JOURNAL OF Evolutionary Biology eseb European Society for Evolutionary Biology

doi: 10.1111/jeb.13062

Testing the phenotype-linked fertility hypothesis in the presence and absence of inbreeding

W. FORSTMEIER*, M. IHLE*, P. OPATOVÁ†‡, K. MARTIN*, U. KNIEF*, J. ALBRECHTOVÁ‡§, T. ALBRECHT‡§ & B. KEMPENAERS*

*Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Seewiesen, Germany
†Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic
‡External Research Facility Studenec, Institute of Vertebrate Biology, Czech Academy of Sciences, Brno, Czech Republic
§Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic

Keywords:

display behaviour;
mate choice;
phenotype-linked fertility hypothesis;
precopulatory traits;
sexual selection;
sperm abnormalities;
sperm quality;
sperm velocity.

Abstract

The phenotype-linked fertility hypothesis suggests that females can judge male fertility by inspecting male phenotypic traits. This is because male sexually selected traits might correlate with sperm quality if both are sensitive to factors that influence male condition. A recent meta-analysis found little support for this hypothesis, suggesting little or no shared condition dependence. However, we recently reported that in captive zebra finches (*Taeniopygia guttata*) inbreeding had detrimental effects both on phenotypic traits and on measures of sperm quality, implying that variation in inbreeding could induce positive covariance between indicator traits and sperm quality. Therefore, we here assess empirically the average strength of correlations between phenotypic traits (courtship rate, beak colour, tarsus length) and measures of sperm quality (proportion of functional sperm, sperm velocity, sperm length) in populations of only outbred individuals and in mixed populations consisting of inbreds (F = 0.25) and outbreds (F = 0). As expected, phenotype sperm-trait correlations were stronger when the population contained a mix of inbred and outbred individuals. We also found unexpected heterogeneity between our two study populations, with correlations being considerably stronger in a domesticated population than in a recently wild-derived population. Correlations ranged from essentially zero among outbred-only wild-derived birds (mean Fisher's *Zr* ± SE = 0.03 ± 0.10) to moderately strong among domesticated birds of mixed inbreeding status (*Zr* ± SE = 0.38 ± 0.08). Our results suggest that, under some conditions, the phenotype-linked fertility hypothesis might apply.

Introduction

When individual males differ in fertility (i.e. in the ability of their ejaculates to fertilize eggs), females may benefit from avoiding low-fertility males as partners to ensure that all their eggs will be fertilized. Yet, how could females judge male fertility during the mating period? The 'phenotype-linked fertility' hypothesis (Sheldon, 1994) suggests that sexually selected ornaments and displays could function as indicators of male fertility under the condition that both types of traits (indicators and fertility) are sensitive to the same stressors. Males in poor phenotypic condition may show both reduced ornamentation and low fertility, whereas males in good condition may perform well in terms of both strong sexual signalling and high-quality ejaculates.

Empirical tests of the phenotype-linked fertility hypothesis have produced mixed results. A recent meta-analysis (Mautz *et al.*, 2013) found that the average correlation between measures of ornamentation and measures of sperm quality was only weakly

Correspondence: Wolfgang Forstmeier, Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Eberhard-Gwinner-Strasse 7, 82319 Seewiesen, Germany.
Tel: +49-8157-932346; fax: +49-8157-932400; e-mail: forstmeier@orn.mpg.de

positive (mean $r = 0.06$) and not significantly larger than zero. This implies that the practical utility of ornaments or displays as indicators of sperm quality traits is very limited. A weak or absent correlation may arise if the two types of traits are insufficiently affected by the same kinds of stressors. In the extreme case of no shared condition dependence, the correlation between ornamental traits and sperm quality could even become negative, for example due to trade-offs in allocation (Parker, 1998; Tazzyman *et al.*, 2009; Evans, 2010; Mautz *et al.*, 2013).

Shared condition dependence can be studied experimentally (e.g. Bonduriansky *et al.*, 2015), for instance by targeted inbreeding. In many taxa, inbreeding causes substantial stress to the inbred organism, because when recessive deleterious mutations become homozygous, the organism may suffer for instance from nonfunctional gene products reducing the effectiveness of physiological processes. In line with this, we found that captive zebra finches suffer from such inbreeding depression in terms of reduced sexual ornamentation and display (Bolund *et al.*, 2010), reduced sperm quality (Opatová *et al.*, 2016) and lower overall reproductive success (Forstmeier *et al.*, 2012). Hence, by adding an artificial stressor under controlled captive conditions, one can increase the total range of variation in male phenotypic condition. Here, we make use of inbreeding to manipulate the condition of birds in order to investigate the degree of shared condition dependence of sperm traits and of putative indicator traits. This allows us to assess the extent to which phenotypic indicators reflect sperm quality traits both in the presence and absence of inbreeding.

Previously, we studied a set of 24 male phenotypic traits as potential indicators of quality by comparing trait values between outbred males (F = 0) and males that had originated from the mating of full sibs (F = 0.25). We identified three indicator traits that were most strongly affected by inbreeding (Bolund *et al.*, 2010): inbred males showed a lower rate of courtship (Cohen's d = −1.18) and a more orange (less red) coloration of the beak (d = −1.04), and were smaller in size as measured by tarsus length (d = −0.90). Here, we focus on these three, potentially best phenotypic indicator traits. However, one should keep in mind that the above estimates contain sampling noise and estimation error: the three traits were selected from a long list for showing maximal effect size estimates, and we therefore expect the effects to be overestimated ('the winner's curse' (Forstmeier & Schielzeth, 2011) or 'regression towards the mean' (Barnett *et al.*, 2005; Kelly & Price, 2005)).

