doi:10.1111/evo.14

Quantifying the causal pathways contributing to natural selection

Jonathan M. Henshaw,[1,2,3] Michael B. Morrissey,[4] and Adam G. Jones[2]

[1]*Institute of Biology I, University of Freiburg, Freiburg im Breisgau 79104, Germany*

[2]*Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844*

[3]*E-mail: jonathan.henshaw@biologie.uni-freiburg.de*

[4]*School of Biology, University of St Andrews, St Andrews KY16 9TF, United Kingdom*

Received March 23, 2020

Accepted August 7, 2020

The consequences of natural selection can be understood from a purely statistical perspective. In contrast, an explicitly causal approach is required to understand *why* trait values covary with fitness. In particular, key evolutionary constructs, such as sexual selection, fecundity selection, and so on, are best understood as selection *via* particular fitness components. To formalize and operationalize these concepts, we must disentangle the various causal pathways contributing to selection. Such decompositions are currently only known for linear models, where they are sometimes referred to as "Wright's rules." Here, we provide a general framework, based on path analysis, for partitioning selection among its contributing causal pathways. We show how the extended selection gradient—which represents selection arising from a trait's causal effects on fitness—can be decomposed into path-specific selection gradients, which correspond to distinct causal mechanisms of selection. This framework allows for nonlinear effects and nonadditive interactions among variables, which may be estimated using standard statistical methods (e.g., generalized linear [mixed] models or generalized additive models). We thus provide a generalization of Wright's path rules that accommodates the nonlinear and nonadditive mechanisms by which natural selection commonly arises.

KEY WORDS: **Causal derivative, causality, path analysis, structural equation modeling (SEM).**

Causal thinking is essential to understanding natural selection and its evolutionary consequences (Sober 1984; Wade and Kalisz 1990; Arnold and Duvall 1994; Godfrey-Smith 2007; Morrissey 2014b; Okasha 2016; Shipley 2016; Walsh and Lynch 2018; Uller and Laland 2019; Okasha and Otsuka 2020; Queller 2020). At the coarsest explanatory level, selection and its consequences can be treated as purely statistical patterns. However, explaining *why* selection operates the way it does requires causal analysis of the interrelationships among traits, fitness components, and fitness. For instance, covariation among traits due to shared genetic or environmental underpinnings can lead a trait distribution to change under selection, even when the trait has no direct influence on fitness (Lande and Arnold 1983; Godfrey-Smith 2007). Sober (1984) accordingly distinguished between "selection for" and "selection of" a trait. Causality is also inherent to a series of key evolutionary concepts—sexual selection, fecundity selection, longevity selection, and so on—that can be understood as selection *via* particular fitness components (Andersson 1994; Arnold and Duvall 1994; Henshaw et al. 2016, 2018). Our ability to explain and generalize from observations of trait-fitness relationships depends crucially on pinning down such causal mechanisms.

Classical approaches to quantifying the causal structure of selection are based on regression (Lande and Arnold 1983; Arnold and Wade 1984) and path analysis (Wright 1921, 1934; Arnold and Duvall 1994; Scheiner et al. 2000) (reviewed in Walsh and Lynch 2018). Regression-based approaches, as pioneered by Lande and Arnold (1983), can quantify the direct effects of traits on fitness in the absence of unmeasured confounders (Rausher 1992; Walsh and Lynch 2018). However, such approaches are constrained to view selection as a single-step causal process (Henshaw et al. 2018; Grace and Irvine 2020), without causal intermediaries between traits and fitness, and so give limited insight as to *why* traits covary with fitness.

