MOLECULAR ECOLOGY

Molecular Ecology (2017) **26**, 2410–2425

NEWS AND VIEWS

OPINION

Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses

DANIEL W.A. NOBLE,*[1] MALGORZATA LAGISZ,*[1] ROSE E. O'DEA*† and SHINICHI NAKAGAWA*†

*Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Kensington, NSW, Australia; †Diabetes and Metabolism Division, Garvan Institute of Medical Research, Sydney, NSW, Australia

Abstract

Meta-analysis is an important tool for synthesizing research on a variety of topics in ecology and evolution, including molecular ecology, but can be susceptible to nonindependence. Nonindependence can affect two major interrelated components of a meta-analysis: (i) the calculation of effect size statistics and (ii) the estimation of overall meta-analytic estimates and their uncertainty. While some solutions to nonindependence exist at the statistical analysis stages, there is little advice on what to do when complex analyses are not possible, or when studies with nonindependent experimental designs exist in the data. Here we argue that exploring the effects of procedural decisions in a meta-analysis (e.g. inclusion of different quality data, choice of effect size) and statistical assumptions (e.g. assuming no phylogenetic covariance) using sensitivity analyses are extremely important in assessing the impact of nonindependence. Sensitivity analyses can provide greater confidence in results and highlight important limitations of empirical work (e.g. impact of study design on overall effects). Despite their importance, sensitivity analyses are seldom applied to problems of nonindependence. To encourage better practice for dealing with nonindependence in meta-analytic studies, we present accessible examples demonstrating the impact that ignoring nonindependence can have on meta-analytic estimates. We also provide pragmatic solutions for dealing with nonindependent study designs, and for analysing dependent effect sizes. Additionally, we offer reporting guidelines that will facilitate disclosure of the sources of nonindependence in meta-analyses, leading to greater transparency and more robust conclusions.

Keywords: hierarchical structure, meta-analysis, meta-regression, mixed models, multilevel models, quantitative research synthesis, random effects

Received 3 July 2016; revision received 8 January 2017; accepted 10 January 2017

Correspondence: Daniel W.A. Noble and Shinichi Nakagawa, Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Kensington, NSW 2052, Australia,
E-mails: daniel.wa.noble@gmail.com, s.nakagawa@unsw.edu.au

[1]Authors contributed equally to the preparation of this manuscript.

Introduction

Meta-analyses have become indispensable research tools for synthesizing primary research studies on a variety of topics in ecology and evolution (Nakagawa & Poulin 2012; Koricheva *et al.* 2013; Vetter *et al.* 2013; Koricheva & Gurevitch 2014; ArchMiller *et al.* 2015). Their application is far reaching. They can, for example, inform policy, permit researchers to revise and refine current theoretical paradigms, establish more rigorous empirical tests of theory and help determine new research directions (Borenstein *et al.* 2009; Vetter *et al.* 2013; Koricheva & Gurevitch 2014). The field of molecular ecology has already benefited greatly from meta-analyses over the last 10 years. Some examples include helping clarify the usefulness of G_{ST} for estimating genetic differentiation (Heller & Siegismund 2009), the strength of heterozygosity–fitness correlations across animal taxa (Chapman *et al.* 2009) and the effects of habitat fragmentation on plant genetic diversity (Aguilar *et al.* 2008). Given that meta-analyses differ from traditional qualitative (narrative) reviews by providing a quantitative overview of a specific research question, they hold a privileged place in various research fields (Koricheva & Gurevitch 2014).

Numerous reviews have helped establish fundamental guidelines on how to conduct and report meta-analyses (Gates 2002; Borenstein *et al.* 2009; Cooper *et al.* 2009; Koricheva *et al.* 2013), outlining the steps for transparent data extraction and collection protocols (Liberati *et al.* 2009), the importance of moderator variables (predictor variables in a normal linear model) in explaining effect size *heterogeneity* (Thompson & Higgins 2002—see Table 1 for definitions of italicized words) and some ways in which meta-analysts can effectively deal with sources of *nonindependence* (Lajeunesse 2011; Nakagawa & Santos 2012; Mengersen *et al.* 2013). While there have been excellent discussions on how to correct for many sources of nonindependence (Borenstein *et al.* 2009; Cooper *et al.* 2009; Lajeunesse 2011; Nakagawa & Santos 2012; Mengersen *et al.* 2013; Koricheva & Gurevitch 2014), these are often not comprehensively

Table 1 Glossary of terms

Term	Definition
Effect size statistics (effect statistics)	Quantitative measure of the magnitude, or size of an effect, for example difference between two groups, association between two variables, risk of event happening; for more details, see Nakagawa & Cuthill (2007)
Effect size weights (weight)	Quantitative measure of (un)certainty for each effect size. For standardized effect statistics, such as Hedges' *g* or correlation coefficients, weights (or the inverse of weights, i.e. variance) are calculated via formulas based on sample size
Fixed-effect model	Statistical approach for combining effect sizes assuming that study outcomes vary across studies purely due to sampling error, and observed effect sizes share a common underlying effect. The fixed-effect model can have a moderator(s) in the form of meta-regression when observed variation across effect sizes is due to the moderator, for example temperature (yet, all the effect sizes share a common effect). The term 'fixed effect' should not be confused with 'fixed effects' used in linear (mixed) modelling. This is a misnomer and 'fixed effect' should have probably been named 'common effect' (Borenstein *et al.* 2009)
Heterogeneity	Variation not explained by sampling error (often referred to as within-study variance) among effect sizes. In traditional models (fixed-effect and random-effects models), a test of heterogeneity can be conducted by calculating the *Q* statistic. In a (non-multilevel) random-effects model, heterogeneity is the proportion of between-study variation (often referred to as τ^2) to the total variance (i.e. within-study variance and between-study variance); this proportion is named I^2. For more details on I^2, see Thompson & Higgins (2002), and for an extension of I^2, see Nakagawa & Santos (2012) and Cheung (2014)
Meta-regression	Statistical approach for combining effect sizes which is analogous to multiple regression models. Continuous and categorical variables (moderators or fixed effects) are included in a meta-analytic model (i.e. fixed-effect, random-effects or multilevel model) to understand hypothesized drivers of heterogeneity in effect size estimates across studies (or across different levels, such as species)
Multilevel/hierarchical (meta-analytic) model	Statistical models with multiple random effects/factors such as species, population and experiments (see 'random-effects model' below). For more details, see Nakagawa & Santos (2012) and Cheung (2014)
Nonindependence	Violation of the assumption of data independence, for example when multiple effect sizes are taken on the same individuals, or the same control group is used for comparisons with more than one treatment group (for more details, see the main text)
Publication bias (analysis)	Publication bias in meta-analyses can have numerous causes (Rothstein *et al.* 2005), but most famously is due to overrepresentation of positive (statistical significant) results over negative (nonsignificant) results in the literature. Publication bias analyses can be categorized into two groups: (i) methods identifying publication bias and (ii) methods correcting for publication bias (Sutton 2009)
Random-effects model	Statistical approach for combining effect sizes when there are differences between studies that affect estimation of overall effect size in addition to sampling error within studies. Random-effects models assume that effect sizes in each study have different underlying effects. The model contains both between- and within-study variance, which allows the quantification of heterogeneity (i.e. I^2). The random-effects model does also include a random effect, which is the study-specific random effect
Sensitivity analysis	A set of alternative statistical analyses, which investigate the robustness of results from an original analysis (for more details, see the main text)

examined and connections between various statistical approaches can be unclear (but see Koricheva *et al.* 2013 for a notable exception). Furthermore, there is little advice on ways to deal with nonindependence between effect sizes if appropriate modelling is not possible or when a meta-analysis contains studies with nonindependent experimental designs.

In this article, we use examples to discuss why nonindependence is a problem and suggest pragmatic solutions to ameliorate these problems, particularly if data limitations exist. While our examples are framed around questions and data that are likely to be encountered in the field of molecular ecology, the concepts and sources of nonindependence apply to meta-analyses in ecology and evolution more generally. We argue that there is a need for greater transparency about sources of nonindependence in published papers, and the use of *sensitivity analyses* to thoroughly explore the consequences of violating independence assumptions on study conclusions.

Nonindependence and the role of sensitivity analyses

Formal meta-analyses involve a systematic search for relevant literature (for studies testing a particular question), extraction of data from these studies (those meeting a set of inclusion criteria), generation of *effect size statistics* from these data and statistical analysis of effect size estimates. Each effect size is weighted by the inverse of their sampling error variance (i.e. *effect size weight* or *weight*), to estimate an overall mean effect size. The pooled (i.e. mean) effect size estimate takes into consideration that more precise effect size estimates have greater influence on results. Moderator variables can also be included in analyses to understand whether hypothesized differences between studies (or effect sizes) explain heterogeneity (variance) among effects (using a so-called *meta-regression* model).

Nonindependence among effect size estimates is probably the most widespread statistical problem faced by meta-analysts in this process, and can have important consequences on study conclusions (Hedges 2009a; Cheung 2014; for an excellent recent review of statistical nonindependence in a more general context, see Forstmeier *et al.* 2016). Ideally, a meta-analysis would involve a single effect size estimate being derived for each study, making every effect size within the meta-analysis statistically independent. However, for meta-analyses in ecology and evolution, it is often the case that effect size estimates are related to each other (i.e. are correlated) at various hierarchical levels, possibly because they come from the same study, are derived through comparisons with the same control group or are from correlated traits (Nakagawa & Santos 2012; Mengersen *et al.* 2013). While discussions of nonindependence often focus on nonindependence between effect sizes, it is often forgotten that nonindependence in primary studies (i.e. the specific study design) can also affect the calculation of effect size statistics (Hedges 2007, 2009a) further impacting results from a meta-analysis. It is therefore not surprising that problems related to nonindependence have been highlighted in ecological and evolutionary reviews of meta-analyses (Chamberlain *et al.* 2012; Koricheva & Gurevitch 2014; ArchMiller *et al.* 2015), suggesting a major under appreciation of its potential impact on study conclusions.