Previously, we also compared 10 sperm characteristics between inbred and outbred males (Opatová *et al.*, 2016) and found that inbred males had a higher 'proportion of abnormal sperm' within ejaculates (d = 1.40), lower curvilinear 'sperm velocity' (d = −0.74) and smaller total sperm length (d = −0.55). In the following, we use the 'proportion of functional sperm' rather than 'proportion of abnormal sperm', such that for all measures large trait values are associated with high fertility. The three sperm traits were the ones that showed maximal inbreeding depression (among the 10 traits studied), but were also picked because they predicted fertilization success (Mautz *et al.*, 2013; Bennison *et al.*, 2015). The latter is important because we are interested in estimating sperm quality as reflected in fertilization success (see Mautz *et al.*, 2013 for a detailed discussion). Hence, for sperm quality indicator traits, the issue of trait selection and overestimation of inbreeding depression are less important, because most of the nonfocal sperm traits ($n = 7$ morphometric traits) are unlikely indicators of sperm quality (see Mautz *et al.*, 2013; Opatová *et al.*, 2016).

In summary, we here focus on nine pairwise correlations between three selected phenotypic indicator traits and three putative measures of sperm quality. All nine correlations are expected to be positive under the phenotype-linked fertility hypothesis. We examine these correlations in three different sets of experimental birds: (1) birds from a domesticated population of zebra finches consisting of both inbred and outbred males, (2) birds from a recently wild-derived population of zebra finches, also consisting of a mix of inbreds and outbreds and (3) a larger sample of outbred males from the same recently wild-derived population. This allows us to examine the average strength of correlations between phenotypic indicators and sperm quality measures under two conditions: (1) a population of only outbred birds that contains less variation in male quality and (2) a mixed population of inbred and outbred males where experimental inbreeding led to increased variation in male quality. For the purpose of this study, using fully inbred (F = 0.25) vs. outbred (F = 0) males is ideal, because it maximizes differences in individual quality and should thus induce a clear positive correlation between indicator traits and sperm quality measures. Alternatively, one could use populations where males vary continuously in their inbreeding levels, but then weaker correlations are expected. Inbreeding is almost entirely absent in zebra finches that live in the wild in Australia (Knief *et al.*, 2015), but populations could experience increased inbreeding when colonizing islands (as happened in the zebra finches on Timor (Balakrishnan & Edwards, 2009)) or when going through population bottlenecks. Furthermore, other stressors such as food shortage, parasites or diseases can also increase between-individual variance in condition in the wild.

The aim of our study was to quantify the strength of a correlation rather than to test it against the null hypothesis of $r = 0$. The problem with such hypothesis testing is that, in practice, it often leads to biased reporting by leaving out traits or entire data sets that produce only nonsignificant relationships. We specifically aimed at avoiding any such bias that would inflate the strength of the reported correlation. We therefore report comprehensively on all three phenotypic traits, all three sperm traits and all three groups of birds where sperm traits have been measured (despite some tests having missing data, small sample sizes or inconsistencies in sampling). To reduce researcher flexibility in analysis and presentation (Simmons *et al.*, 2011; Forstmeier *et al.*, 2016), all arbitrary choices of how to analyse the data and what to present were made without considering obtained effect sizes or *P*-values.

Materials and methods

Study subjects and sampling design

We studied two independent captive populations, one domesticated and one recently wild-derived, held at the Max Planck Institute for Ornithology in Seewiesen, Germany (birds originated from populations #4 and #18 described in (Forstmeier *et al.*, 2007)). All birds were raised and housed in standardized captive conditions with standardized *ad libitum* food. The domesticated population was housed in a more constant environment (indoors; Bolund *et al.*, 2007; Atagan & Forstmeier, 2012), whereas the recently wild-derived population was housed in outdoor aviaries with greater temporal fluctuations in humidity and temperature (Ihle *et al.*, 2013).

For the purpose of studying inbreeding depression in sperm traits (Opatová *et al.*, 2016) and/or the relationship between sperm traits and fitness traits, we attempted to take sperm samples from 143 males ($n = 41$ domesticated birds; $n = 102$ recently wild-derived birds). Here, we examine this data set of sperm traits (available on Dryad https://doi.org/10.5061/dryad.4h245) in relation to phenotypic indicator traits.

The first set of birds consisted of 41 males from the domesticated population. This included 16 inbred males (pedigree inbreeding coefficient F = 0.25) originating from full-sib matings of 11 different families, and 25 outbred males (F = 0-0.016, median = 0) from 16 different families. Sperm was sampled only once in July 2011 when inbred males were 630–1375 days of age (median = 1311 days) and outbred males were 633–1373 days of age (median = 1333 days). This set of birds comprised all the survivors (7 of 18 inbreds and 10 of 18 outbreds) of a breeding experiment carried out in 2009 during which courtship rates had been quantified (see Forstmeier *et al.*, 2011).

The second set of birds consisted of 43 males from the recently wild-derived population. These birds had been bred for the study of inbreeding depression and consisted of 23 inbreds (F = 0.25) from seven different families and 20 outbreds (F = 0–0.031, median = 0) from 14 different families, matched for rearing conditions and age. Thirty-six of the 43 males took part in a breeding experiment in 2012 (where courtship rate was measured), and those were sampled for sperm three times (in April 2012, August 2012 and April 2013; three birds died before the third measurement). The other seven birds (five inbreds, two outbreds) were sampled for sperm only once in April 2012. At first sampling in April 2012, inbred males were 196-367 days of age (median = 286 days) and outbred males were 184–366 days of age (median = 319.5 days).

