak. Hence, the number of fledged offspring, the number of recruits and the probability of survival between years are only to a small extent dependent on body size, but the vast proportion of variation in these fitness measures is not. This result is exactly what can be expected on theoretical grounds if the adaptive peak is large compared to the phenotypic distribution, a pattern that might be common in nature and can help to explain the prevalence of stasis over time (Haller & Hendry, 2013).

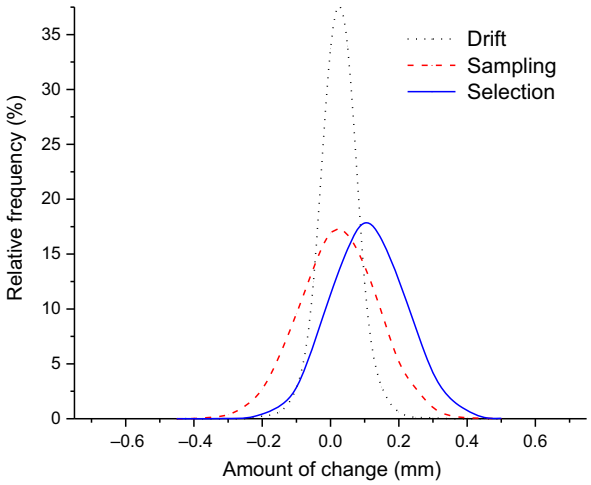

Fig. 6 Result from simulations of relative importance of three processes for the change in wing length in males over years.

By analysing the difference between uni- and multivariate gradients, we can get an understanding of the importance of the phenotypic correlations between traits. The fact that uni- and multivariate gradients only sometimes show the same thing emphasizes the difference between these two ways of measuring selection, and the fact that univariate measures cannot be used as a substitute for multivariate gradients. Univariate selection gradients relate to the interaction between a trait and a measure of fitness. Thus, this is a measure of the ecological impact on the variance in fitness mediated by the trait in question. This is a very important use of

the univariate selection gradients in order to understand ecological processes. On the other hand, the evolutionary impact of selection can only be understood in terms of the multivariate selection gradients because the effect of selection on correlated traits is taken into account, and as can clearly be seen in this study, these have a major importance in enhancing, channelling and directing the selection trajectories (e.g. Gould, 1989; Björklund, 1996; Hansen & Houle, 2008). The difference between the uni- and multivariate gradients shows the dual effects of 'constraints' like phenotypic correlations (Gould, 1989), insofar that the vectors of multivariate directional selection gradients generally were longer than the corresponding vectors of univariate selection gradients. This means that the correlations between traits act to enhance the overall strength of selection, which will result if correlations between traits are positive. On the other hand, the variance in the direction of the selection vectors was clearly reduced in the gradients, which means that the other effect of the correlations is to constrain selection into certain directions determined by the pattern of correlations (Gould, 1989; Björklund, 1996; Hansen & Houle, 2008).

One reason for the finding of low levels of selection could be the fact that we are dealing with a population in a patchy habitat. We know that in some years there is a significant variance among patches in the number of fledglings and recruits (M. Björklund & L. Gustafsson, unpublished), and there might be a spatial autocorrelation of selection (Marrot *et al.*, 2015). Thus, a positive selection in one patch might be offset by a negative selection in another patch, resulting in a net zero selection. However, even though this is a possibility, the likelihood that this would occur in exactly the same way over 33 years is low given the large interannual differences in breeding conditions. Another factor might be age and we have controlled for age in the analysis on wing length. This might not be correct if selection is different in different age classes, and thus, these differences cancel out when we combine data. On average, the proportion of first-year birds was 27% (range 11–46%), so even if this can be of importance in single years, it is unlikely to be a general explanation. Another, maybe more plausible explanation, is the one put forward by Morrissey (2014), arguing that morphology rarely interacts directly with fitness but indirectly through a life-history parameter. Indirect relations diminish correlations for each step between the trait and fitness, and the net result will be a weak correlation between morphology and fitness. An example might be wing length that was found to be under positive selection. A longer wing can result in earlier arrival at breeding grounds, and earlier breeding, which is known to relate positively to clutch size in this species (Sheldon *et al.*, 2003) and in other birds (Hahn *et al.*, 2016). Hence, due to the indirect links between wing length, arrival and clutch size, we find the weak but consistent correlation between wing length and the number of fledged offspring.

