MOLECULAR ECOLOGY

Molecular Ecology (2009) **18**, 2746–2765 doi: 10.1111/j.1365-294X.2009.04247.x

INVITED REVIEW

A quantitative review of heterozygosity–fitness correlations in animal populations

J. R. CHAPMAN,* S. NAKAGAWA,† D. W. COLTMAN,‡ J. SLATE§ and B. C. SHELDON*

*Edward Grey Institute, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK, †Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand, ‡Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E9, §Department of Animal and Plant Sciences, University of Sheffield, Alfred Denny Building, Sheffield S10 2TN, UK

Abstract

The ease of obtaining genotypic data from wild populations has renewed interest in the relationship between individual genetic diversity and fitness-related traits (heterozygosity–fitness correlations, or HFC). Here we present a comprehensive meta-analysis of HFC studies using powerful multivariate techniques which account for nonindependence of data. We compare these findings with those from univariate techniques, and test the influence of a range of factors hypothesized to influence the strength of HFCs. We found small but significantly positive effect sizes for life-history, morphological, and physiological traits; while theory predicts higher mean effect sizes for life-history traits, effect size did not differ consistently with trait type. Newly proposed measures of variation were no more powerful at detecting relationships than multilocus heterozygosity, and populations predicted to have elevated inbreeding variance did not exhibit higher mean effect sizes. Finally, we found evidence for publication bias, with studies reporting weak, nonsignificant effects being under-represented in the literature. In general, our review shows that HFC studies do not generally reveal patterns predicted by population genetic theory, and are of small effect (less than 1% of the variance in phenotypic characters explained). Future studies should use more genetic marker data and utilize sampling designs that shed more light on the biological mechanisms that may modulate the strength of association, for example by contrasting the strength of HFCs in mainland and island populations of the same species, investigating the role of environmental stress, or by considering how selection has shaped the traits under investigation.

Keywords: fitness, inbreeding, mean d^2, multilocus heterozygosity, multivariate meta-analysis, population demography

Received 17 March 2009; revision received 24 April 2009; accepted 25 April 2009

Introduction

Evolution occurs when variation in fitness is transmitted from one generation to the next, and when that variation has a genetic basis. Hence, understanding how genetic variation underpins individual fitness in populations is crucial to our understanding of evolution (Merilä & Sheldon 1999; Merilä & Crnokrak 2001; Barton & Keightley 2002; Hansson & Westerberg 2002; Ellegren & Sheldon 2008; Leinonen *et al*. 2008). There have been many approaches taken to try to understand the genetic causes of variation in fitness in populations, but an increasingly popular approach has been to seek to relate within-individual variation (some measure related to heterozygosity) at genetic marker loci to variation in fitness, or variation in characters that are potentially related to fitness. This increasing popularity is probably driven by two developments. First, the increased ease in obtaining genetic data, and second the increased focus on long-term population studies that allow characterization of selection in wild populations (see Kruuk & Hill 2008). The primary aim of this quantitative review is to assess the evidence for relationships between within-individual variation and fitness traits in animal populations. We begin with a brief review of the theoretical framework before the analysis of existing patterns.

Correspondence: J. R. Chapman, Fax: +44(0) 1865 271168; E-mail: joanne.chapman@zoo.ox.ac.uk

A brief review of hypotheses explaining heterozygosity-fitness correlations

Estimates of the relationship between individual genetic diversity and individual measures of life-history, morphological and physiological traits have become known collectively as heterozygosity-fitness correlations (HFC). Three hypotheses have been proposed to account for the existence of significant HFCs (reviewed in Hansson & Westerberg 2002), all of which assume that genetic diversity at marker loci reflects genetic diversity at loci that affect trait variation. First, functional overdominance may occur when the markers used to estimate genetic variation are themselves expressed (e.g. allozyme loci, MHC loci), have an effect on fitness, and are thus under direct selection (reviewed in David 1998). This is often referred to as the 'direct effect' hypothesis (Mitton 1997; David 1998). While this hypothesis was important for early studies of HFCs, when the use of allozyme markers was widespread, it does not readily explain why fitness is often found to be correlated with genetic diversity measured with markers assumed to be selectively neutral such as microsatellites (Goldstein & Schlötterer 1999), although see Price *et al*. (1997), Ranum & Day (2002), Streelman & Kocher (2002) and Westgaard & Fevolden (2007) for examples of non-neutral microsatellites.

Two alternative hypotheses have been proposed to account for associations between neutral markers and fitness traits. The first, the 'local effect' hypothesis (David *et al*. 1995; Lynch & Walsh 1998; pp. 288–290; Hansson & Westerberg 2002), invokes associative overdominance (Ohta 1971) as the explanation: the apparent fitness increase with increasing heterozygosity at marker loci is due to nonrandom association of these loci with loci affecting fitness — that is, the marker and fitness loci are in linkage disequilibrium (LD, David 1998). A classic model of this is repulsion phase disequilibrium, where two linked loci with segregating deleterious recessives produce gametes which carry the recessive at only one locus; if a marker is linked to these two loci, heterozygotes at this locus will have higher fitness because they are homozygous for neither deleterious recessive. One common misconception is that these loci need to be in close chromosomal proximity (i.e. physically linked). While this may often be the case, it is possible for physically unlinked loci to be in linkage disequilibrium due to a wide range of demographic processes (Briscoe *et al*. 1994; Bierne *et al*. 2000; Hedrick 2005), and it has thus been suggested that gamete phase disequilibrium would be a better term for this phenomenon (Crow & Kimura 1970; Hedrick 2005). Local effects have been increasingly implicated in HFCs (e.g. Hansson *et al*. 2001; Acevedo-Whitehouse *et al*. 2006; Lieutenant-Gosselin & Bernatchez 2006; von Hardenberg *et al*. 2007). While it might seem surprising that many studies have chosen markers that just happen to be in LD with the fitness trait under investigation, recent empirical studies have shown that LD can extend over many hundreds of kilobases, and be maintained for hundreds of generations (Laan & Pääbo 1997; Wright *et al*. 1999; Reich *et al*. 2001; Nordborg & Tavaré 2002; Hansson *et al*. 2004; although see also Bierne *et al*. 2000; Dunning *et al*. 2000; Edwards & Dillon 2004).

The final hypothesis invoked to explain HFCs detected with neutral markers is the 'general effect' hypothesis (David *et al*. 1995; Lynch & Walsh 1998; Hansson & Westerberg 2002). Here the apparent fitness increase with increasing heterozygosity at marker loci is due to the nonrandom association of diploid genotypes in zygotes and reflects the fitness cost of homozygosity at loci throughout the genome — that is the marker and fitness loci are in identity disequilibrium (ID, David 1998). ID is caused by variance in the inbreeding coefficient, f, of individuals within a population, because inbred individuals will be relatively homozygous throughout their genome due to recent allelic co-ancestry, and as such will also be homozygous at marker loci (Weir & Cockerham 1973; Hansson & Westerberg 2002; Coltman & Slate 2003), whereas in relatively outbred individuals the coupling of heterozygosity at marker and fitness loci will be weaker. Populations suffering outbreeding depression would be expected to exhibit negative HFCs, whereas inbreeding depression will result in positive HFCs. The general effect hypothesis is the only hypothesis invoking variance in inbreeding as an explanation for HFCs (Slate *et al*. 2004). General effects, and by extension inbreeding depression, are commonly suggested to explain HFCs in natural populations (e.g. Rossiter *et al*. 2001; Bensch *et al*. 2006; Fossøy *et al*. 2008; Rijks *et al*. 2008) although the extent to which heterozygosity measured at a few loci can truly reflect levels of genome-wide diversity is a contentious issue (Balloux *et al*. 2004; Slate *et al*. 2004; Hansson & Westerberg 2008; although see Aparicio *et al*. 2007).

As the general effect hypothesis relies on variance in inbreeding within populations, if this hypothesis is correct, the strength of the relationship between heterozygosity and fitness will depend on the variance of f in the population; hence, highly inbred populations with low variance in f would not be expected to exhibit strong HFCs due to general effects any more than outbred populations would (Balloux *et al*. 2004; Slate *et al*. 2004; Overall *et al*. 2005). Similarly, populations that have undergone historical or prolonged bouts of inbreeding may have successfully purged much of their mutational load (Charlesworth & Charlesworth 1987; Crnokrak & Barrett 2002), and the relationship between heterozygosity and fitness attributable to general effects might thus be lower (Reed & Frankham 2003), although the degree of purging will be dependent on the extent to which deleterious alleles result in lowered fitness (Wang *et al*. 1999; Crnokrak & Barrett 2002), and whether the purging occurs via drift or nonrandom mating (Glémin 2003).

Clearly then, the demographic history of populations is potentially an important factor in generating HFCs. Local effects are more likely to be detected when levels of LD are high, hence local effect HFCs might be expected in small populations due to genetic drift (Hill & Robertson 1968; David 1998), populations that have recently expanded in numbers, for example after release from a genetic bottleneck or expansion into a new range (Wall *et al*. 2002; Gaut & Long 2003) and in recently admixed populations (Briscoe *et al*. 1994). General effects are more likely to be detected in those populations with a high variance in the inbreeding coefficient *f*, as levels of ID will be elevated in such populations, although genetic purging in some populations may act to obscure this relationship somewhat. It is thus surprising that many studies have pooled samples from multiple populations or subpopulations (e.g. Acevedo-Whitehouse *et al*. 2005; Gage *et al*. 2006; Ortego *et al*. 2007; Välimäki *et al*. 2007). Demographic impacts on the strength of HFCs may be obscured in such studies. Even worse, spurious relationships may be generated, if combining results from several populations reveals HFCs that are not present within the component populations (Slate & Pemberton 2006). Previous meta-analyses in vertebrate populations have also largely failed to take population demographic history into account (although see Reed & Frankham 2001). Strong effects in small and/or inbred populations may have been diluted by weak or nonexistent effects in large and/or outbred populations when results are pooled for meta-analysis, or alternately could have spuriously inflated global estimates of mean effect size, depending on the overall contribution in the literature of studies from inbred and outbred populations.

Quantitative genetics of fitness and nonfitness traits

An alternative approach to understanding genetic influences on fitness, and related traits, in wild populations has involved the use of the tools and conceptual framework of quantitative genetics. Several comparative studies in the 1980s (Gustafsson 1986; Mousseau & Roff 1987; Roff & Mousseau 1987) confirmed the expectation that traits closely associated with individual fitness would have lower heritabilities than traits less tightly linked with fitness. More recent work has shown that the low heritability of fitness traits can be accounted for by high levels of residual variance (especially environmental variance) rather than low levels of additive genetic variance (Price & Schluter 1991; Kruuk *et al*. 2000; Wilson *et al*. 2006), and that, when scaled correctly, fitness traits can have higher additive genetic variance than nonfitness traits (Houle 1992; Kruuk *et al*. 2000; Merilä & Sheldon 2000; Barton & Keightley 2002; McCleery *et al*. 2004; Coltman *et al*. 2005; Blomqvist 2009). Since fitness traits are expected to have a more complex polygenic architecture than morphological or physiological traits (Merilä & Sheldon 1999), this may also result in greater opportunity for correlations between fitness traits and neutral marker loci. Given that inbreeding depression is due to nonadditive genetic effects (Charlesworth & Charlesworth 1999; Slate *et al*. 2000; Roff & Emerson 2006; but see Hill *et al*. 2008), and the evidence that fitness traits are more strongly influenced by inbreeding (e.g. Crnokrak & Roff 1995; Van Buskirk & Willi 2006), then we may predict that (i) the relatively higher ratio of dominance to additive variance, and polygenic architecture, of fitness traits should result in stronger HFCs for fitness traits when compared with nonfitness traits, and (ii) that this difference should be especially pronounced in inbred populations.