The third set of birds, also from the recently wild-derived population, consisted of 59 outbred males from 23 different families (F = 0–0.016, median = 0). These birds were all part of a breeding experiment during which courtship rate was measured (see Ihle *et al.*, 2015). All of them bred once in 2012, and 41 males bred a second time in 2013. These birds were sampled for sperm up to four times, in April 2012, August 2012, April 2013 and August 2013, always a few days before and after their breeding season. The birds were 230–367 days of age (median = 264 days) at first sperm sampling in April 2012.

Measurement of sperm traits

Sperm samples (~0.5–3 μl) obtained by massaging the cloacal protuberance were immediately diluted in preheated (40 °C) Dulbecco's modified Eagle's medium solution (Advanced D-MEM, Invitrogen, Carlsbad, CA, USA). An aliquot was pipetted onto a standard count slide (depth: 20 μm, two chambers; Leja, the Netherlands) that was placed on a heating table kept at a constant temperature of 40 °C for analysis of velocity. The rest of the sperm sample was fixed in 250 μl ~5% formalin for later analyses of morphology.

Sperm velocity was recorded for 45 sec simultaneously at eight different fields of the slide with a digital camera (UI-1540-C, Olympus) mounted on a microscope (CX41, Olympus) with a 100× magnification. Each field of recording was later analysed by the CEROS computer-assisted sperm analysis (CASA) system (Hamilton Thorne, Inc., Beverly, MA, USA). All tracked objects were visually inspected by a single person (JA) to manually exclude nonsperm objects as well as spermatozoa with a straight-line velocity under 20.5 μm s^{-1} (see also Laskemoen *et al.*, 2010; Cramer *et al.*, 2016). Thus, strongly abnormal immotile sperm were excluded from the velocity measurements. As the medium (D-MEM) does not contain any component to guide the spermatozoa towards one direction, curvilinear velocity (VCL) rather than straight-line velocity was used as our measurement of sperm swimming speed (Laskemoen *et al.*, 2010).

For each sample fixed in 5% formalin, a droplet was placed on a slide, air-dried and inspected under a light microscope (BX51, Olympus) under 400 × magnification. A picture was taken with a digital camera system (DP71, Olympus) and analysed with the software QuickPHOTO Industrial 2.3 (Olympus). The total length of 10 intact (normal) spermatozoa was measured, to yield an estimate of average sperm length for each sperm sample. In addition, 100 randomly chosen sperm were inspected by a single person (PO) and categorized as abnormal (deformities in head shape, mid-piece or tail). The counts of abnormal sperm (among 100 sperm) were log(n + 1)-transformed to approach normality and were then multiplied by −1 to yield a measure of sperm functionality (rather than abnormality).

Five domesticated males (two inbred and three outbred) and three males from the recently wild-derived population (two inbred and one outbred) never yielded sperm. In addition, 16 of a total of 353 sampling attempts led to incomplete data (missing data on either morphology or velocity).

Measurement of phenotypic indicator traits

Beak colour was scored using the Munsell colour scale (Forstmeier & Birkhead, 2004) when birds had reached sexual maturity (before they bred for the first time). On the same occasion, right tarsus length was measured to the nearest 0.1 mm (one male had an old leg fracture and could not be measured). Within each population, measurements were made by one person (domesticated: WF, wild-derived: UK).

In each set of birds, measurements of courtship rate were taken when birds participated in breeding experiments. Most males (112 of 143) participated in such breeding experiments, but the ones who did not experienced similar housing conditions until they were sampled for sperm once. Communal breeding aviaries were equipped with a video-surveillance system and contained six males and six females (half of them were inbred in two of three experiments). All breeding seasons lasted 4 months during which video-recording was continuous, covering the entire day. For the domesticated birds, we only recorded an artificial tree in the centre of each aviary on which, according to live observations, birds performed 35% of all courtships (Forstmeier *et al.*, 2011). For this population, all recordings were analysed, which sums to on average 1,639 h per male, yielding on average 344 courtships per male (range 2–1,108; see Forstmeier *et al.*, 2011 for further details). The wild-derived birds were recorded at four different positions: an artificial tree on which 69% of all courtships took place, two sets of 3–4 nest boxes and a set of perches (see Ihle *et al.*, 2015 for more details). For the breeding periods involving wild-derived birds, we only analysed the first hour of each day, because this is when copulations are most frequent (Forstmeier *et al.*, 2011). For this population, on average 82 h of recordings were analysed per male per season, yielding on average 58 courtships per male per season (range 4-380). For both domesticated and recently wild-derived birds, the duration of each male display to any female observed from a given camera position was timed (in seconds) by a single person (KM). This duration was divided by the total observation time from this camera position. The courtship rates obtained for each male were then summed across all camera positions. In this way, we avoided creating a bias for individuals with a preference for a certain courting location. Courtship rates were square-root-transformed to approach normality.

All sperm and phenotypic measurements were taken blindly with respect to male inbreeding status, and sperm measures were taken blind to phenotypic indicator traits and vice versa.