We argue that sensitivity analyses are an integral part of the solution to deal with problems associated with nonindependence, and provide some examples of common and useful approaches to explore the impact of nonindependence on meta-analytic results. While we often think about sensitivity analyses as a way of determining how individual data points impact results, it need not be restricted to these situations. More generally, sensitivity analyses are additional analyses that evaluate the effect of procedural decisions made by the meta-analyst (e.g. choice of effect size, inclusion or exclusion of low-quality data) and statistical assumptions (e.g. no phylogenetic covariance) on conclusions. They can include analysing subsets of data, or using alternative statistical models that make different assumptions, and presenting these analyses in the same paper (Koricheva & Gurevitch 2014). Indeed, *publication bias* analysis, a type of sensitivity analysis, is already commonly used to explore the impact of missing unpublished studies (i.e. overrepresentation of certain research results in the meta-analytic data set) on meta-analytic results (Rothstein *et al.* 2005). Despite their importance, sensitivity analyses are seldom utilized to deal with problems of nonindependence, even though they provide greater confidence in meta-analysis results (Chamberlain *et al.* 2012; Koricheva & Gurevitch 2014).

Dealing with nonindependence when calculating effect sizes: considering study design

Effect size calculations make important assumptions about the data extracted from studies. Common effect size equations (Table 2) assume that the individual samples (replicates) collected during a study are independent and, if comparing two groups, that these groups are independent from each other (i.e. a replicate is not present in both groups). Formulas for effect size and corresponding sampling error variance have been derived to ameliorate the problems of nonindependence when correlations exist between groups for specific study designs (e.g. for matched-pairs designs; See Dunlap *et al.* 1996; Lajeunesse 2011; Borenstein *et al.* 2009). However, these only apply to study designs where each replicate is present in both groups, for example in treatment and control groups (Fig. 1A). Such balanced designs are unlikely to be generally applicable to experiments in ecology and evolution. It is more common to have a mixture of nested and crossed sampling designs with clustered replication (Fig. 1B, C). Sampling design can cause a number of problems for effect size calculations, and these are often underappreciated in ecological and evolutionary meta-analyses. We describe some of these problems below, their impact on effect size calculations and meta-analytic results, and provide some possible solutions.

Problem 1: nonindependence affects the sampling error variance of an effect size. Nonindependence will artificially inflate sample size, increasing the magnitude of the denominator in variance equations and, thus, decreasing the sampling error variance for an effect size (see Table 2). This is most easily demonstrated by considering the calculation of the sampling error variance for Fisher's *Z*-transformed correlation coefficient, V_z (Table 2). Here, V_z is inversely proportional to the sample size, n. Increased sample size will therefore decrease the value of V_z, making the effect size estimate (z) appear more precise. Nonindependence will also decrease variances of other effect size statistics. For example, studies that present statistics based on samples taken from clutches or broods where more than one sibling was included in the sample (clustered designs —Fig. 1B, C) would result in artificially inflated sample sizes and reduced sampling variance for comparisons between control and experimental groups.

An accurate calculation of sampling error variance is also important because the amount of sampling error variance influences the estimation of heterogeneity, the absolute

Table 2 Common effect sizes used in ecology and evolution, their calculation, the effects of nonindependence on their values and suggestions on dealing with nonindependence. Symbols are defined as: M_1 = mean of treatment 1, M_2 = mean of treatment 2, n_1 = sample size of treatment 1, n_2 = sample size of treatment 2, SD_1 = standard deviation of treatment 1, SD_2 = standard deviation of treatment 2, SD_p = pooled or weighted standard deviation, r = correlation coefficient, n = total sample size, t = t-value from two-group statistical tests (e.g. t-test) assuming independence, F = F-value from a two-group ANOVA test assuming independence, d.f. = degrees of freedom of the statistical test

Effect size	Calculation of mean and variance	Effect of nonindependence	Suggestions
Based on descriptive statistics			
Standardized mean difference (Cohen's d; Hedges' d or g)	$SD_p = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{(n_1+n_2-2)}}$ $d = \frac{M_1-M_2}{SD_p}$ $V_d = \frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}$ Hedges' g: correction factor, J, can be multiplied by d to get g and V_d corrected by multiplying by J^2 to get V_g $J = 1 - \frac{3}{4(n_1+n_2-2)-1}$	Value of d (or g) is influenced by changes in the pooled standard deviation (SD_p) resulting from effects of nonindependence on n and poor estimation of SD_1 and SD_2. Increased d and n will directly affect its variance (V_d). Nonindependence can arise from lack of random sampling within each treatment or nonindependent 'replicates' across treatment groups	**1** Use an alternative effect size that is *less* susceptible to nonindependence (e.g. lnRR) **2** Use 'effective sample size' for n by dividing the original sample size by the design effect or by multiplying V_d by the square root of the design effect. The 'design effect' is equal to 1 + (M−1) × ICC, where M is the average cluster size (See Higgins & Green 2009, p. 496) **3** Use an estimate of ICC and modified effect size equations for d and V_d for clustered samples (Hedges 2009b, pp. 341–345) to account for simple clustering situations **4** For nonindependence between groups, use equations for matched-pairs designs (balanced—see Borenstein 2009) and ICC for unbalanced nonindependent treatment groups **5** Conduct sensitivity analysis using alternative effect size statistics, coding a study design moderator and/or varying values of ICC (e.g. 0, 0.5, 1). Report on how analysis results were affected
log response ratio (lnRR)	$\mathrm{lnRR} = \ln\left(\frac{M_1}{M_2}\right)$ $V_{\mathrm{lnRR}} = \frac{SD_1^2}{n_1 M_1} + \frac{SD_2^2}{n_2 M_2}$	Value of lnRR is not affected by nonindependent samples; however, the variance in lnRR (V_{lnR}) is affected by changes in n resulting from lack of independent random sampling and changes in SD resulting from lack of independence among replicates	**1** Use 'effective sample size' for n by dividing the original sample size by the design effect or by multiplying V_d by the square root of the design effect. The 'design effect' is equal to 1 + (M − 1) × ICC, where M is the average cluster size (See Higgins & Green 2009, p. 496) **2** Not possible to convert to other effect sizes, so conduct sensitivity analysis by varying values of ICC (e.g. 0, 0.5, 1) and/or by coding study design type for each effect size and include these codes as a moderator in analysis. Report on how results are affected
Fisher's Z-transformed correlation coefficient (r)	$z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$ $V_z = \frac{1}{n-3}$	Values of z-transformed correlation coefficients are not affected by nonindependence, but sample size (n) is inversely proportional to the variance in z (V_z). Inflated sample sizes resulting from pseudoreplication will decrease V_z	**1** Use 'effective sample size' for n by dividing the original sample size by the design effect or by multiplying V_d by the square root of the design effect. The 'design effect' is equal to 1 + (M−1) × ICC, where M is the average cluster size (See Higgins & Green 2009, p. 496)

Table 2 *Continued*

Effect size	Calculation of mean and variance	Effect of nonindependence	Suggestions
			2 Conduct sensitivity analysis by coding study design effect for each effect size, using alternative effect size statistics and/or varying values of ICC (e.g. 0, 0.5, 1). Report on how analysis results are affected
Based on inferential statistics			
t-statistics (independent *t*-tests)	*d*-type $d = \frac{t(n_1 + n_2)}{\sqrt{n_1 n_2}\sqrt{\text{d.f.}}}$ $V_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$ *r*-type $r = \frac{t}{\sqrt{t^2 + \text{d.f.}}}$ $z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$ $V_z = \frac{1}{n-3}$	Values of *d* and *r* calculated from *t*-statistics are affected by nonindependence via incorrect degrees of freedom (d.f.) and large *t*-values because of increased sample size. Variance estimates of both *d* and *r* will be affected in the same way by inflated samples sizes as indicated above	**1** Extract *t* from two-group tests. Ensure correct d.f. and *n* used for statistical test are extracted. Do not use conservative d.f. or *n*. Use 'effective sample size' to calculate V_d or V_z by dividing the original sample size by the 'design effect', $1 + (M - 1) \times$ ICC, where *M* is the average cluster size (See Higgins & Green 2009, p. 496). Alternatively, multiply V_d or V_z by the square root of design effect **2** Use an estimate of ICC and modified effect size equations for *t*-statistics and V_d for clustered samples (Hedges 2009b, p. 341–345) **3** Obtain raw data and calculate descriptive statistics to calculate effect sizes. Correct nonindependence problems as recommended above **4** Conduct sensitivity analysis by coding study design effect for each effect size and/or varying values of ICC (e.g. 0, 0.5, 1). Report on how analysis results are affected
F-statistics (one-way ANOVA)	*d*-type $d = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}$ *r*-type $r = \sqrt{\frac{F}{F + n_1 + n_2 - 2)}}$	See *t*-statistics. Note that when the numerator d.f. = 1 for *F*, $t^2 = F$	See *t*-statistics

value of which is the between-study variance, or tau-squared τ^2, and the relative value of which is I^2 (see also Table 1). The estimation of heterogeneity could, in turn, affect decisions about the type of model to use in a meta-analysis (i.e. the *fixed-effect model* vs. the *random-effects model*). This is because meta-analysts can choose the fixed-effect model over the random-effects model when little to no heterogeneity among effect sizes exists (e.g. $\tau^2 = 0$ and $I^2 = 0$). It is notable, however, that Senior *et al.* (2016) recently showed total heterogeneity in ecological and evolutionary meta-analyses to be very high on average (~92%), which indicates that random-effects models are more appropriate for typical meta-analyses in ecology and evolution. Nonetheless, we should be aware that incorrect calculations of sampling error variance can lead to the incorrect estimation of heterogeneity in a meta-analysis (i.e. τ^2, and I^2; for *multilevel/hierarchical models*, it affects every variance component in the model). The estimation of among-study consistency or heterogeneity is as important as the overall mean effect size because its interpretation will depend on how much variation among effect sizes exist (Nakagawa *et al.* 2017).