Scope of this review

Here we report a multivariate meta-analysis of studies that have assessed the association between heterozygosity (also commonly referred to as marker-based diversity, MBD) and variation in measures of fitness in animal populations. Meta-analysis provides a quantitative approach to synthesize research findings addressing a common research question; this quantitative approach allows the assessment of the importance of potential causes of variation in the strength of relationships (e.g. methodology, or in the case of HFCs, population history, type of fitness measures), and also allows for the exploration of publication bias. Meta-analysis has been used several times to address the importance of individual genetic diversity for individual fitness among animal (Britten 1996; Reed & Frankham 2001; Coltman & Slate 2003; Reed & Frankham 2003) and plant (Leimu *et al*. 2006) species. However, only one study (Coltman & Slate 2003) concentrated on neutral microsatellite markers, which have now largely replaced other types of markers such as allozymes and restriction fragment length polymorphisms (RFLP) as the marker of choice in such studies, as direct effects of the scored loci are assumed to be negligible. Apart from the very large number of studies published since, there is also growing acceptance of hypotheses relating population demography to the existence of HFCs (e.g. Rowe *et al*. 1999; Hansson & Westerberg 2002; Cena *et al*. 2006; Leimu *et al*. 2006), although there has been no formal assessment of the explanatory power of these hypotheses. All previous HFC meta-analyses have been conducted using univariate methods. Recently, an accessible implementation of multivariate methods which use a mixed-effects model framework has become available (Hedges 1983; Nam *et al*. 2003; Nakagawa *et al*. 2007). Such methods help to solve an important and common problem in meta-analysis: that of nonindependence of effect size estimates, for example when a disproportionately large number of effect size estimates are provided for one study population, or species (Nakagawa *et al*.

2007). These methods also allow for the nesting of studies by taxonomic grouping to correct a further possible source of nonindependence in data sets: that of shared evolutionary history. Furthermore, these methods also allow simultaneous consideration of the effect of multiple moderator variables.

This paper aims to review the current literature on HFCs by collating all HFC estimates published using microsatellite markers in animal populations up to May 2008, along with any unpublished studies we could identify. We had seven principal objectives: (i) to compare and contrast results from univariate and multivariate meta-analyses; (ii) to assess the evidence for a continued publishing bias in this field in favour of large effect sizes; (iii) to identify temporal trends in HFC studies; (iv) to assess whether mean effect sizes depend upon the genetic metric used; (v) to test whether fitness, morphological and physiological traits differ in the strength of effect sizes reported; (vi) to test whether population demographic history has a detectable impact on the strength of HFCs; and (vii) to suggest which avenues of future HFC research might be most fruitful.

Methods

Data collection

In May 2008, we used keyword searches in Web of Knowledge (http://apps.isiknowledge.com) using a combination of the following search terms: heterozygosity, heterosis, inbreeding, genetic diversity, marker, MLH, SH, IR, d^2, HFC and fitness. These methods were successful in retrieving all studies included in Coltman & Slate (2003; hereafter C&S), and we were therefore confident that our keyword searches were appropriate. A cited reference search was also conducted of all papers citing C&S, or Reed & Frankham (2001). In addition, we solicited unpublished results by emailing rquests to two widely subscribed mailing lists: EvolDir (http://evol.mcmaster.ca/evoldir.html, 6784 subscribers) and the Animal Gene Mappers Discussion Group (www.animalgenome.org/community, 1559 subscribers). Very few unpublished results were received via this route, and of these only two satisfied the criteria for inclusion listed below. Some of the studies listed as unpublished in C&S have now been published, while others remain unpublished. We did not include studies of plants or fungi in the scope of this review (see Leimu *et al*. 2006 for a recent appraisal of HFCs in plants).

To be included in the meta-analysis, studies (both published and unpublished) had to satisfy the following criteria: (i) having been conducted on a single population of animals from a distinct geographical area (this included wild, domestic and captive populations); (ii) genetic variation had been quantified using microsatellites and expressed as multilocus, rather than single locus, heterozygosity; (iii) at least one of the following genetic metrics had been used to quantify individual variation: MLH, SH, IR, mean d^2, or St d^2 (standardized d^2 see Appendix S1 Supporting information for further details); (iv) relationships had been quantified using a statistical measure that could be converted to an estimate of effect size — namely r, t, F, χ^2, or the exact P value; (v) HFCs were quantified between individuals within the population, rather than groups (such as family groups, e.g. Seddon *et al*. 2004); (vi) it was possible to determine the direction of the effect (see Appendix S1, Supporting information).

The two types of univariate meta-analysis described below where chosen to facilitate comparisons with earlier meta-analyses of HFCs (e.g. Britten 1996; Reed & Frankham 2001; C&S). We then describe a more comprehensive multivariate analysis that allows the modelling of multiple predictor variables simultaneously, allowing us to compare and contrast the results of these meta-analytical approaches.

Univariate meta-analyses

We considered three categories of trait type, following other authors in this general field (e.g. Mousseau & Roff 1987; C&S): morphological (M) traits, such as size and shape; life-history (LH) traits, such as survival and breeding success; and physiological (P) traits, such as parasite load or infection intensity. We also categorized traits as fitness and nonfitness traits (see Appendix S1, Supporting information). We converted the results of each reported HFC statistical analysis to r, the equivalent of the Pearson product moment correlation coefficient, a common measure of effect size. This measure of effect size was chosen to be consistent with the C&S meta-analysis, and because it is possible to convert most commonly reported statistical metrics to r (Rosenthal 1991). We estimated r for each statistical analysis reported based on MLH, SH, IR, d^2 or St d^2 separately, using the statistical calculator provided in the MetaWin software package version 2.0 (Rosenberg *et al*. 2000). Details of how each statistical metric was converted to r can be found in Rosenthal (1991, 1994) and C&S. We used Fisher's transformation, Zr, for statistical calculations, and then back-transformed to r for presentation. The homogeneity of weighted mean effect sizes was testing with the Q_T-statistic (Hedges & Olkin 1985; Rosenthal 1991; Shadish & Haddock 1994).

It is important to account for psuedoreplication of results in meta-analyses. We used three approaches to do this: first, we treated all effects as independent data in a univariate analysis (i.e. psuedoreplication was ignored); second, we repeated the univariate study unit average analysis of C&S; and third, we used a linear mixed-effects model with random factors to account for levels of psuedoreplication. Further details of these methods can be found below, and in Appendix S1, Supporting information.

Assessing the impact of sample size and evidence for publication bias

The relationship between sample size and effect size was investigated with funnel plots. We assessed temporal trends by regressing the number of individuals, markers and genotypes assayed against time. We used a $\log_{10}$ scale for number of individuals, markers and genotypes, as one paper (Slate *et al*. 2004) had sample sizes for these variables approximately an order of magnitude larger than most other studies. Tests for publication bias were conducted by comparing the mean effect sizes of published and unpublished studies, inspection of funnel plots, and via the trim and fill technique (Duvall & Tweedie 2000), which estimates the number of studies missing due to publication bias against nonsignificant results of small sample size, the mean effect size of such studies, and the influence this has on the overall mean effect size estimated by the meta-analysis. To determine whether publication bias has reduced since the C&S meta-analysis suggested this as a potential problem in this field, we conducted a second trim and fill analysis of papers published after May 2004, 1 year after the publication of C&S, to allow for dissemination of the findings of C&S and to remove any bias due to papers in press.

Multivariate meta-analysis

For the mixed-effects meta-analysis, transformed effect sizes were weighted by the variance, calculated by the reciprocal of the sum of their conditional variance:

$$Var = \frac{1}{n-3},$$

where n is the number of individuals included in the study (Raudenbush 1994; Nakagawa *et al*. 2007). Ninety five per cent confidence intervals were calculated for each mean effect size as,

$$\bar{Z}_r \pm \frac{1.96}{\sqrt{N-3k'}}$$

where N is the sum of all effect size sample sizes and k is the number of effect sizes included (Hedges & Olkin 1985). Testing for homogeneity of effect sizes using the Q_T statistic may not accurately reflect data heterogeneity in linear mixed effects model (LMM, Nakagawa *et al*. 2007), so for the multivariate meta-analysis we also calculated Q_{REML}, the residual heterogeneity in random-effects models which takes into account random variation in effect sizes between studies (Nakagawa *et al*. 2007). More specific details of the mixed-effects meta-analysis can be found as Supporting information, Appendix S1.

For all three of the meta-analytical approaches listed above, we calculated weighted mean effect size for each combination of genetic metric and trait type separately, to ensure that statistically conservative effect sizes were estimated for biologically meaningful groupings of data. We assumed a negligible effect of outbreeding depression, given that this has rarely been shown in wild animal populations (Frankham 1995a; Pusey & Wolf 1996), and has been shown to be of only one-tenth the magnitude of inbreeding depression in a captive population of Goeldi's marmosets (Lacy *et al*. 1993; Frankham 1995a).

Assessing the importance of demographic history and inbreeding

Demographic history might have a large, and often undetected, impact on the strength of HFCs in populations. We thus attempted to account for this variation by scoring populations on three demographic parameters for which information was available for most populations. To test for the influence of population demography, we scored each population based on three demographic criteria, with each study assessed for whether or not the population did or did not fit into that category. The three criteria we used to score populations were as follows: (i) whether the population had passed through a genetic bottleneck or founding event within the previous 20 generations. Such events cause an increase in inbreeding variance due to small instantaneous population size, but this effect weakens over time as the equilibrium associative overdominance for small populations is recovered (Weir *et al*. 1980; Bierne *et al*. 2000). Twenty generations was chosen as it has been shown in *Drosophila* that 20 generations after an inbreeding event is sufficient to significantly reduce the amount of inbreeding depression, which is likely mediated via purging of recessive deleterious alleles of strong effect (Fowler & Whitlock 1999). (ii) Whether the effective population size was between 50 and 500. These values were chosen based on Franklin's 50/500 rule (Franklin 1980) which suggests that an effective population size (*Ne*) of less than 50 is unlikely to maintain typical levels of heritable variation even in the short term, while populations with effective sizes larger than 500 are likely to maintain sufficient heritable variation in the long term. Thus, we grouped together populations with effective sizes lower than 50 and greater than 500 as being those populations likely to have low variance in inbreeding, and populations with effective sizes of 51–499 as being those populations most likely to show elevated levels of inbreeding variance. When an accurate estimate of effective population size was not available, but an estimate of total population size was, we divided *Ne* by 0.11, following a review by Frankham (1995b) showing that in wild populations, the average ratio of *N*/*Ne* is 0.11. (iii) Whether the population had a highly skewed mating system, a high degree of natal philopatry and/or was a small enclosed system with very little

immigration. These factors are all likely to increase inbreeding variance by reducing the pool of available mates. For each category, the population was scored as a one if the answer was 'yes', and a zero if the answer was 'no' or 'unknown'. Thus, we produced a four-point ordinal scale ranging from 0, being those populations likely to exhibit lowest variance in inbreeding, to 3, being those populations likely to exhibit the highest variance in inbreeding. This was a conservative analysis, as populations for which this information was unknown were scored as a 0 in that category. We excluded all captive and domestic populations from these analyses, as demographic processes are likely to be quite different in nonwild settings. More sophisticated analyses, such as using estimates of genetic load, were not possible due to the fragmentary nature of this information for most populations.

Assessing the impact of genetic purging

It is possible that a history of inbreeding will have created increased opportunities for populations to have purged their genetic load. Furthermore, linkage disequilibrium is elevated after founding and bottlenecking events and decays over time. We therefore compared populations known to have undergone historic bottlenecking or founding events (defined as occurring more than 20 generations ago), with those populations suffering recent declines, bottlenecks or founding events (occurring within the last 20 generations). Because genetic purging and decay of linkage disequilibrium are slow processes, we predicted that mean effect sizes should be smaller in historically bottlenecked populations than recently bottlenecked populations.

Model averaging

We used a model averaging approach based on Akaike's information criteria (AIC), corrected for small sample size (AICc, Burnham & Anderson 2002) using the dRedging library implemented in R (version 2.7.1) to determine which factors had important influences on the overall mean effect size, following the methods described in Knowles *et al*. (2009). Model selection from the entire pool of MLH_{inc} effect sizes was conducted using maximum likelihood with the following four fixed factors: trait type (LH/M/P), sex, publication status of the study and ecological setting of the population (wild/domestic/captive), with all possible interactions, resulting in 16 candidate models. We also repeated the model averaging method for the reduced pool of effect sizes from wild populations, in order to determine the relative importance of demographic history. Model selection here was conducted with the parameters listed above with the exception of ecological setting, which was replaced by the demographic variable, thus we again had 16 candidate models for this analysis. In all models, the random factors of exact trait measured nested within study population were included. We used the AIC weights for each model as an indication of how well that model was supported, and model averaging to determine the relative importance of each fixed factor in the model, $\Sigma\omega_i$, which is determined by the sum of Akaike weights from all models in which that factor was included (Burnham & Anderson 2002).