Timing of measurements

The timeline of all sampling events is schematically shown in Fig. 1. Each of the six measured traits is significantly repeatable over time, but the degree of repeatability varies among traits (Opatová *et al.*, 2016). Traits also differ in the time window of condition dependence: tarsus length mainly reflects growth

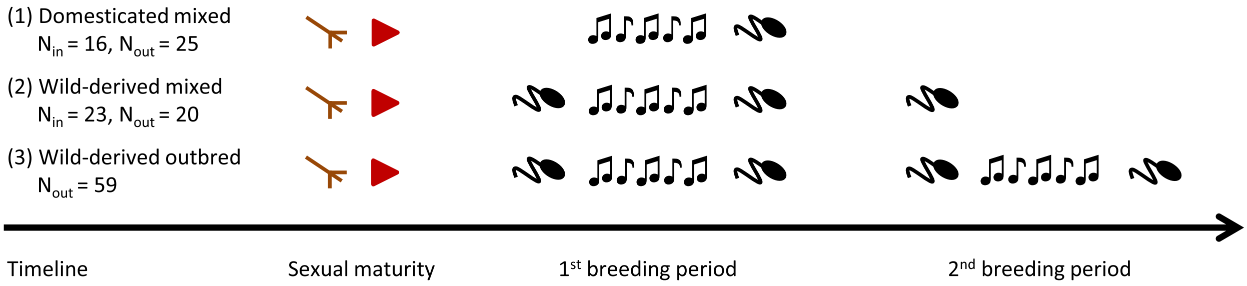

Fig. 1 Schematic timeline of sampling events for the three sets of zebra finches under study. Beak colour (red triangle) was scored, and tarsus length (schematic foot) was measured when the bird had reached sexual maturity. Courtship rate (musical notes) was assessed during periods of 4 months in communal breeding aviaries. Sperm samples (schematic sperm) were taken before and/or after breeding. N_{in} = number of inbred males (F = 0.25), N_{out} = number of outbred males (F = 0–0.031).

conditions during the nestling phase, whereas beak colour mostly reflects conditions 2–4 weeks prior to its measurement, whereas courtship rate signals current condition. However, each trait will also carry some information about permanent aspects of male condition or quality (like inbreeding depression; Bolund *et al.*, 2010).

Female mate choice can take place at multiple occasions. In zebra finches, the most important choice is that of the social partner, which usually takes place soon after birds reach sexual maturity. At this time, females might assess tarsus length, current beak colour and courtship intensity to predict the fertility of the partner which – in this species – they typically keep for a lifetime. Female mate choice also occurs during each breeding period: females then decide how often and with whom to copulate, which includes their social mate and potential extra-pair males. At this stage, short-term predictions of fertility are most relevant, because this is when sperm is transferred and eggs need to be fertilized. Our schedule of measuring targets both choice episodes. Beak colour and tarsus length were measured at the time when social pairing would normally take place and courtship rate was measured across each breeding period to minimize measurement error.

Sperm were sampled before and after the breeding period, because we did not want to interfere with reproduction. We assume that the mean trait value we measured is close to the sperm-trait value during the period when copulations take place.

Statistical analyses

Repeated measures of sperm traits and courtship rate from the same male were averaged to obtain one phenotypic value per individual. We then examined the Pearson correlation coefficient between each of the three phenotypic indicator traits and each of the three measures of sperm quality (nine correlations) for various sets and subsets of birds. This yielded a total of 63 pairwise Pearson correlation coefficients (outbreds: 9 correlations × 3 sets of birds, inbreds: 9 correlations × 2 sets of birds, correlations across outbreds and inbreds: 9 correlations × 2 sets of birds; Table S1). Correlations across inbreds and outbreds were calculated without statistically controlling for inbreeding status, because the aim is to capture the additional covariance induced by the shared condition dependence of the indicator traits and the sperm quality measures. Correlation coefficients r were subjected to Fisher's z-transformation ($Zr = 0.5 \times \ln((1 + r)/(1\text{-}r))$). This transformation leads to normally distributed values by 'stretching out' the values as they approach the boundaries of −1 and +1. Coefficients that are close to zero are practically unchanged (e.g. $r = 0.3$ becomes $Zr = 0.31$). The Zr values were then analysed as the dependent variable in mixed-effect models to estimate an overall mean value (intercept) or to estimate different intercepts for different sets of birds (using 'set of birds' as a fixed effect). In these mixed-effect models, we weighted each Zr value according to sample size (number of males n; weight = $(n-3)^{0.5}$ (Nakagawa & Cuthill, 2007)). All mixed-effect models were run using the lmer function of the lme4 package (Bates *et al.*, 2014) in R 3.1.3 (R Core Team, 2015). We also summarized Zr values using the meta.summaries function of the package rmeta (Lumley, 2012), which yielded practically identical results (not shown for brevity).

The characteristics of individual males from the two populations were not measured with equal precision. First, in the wild-derived population, we used average sperm traits based on repeated measures over a longer period (Fig. 1), whereas domesticated males were measured only once. Second, in the wild-derived population, courtship rate was measured only within the first hour of each day, whereas in the domesticated birds, it was measured throughout the day. To explore how this heterogeneity in measurement precision affected our conclusions, we repeated all analyses based on more standardized individual trait values. For this purpose, we used only the last sperm measurement available for each male from the wild-derived population, which also helped reducing the difference in male age between populations (67% of all sperm measurements discarded). Similarly, for the domesticated population, we only included courtship observations from the first hour of observation on each day (84% of courtships discarded). Individual phenotype values calculated with the full and the reduced data set were strongly correlated (sperm length: $r = 0.96$, sperm velocity: $r = 0.71$, sperm functionality $r = 0.70$, $n = 97–99$ wild-derived birds; courtship rate: $r = 0.92$, $n = 17$ domesticated birds).