Solution 1: using effective and conservative sample sizes. Sampling error variance for effect sizes can be corrected by either calculating 'effective sample sizes', or using conservative sample sizes for n in sampling variance equations (i.e. V equations in Table 2). When nonindependent sampling designs are present, Higgins & Green (2009) suggest collecting additional data on the number of clusters in each sample. For example, if we have two groups, G_1 and G_2, and each group contains 10 different clutches each composed of 10 siblings (total sample size = 100 per group), we can use this information, along with an estimate of the intraclass correlation coefficient (ICC; also known as 'repeatability' in the field of ecology and evolution; Nakagawa & Schielzeth 2010), to calculate an 'effective sample size' (ICC in this example is the ratio of between-clutch variance to the sum of between- and within-clutch variance). We can then obtain the effective sample size, by dividing the original sample size for each group (i.e. $n = 100$ per group) by the 'design effect' [i.e. $1 + (M - 1) \times \text{ICC}$, where M is the average cluster size; in this case $(100 + 100)/(10 + 10) = 10$]. We then use effective sample sizes, rather than the total sample size, in the calculations of effect size statistics to deal with nonindependence (Higgins & Green 2009). We can also multiply the sampling error variance of the effect estimate itself by the square root of the design effect, and this will result in similar corrections (Higgins & Green 2009). The benefit of using 'effective sample size' is that the largest possible sample size can be used to obtain more precise meta-analytic estimates including both the overall mean and heterogeneity statistics, while also dampening the impact of nonindependence on overall results. The weakness is that we require an estimate of ICC, which is often difficult to obtain. As an example of this approach, Rutkowska *et al.* (2014) used effective sample sizes to deal with a situation where more than one egg in their study came from a single clutch, artificially inflating sample size. In this study, ICC values were estimated from three published data sets, but because only three estimates were available, effect size statistics were

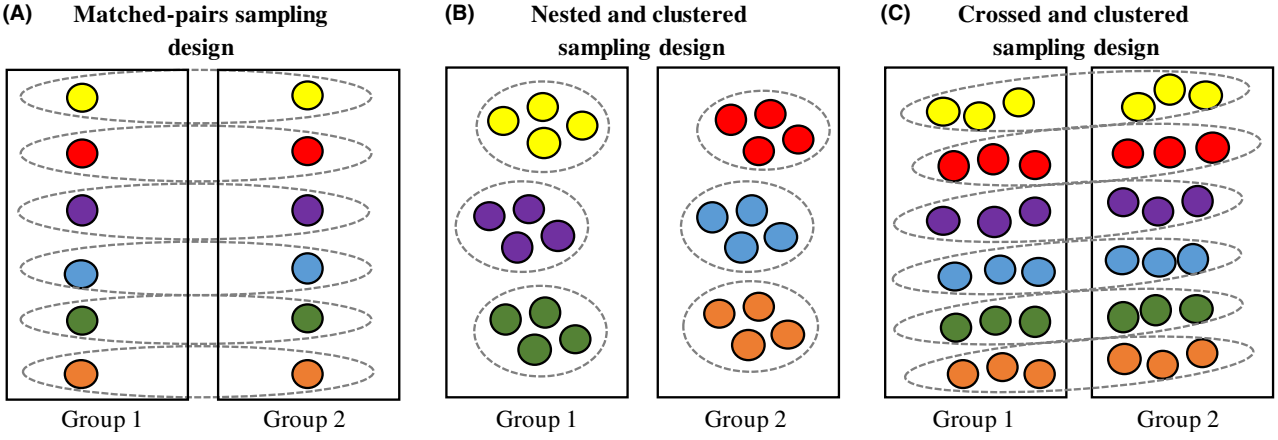

Fig. 1 Examples of three common nonindependent sampling designs. Rectangles represent a single sample or group composed of 'replicates' (each represented as a circle). Similar coloured circles connected by dashed ellipses represent nonindependent replicates. For example, they could be siblings or exactly the same individuals measured more than once. (A) A matched-pairs design where the same individual is present in both group 1 and group 2 creating nonindependence between groups but independence within groups. For example, a situation where data on unrelated individuals was collected before and after some treatment. (B) Two groups containing independent replicates between groups, but clustered nonindependent replicates within groups. For example, a situation where different insect broods or forest patches were randomized to either a control or treatment group, but these manipulations were applied to the entire brood/patch. (C) Two groups that are not independent of each other, but also contain clustered replicates. For example, eggs from clutches that are equally split across two treatments; each contains multiple eggs from a single clutch. [Colour figure can be viewed at wileyonlinelibrary.com]

calculated using ICC values of 0.5 and 0.8 (encompassing the three estimated values) to derive two data sets that were subsequently used in analyses.

If it is difficult to calculate the effective sample size for each group, using a conservative sample size estimate is a more practical approach. For our example above, instead of using $n = 100$, we can use $n = 10$ (number of clusters) to calculate effect size statistics. Conservative sample sizes will reduce the impact of inflated sample sizes on the effect size sampling error variance, but it will also lead to less precise estimates overall, which could increase type II error rates in a meta-analysis. In a similar vein, increased sampling error variance via the use of conservative sample size would result in less accurate heterogeneity estimation when a random-effects model (or a multilevel model) is used. When a fixed-effect model is used, the influence of increased sampling error variance on model estimation could be considerable (see also 'Problem 3' below for more on this point).

If ICC is unavailable, and we are concerned with increased type II error rates and inaccurate heterogeneity estimation, then sensitivity analyses can be presented. To deal with uncertain ICC, we might run meta-analytic models using effect size statistics calculated using effective sample sizes, with ICC values of 0, 0.5 and 1 (or sensible values from the literature). This would be a useful approach in ascertaining the effect nonindependent study designs have on conclusions. ICC values of 0 would cancel out the clustering effect and the design effect would simply be 1, and we would use the total sample size ($n = 100$) for our calculations. In contrast, when ICC = 1, the effective n equals M and we would use the most conservative sample size estimate in all our calculations. In some situations, such as crossed and clustered sampling designs (Fig. 1C), more complex nonindependence will be evident. These situations are more complicated, especially for unbalanced designs; however, ICC can also be used to correct effect size variance estimates when there is dependence between groups. An example of this kind of sensitivity analysis, where multiple meta-analytic models were run with different data sets, can be found in Booksmythe *et al.* (2015).

Problem 2: nested structure can influence effect size (point) estimates. Nested sampling designs (e.g. Fig. 1B) pose additional problems by influencing not only the calculation of sampling error variance, but also point estimates for some effect size statistics (e.g. standardized mean difference, d—including Cohen's d and Hedges' g and d). Standardized mean differences can be affected because they are calculated by dividing the mean difference between two groups by the pooled standard deviation (SD_p—Table 2). As a result, the value of d will be affected by anything that influences SD_p. Samples from nested designs contain 'replicates' from a mixture of biological levels (e.g. within-brood vs. between-brood levels—Fig. 1B). Inadequate sampling at any one of these levels will lead to an underestimation of sample variability at that level, and affect the magnitude and meaning of a samples total variability, SD_T^2 (Hedges 2007, 2009b). As a consequence, this will impact SD_p and, subsequently, the value of d.

To better understand how nested sampling designs can affect point estimates, assume we are interested in the effect habitat fragmentation has on genetic diversity of a single, globally distributed bird species. Genetic diversity estimates are extracted from studies that collected samples from forested and control (fragmented) sites (Fig. 2), and using the standardized mean difference (d), the average number of alleles across microsatellite loci between sites are compared. Using d, we are assuming that the sites sampled in a study contain independent eggs from the two sites (Fig. 2A) and not, for example, samples containing groups of related eggs (Fig. 2B). In a nested sampling design, the total variance of each sample (SD_T^2) can be partitioned into a within-cluster/group variance (SD_{WG}^2) and a between-cluster/group variance (SD_{BG}^2). Considering our bird example, total sample variance within a treatment, SD_T^2, is composed of a within-family variance (SD_{WF}^2) and a between-family variance (SD_{BF}^2). However, the magnitude of variance and number of replicates at each of these levels will affect SD_T^2. Imagine that we were able to estimate the paternity of eggs and it showed that only two different fathers sired eggs from each of the sites (Fig. 2B). We now have only two independent families at each site. Inadequate sampling of independent families (as is the case in Fig. 2B) will lead to an underestimation of SD_{BF}^2, and SD_T^2 will be dominated by SD_{WF}^2. In this example, this leads to an overestimation of d compared to the situation with independent sample replicates.

The effect that nested sampling designs have on d will depend on the amount of replication and magnitude of variance at each hierarchical level. However, while we demonstrate the effect of nested designs on d, given its common use in published meta-analyses in ecology and evolution (~25–30% of studies—Nakagawa & Santos 2012; Koricheva *et al.* 2013), they can also pose problems for correlation coefficients (r). A sample containing a mixture of within- and between-group replicates can affect the direction and magnitude of correlation coefficients (van de Pol & Wright 2009; Dingemanse *et al.* 2012). Therefore, when considering the effects of nested designs on effect size estimates and their sampling variability, we must consider the hierarchical level of our question and how sampling design might impact relevant statistics.