Results

Data collection

We collected 628 reported effect sizes, which comprised 223 effect sizes included in the C&S meta-analysis, and 405 effect sizes published since C&S (Table S1, Supporting information); hence, the amount of work in this area has almost tripled in the 5 years since 2003. A total of 211 effect sizes were reported using MLH, 144 using SH, 76 using IR, 183 using mean d^2, and 14 using St d^2 (Table 1, Fig. S1, Supporting information). The most commonly studied class was mammals ($n = 347$ effect sizes from 29 species), followed by birds ($n = 230$ effect sizes from 23 species), fish ($n = 34$ effect sizes from six species), invertebrates ($n = 11$ effect sizes from two species), and reptiles ($n =$ six effect sizes from one species), hence a total of 61 different species were represented. Most effect sizes came from published studies ($n = 481$ effect sizes from 58 papers) with the remainder coming from unpublished work such as MSc and PhD theses and results in preparation for publication ($n = 147$ effect sizes from nine studies; Table S1). Most studies had been conducted on wild populations (89.8% of study units and 74.5% of effect sizes) with the rest coming from studies of domestic (7.1% of study units and 18.3% of effect sizes), or captive (3.1% of study units and 7.2% of effect sizes) populations. Most studies considered the effects of genetic metrics on morphometric traits ($n = 323$, 51.4%), followed by life-history traits ($n = 240$, 38.2%) and physiological traits ($N = 65$, 10.4%); this pattern was consistent with that found by C&S. Hence, while there has been a huge increase in work in this area in recent years, it is broadly similar in focus.

Univariate meta-analyses

When all reported effect sizes were treated as independent data points, the overall weighted mean effect sizes for the different genetic metrics were low, but significantly greater than zero for all metrics except St d^2, for which the sample size was smallest. The mean effect size for all studies pooled was $r = 0.036$. Weighted mean effect sizes were largest for MLH and IR (both $r = 0.048$), followed by SH ($r = 0.040$, Table 1). The two measures of mean d^2 had the

Observations	Mean r	95% CI	k	Q_T	P_Q
All inclusive meta	0.0363	**0.0302–0.0425**	628	1188.3285	< 0.0001
MLH					
All traits	0.0476	**0.0355–0.0595**	211	333.0781	< 0.0001
Life history	0.0671	**0.0462–0.0880**	55	95.7653	0.0004
Morphometric	0.0324	**0.0166–0.0482**	124	154.2260	0.0297
Physiological	0.0784	**0.0319–0.1246**	32	74.1069	< 0.0001
Published effects	0.0506	**0.0377–0.0634**	194	319.3055	< 0.0001
Unpublished effects	0.0273	–0.0087–0.0632	17	12.1395	0.7343
SH					
All traits	0.0404	**0.0280–0.0527**	144	319.6163	< 0.0001
Life history	0.0898	**0.0702–0.1094**	73	164.3430	< 0.0001
Morphometric	0.0085	0.0087–0.0258	59	80.0062	0.0294
Physiological	–0.0016	–0.0523–0.0491	12	33.30119	0.0005
Published effects	0.0578	**0.0436–0.0720**	107	244.1002	< 0.0001
Unpublished effects	–0.0137	–0.0393–0.0120	37	51.3193	0.0470
IR					
All traits	0.0477	**0.0303–0.0651**	76	183.3384	< 0.0001
Life history	0.0687	**0.0448–0.0925**	38	123.1475	< 0.0001
Morphometric	0.0133	–0.0145–0.0410	29	33.4587	0.2193
Physiological	0.1172	**0.0173–0.2146**	9	14.5153	0.0693
Published effects	0.0781	**0.0578–0.0985**	52	131.8897	< 0.0001
Unpublished effects	–0.0382	–0.0734–0.0030	24	16.9242	0.8130
Mean d^2					
All traits	0.0235	**0.0127–0.0344**	183	327.4431	< 0.0001
Life history	0.0633	**0.0421–0.0845**	61	142.3076	< 0.0001
Morphometric	0.0076	–0.0059–0.0210	110	148.7218	0.0069
Physiological	0.0215	–0.0232–0.0662	12	16.7902	0.1142
Published effects	0.0384	**0.0232–0.0535**	114	228.7453	< 0.0001
Unpublished effects	0.0077	–0.0081–0.0234	69	90.8873	0.0334
St d^2					
All traits	0.0137	–0.0217–0.0492	14	8.8805	0.7819

Table 1 Effect sizes (r) and their confidence intervals, number of reported effect sizes (k), Q_T-statistics (for homogeneity of effect sizes) for all reported effect sizes for the association between genetic diversity measured as MLH, SH, IR, mean d^2, or St d^2 and phenotypic variation using univariate meta-analysis techniques, and treating all reported effect sizes as independent data points; 95% confidence intervals that do not span zero are in bold

lowest effect sizes (mean d^2 $r = 0.024$; St d^2 $r = 0.014$). Weighted mean effect sizes for life-history traits were consistently higher than for morphological traits (Table 1). Nearly all effect sizes were heterogeneous when grouped by trait type and/or genetic metric, the exceptions being physiological traits measured by IR or mean d^2, and morphological traits measured by IR; this suggests that subclasses of effect sizes occur within most of these groupings (Matt & Cook 1994).

The four genetic metrics MLH, IR, SH and mean d^2 all exhibited mean effect sizes significantly different from zero, but not significantly different from each other (Fig. 1). We also determined the weighted mean correlation coefficients for pairwise combinations of the genetic metrics from published studies (Table 2). This analysis revealed strong, significant, correlations between the metrics MLH, SH, and IR and somewhat weaker significant correlations between these metrics and mean d^2. The metric St d^2 was not significantly correlated with the other four metrics. These analyses show that most of the genetic metrics in common use do not measure different aspects of individual

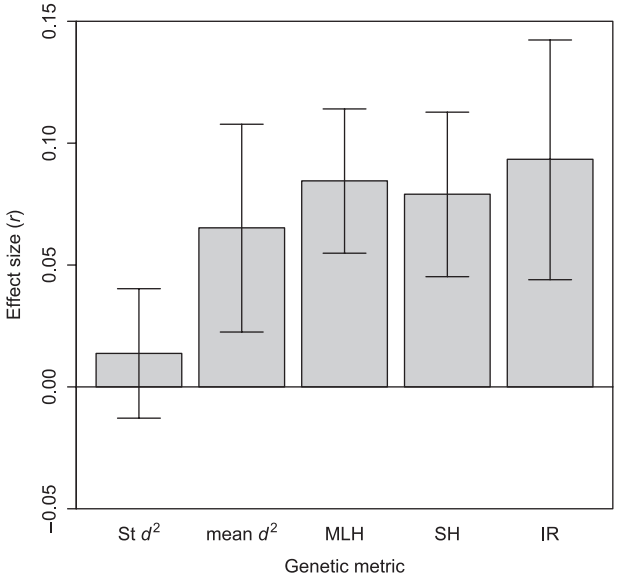

Fig. 1 Mean effect sizes of the different genetic metrics, including 95% confidence intervals.

Table 2 Mean correlation coefficients (r) between the five genetic metrics reported in published studies, weighted by study variance (upper diagonal) and the number of papers reporting this correlation (lower diagonal). Significant correlations (those for which the 95% CIs do not span zero) are in bold. Note that the comparison of MLH and IR was based on only one published correlation, significance in this case was determined directly from the published P value of the correlation

Genetic metric	MLH	SH	IR	Mean d^2	St d^2
MLH	—	**0.971**	**0.94**	**0.395**	0.315
SH	4	—	**–0.967**	**0.235**	**0.347**
IR	1	3	—	NA	0.056
mean d^2	11	3	NA	—	0.069
St d^2	2	5	3	2	—

NA, no published comparison available.

genetic variation; therefore, they can probably be used interchangeably, but should not be used concurrently. As a consequence, and to simplify further models, we pooled results from the three most highly correlated metrics, namely MLH, SH, and IR, for all further analyses. For simplicity, these three pooled metrics will be referred to as MLH_{inc} henceforth.

We repeated the 'all effects independent' univariate analysis using the MLH_{inc} grouping, and compared this to a second, more conservative univariate analysis whereby we took 'study unit averages' for each genetic metric (MLH_{inc} and mean d^2) and trait type (LH, M and P) combination, as in C&S. The univariate analysis treating all effect sizes as independent but combining the results for MLH, SH and IR into one inclusive metric (MLH_{inc}) found broadly similar results (Table 3a) as were found when these metrics were tested individually (Table 1). However, M traits in the study unit average univariate analysis showed a more than twofold increase in weighted mean effect size ('all effects independent' $r = 0.02$, vs 'study unit average' $r = 0.05$; Table 3b). We also found increased weighted mean effect sizes for LH and M traits measured with mean d^2 when using the study unit average approach (Table 3c) compared to when treating all effect sizes as independent data (Table 1). The study unit average approach reduced the amount of heterogeneity in the data, although most data groupings were still significantly heterogeneous (Table 3).

Evidence for publication bias

Other than for St d^2 (all effect sizes were published), mean effect sizes for all genetic metrics were at least twofold greater in published studies, with the greatest disparity being a fivefold difference in the magnitude of effects between published and unpublished studies using the metric mean d^2 (Table 1). The 95% confidence intervals for weighted mean effect sizes from published studies did not span zero for any genetic metric, while the opposite was true for unpublished studies: all 95% CIs spanned zero. For MLH, published and unpublished effect sizes did not differ significantly ($t = 0.10, P = 0.31, n = 211$), whereas the difference was significant for mean d^2 ($t = 2.1, P = 0.038, n = 183$). Unpublished weighted mean effect sizes were statistically homogeneous for MLH and IR but not the other two metrics, while published effect sizes for all four genetic metrics were statistically heterogeneous (Table 1).

Table 3 Univariate meta-analyses of effect sizes (r) and their confidence intervals, number of reported effect sizes (k), and Q_T-statistics (for homogeneity of effect sizes) using (a) the pooled MLH_{inc} metric and treating all effect sizes as independent data points; (b) the pooled MLH_{inc} metric and study unit mean effect sizes; and (c) mean d^2 and study unit mean effect sizes. Ninety-five per cent confidence intervals that do not span zero are in bold

Observations	Mean r	95% CI	k	Q_T	P_Q
(a) MLH_{inc}, all reported effect sizes					
All traits	0.0448	**0.0371–0.0525**	431	836.8545	< 0.0001
Life history	0.0763	**0.0642–0.0884**	166	386.3551	< 0.0001
Morphometric	0.0201	**0.0095–0.0307**	212	272.1239	0.0029
Physiological	0.0484	**0.0179–0.0796**	53	130.6207	< 0.0001
(b) MLH_{inc}, study unit averages					
All traits	0.0737	**0.0566–0.0907**	105	135.0730	0.0219
Life history	0.0926	**0.0691–0.1161**	45	65.4610	0.0195
Morphometric	0.0531	**0.0233–0.0830**	49	42.1068	0.7120
Physiological	0.0468	–0.0076–0.1009	11	21.7535	0.0164
(c) Mean d^2, study unit averages					
All traits	0.0424	**0.0161–0.0687**	57	87.1411	0.0049
Life history	0.0774	**0.0379–0.1168**	26	50.6721	0.0018
Morphometric	0.0117	–0.0326–0.0559	27	25.9100	0.4681
Physiological	0.0113	–0.0931–0.1155	4	4.3017	0.2307

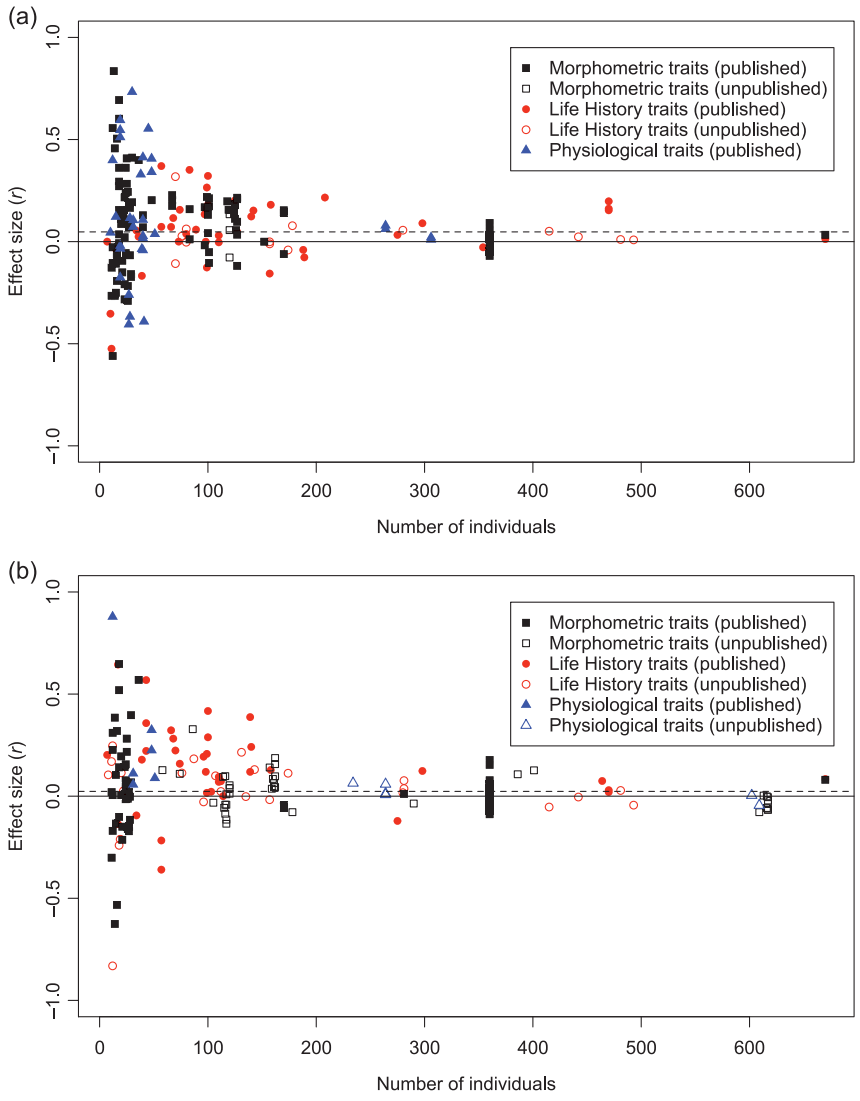

Fig. 2 Variation in effect sizes as a function of sample size (a: MLH; b: mean d^2). Published studies are represented by closed symbols, while unpublished studies are represented by open symbols.