Results

Correlations among outbred individuals

We examined the 27 correlation coefficients from outbred individuals (9 pairs of traits × 3 sets of birds) after Fisher's z-transformation in a weighted mixed-effect model without a common intercept but with 'set of birds' as a fixed effect (thereby estimating separate intercepts for each of the three sets of birds). To control for all levels of nonindependence, we fitted three random effects, namely the involved sperm trait (three levels), the indicator trait (three levels) and the pair of traits (nine levels). All random effect variance components were zero or negligible (see also below). However, correlations differed significantly between the three sets of birds (tested by comparing models with and without fixed effect: $\chi^2 = 6.6$, df = 2, $P = 0.037$). In the domesticated population (Fig. 2a), correlations

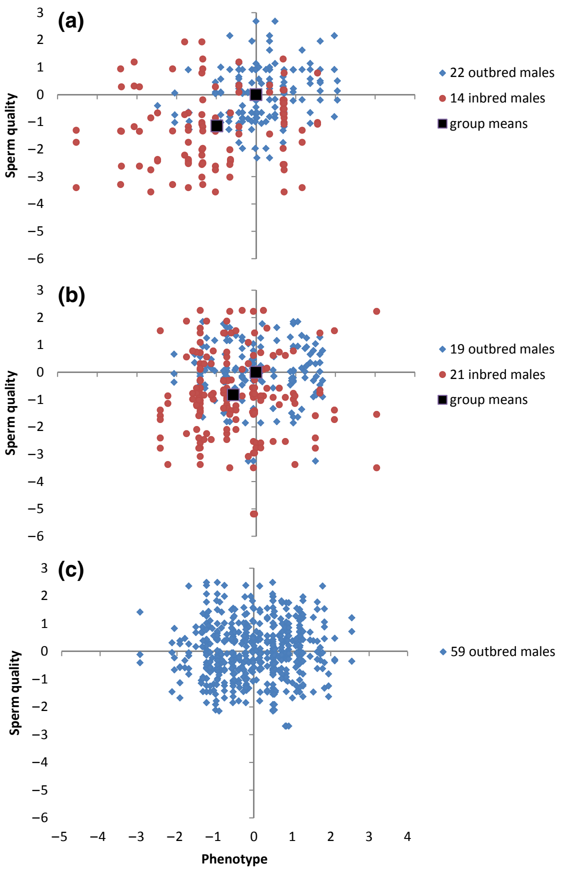

Fig. 2 Scatter plot showing measurements of sperm quality of individual males in relation to trait values for phenotypic indicators of quality of these males in three sets of birds (A: domesticated mixed outbred-inbred population, B: wild-derived mixed, C: wild-derived outbred only). Note that each male (total $n = 135$) is represented by up to nine data points, one for each combination of three sperm traits (proportion of functional sperm, sperm velocity, sperm length) with three phenotypic indicators (courtship rate, beak colour, tarsus length). For simultaneous illustration of nine trait combinations, all trait values were z-standardized to the between-individual standard deviation observed in the subset of outbred males. Hence, approximately 95% of data points from outbred males fall into the range from −2 to +2 standard deviations and their mean is centred on the origin. Note how trait values from inbred males are shifted towards lower trait values for measures of sperm quality and for measures of phenotypic indicators (inbreeding depression). Black squares mark the group mean values for inbred and outbred males, respectively. For exact sample sizes for each pair of traits in each subset of birds, see Table S1.

were significantly larger than zero (mean $Zr \pm SE = 0.29 \pm 0.11$, $z = 2.7$, $P = 0.007$), whereas correlations were close to zero in both sets of birds from the recently wild-derived population (Fig. 2b,c, Table 1), yielding an estimated weighted average of $Zr \pm SE = 0.03 \pm 0.10$ for the recently wild-derived population (intercept from a model on the subset of 18 correlations).

Correlations among inbred individuals

A mixed-effect model on the 18 correlation coefficients (*Zr* values) from inbred males (red data points in Fig. 2a,b) yielded qualitatively similar results to those from outbred birds. Random effects were again zero or negligible (see also below), and the two sets of birds differed significantly in the strength of correlations ($\chi^2 = 4.2$, df = 1, $P = 0.039$). Correlations were stronger in the domesticated population than in the recently wild-derived population (Table 1).

Correlations across inbred and outbred males

Inbreeding depression contributed to a stronger positive correlation between sperm traits and indicator traits (Table 1). This is because the downwards-shifted trait means for inbred males along both the x- and the y-axes in Fig. 2a,b induce an overall positive correlation across a mixed population of inbred and outbred individuals. In the domesticated population, the group mean value for inbred males was −0.99 (outbred standard deviations) for the phenotypic indicator traits (mean of three traits) and −1.14 for the sperm quality traits (mean of three traits), whereas outbred males are centred on zero (see 'group means' in Fig. 2a). These shifts were somewhat smaller for the recently wild-derived population (−0.57 and −0.82; Fig. 2b). A mixed-effect model on the 18 correlation coefficients (*Zr* values) measured across inbred and outbred males (i.e. not statistically controlling for inbreeding status) yielded similar results as those described above.