Solution 2: choosing appropriate effect size statistics. One approach to dealing with nested designs is to obtain estimates of sampling variance for each hierarchical level, and use these to calculate effect size statistics with alternative effect size equations proposed by Hedges (2009b). Partitioned sampling variance from mixed models means that one can choose the hierarchical level (i.e. SD_T^2, SD_{BG}^2, SD_{WG}^2) that is most relevant to the question at hand and the types of study designs most prevalent in the data set. Using a common variance ensures that effect sizes are calculated using sample variances with the same meaning. However, obtaining variance estimates of these different

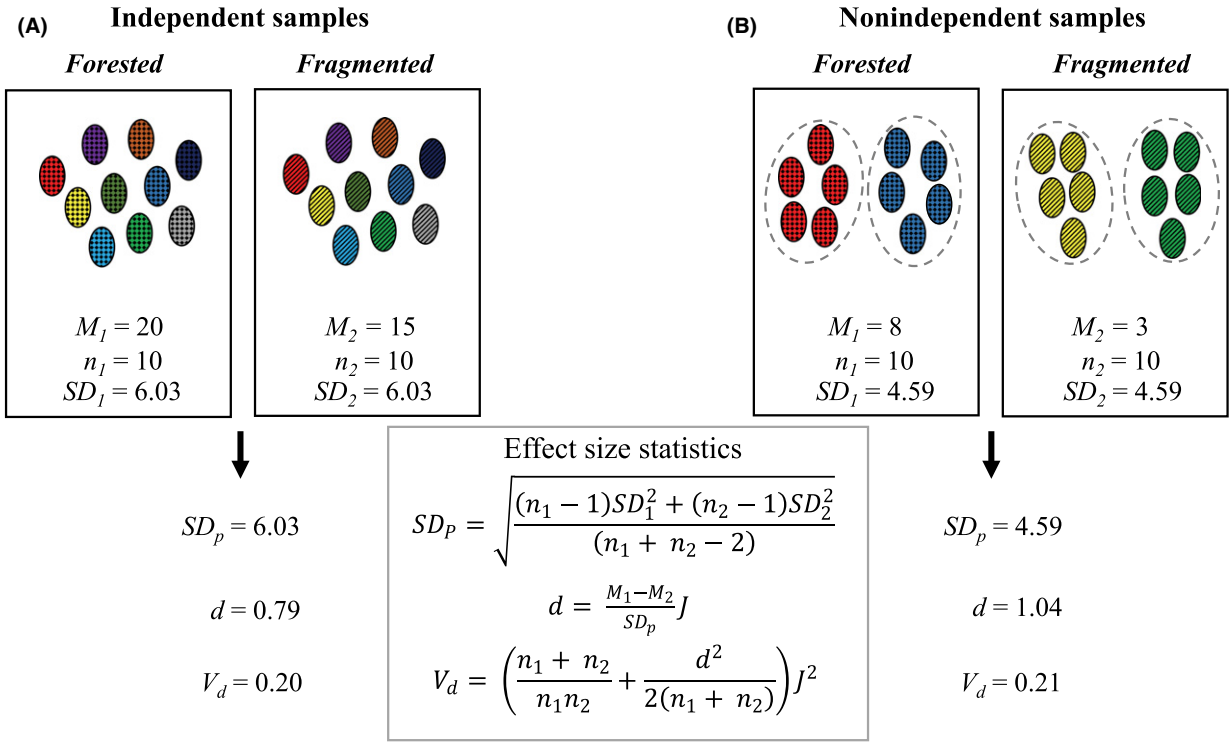

Fig. 2 Effect of nonindependence on the calculation of effect size point estimate. Bird eggs were collected from forested and fragmented habitat, and the genetic diversity (mean number of microsatellite alleles) of this sample was estimated. A total of one egg was collected per nest across 10 different nests in each sample (square box) under the assumption that eggs from different breeding pairs were independent from one another. For simplicity, eggs collected in forested sites are assumed to be independent from eggs in fragmented sites (indicated by different egg patterns) and different coloured eggs within a group are also assumed to be independent. Here, total sample variance, SD_T^2, equals $SD_{BF}^2 + SD_{WF}^2$, where SD_{BF}^2 is the between-family variance and SD_{WF}^2 is the within-family variance in genetic diversity. (A) In the first scenario, eggs sampled from forested and fragmented sites are assumed to be independent, allowing for more effective estimation of SD_{BF}^2 leading to a larger SD_T. (B) Paternity analyses later showed that only two males in each group sired eggs, so that eggs could no longer be considered independent sample replicates. In other words, samples contain related eggs (related eggs are one colour encircled by a dashed ellipse) in the forested and fragmented sites, leading to poorly estimated SD_{BF}^2 and a decrease in SD_T. Below these two scenarios, we show calculations of standardized mean difference, d ('Effect size statistics'—in this case Hedges' g; see Table 2). These calculations illustrate how effect size statistics can be affected despite the same mean difference between forested and fragmented samples ($M_1 - M_2 = 5$ for both scenarios A and B). In this example, $SD_T^2 = 6^2 + 1^2 = 37$ and the 'true' total sample SD is approximately equal to 6.08 (without sampling variability). Effect size is affected by nonindependence, but sampling error variance, V_d, is not strongly affected because the samples sizes for scenarios are the same. [Colour figure can be viewed at wileyonlinelibrary.com]

levels will be difficult without access to the raw data for a given study. Therefore, this approach is unlikely to be generally applicable in ecological and evolutionary studies.

A more practical solution to somewhat account for the effects on nonindependence would be to make use of other effect size statistics (Mengersen *et al.* 2013). For example, both d and its variance can be influenced by study design. We might therefore use the log response ratio, lnRR (Hedges *et al.* 1999), or simply the (unstandardized) mean difference instead. The point estimates of these effect sizes would at least be unaffected by study design (Table 2). The limitations of the standardized mean difference have been recognized in the literature (Osenberg *et al.* 1997; Hedges *et al.* 1999) and this has led to the development and use of lnRR as an alternative (Nakagawa & Santos 2012; Koricheva & Gurevitch 2014). However, using lnRR may limit the type of data that can be used (Borenstein *et al.* 2009), as lnRR requires ratio scale data where measurements are bounded at zero (e.g. height and weight). Additionally, lnRR cannot easily be converted to the other effect sizes statistics like the correlation coefficient (r), standardized mean difference (d) and log odds ratio (lnOR), all of which are convertible to each other. Importantly, these 'solutions' will also not completely solve nonindependence issues, because the variance estimates may still be affected by inflated sample sizes.

Given these limitations, we might decide to keep the use of the standardized mean difference, and instead conduct sensitivity analyses using another effect size statistic (to correct the point estimate) with effective sample sizes (to correct effect size variances), and then compare results.

Multiple effect size statistics have been used in meta-analyses in the past (Isaksson 2010; Besson *et al.* 2016) for other reasons than to deal with nonindependence. If authors were concerned that study design might have important consequences on conclusions, then such an approach would be a useful way to explore the robustness of results (Koricheva & Gurevitch 2014).

Problem 3: effect size from inferential statistics with nonindependent sampling designs. When descriptive statistics (e.g. means, standard deviation, correlation coefficient) cannot be obtained from a paper, effect size statistics (e.g. d) have to be derived from inferential statistics (e.g. t, F, χ^2), assuming that the degrees of freedom (d.f.) and/or sample sizes can also be extracted (Table 2). Often this leads to data sets containing a mixture of effect sizes that have been derived from both descriptive and inferential statistics. Additionally, these estimates must be taken from independent, two-group tests (note that t-values from paired designs—Fig. 1A—have different formula; see Nakagawa & Cuthill 2007 and Borenstein 2009). Nonindependent sampling designs that are incorrectly analysed using statistical tests that assume independence (e.g. independent t-test) will affect the calculation of effect size statistics because t and F values are inflated (due to smaller standard error), leading to larger effect size magnitude. It is, therefore, important that the correct sample size and/or degrees of freedom are extracted from the paper to prevent incorrect calculations of effect size magnitude. In addition, larger sample sizes will also contribute to decreased sampling error variance for these effect sizes (see 'Problem 1'). Larger effect size estimates, in combination with smaller sampling error variance (e.g. variance of d, V_d), can create major problems in the final analysis. This is because effect size estimates themselves are usually weighted by the inverse of their sampling error variance (i.e. a weighted meta-analysis). Importantly, the influence of these problems on results will be larger for fixed-effect than random-effects models because in random-effects models weights are calculated using the sampling error variance along with the between-study variance (i.e. σ_s^2 or more commonly called tau-squared, τ^2).

To better understand the consequences nonindependent study designs can have on the overall results of a meta-analysis, two hypothetical scenarios are presented (Fig. 3). In the first, t-statistics were extracted from six independent studies and were used to calculate, and weight, effect sizes (in this case d—Fig. 3—black circles). We use a random-effects meta-analysis for this example because each study was conducted by independent research groups and are unlikely to share a common underlying effect size (Borenstein *et al.* 2009). From this analysis, we get the pooled effect size estimate and standard error (Fig. 3—black square). However, assume that in reality, we realized that the t-statistics used to calculate d for studies 2, 4 and 6 incorrectly came from independent t-tests that did not contain independent sample replicates. The calculations of d and V_d from these t-statistics would be incorrect because

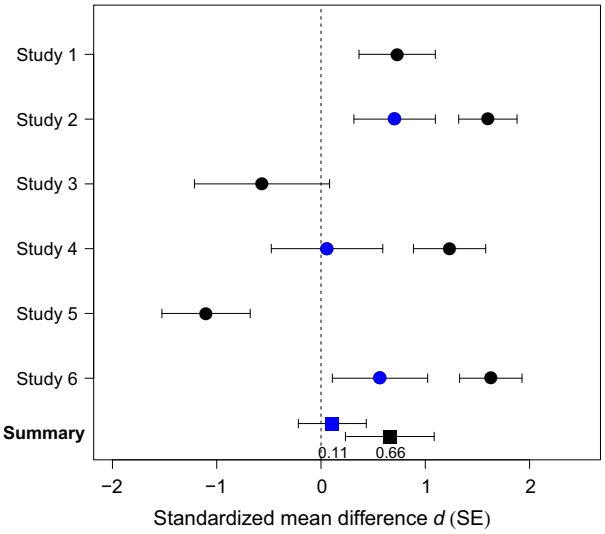

Fig. 3 The consequences of nonindependent study designs on a meta-analysis when effect sizes and their variance estimates are calculated from t-statistics. The standardized mean difference, d, and its sampling error (in this case standard error—SE_d) were derived from t-statistics for six studies (black circles). A random-effects meta-analysis was used to estimate an overall meta-analytic (pooled) mean and its standard error (black square). However, after the analysis, it was recognized that studies 2, 4 and 6 were incorrectly analysed with an independent t-test because they contained groups that had nonindependent replicates (e.g. siblings). If we use correctly calculated t-statistics (blue circles), we get the correct meta-analytic mean and error (blue square) for the six studies. [Colour figure can be viewed at wileyonlinelibrary.com]

large t-values and sample sizes lead to increased d values and a decrease in their sampling variance, V_d (Fig. 3). As most meta-analyses weigh effect sizes based on their precision, if we used correctly calculated t-values (Fig. 3—blue dots), we could estimate the correct pooled estimate and its standard error for the six studies (Fig. 3—blue square). It is not difficult to imagine how this could have more dramatic effects on an analysis, depending on the proportion of studies influenced, and the overall sample size.