Visual inspection of funnel plots (Fig. 2) also shows that unpublished studies (represented by open symbols) tend to have smaller effect sizes than published studies (closed symbols). We conducted a trim and fill analysis on the 481 published HFC effect sizes to test for publication bias. The mean r of these published effects was 0.0515. The trim and fill analysis suggested there were 48 effect sizes missing from our meta-analysis. Adding these 'missing' studies reduced the mean effect size to $r = 0.0431$, with 95% confidence intervals of 0.0359–0.0503. We then repeated this trim and fill analysis including only those papers published since C&S. This analysis comprised 278 effect sizes from 35 papers, and the mean effect size among these papers was $r = 0.065$. The trim and fill analysis identified a potential 11 missing studies, and when these 'missing' 11 studies were added, the mean effect size was $r = 0.0612$ (95% CIs 0.0495–0.0730).

Influence of sample size

Visual inspection of funnel plots (Fig. 2) confirms the expectation that small sample sizes can produce large fluctuations around the predicted true mean effect sizes; the ranges of sample size (7–1055), markers typed (3–101) and total genotypes assayed (40–57873) were all very large. We found no evidence that these parameters have increased over time (number of individuals $F_{1,57} = 3.04$, $P = 0.09$, Fig. 3a; number of markers $F_{1,57} = 0.92$, $P = 0.34$, Fig. 3b; number of genotypes $F_{1,57} = 0.68$, $P = 0.41$; Fig. 3c).

Multivariate meta-analyses

We next ran the meta-analyses in an LMM framework with the random effects of study population, and, where appropriate, the exact trait measured, to help control for

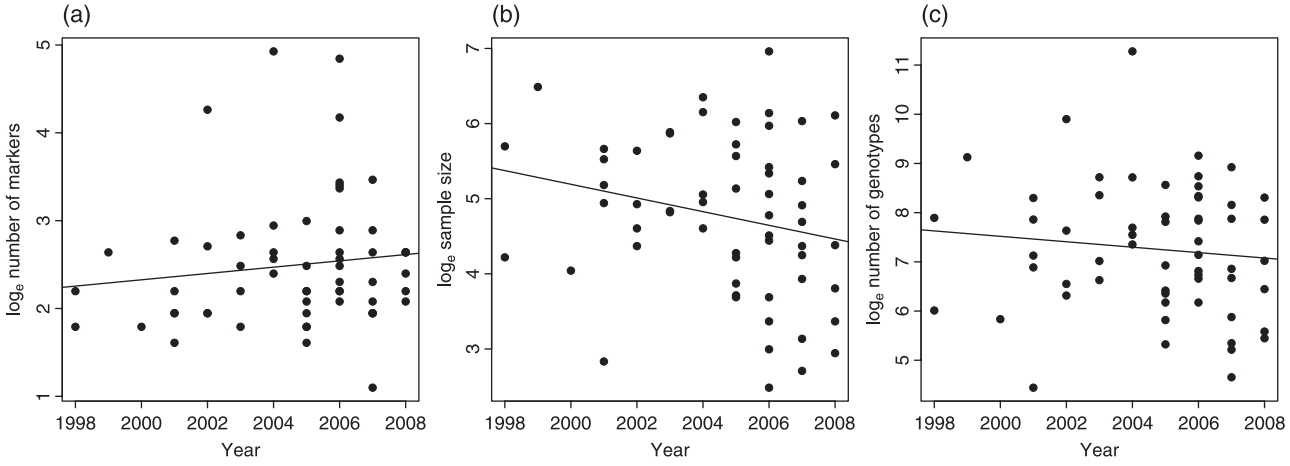

Fig. 3 The relationship between number of individuals assayed (a), markers typed (b), and total genotypes assayed (c) and year of publication for all published HFC studies included in these analyses.

Table 4 Linear mixed effects model of effect sizes (r) and their standard errors and 95% confidence intervals, number of reported effect sizes (k), t-tests and their associated P values, and Q_T and Q_{REML} statistics (for homogeneity of effect sizes) and their associated P values, for all reported effect sizes for the association between genetic diversity measured as MLH, SH, IR, mean d^2, or St d^2; 95% confidence intervals that do not span zero are in bold

Genetic metric	Mean r	Standard error	95% CI	t-value	P	k	Q_T	P_Q	Q_{REML}	P_{QREML}
All metrics pooled	0.0826	0.0110	**0.0610–0.1038**	7.53	< 0.0001	628	1185.2	< 0.0001	822.95	< 0.0001
MLH	0.0845	0.0152	**0.0549–0.1140**	5.57	< 0.0001	211	333.08	< 0.0001	209.55	0.4763
SH	0.0791	0.0173	**0.0452–0.1127**	4.57	< 0.0001	144	319.62	< 0.0001	228.12	< 0.0001
IR	0.0934	0.0253	**0.0440–0.1423**	3.70	< 0.0001	76	183.34	< 0.0001	110.48	0.0038
mean d^2	0.0653	0.0219	**0.0225–0.1078**	2.99	0.0033	183	327.44	< 0.0001	185.34	0.3969
St d^2	0.0137	0.0136	–0.0128–0.0403	1.01	0.3403	14	8.8805	0.7131	8.8805	0.7131

nonindependence of data points and taxonomy. Including further random factors was not supported by the data (see Appendix S2, Supporting information).

In the null model, the mean effect size across all studies, trait types and genetic metrics was $r = 0.08$. The 95% CIs for r (0.061–0.104) did not span zero, thus the weighted mean effect size for all studies combined was significantly different from zero ($t_{530} = 7.53, P \leq 0.0001$). The effect sizes were significantly heterogeneous ($Q_T = 1185.2$, $P \leq 0.0001$; $Q_{REML} = 822.95, P \leq 0.0001$, both $d.f. = 626$).

For comparison with the initial univariate analysis (Table 1), we assessed weighted mean effect sizes for the five genetic metrics individually (the three trait types were pooled for this analysis), and found the same broad pattern as in the univariate analysis, although weighted mean effect sizes were two- to threefold larger in the multivariate analysis (Table 4). We then repeated the multivariate analysis using the more inclusive MLH_{inc} grouping of metrics. Here, we found that life-history traits exhibited the strongest mean effect sizes ($r = 0.10$, Table 5), and that this result was similar to that found by both types of univariate analysis (all effect sizes independent $r = 0.08$, study unit average $r = 0.09$). In contrast, we found that the weighted mean effect size of physiological traits was larger here ($r = 0.08$, Table 5) than either univariate analysis (all effect sizes independent $r = 0.05$, study unit average $r = 0.05$), and that the weighted mean effect size of morphological traits ($r = 0.06$, Table 5) agreed more closely with that found by the study unit average method ($r = 0.05$) than with the nonconservative independent effect size analysis ($r = 0.02$). Indeed, for all three traits, the study unit average and multivariate estimates of weighted mean effect size had largely overlapping confidence intervals (Fig. 4). A similar pattern was found for the metric mean d^2: life-history traits exhibited the largest weighted mean effect size, physiological traits were much larger in the multivariate analysis than either univariate method, and morphological traits were in slightly closer agreement with the study unit average analysis than the independent effect sizes analysis (Table 5). Data groupings in the multivariate analysis were again mostly significantly heterogeneous when measured with Q_T, however, the Q_{REML} approach showed much lower levels

Table 5 Linear mixed effects model of effect sizes (r) and their standard errors and 95% confidence intervals, number of reported effect sizes (k), t-tests and their associated P values, and Q_T and Q_{REML} statistics (for homogeneity of effect sizes) and their associated P values, for all reported effect sizes for the association between genetic diversity (MLH_{inc}, mean d^2 and St d^2) and trait type (LH vs M vs P), ecological setting of population (wild vs captive vs domestic) and publication status of study; 95% confidence intervals that do not span zero are in bold

Factors	Mean r	Standard error	95% CI	t-value	P	k	Q_T	P_Q	Q_{REML}	P_{QREML}
MLH_{inc}										
Life-history traits	0.0984	0.0159	**0.0674–0.1293**	6.19	< 0.0001	166	386.36	< 0.0001	138.48	0.9268
Morphometric traits	0.0611	0.0158	**0.0302–0.0919**	3.87	< 0.0001	212	272.12	0.0025	168.07	0.9849
Physiological traits	0.0809	0.0389	**0.0048–0.1560**	2.08	0.0459	53	130.62	< 0.0001	96.04	0.0001
Population wild	0.0921	0.0135	**0.0657–0.1184**	6.82	< 0.0001	317	549.02	< 0.0001	86.38	1.0000
Population captive	0.1351	0.0686	**0.0014–0.2640**	1.98	0.0598	45	154.79	< 0.0001	146.50	< 0.0001
Population domestic	0.0593	0.0287	**0.0032–0.1150**	2.07	0.0426	69	96.56	0.0105	65.84	0.5172
Published effects	0.0957	0.0119	**0.0725–0.1187**	8.06	< 0.0001	353	700.56	< 0.0001	390.21	0.0731
Unpublished effects	0.0173	0.0235	–0.0287–0.0632	0.74	0.4656	78	88.03	0.1631	1.98	1.0000
Mean d^2										
Life-history traits	0.0931	0.0302	**0.0341–0.1514**	3.09	0.0039	61	142.31	< 0.0001	80.02	0.0357
Morphometric traits	0.0229	0.0198	–0.0159–0.0616	1.16	0.2501	110	148.72	0.0057	112.56	0.3628
Physiological traits	0.0946	0.0727	–0.0475–0.2330	1.31	0.2280	12	16.79	0.0791	4.99	0.8915
Population wild	0.0702	0.0251	**0.0211–0.1189**	2.80	0.0063	137	233.13	< 0.0001	135.20	0.4789
Population domestic	0.0293	0.0378	–0.0448–0.1030	0.77	0.4431	46	64.31	0.0245	47.03	0.3497
Published effects	0.0893	0.0282	**0.0343–0.1438**	3.18	0.0022	114	228.75	< 0.0001	125.77	0.1765
Unpublished effects	0.0201	0.0194	–0.0179–0.0580	1.03	0.3054	69	90.89	0.0277	57.77	0.7820
St d^2										
Life-history traits	0.0081	0.0164	–0.0240–0.0402	0.50	0.6336	13	8.58	0.6604	8.58	0.6604

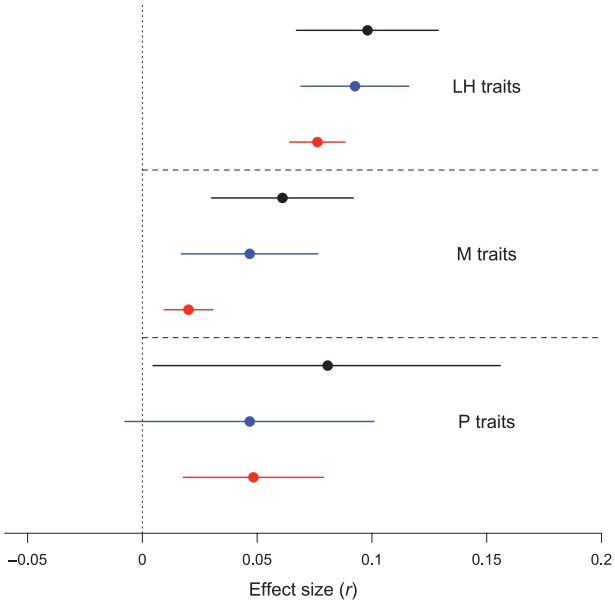

Fig. 4 Weighted mean effect sizes of life history (LH), morphometric (M) and physiological (P) traits detected using the three meta-analysis methods: all effect sizes independent univariate (red lines), study unit average univariate (blue lines) and multivariate (black lines).

of data heterogeneity in this analysis (Table 5). We tested whether the general pattern of LH and M traits being of a similar order of magnitude was also found in the subset of studies that had measured both LH and M traits, using any of the MLH_{inc} metrics, in the same population, over the same time-span and with the same set of markers (i.e, LH and M HFCs were both reported in the same paper). We identified 11 such studies, and again found no difference in the weighted mean effect size reported for M and LH traits (paired t-test $= 0.23$, $P = 0.82$, Fig. S2, Supporting information).