It can be argued that as there are many studies actually showing significant selection on body size, or components of it, this study is an outlier and stronger selection is the norm (Hereford *et al.*, 2004; Morrissey & Hadfield, 2012). Indeed, the strength of selection on the morphological traits in this study is substantially lower than in the data presented by Hereford *et al.* (2004), where the median bias-corrected and mean-standardized estimates of selection were 0.38 and 0.89 for multivariate and univariate gradients, respectively, about 3–7 times larger than in our study. The variance-standardized values were, on the other hand, in the same range as in our study. However, extrapolation between populations of different species should always be made with great care. Each population has a history of selection and adaptation to its particular environment. Lack of selection in one population does not imply anything of what we can expect in other populations. Viewed over a number of populations and years, we can get an idea of the relative importance of selection of different kinds, but this information cannot be used to extrapolate to a particular population more than in terms of probabilistic expectations. In our studied population, selection on body size is weak and that is a pattern that is recurrent over three decades.

The results presented here should not be surprising if we assume that populations in general are adapted to their environment. However, we need to define environmental stability in relation to the traits under analysis. We looked at body size, and in relation to body size, we do have relatively stable environment even if measures of yearly climatic variables might change over years. Thus, the part of the phenotype that consists of body size is well adapted and robust against changes in the environment, whereas other traits might be more affected. In other words, the relative importance of body size in determining fitness in this population is weak compared to other traits such as laying date and clutch size (Sheldon *et al.*, 2003), MHC (Radwan *et al.*, 2012), fledgling mass (Linden *et al.* 1992), hormones (Tschirren *et al.*, 2014), mate choice (Robinson *et al.*, 2012) and the forehead patch (Evans & Gustafsson, 2017). Identifying the traits that are important for fitness is a major challenge as we cannot expect that all aspects of the phenotype are under selection at any given time.

Acknowledgments

This work was made possible by numerous grants from the Swedish Research Council (VR, formerly NFR), the Swedish Research Council Formas, Olle Engkvist Foundation and numerous smaller grants over the years. We thank Michael Morrissey for very insightful help with the analyses, Mark Kirkpatrick for important discussions and all the researchers, PhD students, post-docs

and field assistants who have contributed to data collection over the years.

Conflict of interests

The authors declare no conflicts of interest.

Data accessibility

Mean and standard deviations for each trait and sex over years, and all (variance-standardized) selection coefficients with standard errors are available in Supplementary file 1.

References

Arnold, S.J. 1986. Limits on stabilizing, disruptive and correlational selection set by the opportunity for selection. *Am. Nat.* **128**: 143–146.

Björklund, M. 1996. The importance of evolutionary constraints in ecological time scales. *Evol. Ecol.* **10**: 423–431.

Björklund, M. & Gustafsson, L. 2013. The importance of selection at the level of the pair over 25 years in a natural population of birds. *Ecol. Evol.* **3**: 4610–4619.

Björklund, M., Husby, A. & Gustafsson, L. 2013. Rapid and unpredictable changes of the G-matrix in a natural bird population over 25 years. *J. Evol. Biol.* **26**: 1–13.

Björklund, M., Borras, A., Cabrera, J. & Senar, J.C. 2015. Increase in body size is correlated to warmer winters in a passerine bird as inferred from time series data. *Ecol. Evol.* **5**: 59–72.

Blows, M.W. 2007. A tale of two matrices: multivariate approaches in evolutionary biology. *J. Evol. Biol.* **20**: 1–8.

Blows, M.W. & Brooks, R. 2003. Measuring nonlinear selection. *Am. Nat.* **162**: 815–820.

Chatfield, C. 2004. *The Analysis of Time Series*. Chapman and Hall/CRC, New York.

Endler, J.A. 1986. *Natural Selection in the Wild*. Princeton University Press, Princeton.

Evans, S.R. & Gustafsson, L. 2017. Climate change upends selection on ornamentation in a wild bird. *Nat. Ecol. Evol.*, **1**: 0039. https://doi.org/10.1038/s41559-016-0039.

Gould, S.J. 1989. A developmental constraint in *Cerion*, with comments on the definition and interpretation of constraint in evolution. *Evolution* **43**: 516–539.

Hahn, S., Korner-Nievergelt, F., Emmenegger, T., Amrheim, V., Csõrgõ, T., Gursoy, A. *et al.* 2016. Longer wings for faster springs – wing length relates to spring phenology in a long-distance migrant across its range. *Ecol. Evol.* **6**: 68–77.

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

Hamed, K.H. & Rao, A.R. 1998. A modified Mann-Kendall trend test for autocorrelated data. *J. Hydro.* **204**: 182–196.

Hansen, T.F. & Houle, D. 2008. Measuring and comparing evolvability and constraint in multivariate characters. *J. Evol. Biol.* **21**: 1201–1219.

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

Hersch, E.I. & Phillips, P.C. 2004. Power and potential bias in field studies of natural selection. *Evolution* **58**: 479–485.

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

Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E. *et al.* 2001. The strength of phenotypic selection in natural populations. *Am. Nat.* **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. *Evol. Ecol.* **26**: 1101–1118.