Test statistics from studies with nonindependent experimental designs can take into account dependence by applying mixed modelling approaches to account for nested or crossed structure in the data (Pinheiro & Bates 2000). Mixed-effects models are now commonly applied in empirical studies; however, effect size statistics are challenging to calculate for test statistics taken from these models for two reasons. First, it is unclear what the correct degrees of freedom would be for such effect size statistics (Nakagawa & Cuthill 2007; Bolker *et al.* 2009), despite there being a number of approximations that are probably sensible in many cases. Second, it is difficult to deal with different covariates used in mixed-effects models across empirical studies (Nakagawa & Cuthill 2007; Aloe 2015). Note that this second problem is a general issue when

taking test statistics from any type of linear model with covariates. Highly correlated predictors (covariates) in a model can cause the magnitude and sign of the test statistic to differ (Aloe 2015), making it difficult to compare effect size statistics generated from different linear models. The combination of the above problems means that one needs to carefully consider inferential statistics taken from a study with respect to study design and analysis.

Solution 3: resorting to mixed strategies. A number of possible solutions, not without their limitations, exist to mitigate the impact of nonindependent sampling designs on calculations of effect size statistics generated from inferential statistics. These include the following: (i) calculating effect size estimates normally from two-group tests, using incorrect test statistics (e.g. *t* or *F*) and correcting the sampling variance; (ii) resorting to descriptive statistics whenever possible and (iii) coding effect sizes as coming from inferential statistics or nonindependent study designs in meta-regression models.

The first approach is to simply acknowledge study design effects (e.g. in methods section), extract incorrect *F*- and *t*-values from two-group tests and compute effect size statistics using standard methods (Table 2). If we proceed as normal, it is important to extract the correct d.f. used for the original statistical test, because using 'conservative' d.f. instead would further exacerbate the problem by inflating the magnitude of the effect size even more. We can then use an 'effective sample size' to correct the sampling error variance for effect sizes affected by nonindependence (see 'Solution 1'). This approach might be sensible if nonindependent study designs, or the use of inferential statistics, are uncommon in the meta-analytic data set. It is also important to note that, if test statistics are taken from models (even with one predictor), this approach does not deal with the fact that they may be conditioned on other covariates that differ between studies (Aloe 2015).

A second solution might be to only use descriptive statistics from a paper, because these are less impacted by nonindependence and are more easily corrected (see 'Problem 1' and 'Problem 2' along with their solutions). Descriptive statistics can most commonly be obtained by extracting these statistics from figures. Alternatively, the necessary statistics could be calculated from raw data that have been extracted from figures, obtained from authors of the study or downloaded from repositories. The benefit of this approach is that the data structure and its limitations are clear. The downside is that it is difficult to gain access to raw data if these are not already put in repositories, which can be a common problem (Nekrutenko & Taylor 2012; Roche *et al.* 2014; White *et al.* 2015), and it can be time-intensive to extract from figures and/or reanalyse data.

All the above solutions operate on the effect size itself. We could also assess the impact of nonindependence on effect size calculations at the statistical analysis stage (see also 'Dealing with nonindependence after effect size calculations' below). Effect sizes could be coded based on their suspected level of nonindependence, or whether these were derived from inferential statistics. This 'study design' moderator variable can then be included in a meta-regression to explore how its levels impact the mean effect size. If nonindependent study designs have a strong influence on effect sizes and/or their variance, then we would expect this moderator variable to explain heterogeneity in effect sizes, leading to different pooled estimates and standard errors that depend on the type of study design. Alternatively, given that nonindependence can impact results by weighting effect sizes differently, one might conduct a sensitivity analysis that compares both weighted and unweighted meta-analytic models, and then discusses any differences in the results (Koricheva & Gurevitch 2014). Meta-analysts are often not able to calculate effect size weights because essential information (e.g. sample size or SD) is missing. Therefore, an unweighted meta-analysis could be conducted using all effect sizes while a weighted meta-analysis can be carried out using a partial data set. It may be useful to know that unweighted meta-analysis is expected to produce unbiased estimates and it can be more powerful than weighted meta-analysis when many effect sizes are without weights (Morrissey 2016).

Dealing with nonindependence after effect size calculations

Problem 4: nonindependence among effect sizes. A second major way nonindependence can affect a meta-analysis is when effect size estimates themselves are correlated. For example, effect size estimates could be calculated for different species that are phylogenetically related (Adams 2008; Lajeunesse 2009; Hadfield & Nakagawa 2010) or come from the same population, research group or study (Slate & Phua 2003; Borenstein *et al.* 2009; Nakagawa & Santos 2012) (Fig. 4). Multiple effect sizes collected from a single study (i.e. a single research article) may also exhibit more complicated within-study correlation structures. Examples of within-study correlation structures include the calculation of separate effect sizes: (i) that are from correlated traits on the same subjects or treatments (Fig. 4C); (ii) that share a control group (Fig. 4D); (iii) that are spatially related to one another (Fig. 4E) and/or (iv) that have been repeatedly measured (Fig. 4F) (Borenstein *et al.* 2009; Gleser & Olkin 2009; Mengersen *et al.* 2013; Riley *et al.* 2014). Correlated data require careful consideration during a meta-analysis to avoid 'pseudoreplication' (Hurlburt 1984) and ill-informed inferences (Gurevitch & Hedges 1999; Nakagawa & Santos 2012). Treating effect sizes as being independent within a meta-analysis, when in reality they are not, causes standard errors of point parameter estimates to be incorrect (Gurevitch & Hedges 1999) and can impact (but not bias) overall point estimates themselves in weighted meta-analyses (Cheung 2014).

In general, a variety of approaches have been proposed to deal with nonindependence, all with their own strengths and limitations (Borenstein *et al.* 2009; Gleser & Olkin 2009; Lajeunesse 2011; Nakagawa & Santos 2012; Mengersen *et al.* 2013). The usefulness of these approaches will depend on the question of interest, sample size and statistical

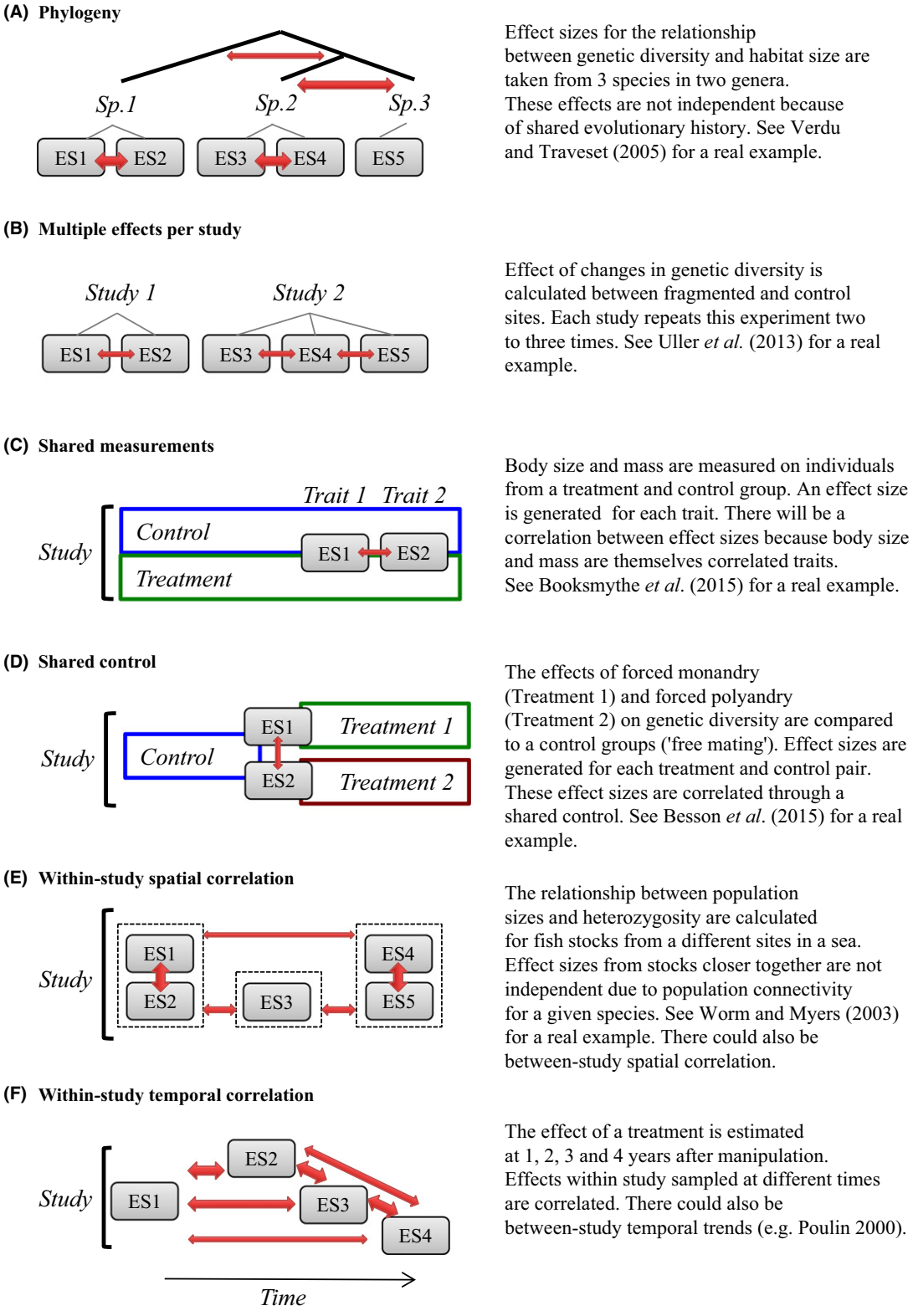

Fig. 4 Common sources of nonindependence in meta-analyses and hypothetical examples of how such nonindependence may arise. References to real examples are provided. 'ES' represents effect size (e.g. log response ratio or lnRR) and double-headed arrows between effect sizes indicate that they are correlated; the size of the arrow indicates the strength of the correlation. [Colour figure can be viewed at wileyonlinelibrary.com]