For the remainder of the multivariate results, we focus on effect sizes obtained with the MLH_{inc} genetic metrics, for three reasons: (i) ease of presentation and comparison of results; (ii) the genetic metric mean d^2 is now widely regarded as relatively uninformative; and (iii) we had the largest set of reported effect sizes for MLH_{inc}.

Impact of demographic history and inbreeding

We classified the subset of wild populations on a continuous ordinal scale ranging from zero to three for a range of demographic factors. We tested whether populations scoring higher on this scale, and thus being more likely to exhibit higher inbreeding variance among individuals, had higher mean effect sizes, but found no evidence this was the case (regression coefficient $= -0.005$, SE $= 0.01$, $t_{76} = -0.40$, $P = 0.69$, $n = 317$, Fig. 5). We then grouped together all populations scoring between 1 and 3 on this ordinal scale (in other words, all populations scored as likely to have a nonzero rate of inbreeding variance), and compared this group with the group of populations scoring zero (and thus likely having very little inbreeding

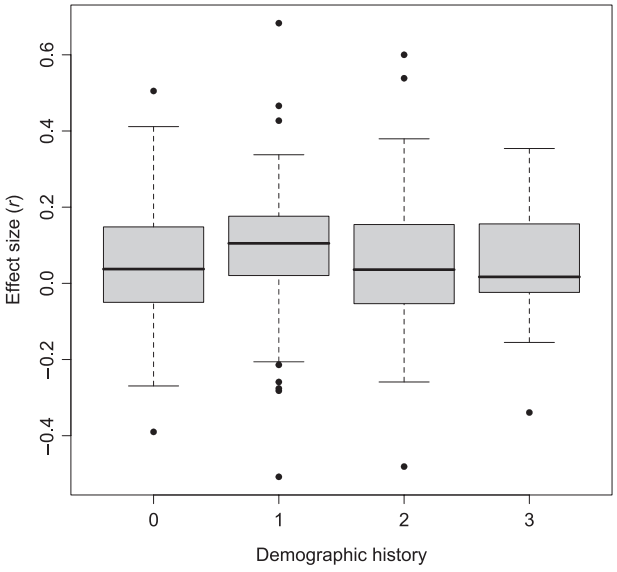

Fig. 5 Relationship between effect size (Z_r) and demographic history scored on an ordinal scale from 0 (populations estimated to have the lowest variance in inbreeding) to 3 (populations estimated to have the highest variance in inbreeding) for studies reporting effect sizes using any of the MLH_{inc} metrics. Boxes are bounded by the 25th and 75th percentiles, with the median value shown inside the box. Whiskers below and above the boxes indicate the 10th and 90th percentiles, with outliers plotted as dots.

Table 6 Model selection of variables influencing the magnitude of overall mean effect size. Candidate models are shown with their relative AICc weight, difference between the candidate model and the best model (Δ_i), and Akaike weight (ω_i). The best model is listed first, and all other models in descending order of support

Model rank	Variables in model	AICc	Δ_i	ω_i
1	1 (null model)	–495.57	0.000	0.192
2	1 + 3	–495.44	0.134	0.179
3	1 + 3 + 5	–494.32	1.259	0.102
4	1 + 5	–494.17	1.408	0.095
5	1 + 4	–493.55	2.025	0.070
6	1 + 3 + 4	–493.42	2.156	0.065
7	1 + 2	–493.21	2.367	0.059
8	1 + 2 + 3	–493.10	2.471	0.056
9	1 + 3 + 4 + 5	–492.24	3.334	0.036
10	1 + 4 + 5	–492.10	3.475	0.034
11	1 + 2 + 5	–491.60	3.972	0.026
12	1 + 2 + 3 + 5	–491.58	3.993	0.026
13	1 + 2 + 4	–491.20	4.371	0.022
14	1 + 2 + 3 + 4	–491.12	4.458	0.021
15	1 + 2 + 4 + 5	–489.53	6.049	0.009
16	1 + 2 + 3 + 4 + 5	–489.50	6.072	0.009

Variables: 1, intercept; 2, ecological seeing of population; 3, publication status of study; 4, sex; 5, trait type (LH/M/P).

variance), but again found no evidence for higher mean effect sizes in populations likely to have a nonzero level of inbreeding variance (regression coefficient = –0.008, SE = 0.03, $t_{76} = -0.29$, $P = 0.77$, $n = 317$); we also found no evidence that effect size depended on the type of fitness trait in these populations (results not shown). It is possible this lack of an increase in effect size in populations predicted to have high variance in inbreeding is due to more complete purging of deleterious recessive alleles in such populations. We identified 48 effect sizes from 17 populations (13 species) in which there is evidence of historic bottlenecking or founding events, and 26 effect sizes from 8 populations (6 species) in which founding or bottlenecking has occurred recently. We found slightly higher mean effect sizes in recently bottlenecked populations ($r = 0.12$) than historically bottlenecked populations ($r = 0.08$), however, the difference was not significant ($t_{23} = 0.65$, $P = 0.51$, $n = 74$).

We tested whether our failure to detect a relationship between demographic history and mean effect size was due to small sample sizes (and thus higher sampling variance) from small populations, but found no relationship between the sample size used in a study and $\log_{10}$ of the estimated annual population size, where this information was available (regression coefficient = –14.9, $P = 0.56$, $n = 57$, Fig. S3a, Supporting information), or between population average heterozygosity and population size (regression coefficient = 0.03, $P = 0.19$, $n = 42$, Fig. S3c). Furthermore, we found no evidence that sample sizes were smaller from populations with higher demographic scores; sample sizes were in fact higher, although the trend was not significant (regression coefficient = 28.0, $P = 0.16$, $n = 76$, Fig. S3b). However we did find that populations with higher demographic scores had significantly lower heterozygosity (regression coefficient = –0.04, $P = 0.0008$, $n = 58$, Fig. S3d), suggesting that our attempt to classify populations on the basis of variance in inbreeding was, at least to some extent, effective, as lower population heterozygosity would be expected in more inbred populations (see also Evans & Sheldon 2008).

Model averaging

We used just the effect sizes reported using any of the MLH_{inc} metrics ($n = 431$), and tested the relative importance of the following variables: trait type (LH, M or P), ecological setting of the population (wild, domestic or captive), whether or not the study had been published, and the sex of the individuals for which the HFC was reported. The best-supported model was the null model; in other words, no combination of the variables listed above produced a better-supported model than the model in which none were included (Table 6). The relative variable importance, $\Sigma\ \omega_i$, was reasonably low for all four fixed factors (trait type $\Sigma\ \omega_i = 0.338$; ecological setting $\Sigma\ \omega_i = 0.227$; publication status $\Sigma\ \omega_i = 0.494$; sex $\Sigma\ \omega_i = 0.265$). Repeating

the model averaging exercise with all 629 effect sizes we gathered, and including the genetic metric used as a fixed factor, assigned a high relative importance to this variable ($\Sigma\ \omega_I \approx 0.9$, data not shown), reflecting the large difference in effect sizes found with the metric St d^2 and all other metrics. We repeated the model averaging analysis in the subset of wild populations ($n = 317$) in order to determine the relative importance of population demographic history among these studies, with broadly similar results (see Appendix S2 and Table S2, Supporting information).

Discussion

Comparison of univariate and multivariate methods for meta-analysis

Multivariate methods of meta-analysis in a mixed-effects model framework represent an improvement in our ability to draw valid conclusions when pooling the results of multiple studies in ecology and evolution, and indeed in all fields of quantitative science, as they help to account for a major problem with meta-analysis, that of nonindependence of data points. Nevertheless, we chose to conduct two types of univariate meta-analysis in addition to a more comprehensive multivariate meta-analysis, both to facilitate comparisons with earlier meta-analyses of such studies (Britten 1996; Reed & Frankham 2001; Coltman & Slate 2003) and to allow direct comparison between the methods.

The 'all effects independent' univariate analysis revealed weak but significantly positive mean effect sizes for the all of the genetic metrics with the exception of St d^2; stronger effect sizes were found for life history than morphological and physiological traits. These results differ in some ways from findings of previous univariate meta-analyses: Britten (1996) found an overall effect size of $r = 0.133$ for HFCs, while Reed & Frankham (2001) found an overall effect size of $r = 0.217$, with LH traits exhibiting a weaker and negative mean effect ($r = -0.110$) when compared with M traits ($r = 0.311$). More recently, Coltman & Slate (2003) found that the overall effect size for life-history traits was $r = 0.086$ and $r = 0.048$ when MLH and mean d^2, respectively, were used as the genetic metric, while effect sizes were lower for morphological and physiological traits ($r = 0.004$–0.008). Our 'study unit average' univariate analysis revealed an increase in the estimated weighted mean effect size for both LH and M traits, but a smaller difference in effect size between these two trait types (LH vs. M: $\Delta r = 0.056$ when all effects treated as independent data, $\Delta r = 0.039$ for study unit average analysis). This was again in broad agreement with previous results: C&S found an increase in mean effect size for both LH ($r = 0.112$) and M traits ($r = 0.052$), but a decrease in the magnitude of difference between LH and M traits when using the study unit average approach ($\Delta r = 0.080$ for independent analysis, $\Delta r = 0.060$ for study unit average analysis). In plants, the weighted mean correlation between heterozygosity and fitness was also significantly positive ($r = 0.306$), and was significantly influenced by mating system, whereby self-incompatible plants showed significant positive HFCs whereas self-compatible plants did not, but not by plant rarity or longevity (Leimu *et al.* 2006). It is interesting to note the much stronger mean effect size found for plants than for animals. This is perhaps due to the fundamental differences in mating system and demography between plants and animals. For example plant populations can exhibit strong demographic population structure (such as age structure) not often seen in animal populations. A lack of published information meant that Leimu *et al.* (2006) were not able to specifically address the impact of demography on HFCs in plants. However, it should also be noted that study sample sizes in Leimu *et al.* (2006) are very small ($n = 2$–14), and the authors recommend their conclusions be considered preliminary. Methodological differences between our study and that of Leimu *et al.* (2006) may also help to account for the large difference in mean effect sizes seen between plants and animals. For example, along with heterozygosity, Leimu *et al.* (2006) also considered percentage of polymorphic loci and the number of alleles as measures of genetic diversity.

The multivariate meta-analysis we carried out here also revealed weak but significantly positive mean effect sizes for the genetic metrics MLH, SH, IR and mean d^2 and a nonsignificant mean effect size for St d^2. This lack of significance for St d^2 may be due to a paucity of studies that have reported this metric (14 effect sizes from 9 study populations); however, it seems likely that standardizing this measure actually results in a loss of genetic signal, and standardized d^2 is increasingly considered to be less informative than other measures (Hoffman *et al.* 2006).

Only one study included in the meta-analysis identified outbreeding depression in their population (Marshall & Spalton 2000), and, given that outbreeding depression is thought to be rare in animal populations (Frankham 1995a; Pusey & Wolf 1996), we felt justified in assuming that overall, any signature of outbreeding depression in our meta-analysis would be negligible. Nevertheless, it is possible that some populations included here actually had high rates of undetected outbreeding depression. Such populations would be expected to exhibit strong negative HFCs, and as such may have lowered our global estimates of weighted mean effect size. Negative effects were less common than positive effects among the studies in our meta-analysis (34% of effects), and strong negative HFCs were especially rare (3.8% of effects were $Zr \leq -0.25$), but were detected using both MLH_{inc} and mean d^2 metrics (Fig. 2), in contrast to the analysis of C&S, where strong negative effects were only detected with mean d^2.