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

Lindén, M., Gustafsson, L. & Pärt, T. 1992. Selection on fledgling mass in the collared flycatcher and the great tit. *Ecology* **73**: 336–343.

Marrot, P., Garant, D. & Charmantier, A. 2015. Spatial autocorrelation in fitness affects the estimation of natural selection in the wild. *Methods Ecol. Evol.* **6**: 1474–1483.

Morrissey, M.B. 2014. Selection and evolution of causally covarying traits. *Evolution* **68**: 1748–1761.

Morrissey, M.B. 2016. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *J. Evol. Biol.* **29**: 1882–1904.

Morrissey, M.B. & Goudie, I.B.J. 2016. Analytical results for directional and quadratic selection gradients for log-linear models of fitness functions. *BioRxiv*. https://doi.org/10.1101/040618

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

O'Hara, R.B. & Kotze, D.J. 2010. Do not log-transform count data. *Methods Ecol. Evol.* **1**: 118–122.

Phillips, P.C. & Arnold, S.J. 1989. Visualizing multivariate selection. *Evolution* **4**: 1209–1222.

Pike, N. 2011. Using false discovery rates for multiple comparisons in ecology and evolution. *Meth. Ecol. Evol.* **2**: 278–282.

Qvarnström, A., Brommer, J.E. & Gustafsson, L. 2006. Testing the genetics underlying the co-evolution of mate choice and ornament in the wild. *Nature* **441**: 84–86.

Radwan, J., Zagalska-Neubauer, M., Chicon, M., Sendecka, J., Kulma, K., Gustafsson, L. *et al.* 2012. MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Mol. Ecol.* **21**: 2469–2479.

Reynolds, R.J., Childers, D.K. & Pajweski, N.M. 2010. The distribution and hypothesis testing of eigenvalues from the canonical analysis of the gamma matrix of quadratic and correlational selection gradients. *Evolution* **64**: 1076–1085.

Rice, W.R. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* **381**: 232–234.

Robinson, M.R., van Doorn, G.S., Gustafsson, L. & Qvarnström, A. 2012. Environment-dependent selection on mate choice in a natural population of birds. *Ecol. Lett.* **15**: 611–618.

Scranton, K., Lummaa, V. & Stearns, S.C. 2016. The importance of the timescale of the fitness metric for estimates of selection of phenotypic traits during a period of demographic change. *Ecol. Lett.* **19**: 854–861.

Sheldon, B.C., Kruuk, L.E.B. & Merilä, J. 2003. Natural selection and inheritance of breeding time and clutch size in the collared flycatcher. *Evolution* **57**: 406–420.
Siepielski, A.M., DiBattista, J.D. & Carlson, S.M. 2009. It's about time: the temporal dynamics of natural selection in the wild. *Ecol. Lett.* **12**: 1261–1276.
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. & Calsbeek, R. 2012. *The Adaptive Landscape in Evolutionary Biology*. Oxford University Press, Oxford, UK.
Teplitsky, C. & Millien, V. 2013. Climate warming and Bergmann's rule through time: is there any evidence? *Evol. Appl.* **7**: 156–168.
Tschirren, B., Postma, E., Gustafsson, L., Groothuis, T.G.G. & Doligez, B. 2014. Natural selection acts in opposite ways on correlated hormonal mediators of prenatal maternal effects in a wild bird population. *Ecol. Lett.* **17**: 1310–1315.
Wolfram Research, Inc. 2016. *Mathematica*, Version 10.4. Wolfram Research, Inc, Champaign, IL.

Appendix 1: Simulation details

We used the estimated gradients and the G-matrix from Björklund *et al.* (2013) and simulated different outcomes by taking deviations from a normal distribution with the mean and variance taken from original data. The response was estimated using the standard equation $\Delta\mathbf{z} = \mathbf{G}\boldsymbol{\beta}$. Drift can be modelled as deviates taken from a normal distribution with a zero mean and a variance of V_a/N_e, where V_a is the additive genetic variance for the trait and N_e is the effective population size. V_a was taken from Björklund *et al.* (2013), and Ne was taken from uniform distribution from 100 to 1000. Sampling was simulated as sampling N individuals from a distribution of wing length at year 1981 (arbitrary choice) and comparing the new mean to the original mean. We used the median sample size of 297 as a value of N.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:
Table S1 Sample sizes for males (Nm) and females (Nf), mean fitness (x) and opportunity for selection (I).
Data S1 Means, standard deviations of all traits and all years, univariate and multivariate selection coefficients and their standard errors.

Data deposited at Dryad: https://doi.org/10.5061/dryad.24tm0.

Received 29 December 2016; revised 3 May 2017; accepted 9 May 2017