Table 1 Summary of correlation coefficients (average ± SE, Fisher-transformed *Zr* values) between phenotypic indicator traits and sperm quality measures in different subsets of birds. Bold print highlights correlations that are more than 1.96 SE away from zero. Sets of birds that comprise both inbred and outbred individuals are designated as 'mixed'. Average values for the three sets of birds combined ('overall') were obtained from another mixed model that contained 'set of birds' (three levels) as an additional random effect. For orientation, back-transformation of the highest *Zr* value (0.382) to a Pearson correlation coefficient yields 0.364. Correlations 'across groups' were calculated without statistically controlling for inbreeding status.

Set of birds	Within outbreds	Within inbreds	Across groups
(1) Domesticated mixed	**0.293 ± 0.108**	0.236 ± 0.139	**0.382 ± 0.077**
(2) Wild-derived mixed	−0.011 ± 0.104	−0.088 ± 0.122	0.056 ± 0.072
(3) Wild-derived outbred	0.055 ± 0.086	NA	NA
Overall	0.107 ± 0.106	0.066 ± 0.178	0.217 ± 0.167

Random effects were negligible (see also below) and the two populations differed significantly in the strength of correlations ($\chi^2 = 9.7$, df = 1, $P = 0.002$). Correlations were remarkably strong in the domesticated population (Z*r* ± SE = 0.38 ± 0.08, z = 4.9, $P < 0.001$) and again weak in the recently wild-derived population (Z*r* ± SE = 0.06 ± 0.07, z = 0.8, $P = 0.43$).

Controlling for measurement error

When phenotypes of males were calculated from a more limited data set that ensured approximately equal measurement error in the two populations, correlation coefficients were slightly lower (average of 63 Z*r* values was 0.11 (Table S2) compared to 0.14 (Table S1)), possibly due to a somewhat reduced data quality. However, overall these correlation coefficients were very similar to those based on the full data set described above (Pearson $r = 0.95$, $n = 63$ pairs of Z*r* values), and all above conclusions remain unchanged (see Table S3 for a modified version of Table 1).

Do some traits show stronger correlations than others?

Above we repeatedly report that the random effect estimates for 'sperm trait' and 'phenotypic trait' were zero or negligible. The maximum power for detecting nonzero random effects, which would imply differences in the correlations depending on which traits are considered, should be obtained when combining the nine 'across' correlations (Fig. 2a,b) with the 'within outbreds' correlations (Fig. 2c). We examined these 27 correlation coefficients jointly in a mixed-effect model with three random effects: 'set of birds', 'sperm trait', and 'phenotypic trait', each having three levels. Whereas 'set of birds' explained 10% of the total variance in Z*r* values, both other variance components were estimated to be zero, indicating smaller differences between traits than expected from sampling noise alone. Hence, we found no evidence that correlations differed depending on the particular sperm trait or phenotypic trait used.

Discussion

When males face a trade-off between investing in sexual signalling and sperm production, a negative correlation between signal intensity and fertility could result (Parker, 1998; Evans, 2010). However, when individual males differ with regard to how much they can invest overall, a positive relationship can emerge, such that individuals in good condition show both high levels of sexual signalling and high fertility. Only under the latter scenario can females use sexual signals as indicator traits of high male fertility. The sign and strength of correlations between sexually selected traits and fertility traits hence critically depends on the amount of variation in condition between males.

In line with the expectation that phenotypic indicator traits convey more information about male fertility when the differences in condition between males are experimentally increased, we found stronger correlations between indicator traits and sperm traits when examining populations that consisted of a mixture of inbred and outbred males. The correlation arose because inbred males showed both reduced expression of indicator traits and lower measures of sperm quality (Fig. 2). However, this effect of inbreeding depression on correlation coefficients was relatively small: Z*r* increased by 0.12 in the domesticated population (difference between 'across' and 'within-group' correlations, Table 1) and by 0.11 in the recently wild-derived population. Weak correlations would limit the fertility benefits that individual females would gain from paying attention to these indicator traits, but even small benefits could be sufficient for the evolution of female preferences for indicators of high condition. In zebra finches, the benefits of choosing a high-fertility partner would be particularly high, because pair bonds last for a lifetime, males vary in fertility, and females do not seem to alleviate fertility problems by seeking extra-pair copulations (Ihle *et al.*, 2013).

Inbreeding depression for phenotypic indicator traits was somewhat weaker in the recently wild-derived population (reduction by 0.57 outbred SDs) compared to the domesticated population (reduction by 0.99 SDs). This difference in effect size was expected for the following reason. Selecting the strongest and most significant effects in a long list of potential indicator traits should lead to overestimation of the true effects (winner's curse (Forstmeier & Schielzeth, 2011) or regression to the mean (Barnett *et al.*, 2005; Kelly & Price, 2005)). The study from which the three best indicator traits were selected (Bolund *et al.*, 2010), was partly based on the same individuals (31 of the 41 birds from the domesticated population used here), so the effect size estimate for this population is likely to be inflated. The effect size in an independent study should then be smaller and closer to the true mean, which may explain why inbreeding depression for these phenotypes was less strong in the wild-derived population. Note also, however, that we measured courtship rate in a more labour-intensive way (long-term courtship effort in communal breeding aviaries) than in the original study of Bolund *et al.* (2010), where courtship rate was measured in staged 5-min encounters. The change in method was necessary because we found that recently wild-derived birds did not readily court in such experimental test (wild-derived birds appeared more stressed when put in small cages and were therefore kept in large aviaries). We expected that our new measurement of long-term courtship effort would be a particularly sensitive indicator of male condition, potentially leading to stronger correlations with sperm quality, but this was not the case (weighted averages of Zr across Table S1: beak colour: 0.17, courtship rate: 0.01, tarsus length: 0.17). Overall, differences between phenotypic traits in their indicator value for variation in sperm quality were smaller than expected from sampling noise alone (i.e. the variance component due to trait was estimated as zero), so we refrain from further discussion.