While the study unit average analyses suggested that effect sizes are strongest for life-history traits, this pattern was not mirrored in the multivariate meta-analysis; here mean effect sizes were much more similar between the three trait types. Thus, the large difference in mean effect size for life history vs. morphological traits found in C&S was only partially supported here, and was dependent upon the genetic metric employed and the type of meta-analysis conducted. The change in effect size dependent on the method used suggests that the low value found for M traits in the initial independent effects univariate analysis, was due, at least in part, to psuedoreplication of effect sizes. Sequentially dropping each of the higher order nested random factors (class, family, and species) in the multivariate analysis revealed increasingly large differences between life-history traits and morphological and physiological traits (data not shown). This again suggests that earlier meta-analyses have not adequately controlled for pseudoreplication within and between studies, and that the 'study-unit average' approach used here, and taken by some authors, does not fully account for replicated results.

The model averaging analysis showed that the main factors that appear to be influencing the magnitude of reported effect sizes in this meta-analysis are properties intrinsic to the study design; namely (i) whether Std^2 or any of the other four genetic metrics was used to determine genetic variability; and (ii) whether the study had been published or not. Intrinsic properties of the individuals and populations studied, such as the type of trait measured (LH/M/P) and the ecological setting of the population (wild/captive/domestic) had much lower influence on mean effect sizes, as models incorporating these factors were less well supported (Table 6) and these variables had lower relative importance. In general, this suggests a rather poor fit between the results of HFC analyses and expectation based on population genetic theory. Below we discuss how these results relate to the objectives of the paper listed in the introduction.

Evidence for publication bias

We found three lines of evidence for a bias towards publishing significant effects. First, unpublished studies had smaller effect sizes than published studies. Indeed, in the univariate analysis, unpublished effect sizes for all genetic metrics were not statistically different from zero. Second, the funnel plots show clear evidence for missing studies with small effects and low sample size. Third, the trim and fill analysis suggested there were 48 missing effects from the pool of 481 published effects used in the analysis, suggesting that around 10% of all HFC effect sizes recorded by researchers go unreported in the literature. The analysis suggested that if these 'missing' effect sizes had been published, weighted mean effect sizes would be weaker, albeit still significantly positive. The trim and fill analysis of papers published since C&S suggested that only around 4% of detected HFC effect sizes now go unpublished, suggesting a reduction in bias since the publication of C&S.

It has been suggested that meta-analyses are biased towards finding positive effects, especially if care is not taken to identify and include 'missing' nonsignificant results (Kotiaho & Tomkins 2002; Jennions *et al*. 2004; Tomkins & Kotiaho 2004). However, our meta-analysis suggests that publication bias in this field does not necessarily result in spurious conclusions being reached with regard to the existence of HFCs, as the trim and fill analysis suggested that including 'missing' studies would still reveal weakly significant positive effects. One reason for this may be that HFC studies generally report many correlations, and often many of these are nonsignificant; indeed, of the 481 effect sizes we collected from published studies, only 115 (24%) were significant at the $P = 0.05$ level. This indicates that there is no dearth of nonsignificant results in this field. However, many of these nonsignificant results were published in papers also reporting significant results. We have no way of knowing how many studies that fail to detect any significant results at all remain unpublished, although the trim and fill analysis suggests that this may not be a serious problem (although see Koricheva 2003). The publication bias seems to be strongest for negative results based on small sample sizes (Fig. 2). We also found evidence to suggest that publication bias in this field may have lessened since the publication of C&S, where this problem was first identified as a potential issue for microsatellite HFC studies.

Temporal trends in HFC studies

If it is reassuring to find a recent reduction in publication bias in this field (suggesting heeding of advice in C&S), other temporal trends inspire less confidence. Coltman & Slate (2003) suggested that sample sizes were often too small in this field, but there is little evidence that the number of individuals sampled, markers assayed, or total genotypes scored have been increasing since the publication of C&S (Fig. 3). As can be seen by visual inspection of funnel plots (Fig. 2), studies with small sample sizes exhibit large fluctuations around the estimated 'true' effect size when all studies are pooled, which may potentially have resulted in an overemphasis of the importance of this field of research. Certainly in small populations, there may be a limit to the number of individuals that can be realistically sampled, but we found no relationship between sample size and population size (Fig. S3). Furthermore, even when few individuals are available for sampling, researchers should aim to maximize the number of markers assayed, and ultimately genotypes scored, in order to have confidence in their measure of genetic diversity. Given recent developments aimed at reducing laboratory costs (e.g. Schuelke 2000;

Wang *et al*. 2003; Symonds & Lloyd 2004; Guicking *et al*. 2008), and theoretical and empirical results showing that accurate estimates of genome-wide heterozygosity require large numbers of genotypes (Balloux *et al*. 2004; DeWoody & DeWoody 2005; Väli *et al*. 2008; Alho *et al*. 2009), it is somewhat surprising that this does not (yet) seem to have resulted in a concurrent increase in the amount of molecular data gathered for HFC studies.

Comparison of different genetic metrics

There has been a great deal of debate in the literature as to which genetic metric most accurately acts as a surrogate of the true inbreeding coefficient of an individual (see, for example, Aparicio *et al*. 2006). Here we show that the metrics MLH, SH, and IR are highly correlated and nonindependent; we would encourage researchers to report only one of these genetic metrics when publishing HFC studies in the future, since reporting the correlation between a fitness measure and multiple highly correlated genetic measures is pseudoreplication. It is important that this choice is made before data analysis begins, rather than driven by posthoc choices based on statistical significance, as this will result in another layer of bias in this field. In the multivariate meta-analysis, we found similar mean effect sizes with the three MLH_{inc} genetic metrics, namely MLH, IR and SH (Table 4); mean effect size for mean d^2 was smaller, and that for St d^2 was not significantly different from zero. The rationale for mean d^2 as a metric has been called into question (Hedrick *et al*. 2001; Tsitrone *et al*. 2001; Goudet & Keller 2002; Slate & Pemberton 2002; although see Neff 2004a; Kretzmann *et al*. 2006); and both d^2 measures correlate only weakly with other, more direct, measures of heterozygosity. Until microsatellite mutational processes are more accurately elucidated, it seems likely that the relevance of these measures will continue to be debated (Slate *et al*. 2000), and as such we advocate the use of the simplest metric, MLH, in future HFC studies. Use of more than one metric is likely only justified when certain population demographic histories exist, where the use of MLH and mean d^2 in tandem might actually provide insight into evolutionary processes such as stabilizing selection (Neff 2004b), or in populations where all individuals are highly heterozygous, and thus more traditional measures of heterozygosity fail to differentiate between individuals (Hedrick *et al*. 2001; Tsitrone *et al*. 2001; Goudet & Keller 2002; Slate & Pemberton 2002).

Comparison of different trait types: what constitutes an HFC study?

Studies reporting correlations between measures of individual variation and genetic diversity are collectively known as HFC studies. To some extent, this is a misnomer — in the published literature to date, the majority of traits for which relationships with genetic diversity have been reported (e.g. morphological traits) are likely to have little or no linear relationship with fitness. For instance, many morphological and physiological traits are more plausibly under stabilizing selection around an optimum, and this may also be true for other traits as well. For example, it is often assumed that life-history traits such as timing or age of first breeding and clutch or litter size are under strong directional selection, but these traits evolve in concert with selection on other life-history traits, with which they may exhibit genetic covariance (e.g. between clutch size and offspring size); only fitness itself can always be assumed to be under positive directional selection. In the absence of directional selection on a character, there is no clear reason to expect a relationship between that trait and heterozygosity, however extensively this is measured, and however high the variance in inbreeding within the population. In this light, the expectation about the strength of effect size for HFCs with respect to different classes of trait is perhaps unrealistic. Interestingly, we found effect sizes to be of a similar magnitude when classed as fitness or nonfitness traits (see Appendix S2, Supporting information). While it is possible we were too generous with our classification of fitness traits and too conservative with our classification of nonfitness traits, given that we also found reasonably similar effect sizes for life-history and morphological traits, our results suggest that such broad classification of traits does little to enhance or understanding of the underlying causes of HFCs in animal populations. Thus, while we do not advocate renaming HFC studies, as use of this term is now widespread, we do advocate consideration of the likely form of selection on characters, and it might be illuminating to explore the relationship between the HFC effect size for traits and the form of selection on those traits for which empirical estimates of selection intensity are available.

Evolutionary theory would suggest that we should only expect correlations of genetic diversity with fitness-related traits because dominance variance is expected to be high for traits with a direct effect on fitness, and such traits have a more complex genetic architecture (Crnokrak & Roff 1995; DeRose & Roff 1999; Merilä & Sheldon 1999). It is thus surprising to find that mean effect sizes for fitness and nonfitness traits were of a similar (positive) magnitude. This might be due to publication bias if papers reporting correlations with nonfitness measures are more likely to be published if effects are large, whereas papers assessing the relationship with fitness are equally likely to be published regardless of effect size. Additionally, measurement error may be greater for life history than morphological and physiological traits; such an explanation has been invoked to explain why morphological traits appeared to exhibit stronger directional selection than life-history traits (Kingsolver *et al*. 2001).

Does demographic history influence the strength of HFCs?

We found no evidence that populations likely to have higher inbreeding variance exhibited stronger HFCs than populations likely to have low inbreeding variance (highly inbred or outbred populations, Fig. 5, Table S2). While our measures of demographic structure were quite crude, it is perhaps surprising that they did not reveal a coarse relationship in the expected direction. This may suggest that a majority of studies are actually detecting local, rather than general, effects (Balloux *et al*. 2004), or alternatively that publication bias strongly clouds any pattern in the data. A study by Hansson *et al*. (2004) employing within brood comparisons found that even when the inbreeding coefficient is held constant, more heterozygous individuals were more likely to recruit to the local breeding population — strong evidence for local effects. Testing for local effects by regressing each individual marker with fitness is becoming standard practice; such tests are valuable in allowing us to understand the mechanisms underlying HFCs, and may also allow future identification of functionally important loci (e.g. Acevedo-Whitehouse *et al*. 2006; Luikart *et al*. 2008). However, caution must be exercised here: such multiple tests for significance will result in spurious significant results unless authors are careful to adjust the critical α level and thus guard against inflated type I errors (Simes 1986; Aiken & West 1991). It should also be remembered that single loci correlations are not independent because heterozygosity is correlated across loci (P. David, personal communication), and that, even under the general effect hypothesis, we still expect more than 5% of loci to show single locus HFCs. The standard approach should be to examine the distribution of effect sizes and identify outliers as being those effect sizes that may be statistically, and biologically, significant.

A fruitful direction for future studies of HFC would be to specifically address the impact of demographic factors by sampling from multiple populations such as island and mainland populations, populations with varying levels of habitat disturbance, or populations from a continuum of bottlenecking or founder events. While this approach will not be possible in small, endangered populations with limited range distributions, studies in more widespread species may help provide insight into the demographic processes important in endangered populations and thus help to inform conservation decisions (Reed & Frankham 2003; Grueber *et al*. 2008).

The future of HFC studies — where to from here?

The results of this meta-analysis indicate that while heterozygosity-fitness correlations may well be a general phenomenon in many wild vertebrate populations, these effects are very weak, equivalent in strength to correlations that explain < 1% of the variance in traits. As discussed above, the various measures of heterozygosity now in common use are not statistically independent, and should not be used in concert, as this will result in psuedoreplication. We would also encourage researchers to base future studies of HFCs in wild populations on the measurement of large numbers of individuals with larger marker panels. Furthermore, we would argue that the goal of such studies should ultimately be to infer evolutionary processes in populations (e.g. Slate *et al*. 2000), and as such an increase in the number of studies reporting HFCs in populations with known individual inbreeding coefficients would be beneficial (e.g. Coulson *et al*. 1998; Slate *et al*. 2004; Bensch *et al*. 2006; Olafsdottir & Kristjansson 2008), as would a more explicit investigation of the role of other demographic processes such as bottlenecks, admixture and the role of genetic purging. Another avenue of research that may well prove fruitful is to investigate the role of environmental stress on influencing the magnitude and direction of HFCs detected with microsatellites. Stress, such as periods of low food availability, high predation or increased environmental disturbance, is a key factor in reducing the fitness of populations, and individuals will vary in their response to stress (Hoffmann & Hercus 2000). This can result in an increase in genetic variance at the population level, for example due to the expression of genetic variance that was neutral under normal environmental conditions (Badyaev 2005). Individuals with increased heterozygosity may well possess the necessary diversity of alleles required to adequately cope with environmental stochasticity, this has been termed episodic heterozygote advantage (Samollow & Soulé 1983). This avenue of research has received limited attention to date, however, the magnitude of HFC effects has been shown to correlate positively with habitat fragmentation in Taita thrush (Lens *et al*. 2000), salinity tolerance in guppy at the population, but not individual, level (Shikano & Taniguchi 2002), and food limitation in common frogs (Lesbarreres *et al*. 2005). Studies using allozyme variation have revealed similar patterns (see, for example, Scott & Koehn 1990; Audo & Diehl 1995; Myrand *et al*. 2002).