practicality (i.e. analyst's expertise, effectiveness in estimating parameters). Practical considerations may limit the ability to deal with nonindependence, and in judging the final analysis, it is important to consider the strength of the overall effect being estimated. If the effect is strong, then inferences from an analysis will likely be robust, whether or not all sources of nonindependence have been controlled for. In contrast, more cautious interpretations are needed when overall effect size estimates and confidence intervals are close to zero and not all sources of nonindependence could be accounted for. In these cases, sensitivity analyses will be particularly important in providing readers with a sense of the robustness of inferences drawn. Below, we describe a realistic example of how nonindependence can arise between effect sizes calculated with data from primary literature. We then discuss some practical approaches a meta-analyst might take to eliminate or reduce the impact of such nonindependence on inferences, along with suggestions on useful sensitivity analyses that can be conducted.

Hypothetical example: the role of fragmentation on genetic diversity in tree species

Let us assume we are interested in testing whether fragmentation affects observed heterozygosity in forest tree species (Fig. 5). After a thorough literature search, we found studies reporting results from experiments in which a proportion of forest was cleared in one plot (fragmented) and another plot was left untouched (control). To be included, these studies needed to report experiments in which researchers had collected leaf samples from germinating trees (across a number of different species) from manipulated sites after 10 or more years since the manipulation, as well as from control sites. Exact methodological approaches differed across studies, but in all cases molecular markers were used to estimate the average observed heterozygosity across loci for both forest plot types (fragmented and control). Studies varied in the type of markers used, but all of them estimated observed heterozygosity using SNPs, microsatellite DNA loci or both types of markers. From the extracted data, a log response ratio (lnRR) effect size was generated (Hedges *et al.* 1999) using the average observed heterozygosity across loci for the two plot types (i.e. from fragmented and control plots). While this is a contrived example, it demonstrates many of the realistic data complexities that often arise in ecological and evolutionary meta-analyses. Effect size estimates are not independent of one another due to a number of common sources of nonindependence (see common letters in Figs 4 and 5). In particular, we have multiple effect sizes that are derived from the same species, and a set of species that share a known phylogenetic history (Fig. 5, 'A'). We also have groups of effect sizes that come from the same study (Fig. 5, 'B'). More complicated within-study correlations arise from the fact that in many cases, effect sizes were calculated from microsatellite and SNP markers on the same sample ('Shared measurement'—Fig. 5, 'C'), and effect sizes were derived using a shared control group in two studies (for S2 and S9—Fig. 5, 'D'). How might we circumvent these problems?

Solution 4A: Eliminating nonindependence prior to analysis. Nonindependence among effect size estimates could be dealt with prior to analyses by either: (i) only selecting the most appropriate measure/trait that closely addresses the question (Marín-Martínez & Sánchez-Meca 1999; Cheung 2014); (ii) carrying out separate meta-analyses, one for each type of effect size or species (Borenstein *et al.* 2009; Mengersen *et al.* 2013) or (iii) averaging effect sizes in each study to generate a composite effect size that can be used in a meta-analysis (Marín-Martínez & Sánchez-Meca 1999; Borenstein *et al.* 2009; Mengersen *et al.* 2013; Cheung 2014).

To help deal with nonindependence in our data set above we might (i) only use effect sizes generated with SNP genotype data because they provide more robust estimates of genetic diversity given the larger number of loci; (ii) limit each study to one treatment–control comparison (i.e. randomly sample one effect size from each of the studies 2 and 9; and (iii) analyse each set of collected effect sizes separately for each species. These approaches would remove nonindependence caused by 'shared marker' and 'shared control' comparisons and remove the need to incorporate phylogenetic effects while still allowing us to understand how genetic diversity is affected by fragmentation across all five species. However, such an approach would exclude Study 2 and Study 10 (if only SNP data was used), as well as remove a number of effect size estimates for the other 'treatment–control' comparisons (i.e. those from Study 2 and Study 9) dropping analytical power. Additionally, analysing each species separately assumes each species is independent (an unlikely assumption) and would increase the number of analyses resulting in inflated chances of detecting spurious effects (i.e. type I error). From the perspective of overall study conclusions, selecting one marker type would limit inferences to only that type of marker (i.e. SNP markers). If there was clear justification that, say, SNP markers were superior to microsatellite loci (i.e. had more power), or if both markers were strongly correlated in any case, this may not be a problem. Nonetheless, using only SNPs would not be an ideal approach if we were interested in the differences between microsatellite and SNP markers. For example, microsatellite markers may detect a weaker overall effect compared with SNP markers, informing what markers might be most useful in future studies.

Given our question above, we are not so much interested in the differences between the markers, but in obtaining an overall estimate of genetic differentiation between fragmented and control sites. It would therefore be best to include all markers and simply get a single overall effect size for each species, for both markers, by creating a composite (i.e. average) effect size that combines estimates from the SNP and microsatellite markers from the same study (Marín-Martínez & Sánchez-Meca 1999; Borenstein *et al.* 2009; Mengersen *et al.* 2013; Cheung 2014). Creating a composite effect size will ensure that effects remain independent in the analysis, and provide a more thorough coverage of markers. This will allow us to make inferences