Whereas differences in correlations depending on which indicator or sperm traits we used appeared negligible, we found unexpected differences between our two study populations. The aim of our study was to produce robust results by examining more than one set of birds and by following a strategy of reporting unconditional on reaching significance to avoid the inflation of effect sizes. However, these three 'replicates' yielded correlation coefficients that differed significantly in strength (tested using two degrees of freedom), and the correlations were clearly stronger in the domesticated than in the wild-derived population (see Fig. 2a vs. 2b, c). We have no plausible explanation for this difference, but note that it may be independent of domestication as such (comparison of birds from one domesticated population vs. two groups of birds from one wild-derived population). Yet, since the difference appeared repeatedly in the comparison among outbred males (column 2 in Table 1) and among inbred males (column 3 in Table 1), we still present some thoughts in the following paragraph.

First, the difference in the strength of phenotype–sperm quality associations between the two captive populations might be due to differences in baseline levels of inbreeding (not captured by the pedigree). However, this seems unlikely, because the baseline level of inbreeding is probably similar between the two populations, judging from the expected heterozygosity (domesticated He = 0.82, wild-derived He = 0.84) compared to birds from the wild (He = 0.94; data from Forstmeier *et al.*, 2007) where inbreeding is practically absent (Knief *et al.*, 2015). Second, other sources of stress with pleiotropic effects on sperm and quality indicators could be present in the domesticated but not in the recently wild-derived population. This could then lead to positive correlations even within groups of males with the same inbreeding status in the domesticated population. If that were the case, one might expect increased variation in phenotypic traits in the domesticated relative to the wild-derived population. Among outbred males, within-group SDs of sperm traits were indeed larger (on average 31%) in the domesticated population, but the pattern was the opposite for the indicator traits (15% smaller SD). Our measurements of male sperm quality should have been more reliable for the wild-derived population (where we averaged 3–4 measurements per male) than for the domesticated population (only a single measurement). Hence we would have expected a stronger, not a weaker, correlation in the recently wild-derived population.

Correlations between indicator traits (including beak colour and courtship rate) and sperm traits (including sperm length, velocity and proportion of abnormal sperm) have been reported previously (Birkhead & Fletcher, 1995; Birkhead *et al.*, 1998) for the zebra finch population from which our domesticated birds were derived. For comparison, we calculated average correlation coefficients across all pairs of traits that we also measured (again using Fisher's transformation, and inverting the proportion of abnormal sperm to a positive measure of quality). The first study (Birkhead & Fletcher, 1995) was based on 10 outbred males and yielded an average correlation of Zr ± SE = 0.24 ± 0.18 (mean across four pairs of traits), whereas the second study (Birkhead *et al.*, 1998) was based on two groups of 31 outbred males each and yielded an average correlation of Zr ± SE = 0.08 ± 0.05 (n = 12 estimates). The weighted average for those two earlier studies was Zr ± SE = 0.100 ± 0.052, and hence almost significantly larger than zero (P = 0.055). Although this value is lower than the one we found among our domesticated outbred birds (Zr = 0.29 ± 0.11), the difference is not significant (P = 0.11). As in our study, the variance components for the random effects of 'indicator trait' and 'sperm trait' were estimated as zero.

More studies are necessary to elucidate whether and under which conditions indicator traits may be indicative of sperm quality to an extent that it would pay females to select males on the basis of these traits. Our data show that the indicator function of the phenotypic traits becomes more reliable when variation in inbreeding increases the differences in condition between males. Other sources of stress (e.g. oxidative stress, social stress, food shortage, parasite load, diseases) could also induce differences in condition among individuals and could cause positive associations between indicators and sperm traits. In general, however, the predictive power of phenotypic indicator traits may remain limited, here explaining maximally 15% of the variance in sperm traits related to fertilization success.

Acknowledgments

We thank Sanja Janker, Johannes Schreiber, Anaïs Edme, Martina Soudková and Oldřich Tomášek for help in data collection. This study was supported by the Max Planck Society and the Czech Science Foundation grant no. P506/12/2472.

References

Atagan, Y. & Forstmeier, W. 2012. Protein supplementation decreases courtship rate in the zebra finch. *Anim. Behav.* **83**: 69–74.

Balakrishnan, C.N. & Edwards, S.V. 2009. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics* **181**: 645–660.

Barnett, A.G., van der Pols, J.C. & Dobson, A.J. 2005. Regression to the mean: what it is and how to deal with it. *Int. J. Epidemiol.* **34**: 215–220.

Bates, D., Maechler, M., Bolker, B. & Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4.

Bennison, C., Hemmings, N., Slate, J. & Birkhead, T. 2015. Long sperm fertilize more eggs in a bird. *Proc. R. Soc. B Biol. Sci.* **282**: 20141897.

Birkhead, T.R. & Fletcher, F. 1995. Male phenotype and ejaculate quality in the zebra finch *Taeniopygia guttata*. *Proc. R. Soc. B Biol. Sci.* **262**: 329–334.

Birkhead, T.R., Fletcher, F. & Pellatt, E.J. 1998. Sexual selection in the zebra finch *Taeniopygia guttata*: condition, sex traits and immune capacity. *Behav. Ecol. Sociobiol.* **44**: 179–191.