We would advocate that the number of genotypes assayed be maximized, and would question the merit of future studies reporting HFCs detected with small numbers of microsatellite loci, given the lack of evidence that small marker sets have any power to infer genome-wide heterozygosity (Balloux *et al*. 2004; Slate *et al*. 2004; DeWoody & DeWoody 2005; Hansson & Westerberg 2008; Väli *et al*. 2008). One of the most hotly debated issues in HFC research at present is the relative contribution of local and genome-wide effects. This can only be resolved by studies assessing HFCs using large sets of markers. Furthermore, we would encourage authors to test the covariance in heterozygosity across markers, using the methods suggested by

either Balloux *et al*. (2004) or Slate *et al*. (2004) in order to assess how well their marker set is likely to infer total genomic heterozygosity. A sobering conclusion is that, despite the very large amount of work in this area, the only factors that we have been able to find that explain variation in the strength of HFCs are methodological. Hence, our understanding of the biological reasons for variation in their strength remains poorly developed.

Acknowledgements

We thank M. Szulkin, D. Reed and two anonymous reviewers for helpful comments on the manuscript, S. Knowles and A. Hinks for advice on analysis, and numerous authors who provided additional and/or unpublished results. Funding was provided by the Tertiary Education Commission of New Zealand (JRC).

References

Acevedo-Whitehouse K, Vicente J, Gortazar C *et al*. (2005) Genetic resistance to bovine tuberculosis in the Iberian wild boar. *Molecular Ecology*, **14**, 3209–3217.

Acevedo-Whitehouse K, Spraker TR, Lyons E *et al*. (2006) Contrasting effects of heterozygosity on survival and hookworm resistance in California sea lion pups. *Molecular Ecology*, **15**, 1973–1982.

Aiken LS, West SG (1991) *Multiple Regression: Testing and Interpreting Interactions*. Sage, Newbury Park, California.

Alho JS, Lillandt B-G, Jaari S, Merilä J (2009) Multilocus heterozygosity and inbreeding in the Siberian jay. *Conservation Genetics*, **10**, 605–609.

Aparicio JM, Ortego J, Cordero PJ (2006) What should we weigh to estimate heterozygosity, alleles or loci? *Molecular Ecology*, **15**, 4659–4665.

Aparicio JM, Ortego J, Cordero PJ (2007) Can a simple algebraic analysis predict markers-genome heterozygosity correlations? *Journal of Heredity*, **98**, 93–96.

Audo MC, Diehl WJ (1995) Effect of quantity and quality of environmental stress on multilocus heterozygosity-growth relationships in *Eisenia fetida* (Annelida: Oligochaeta). *Heredity*, **98**, 98–105.

Badyaev AV (2005) Role of stress in evolution: from individual adaptability to evolutionary adaptation. In: *Variation: A Central Concept in Biology* (eds Hallgrímsson B, Hall BK), pp. 277–302. Elsevier Academic Press, San Diego, California.

Balloux F, Amos W, Coulson T (2004) Does heterozygosity estimate inbreeding in real populations? *Molecular Ecology*, **13**, 3021–3031.

Barton NH, Keightley PD (2002) Understanding quantitative genetic variation. *Nature Reviews Genetics*, **3**, 11–21.

Bensch S, Andrén H, Hansson B *et al*. (2006) Selection for heterozygosity gives hope to a wild population of inbred wolves. *PLoS One*, **1**, e72.

Bierne N, Tristone A, David P (2000) An inbreeding model of associative overdominance during a population bottleneck. *Genetics*, **155**, 1981–1990.

Blomqvist D (2009) Fitness-related patterns of genetic variation in rhesus macaques. *Genetica*, **135**, 209–219.

Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. *Heredity*, **85**, 59–63.

Britten HB (1996) Meta–analyses of the association between multilocus heterozygosity and fitness. *Evolution*, **50**, 2158–2164.

Burnham KP, Anderson DR (2002) *Model Selection and MultiModel Inference: a Practical Information Theoretic Approach*. Springer, New York.

Cena CJ, Morgan GE, Malette MD, Heath DD (2006) Inbreeding, outbreeding and environmental effects on genetic diversity in 46 walleye (*Sander vitreus*) populations. *Molecular Ecology*, **15**, 303–320.

Charlesworth D, Charlesworth B (1987) Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics*, **18**, 237–268.

Charlesworth B, Charlesworth D (1999) The genetic basis of inbreeding depression. *Genetical Research*, **74**, 329–340.

Coltman DW, Slate J (2003) Microsatellite measures of inbreeding: a meta-analysis. *Evolution*, **57**, 971–983.

Coltman DW, O'Donoghue P, Hogg JT, Festa-Bianchet M (2005) Selection and genetic (co) variance in bighorn sheep. *Evolution*, **59**, 1372–1382.

Coulson TN, Pemberton JM, Albon SD *et al*. (1998) Microsatellites reveal heterosis in red deer. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 489–495.

Crnokrak P, Barrett SCH (2002) Purging the genetic load: a review of the experimental evidence. *Evolution*, **56**, 2347–2358.

Crnokrak P, Roff DA (1995) Dominance variance – associations with selection and fitness. *Heredity*, **75**, 530–540.

Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory*. Burgess, Minneapolis, Minnesota.

David P (1998) Heterozygosity-fitness correlations: new perspectives on old problems. *Heredity*, **80**, 531–537.

David P, Delay B, Berthou P, Jarne P (1995) Alternative models for allozyme-associated heterosis in the marine bivalve *Spisula ovalis*. *Genetics*, **139**, 1719–1726.

DeRose MA, Roff DA (1999) A comparison of inbreeding depression in life-history and morphological traits in animals. *Evolution*, **53**, 1288–1292.

DeWoody YD, DeWoody JA (2005) On the estimation of genome-wide heterozygosity using molecular markers. *Journal of Heredity*, **96**, 85–88.

Dunning AM, Durocher F, Healey CS *et al*. (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics*, **67**, 1544–1554.

Duvall S, Tweedie R (2000) Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 455–463.

Edwards SV, Dillon M (2004) Hitchhiking and recombination in birds: evidence from *Mhc*-linked and unlinked loci in red-winged blackbirds (*Agelaius phoeniceus*). *Genetical Research*, **84**, 175–192.

Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. *Nature*, **452**, 169–175.

Evans SR, Sheldon BC (2008) Interspecific patterns of genetic diversity in birds: correlations with extinction risk. *Conservation Biology*, **22**, 1016–1025.

Fossøy F, Johnsen A, Lifjeld JT (2008) Multiple genetic benefits of female promiscuity in a socially monogamous passerine. *Evolution*, **62**, 145–156.

Fowler K, Whitlock MC (1999) The variance in inbreeding depression and the recovery of fitness in bottlenecked populations. *Proceedings of the Royal Society B: Biological Sciences*, **266**, 2061–2066.

Frankham R (1995a) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.
Frankham R (1995b) Effective population size/adult population size ratios in wildlife: a review. *Genetical Research*, **66**, 95–107.
Franklin IR (1980) Evolutionary changes in small populations. In: *Conservation Biology, an Evolutionary Ecological Perspective* (eds Soulé ME, Wilcox BA), pp. 135–149. Sinauer & Associates, Sunderland, Massachusetts.
Gage MJG, Surridge AK, Tomkins JL *et al*. (2006) Reduced heterozygosity depresses sperm quality in wild rabbits, *Oryctolagus cuniculus*. *Current Biology*, **16**, 612–617.
Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. *Plant Cell*, **15**, 1502–1506.
Glémin S (2003) How are deleterious mutations purged? Drift versus nonrandom mating. *Evolution*, **57**, 2678–2687.
Goldstein DB, Schlötterer C (1999) *Microsatellites: Evolution and Applications*. Oxford University Press, Oxford, UK.
Goudet J, Keller L (2002) The correlation between inbreeding and fitness: does allele size matter? *Trends in Ecology & Evolution*, **17**, 201–202.
Grueber CE, Wallis GP, Jamieson IG (2008) Heterozygosity-fitness correlations and their relevance to studies on inbreeding in threatened populations. *Molecular Ecology*, **17**, 3978–3984.
Guicking D, Kroger-Kilian T, Weising K, Blattner FR (2008) Single nucleotide sequence analysis: a cost- and time-effective protocol for the analysis of microsatellite- and indel-rich chloroplast DNA regions. *Molecular Ecology Resources*, **8**, 62–65.
Gustafsson L (1986) Lifetime reproductive success and heritability: empirical support for Fisher's Fundamental Theorem. *American Naturalist*, **128**, 761–764.
Hansson B, Westerberg L (2002) On the correlation between heterozygosity and fitness in natural populations. *Molecular Ecology*, **11**, 2467–2474.
Hansson B, Westerberg L (2008) Heterozygosity-fitness correlations within inbreeding classes: local or genome-wide effects? *Conservation Genetics*, **9**, 73–83.
Hansson B, Bensch S, Hasselquist D, Åkesson M (2001) Microsatellite diversity predicts recruitment of sibling great reed warblers. *Proceedings of the Royal Society B: Biological Sciences*, **268**, 1287–1291.
Hansson B, Westerdahl H, Hasselquist D, Åkesson M, Bensch S (2004) Does linkage disequilibrium generate heterozygosity-fitness correlations in great reed warblers? *Evolution*, **58**, 870–879.
von Hardenberg A, Bassano B, Festa-Bianchet M *et al*. (2007) Age-dependant genetic effects on a secondary sexual trait in male Alpine ibex, *Capra ibex*. *Molecular Ecology*, **16**, 1969–1980.
Hedges LV (1983) A random effects model for effect sizes. *Psychological Bulletin*, **93**, 388–395.
Hedges LV, Olkin I (1985) *Statistical Methods for Meta-Analysis*. Academic Press, New York.
Hedrick PW (2005) *Genetics of Populations*, 3rd edn. Jones and Bartlett, Sudbury, Massachusetts.
Hedrick PW, Fredrickson R, Ellegren H (2001) Evaluation of d^2, a microsatellite measure of inbreeding and outbreeding, in wolves with a known pedigree. *Evolution*, **55**, 1256–1260.
Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.
Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, **4**, 1–10.
Hoffmann AA, Hercus MJ (2000) Environmental stress as an evolutionary force. *Bioscience*, **50**, 217–226.
Hoffman JI, Forcada J, Amos W (2006) No relationship between microsatellite variation and neonatal fitness in Antarctic fur seals, *Arctocephalus gazella*. *Molecular Ecology*, **15**, 1995–2005.
Houle D (1992) Comparing evolvability of quantitative traits. *Genetics*, **130**, 195–204.
Jennions MD, Møller AP, Hunt J (2004) Meta-analysis can 'fail': reply to Kotiaho and Tomkins. *Oikos*, **104**, 191–193.
Kingsolver JG, Hoekstra HE, Hoekstra JM *et al*. (2001) The strength of phenotypic selection in natural populations. *American Naturalist*, **157**, 245–261.
Knowles SCL, Nakagawa S, Sheldon BC (2009) Elevated reproductive effort increases blood parasitaemia and decreases immune function in birds: a meta-regression approach. *Functional Ecology*, **23**, 405–415.
Koricheva J (2003) Non-significant results in ecology: a burden or a blessing in disguise? *Oikos*, **102**, 397–401.
Kotiaho JS, Tomkins JL (2002) Meta-analysis, can it ever fail? *Oikos*, **96**, 551–553.
Kretzmann M, Mentzer L, DiGiovanni R, Leslie MS, Amato G (2006) Microsatellite diversity and fitness in stranded juvenile harp seals (*Phoca groenlandica*). *Journal of Heredity*, **97**, 555–560.
Kruuk LEB, Hill WG (2008) Introduction. Evolutionary dynamics of wild populations: the use of long-term pedigree data. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 593–596.
Kruuk LEB, Clutton-Brock TH, Slate J *et al*. (2000) Heritability of fitness in a wild mammal population. *Proceedings of the National Academy of Sciences, USA*, **97**, 698–703.
Laan M, Pääbo S (1997) Demographic history and linkage disequilibrium in human populations. *Nature Genetics*, **17**, 435–438.
Lacy RC, Petric A, Warneke M (1993) Inbreeding and outbreeding in captive populations of wild animal species. In: *The Natural History of Inbreeding and Outbreeding* (ed. Thornhill NW), pp. 352–374. University of Chicago Press, Chicago, Illinois.
Leimu R, Mutikainen P, Koricheva J, Fischer M (2006) How general are positive relationships between plant population size, fitness and genetic variation? *Journal of Ecology*, **94**, 942–952.
Leinonen T, O'Hara RB, Cano JM, Merilä J (2008) Comparative studies of quantitative trait and neutral marker divergence: a meta-analysis. *Journal of Evolutionary Biology*, **21**, 1–17.
Lens L, Van Dongen S, Galbusera P *et al*. (2000) Developmental instability and inbreeding in natural bird populations exposed to different levels of habitat disturbance. *Journal of Evolutionary Biology*, **13**, 889–896.
Lesbarreres D, Primmer SR, Laurila A, Merilä J (2005) Environmental and population dependency of genetic variability-fitness correlations in *Rana temporaria*. *Molecular Ecology*, **14**, 311–323.
Lieutenant-Gosselin M, Bernatchez L (2006) Local heterozygosity-fitness correlations with global positive effects on fitness in threespine stickleback. *Evolution*, **60**, 1658–1668.
Luikart G, Pilgrim K, Visty J, Ezenwa VO, Schwartz MK (2008) Candidate gene microsatellite variation is associated with parasitism in wild bighorn sheep. *Biology Letters*, **4**, 228–231.
Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer & Associates, Inc, Sunderland, Massachusetts.
Marshall TC, Spalton JA (2000) Simultaneous inbreeding and outbreeding depression in reintroduced Arabian oryx. *Animal Conservation*, **3**, 241–248.
Matt GE, Cook TD (1994) Threats to the validity of research syntheses. In: *The Handbook of Research Synthesis* (eds Cooper H, Hedges LV), pp. 503–520. Russell Sage Foundation, New York.