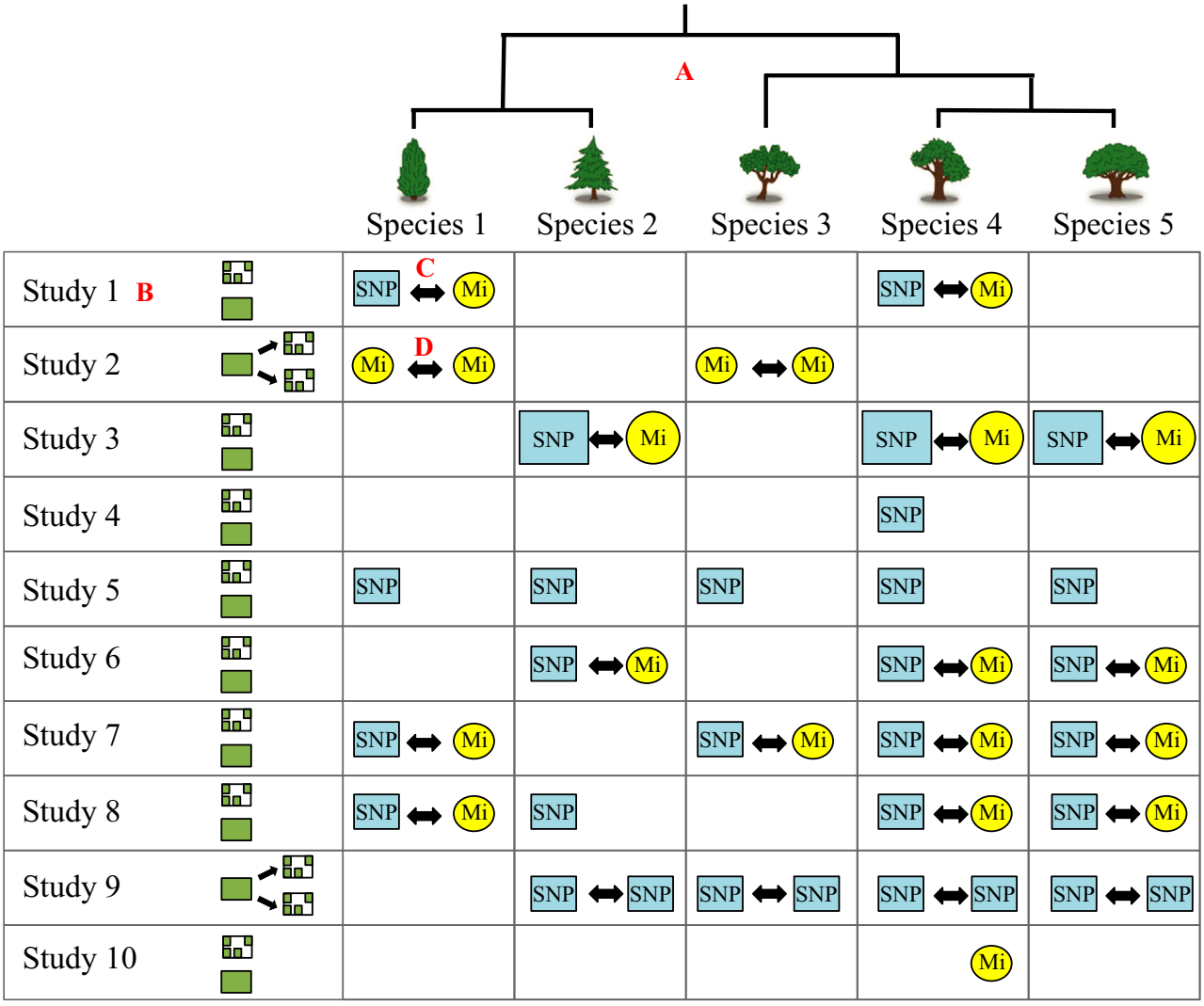

Fig. 5 A hypothetical example of how nonindependence can arise in ecological and evolutionary studies. Assume we are interested in the following question: Does fragmentation lead to a reduction in observed heterozygosity (H_O) in trees? To answer this question, effect sizes were collected from 10 studies that manipulated forest plots and compared fragmented sites to control sites—untouched. Rectangles beside each study name represent each plot (control = filled green; treatment = patchy green) and the number of plots used. Observed heterozygosity (H_O) was estimated using either microsatellite DNA loci, SNP loci or both. The average H_O of estimates from each plot, for each species, was used to calculate an effect size comparing pairs of plots. A total of 50 effect sizes were computed from 10 studies for a total of five different species, represented as tree images on top of the table. The yellow circles represent effect sizes calculated using microsatellite DNA loci (Mi) and the blue rectangles represent effect sizes calculated from SNP loci (SNP). The size of the circle or rectangle represents the size of the effect. Double-headed arrows between some effect sizes indicate that these effect sizes are correlated to varying degrees. Red-bolded letters indicate the types of nonindependence in this example and their descriptions are in Fig. 4 (and see 'Problem 4'). [Colour figure can be viewed at wileyonlinelibrary.com]

about genetic markers more generally (assuming these were the only two markers ever used). However, combining markers in this way will result in, for example, a loss of important information (i.e. we cannot compare markers), lower statistical power and often an unreasonable assumption about no heterogeneity among combined effect sizes (cf. Senior *et al.* 2016). Thus, using a composite effect size could introduce a new set of problems (see Marín-Martínez & Sánchez-Meca 1999; Gleser & Olkin 2009). These limitations should be made clear when reporting such analyses.

Solution 4B: Modelling sources of nonindependence. All effect size estimates might also be retained and the dependence among these explicitly modelled using *multilevel meta-analytic* and meta-regression models (i.e. mixed-effects models) that account for the nonindependence of effect sizes, along with the sampling error variance of individual effects, in a single analysis (Gleser & Olkin 2009; Lajeunesse 2011; Nakagawa & Santos 2012; Mengersen *et al.* 2013). Multilevel meta-regression models have developed rapidly and can now deal with many sources of

nonindependence, while allowing one to explore moderator variables to test a rich set of methodological and biological hypotheses explaining variation among effect sizes (Gleser & Olkin 2009; Lajeunesse 2011; Nakagawa & Santos 2012; Mengersen *et al.* 2013; Cheung 2014). Many of these models can be run in widely available software packages including METAFOR (Viechtbauer 2010), MCMCGLMM (Hadfield 2010) and OPENMEE (Wallace *et al.* 2016). Given we are only interested in the overall pooled effect size estimate (i.e. average effect for all tree species and studies), we might estimate model parameters using the following multilevel meta-analytic model, which accounts for the aforementioned sources of nonindependence, as well as sampling error in each effect size estimate (notation follows, Nakagawa & Santos 2012):

$$z_i = u + a_{k[i]} + \mathrm{sp}_{k[i]} + s_{j[i]} + e_i + m_i,$$

where:

1. z_i is the ith effect size ($i = 1 \ldots N_{ES}$),
2. u the overall mean effect size estimate,
3. $a_{k[i]}$ is the phylogenetic effect for species k applied to effect size i. Phylogenetic effects are assumed to be normally distributed deviates with a mean of 0 and variance $\sigma^2_a\mathbf{A}$ [hereafter denoted ~ $N(0, \sigma^2_a\mathbf{A})$], where $\mathbf{A}$ is a phylogenetic correlation matrix derived from a phylogenetic tree,
4. $\mathrm{sp}_{k[i]}$ is the species-specific effect for the kth species applied to effect size i [assumed to be ~ $N(0, \sigma^2_{sp}\mathbf{I})$, where $\mathbf{I}$ is an identity matrix],
5. $s_{j[i]}$ is the study-specific effect for the jth study applied to effect size i [assumed to be ~ $N(0, \sigma^2_s\mathbf{I})$],
6. e_i is the effect size-specific effect (or residuals) for the ith effect size [assumed to be ~$N(0, \sigma^2_e\mathbf{E})$)] where $\mathbf{E}$ is a correlation matrix derived from the correlation between two measurements on the same samples (e.g. SNPs and microsatellites; see Figs 4C and 5),
7. m_i is the sampling variance around the ith effect as a result of variation in precision among each effect size [assumed to be ~ $N(0, \mathbf{M})$], where $\mathbf{M}$ is a variance–covariance matrix with the diagonal elements the sampling variance and off-diagonal elements the covariance resulting from sharing a control treatment (see Figs 4D and 5; Gleser & Olkin 2009; Lajeunesse 2011).

The meta-analytic model above provides the most powerful and flexible statistical approach to deal with nonindependence in our hypothetical example data. An important, yet often overlooked, statistical aspect of these models is that, if we expect some form of covariance within these random effect groups (i.e. $a_{k[i]}$, $\mathrm{sp}_{k[i]}$, $s_{j[i]}$, e_i and m_i), we can include different correlation matrices by replacing the identity matrix ($\mathbf{I}$—a matrix with all diagonal values as 1 and off-diagonal values as 0) with a matrix that describes the dependence between data, as we do with the phylogenetic and effect size matrices ($\mathbf{A}$ and $\mathbf{E}$). Here, the off-diagonal values of $\mathbf{A}$ and $\mathbf{E}$ contain the correlation between effect size estimates to deal with phylogenetic relationships between species, and the shared covariance between SNPs and microsatellite markers. While the above model is powerful, fitting such a complex model is practically challenging. First, the model may be overparameterized as we only have a total of $n = 50$ effect sizes. Second, including phylogenetic and species-level random effects with only five species is likely to lead to imprecise and unreliable estimates (cf. Bolker *et al.* 2009; Nakagawa & Santos 2012). Additionally, in many cases a phylogenetic tree may not be available or there may be unresolved relationships between taxa. Treating taxa as a series of nested random effects is also unlikely to resolve the above problems and would still require a large data set.

One possible solution might be to run meta-analyses on subgroups (e.g. grouping by different species and different genetic markers), but it is important to recognize that such an approach will lead to a reduction in statistical power and assumes independence between subgroups (as discussed in 'Solution 4A'). Alternatively, we can use a meta-regression model with relevant moderators (e.g. species or even marker type, if we are interested in differences between SNPs and microsatellite markers). Moderator variables may be part of the authors' original question(s), or be covariates that need to be included to control for sources of variation and nonindependence in the data (Thompson & Higgins 2002). For example, while we may not be able to include the phylogenetic effect (Fig. 4—'A') (due to the limited number of species present, or because a robust phylogeny is not available), we could include 'species' as a moderator variable in a meta-regression model to estimate species-specific effect size estimates. The above approach, however, may not necessarily resolve the issues of nonindependence (due to phylogeny) and still requires estimation of four parameters (one for each of the four other species), which may be too many for our small data set. Therefore, it might be more realistic, given our sample size, to estimate only differences for groups of similar species (e.g. coniferous vs. deciduous trees), and we would need to provide justification for our decision on how to analyse these data that incorporate a discussion of the above limitations. The important point here is that there are a number of possible solutions, each with their own strengths and limitations. We should therefore decide on which solutions are most sensible given our data, justify these choices and conduct sensitivity analyses to assess the impact of our choices on our conclusions.

Conclusions and recommendations

Nonindependence is a common problem when conducting meta-analyses in ecology and evolution. Not recognizing and dealing with this issue can have important consequences on study inferences that could affect the development, and direction, of a research field. Recent reviews, however, make it clear that the impact of nonindependence, and its reporting, needs to be taken more seriously (Chamberlain *et al.* 2012; Koricheva & Gurevitch 2014; ArchMiller *et al.* 2015). In our paper, we have described how nonindependence can affect the calculation of effect size statistics, and result in correlations among effect size estimates. We emphasize the important role of sensitivity analyses, and

provide solutions that can be used to better explore the consequences of violating assumptions of independence, making conclusions more robust. While there is no straightforward solution to nonindependence, in all cases, making use of multiple analytical strategies should give greater confidence in meta-analytic results. What is possible in a meta-analysis will always be limited by sample size, data structure, available software and the expertise of authors Nakagawa *et al*. (2017). Therefore, transparency and high-quality reporting are needed to effectively convey to readers the assumptions, limitations and issues inherent to the data at hand. These are points stressed by recent reviews (Gilbert *et al*. 2012; Nekrutenko & Taylor 2012; White *et al*. 2015). Therefore, we recommend that authors report the following information to facilitate an understanding of nonindependence in their meta-analysis:

1 All sources of nonindependence among effect sizes (see Fig. 4) and how study design might influence the chosen effect size, if at all.
2 The number of studies or effect sizes influenced by nonindependence (i.e. types of nonindependent experimental designs and the nested structure of effect sizes).
3 The steps, either during extraction, effect size calculation or analysis, which were taken to ameliorate nonindependence problems.
4 Justification for decisions in the previous point, articulating their assumptions, limitations and potential impact on study conclusions.
5 Presentation and discussion of results from sensitivity analyses that explore the effects of analytical decisions and statistical assumptions on inferences drawn (e.g. Greenhouse & Iyengar 2009).