Bolund, E., Schielzeth, H. & Forstmeier, W. 2007. Intrasexual competition in zebra finches, the role of beak colour and body size. *Anim. Behav.* **74**: 715–724.

Bolund, E., Martin, K., Kempenaers, B. & Forstmeier, W. 2010. Inbreeding depression of sexually selected traits and attractiveness in the zebra finch. *Anim. Behav.* **79**: 947–955.

Bonduriansky, R., Mallet, M.A., Arbuthnott, D., Pawlowsky-Glahn, V., Jose Egozcue, J. & Rundle, H.D. 2015. Differential effects of genetic vs. environmental quality in *Drosophila melanogaster* suggest multiple forms of condition dependence. *Ecol. Lett.* **18**: 317–326.

Cramer, E.R., Ålund, M., McFarlane, S.E., Johnsen, A. & Qvarnström, A. 2016. Females discriminate against heterospecific sperm in a natural hybrid zone. *Evolution* **70**: 1844–1855.

Evans, J.P. 2010. Quantitative genetic evidence that males trade attractiveness for ejaculate quality in guppies. *Proc. Biol. Sci.* **277**: 3195–3201.

Forstmeier, W. & Birkhead, T.R. 2004. Repeatability of mate choice in the zebra finch: consistency within and between females. *Anim. Behav.* **68**: 1017–1028.

Forstmeier, W. & Schielzeth, H. 2011. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behav. Ecol. Sociobiol.* **65**: 47–55.

Forstmeier, W., Segelbacher, G., Mueller, J.C. & Kempenaers, B. 2007. Genetic variation and differentiation in captive and wild zebra finches (*Taeniopygia guttata*). *Mol. Ecol.* **16**: 4039–4050.

Forstmeier, W., Martin, K., Bolund, E., Schielzeth, H. & Kempenaers, B. 2011. Female extrapair mating behavior can evolve via indirect selection on males. *Proc. Natl. Acad. Sci. USA* **108**: 10608–10613.

Forstmeier, W., Schielzeth, H., Mueller, J.C., Ellegren, H. & Kempenaers, B. 2012. Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. *Mol. Ecol.* **21**: 3237–3249.

Forstmeier, W., Wagenmakers, E.-J. & Parker, T.H. 2016. Detecting and avoiding likely false-positive findings - A practical guide. *Biol. Rev*. doi: 10.1111/brv.12315.

Ihle, M., Kempenaers, B. & Forstmeier, W. 2013. Does hatching failure breed infidelity? *Behav. Ecol.* **24**: 119–127.

Ihle, M., Kempenaers, B. & Forstmeier, W. 2015. Fitness benefits of mate choice for compatibility in a socially monogamous species. *PLoS. Biol.* **13**: e1002248.

Kelly, C. & Price, T.D. 2005. Correcting for regression to the mean in behavior and ecology. *Am. Nat.* **166**: 700–707.

Knief, U., Hemmrich-Stanisak, G., Wittig, M., Franke, A., Griffith, S.C., Kempenaers, B. *et al.* 2015. Quantifying realized inbreeding in wild and captive animal populations. *Heredity* **114**: 397–403.

Laskemoen, T., Kleven, O., Fossøy, F., Robertson, R.J., Rudolfsen, G. & Lifjeld, J.T. 2010. Sperm quantity and quality effects on fertilization success in a highly promiscuous passerine, the tree swallow *Tachycineta bicolor*. *Behav. Ecol. Sociobiol.* **64**: 1473–1483.

Lumley, T. 2012. rmeta: Meta-analysis. R package version 2.16. http://CRAN.R-project.org/package=rmeta.

Mautz, B.S., Møller, A.P. & Jennions, M.D. 2013. Do male secondary sexual characters signal ejaculate quality? a meta-analysis. *Biol. Rev.* **88**: 669–682.

Nakagawa, S. & Cuthill, I.C. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* **82**: 591–605.

Opatová, P., Ihle, M., Albrechtová, J., Tomášek, O., Kempenaers, B., Forstmeier, W. *et al.* 2016. Inbreeding depression of sperm traits in the zebra finch *Taeniopygia guttata*. *Ecol. Evol.* **6**: 295–304.

Parker, G. 1998. Sperm competition and the evolution of ejaculates: towards a theory base. *Sperm. Competition. Sexual. Selection.* **3**: 54.

R Core Team. (2015) R: *A language and environment for statistical computing* . R Foundation for Statistical Computing, Vienna, Austria.

Sheldon, B.C. 1994. Male phenotype, fertility, and the pursuit of extra-pair copulations by female birds. *Proc. R. Soc. B Biol. Sci.* **257**: 25–30.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**: 1359–1366.

Tazzyman, S.J., Pizzari, T., Seymour, R.M. & Pomiankowski, A. 2009. The evolution of continuous variation in ejaculate expenditure strategy. *Am. Nat.* **174**: E71–E82.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

Table S1 All 63 Pearson correlations examined in this study.

Table S2 Pearson correlations as in Table S1, but based on a reduced data set (standardized data quality to reach approximately equal measurement error in the different populations).

Table S3 Summary of correlation coefficients as in Table 1, but based on a reduced data set (with equal measurement error in individual phenotypic values in the tree sets of birds).

Data deposited at Dryad: doi:10.5061/dryad.515c6

Received 19 October 2016; accepted 2 March 2017