McCleery RH, Pettifor RA, Armbruster P *et al*. (2004) Components of variance underlying fitness in a natural population of the great tit (*Parus major*). *American Naturalist*, **164**, E62–E72.

Merilä J, Crnokrak P (2001) Comparison of genetic differentiation at marker loci and quantitative traits. *Journal of Evolutionary Biology*, **14**, 892–903.

Merilä J, Sheldon BC (1999) Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity*, **83**, 103–109.

Merilä J, Sheldon BC (2000) Lifetime reproductive success and heritability in nature. *American Naturalist*, **155**, 301–310.

Mitton JB (1997) *Selection in Natural Populations*. Oxford University Press, Oxford, UK.

Mousseau TA, Roff DA (1987) Natural selection and the heritability of fitness components. *Heredity*, **59**, 181–197.

Myrand B, Tremblay R, Sévigny J-M (2002) Selection against blue mussels (*Mytilus edulis* L.) homozygotes under various stressful conditions. *Heredity*, **93**, 238–248.

Nakagawa S, Ockendon N, Gillespie DOS, Hatchwell BJ, Burke T (2007) Assessing the function of house sparrows' bib size using a flexible meta-analysis method. *Behavioral Ecology*, **18**, 831–840.

Nam I-S, Mengersen K, Garthwaite P (2003) Multivariate meta-analysis. *Statistics in Medicine*, **22**, 2309–2333.

Neff BD (2004a) Mean d^2 and divergence time: transformations and standardizations. *Journal of Heredity*, **95**, 165–171.

Neff BD (2004b) Stabilizing selection on genomic divergence in a wild fish population. *Proceedings of the National Academy of Sciences, USA*, **101**, 2381–2385.

Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, **18**, 83–90.

Ohta T (1971) Associative overdominance caused by linked detrimental mutations. *Genetical Research*, **18**, 277–286.

Olafsdottir GA, Kristjansson T (2008) Correlated pedigree and molecular estimates of inbreeding and their ability to detect inbreeding depression in the Icelandic sheepdog, a recently bottlenecked population of domestic dogs. *Conservation Genetics*, **9**, 1639–1641.

Ortego J, Calabuig G, Cordero PJ, Aparicio JM (2007) Egg production and individual genetic diversity in lesser kestrels. *Molecular Ecology*, **16**, 2383–2392.

Overall ADJ, Byrne KA, Pilkington JG, Pemberton JM (2005) Heterozygosity, inbreeding and neonatal traits in Soay sheep on St Kilda. *Molecular Ecology*, **14**, 3383–3393.

Price T, Schluter D (1991) On the low heritability of life history traits. *Evolution*, **59**, 181–197.

Price EA, Bourne SL, Radbourne R *et al*. (1997) Rare microsatellite polymorphisms in the DNA repair genes XRCC1, XRCC3 and XRCC5 associated with cancer in patients of varying radiosensitivity. *Somatic Cell and Molecular Genetics*, **23**, 237–247.

Pusey A, Wolf M (1996) Inbreeding avoidance in animals. *Trends in Ecology & Evolution*, **11**, 201–206.

Ranum LPW, Day JW (2002) Dominantly inherited, non-coding microsatellite expansion disorders. *Current Opinion in Genetics and Development*, **12**, 266–271.

Raudenbush SW (1994) Random effects models. In: *The Handbook of Research Synthesis* (eds Cooper H, Hedges LV), pp. 301–321. Russell Sage Foundation, New York.

Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. *Evolution*, **55**, 1095–1103.

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. *Conservation Biology*, **17**, 230–237.

Reich DE, Cargill M, Bolk S *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Rijks JM, Hoffman JI, Kuiken T, Osterhaus ADME, Amos W (2008) Heterozygosity and lungworm burden in harbour seals (*Phoca vitulina*). *Heredity*, **100**, 578–593.

Roff DA, Emerson K (2006) Epistasis and dominance: evidence for differential effects in life-history versus morphological traits. *Evolution*, **60**, 1981–1990.

Roff DA, Mousseau TA (1987) Quantitative genetics and fitness: lessons from *Drosophila*. *Heredity*, **59**, 103–118.

Rosenberg MS, Adams DC, Gurevitch J (2000) *Metawin: Statistical Software for Meta-Analysis*, Version 2. Sinauer & Associates, Sunderland, Massachusetts.

Rosenthal R (1991) *Meta-Analytic Procedures for Social Research*. Sage, Newbury Park, California.

Rosenthal R (1994) Parametric measures of effect size. In: *The Handbook of Research Synthesis* (eds Cooper H, Hedges LV), pp. 231–244. Russell Sage Foundation, New York.

Rossiter SJ, Jones G, Ransome RD, Barratt EM (2001) Outbreeding increases offspring survival in wild greater horseshoe bats (*Rhinolophus ferrumequinum*). *Proceedings of the Royal Society B: Biological Sciences*, **268**, 1055–1061.

Rowe G, Beebee TJC, Burke T (1999) Microsatellite heterozygosity, fitness and demography in natterjack toads *Bufo calamita*. *Animal Conservation*, **2**, 85–92.

Samollow PB, Soulé ME (1983) A case of stress related heterozygote superiority in nature. *Evolution*, **37**, 646–649.

Schuelke M (2000) An economic method for the flourescent labeling of PCR fragments. *Nature Biotechnology*, **18**, 233–234.

Scott TM, Koehn RK (1990) The effect of environmental stress on the relationship of heterozygosity to growth-rate in the coot clam *Mulinia lateralis* (Say). *Journal of Experimental Marine Biology and Ecology*, **135**, 109–116.

Seddon N, Amos W, Mulder RA, Tobias JA (2004) Male heterozygosity predicts territory size, sone structure and reproductive success in a cooperatively breeding bird. *Proceedings of the Royal Society B: Biological Sciences*, **271**, 1823–1829.

Shadish WR, Haddock CK (1994) Combining estimates of effect size. In: *The Handbook of Research Synthesis* (eds Cooper H, Hedges LV), pp. 261–281. Russell Sage Foundation, New York.

Shikano T, Taniguchi N (2002) Relationships between genetic variation measured by microsatellite DNA markers and a fitness-related trait in the guppy (*Poecilia reticulata*). *Aquaculture*, **209**, 77–90.

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Slate J, Pemberton JM (2002) Comparing molecular measures for detecting inbreeding depression. *Journal of Evolutionary Biology*, **15**, 20–31.

Slate J, Pemberton J (2006) Does reduced heterozygosity depress sperm quality in wild rabbits (*Oryctolagus cuniculus*)? *Current Biology*, **16**, R790–R791.

Slate J, Kruuk LEB, Marshall TC, Pemberton JM, Clutton-Brock TH (2000) Inbreeding depression influences lifetime breeding success in a wild population of red deer (*Cervus elaphus*). *Proceedings of the Royal Society B: Biological Sciences*, **267**, 1657–1662.

Slate J, David P, Dodds KG *et al*. (2004) Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. *Heredity*, **93**, 255–265.

Streelman JT, Kocher TD (2002) Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiological Genomics*, **9**, 1–4.

Symonds VV, Lloyd AM (2004) A simple and inexpensive method for producing fluorescently labelled size standard. *Molecular Ecology Notes*, **4**, 768–771.

Tomkins JL, Kotiaho JS (2004) Publication bias in meta-analysis: seeing the wood for the trees. *Oikos*, **104**, 194–196.

Tsitrone A, Rousset F, David P (2001) Heterosis, marker mutational processes and population inbreeding history. *Genetics*, **159**, 1845–1859.

Väli U, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808–3817.

Välimäki K, Hinten GN, Hanski I (2007) Inbreeding and competitive ability in the common shrew (*Sorex araneus*). *Behavioral Ecology and Sociobiology*, **61**, 997–1005.

Van Buskirk J, Willi Y (2006) The change in quantitative genetic variation with inbreeding. *Evolution*, **60**, 2428–2434.

Wall JD, Andolfatto P, Przeworski M (2002) Testing models of selection and demography in *Drosophila simulans*. *Genetics*, **162**, 203–216.

Wang J, Hill WG, Charlesworth D, Charlesworth B (1999) Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. *Genetics Research*, **74**, 165–178.

Wang D, Shi J, Carlson SR *et al*. (2003) A low-cost, high-throughput polyacrylamide gel electrophoresis system for genotyping with microsatellite DNA markers. *Crop Science*, **43**, 1828–1832.

Weir B, Cockerham CC (1973) Mixed self and random mating at two loci. *Genetical Research*, **21**, 247–262.

Weir BS, Avery PJ, Hill WG (1980) Effect of mating structure on variation in inbreeding. *Theoretical Population Biology*, **18**, 396–429.

Westgaard J-I, Fevolden S-E (2007) Atlantic cod (*Gadus morhua* L.) in inner and outer coastal zones of northern Norway display divergent genetic signature at non-neutral loci. *Fisheries Research*, **85**, 306–315.

Wilson AJ, Pemberton JM, Pilkington JG *et al*. (2006) Environmental coupling of selection and heritability limits evolution. *PLoS Biology*, **4**, 1270–1275.

Wright AF, Carothers AD, Pirastu M (1999) Population choice in mapping genes for complex diseases. *Nature Genetics*, **23**, 397–404.

Supporting information

Additional Supporting information may be found in the online version of this article:

Appendix S1 Further materials and methods.

Appendix S2 Additional results.

Appendix S3 References for the Supporting information.

Fig. S1 Distribution of effect sizes (mean with confidence interval for each estimate) by genetic metric and trait type (A: MLH; B: SH, C: IR; D: mean d^2; E: St d^2).

Fig. S2 Paired plot of weighted mean effect sizes for life-history (LH, left axis) and morphological (M, right axis) traits reported in the same study.

Fig. S3 Relationship between (A) sample size and population size on a $\log_{10}$ scale for those populations for which there were size estimates available ($n = 57$), (B) population average heterozygosity and population size on a $\log_{10}$ scale for those populations for which both estimates available ($n = 42$), (C) sample size and demographic history in wild populations ($n = 76$), and (D) population average heterozygosity and demographic history in wild populations ($n = 58$).

Table S1 Studies included in the meta-analysis of heterozygosity fitness correlations

Table S2 Model selection of variables influencing the magnitude of overall mean effect size in wild populations

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.