We acknowledge that it can sometimes be challenging to diagnose, and predict, the impact that nonindependence might have on a meta-analysis. However, keeping the above points in mind, and carefully considering the extracted data, will often expose potential issues. We hope that an understanding of these various dimensions will lead to greater use of sensitivity analyses, more transparent reporting of nonindependence and more robust conclusions not only from meta-analyses, but also from primary studies (see Forstmeier *et al*. 2016), in ecology and evolution.

Acknowledgements

We thank Alistair Senior, Nolan Kane and six anonymous reviewers for comments that greatly improved this manuscript. D.W.A.N. is funded by an Australian Research Council Discovery Early Career Research Award (DE150101774) and S.N. is funded by an Australian Research Council Future Fellowship (FT130100268).

References

Adams D (2008) Phylogenetic meta-analysis. *Evolution*, **62–3**, 567–572.
Aguilar R, Quesada M, Ashworth L, Herrerias-Diego Y, Lobo J (2008) Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. *Molecular Ecology*, **17**, 5177–5188.
Aloe A (2015) Inaccuracy of regression results in replacing bivariate correlations. *Research Synthesis Methods*, **6**, 21–27.
ArchMiller A, Bauer E, Koch R *et al*. (2015) Formalizing the definition of meta-analysis in molecular ecology. *Molecular Ecology*, **24**, 4042–4051.
Besson A, Lagisz M, Senior A, Hector K, Nakagawa S (2016) Effect of maternal diet on offspring coping styles in rodents: a systematic review and meta-analysis. *Biological Reviews*, **91**, 1065–1080.
Bolker B, Brooks M, Clark C *et al*. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.
Booksmythe I, Mautz B, Davis J, Nakagawa S, Jennions MD (2015) Facultative adjustment of the offspring sex ratio and male attractiveness: a systematic review and meta-analysis. *Biological Reviews*, **92**, 108–134.
Borenstein M (2009) Effect size for continuous data. In: *The Handbook of Research Synthesis and Meta-Analysis* (eds Cooper H, Hedges LV, Valentine JC), pp. 221–235. Russell Sage Foundation, New York.
Borenstein M, Hedges L, Higgens J, Rothstein H (2009) *Introduction to Meta-Analysis*. John Wiley & Sons Ltd, West Sussex, UK.
Chamberlain S, Hovick S, Dibble C *et al*. (2012) Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. *Ecology Letters*, **15**, 627–636.
Chapman JR, Nakagawa S, Coltman DW, Slate J, Sheldon BC (2009) A quantitative review of heterozygosity-fitness correlations in animal populations. *Molecular Ecology*, **18**, 2746–2765.
Cheung MWL (2014) Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods*, **19**, 211–229.
Cooper H, Hedges LV, Valentine JC (2009) *The Handbook of Research Synthesis and Meta-analysis*. Russell Sage Foundation, New York.
Dingemanse N, Dochtermann N, Nakagawa S (2012) Defining behavioural syndromes and the role of "syndrome deviation" in understanding their evolution. *Behavioral Ecology and Sociobiology*, **66**, 1543–1548.
Dunlap W, Cortina JM, Vaslow J, Burke M (1996) Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, **1**, 170–177.
Forstmeier W, Wagenmakers E-J, Parker TH (2016) Detecting and avoiding likely false-positive findings—a practical guide. *Biological Reviews*. [Epub ahead of print]. doi:10.11111/brv.12315
Gates S (2002) Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology*, **71**, 547–557.
Gilbert K, Andrew R, Bock D, Franklin M (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Molecular Ecology*, **21**, 4925–4930.
Gleser LJ, Olkin I (2009) Stochastically dependent effect sizes. In: *The Handbook of Research Synthesis and Meta-Analysis* (eds Cooper H, Hedges LV, Valentine JC), pp. 357–376. Russell Sage Foundation, New York.
Greenhouse J, Iyengar S (2009) Sensitivity analysis and diagnostics. In: *The Handbook of Research Synthesis and Meta-Analysis* (eds Cooper H, Hedges LV, Valentine JC), pp. 417–433. Russell Sage Foundation, New York.
Gurevitch J, Hedges L (1999) Statistical issues in ecological meta-analyses. *Ecology*, **80**, 1142–1149.
Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the *MCMCglmm* R package. *Journal of Statistical Software*, **33**, 1–22.

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies, and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**, 494–508.

Hedges L (2007) Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, **32**, 341–370.

Hedges L (2009a) Statistical consideration. In: *The Handbook of Research Synthesis and Meta-Analysis* (eds Cooper H, Hedges LV, Valentine JC), pp. 38–46. Russell Sage Foundation, New York.

Hedges L (2009b) Effect sizes in nested designs. In: *The Handbook of Research Synthesis and Meta-Analysis* (eds Cooper H, Hedges LV, Valentine JC), pp. 337–355. Russell Sage Foundation, New York.

Hedges LV, Gurevitch J, Curtis PS (1999) The meta-analysis of response ratios in experimental ecology. *Ecology*, **80**, 1150–1156.

Heller R, Siegismund H (2009) Relationship between three measures of genetic differentiation, G_{ST}, D_{EST} and G'_{ST}: how wrong have we been? *Molecular Ecology*, **18**, 2080–2083.

Higgins J, Green S (2009) *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley-Blackwell, Chichester, UK.

Hurlburt S (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

Isaksson C (2010) Pollution and its impact on wild animals: a meta-analysis on oxidative stress. *EcoHealth*, **7**, 342–350.

Koricheva J, Gurevitch J (2014) Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, **102**, 828–844.

Koricheva J, Gurevitch J, Mengersen K (2013) *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton University Press, Princeton, New Jersey.

Lajeunesse M (2009) Meta-analysis and the comparative phylogenetic method. *The American Naturalist*, **174**, 369–381.

Lajeunesse M (2011) On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*, **92**, 2049–2055.

Liberati A, Altman D, Tetzlaff J *et al.* (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine*, **6**, e1000100.

Marín-Martínez F, Sánchez-Meca J (1999) Averaging dependent effect sizes in meta-analysis: a cautionary note about procedures. *The Spanish Journal of Psychology*, **2**, 32–38.

Mengersen K, Jennions M, Schmid C (2013) Statistical models for the meta-analysis of non-independent data. In: *Handbook of Meta-Analysis in Ecology and Evolution* (eds Koricheva J, Gurevitch J, Mengersen K), pp. 255–283. Princeton University Press, Princeton, New Jersey and Oxford.

Morrissey MB (2016) Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology*, **29**, 1882–1904.

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, **82**, 591–605.

Nakagawa S, Poulin R (2012) Meta-analytic insights into evolutionary ecology: an introduction and synthesis. *Evolutionary Ecology*, **26**, 1085–1099.

Nakagawa S, Santos E (2012) Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, **26**, 1253–1274.

Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, **85**, 935–956.

Nakagawa S, Noble DWA, Senior AM, Lagisz M (2017) Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biology*, doi: 10.1186/s12915-017-0357-7.

Nekrutenko A, Taylor J (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics*, **13**, 667–672.

Osenberg C, Sarnelle O, Cooper S (1997) Effect size in ecological experiments: the application of biological models in meta-analysis. *The American Naturalist*, **150**, 798–812.

Pinheiro J, Bates D (2000) *Mixed-effects models in S and S-PLUS*. Springer Verlag, New York.

van de Pol M, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. *Animal Behaviour*, **7**, 753–758.

Poulin R (2000) Manipulation of host behaviour by parasites: a weakening paradigm. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 787–792.

Riley R, Price M, Jackson D *et al.* (2014) Multivariate meta-analysis using individual participant data. *Research Synthesis Methods*, **6**, 157–174.

Roche D, Lanfear R, Binning S *et al.* (2014) Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biology*, **12**, e1001779.

Rothstein H, Sutton A, Borenstein M (2005) *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Wiley, Chichester.

Rutkowska J, Dubiec A, Nakagawa S (2014) All eggs are made equal: a meta-analysis of egg sexual size dimorphism in birds. *Journal of Evolutionary Biology*, **27**, 153–160.

Senior A, Grueber CE, Kamiya T *et al.* (2016) Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology*, **97**, 3293–3299.

Slate J, Phua H (2003) Patterns of linkage disequilibrium in mitochondrial DNA of 16 ruminant populations. *Molecular Ecology*, **12**, 597–608.

Sutton A (2009) Publication bias. In: *The Handbook of Research Synthesis and Meta-Analysis* (eds Cooper H, Hedges LV, Valentine JC), pp. 435–452. Russell Sage Foundation, New York.

Thompson S, Higgins J (2002) How should meta-regression analysis be undertaken and interpreted? *Statistics in Medicine*, **21**, 1559–1573.

Uller T, Nakagawa S, English S (2013) Weak evidence for anticipatory parental effects in plants and animals. *Journal of Evolutionary Biology*, **26**, 2161–2170.

Verdu M, Traveset A (2005) Early emergence enhances plant fitness: a phylogenetically controlled meta-analysis. *Ecology*, **86**, 1385–1394.

Vetter D, Rcker G, Storch I (2013) Meta-analysis: a need for well-defined usage in ecology and conservation biology. *Ecosphere*, **6**, 1–24.

Viechtbauer W (2010) Conducting meta-analysis in *R* with the *metafor* package. *Journal of Statistical Software*, **36**, 1–48.

Wallace BC, Lajeunesse MJ, Dietz G *et al.* (2016) openMEE: intuitive, open-source software for meta-analysis in ecology and evolutionary biology. *Methods in Ecology and Evolution*, doi: 10.1111/2041-210X.12708.

White T, Dalrymple R, Noble DWA *et al.* (2015) Reproducible research in the study of biological coloration. *Animal Behaviour*, **106**, 51–57.

Worm B, Myers RA (2003) Meta-analysis of cod-shrimp interactions reveals top-down control in oceanic food webs. *Ecology*, **84**, 162–173.

All authors conceived and outlined the paper. DWAN and ML wrote the manuscript and SN and REO commented and helped revise the manuscript.

---

doi: 10.1111/mec.14031