© 2020 The Authors. *Evolution* published by Wiley Periodicals LLC on behalf of The Society for the Study of Evolution
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Path analysis allows for more sophisticated, multi-step models of causal processes. Such models can explicitly account for indirect causal effects, such as when a trait's fitness effects are mediated by other variables. Early path-analytic work generally assumes that the relationships between effects and their causes are linear (Wright 1934; Lande and Arnold 1983; Crespi and Bookstein 1989; Kingsolver and Schemske 1991; Arnold and Duvall 1994; Conner et al. 1996; Frank 1997; Latta and McCain 2009; but see Scheiner et al. 2000). This allows total selection to be partitioned across all contributing causal pathways using "Wright's rules" (Wright 1934; Loehlin 2004; Henshaw et al. 2018). Linearity is unlikely to hold in most applications, however, and is a particularly problematic assumption for fitness components, which are often measured as binary or count variables (e.g., survival, fecundity, or the number of matings). Morrissey (2015) addresses this issue by extending the path-analytic approach to more general settings, accommodating nonlinear effects and nonadditive interactions among variables (see also Grace et al. 2012; Lefcheck 2016). Morrissey's "extended selection gradient" allows the total causal effect of a trait on fitness to be quantified for arbitrary smooth relationships among traits and fitness (Morrissey 2014b, 2015). Currently, however, we lack a general nonlinear framework for decomposing total selection into contributions from multiple pathways (Fig. 1). Such pathways may comprise both causal effects of the focal trait on fitness (potentially mediated by other traits) and spurious associations arising from common causes.

Parallel to these developments, there has been a revolution in the mathematical analysis of causal processes (e.g., Pearl 2009, 2018), the fruits of which have gone largely unnoticed and unused in evolutionary biology (but see, e.g., Okasha 2016; Otsuka 2016; Shipley 2016; Fromhage and Jennions 2019; Okasha and Otsuka 2020; Queller 2020). Here, we harness these recent mathematical advances to provide a general framework for quantifying the causal tributaries of selection, combining the intuition behind Wright's rules with the generality of Morrissey's (2015) nonlinear path analysis. We also provide analogous methods for decomposing the predicted evolutionary response to selection between generations. Importantly, our approach can be applied to observational data if appropriate causal assumptions are met. This feature is desirable, because the multi-level causal processes that determine fitness are often difficult or impossible to manipulate experimentally.

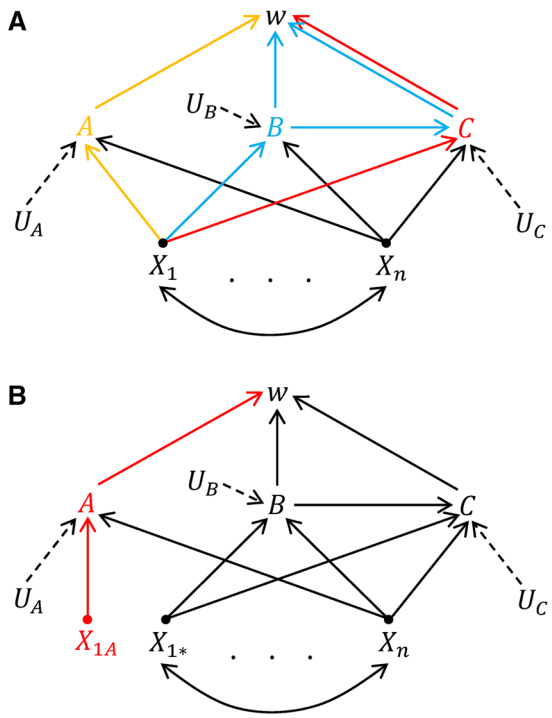

Figure 1. **(A) An example causal diagram, where a set of exogenous traits $X_1, \ldots, X_n$ influences relative fitness w via some endogenous traits A, B, and C. Each endogenous trait Y is also influenced by an unmeasured background variable, denoted U_Y. Single-headed arrows represent potential causal effects, whereas double-headed arrows indicate potential shared influences that are not analyzed explicitly. If an arrow from X to Y is missing, it is assumed that X has no direct causal effect on Y. The subgraphs for the path-specific selection gradients on X_1 via each of its children are shown in color (yellow for A; blue for B; red for C). Note that (i) the path-specific selection gradient $\eta_{X_1|C}$ (shown in red) does not include the path $X_1 \to B \to C \to w$, because by definition it only includes paths with first edge $X_1 \to C$, (ii) the edge $C \to w$ counts toward the path-specific selection gradients via both B and C, indicated here by parallel blue and red arrows, and (iii) the paths $X_1 \to B \to w$ and $X_1 \to B \to C \to w$ do not have identifiable selection gradients, although their union (shown in blue) does. (B) A twin-network model for conceptualizing the path-specific selection gradient on X_1 via A. Effectively, X_1 is split into two variables, X_{1A} and X_{1*}, where A is influenced by X_{1A}, and all other children of X_1 are influenced by X_{1*}. The path-specific selection gradient on X_1 via A is equal to the extended selection gradient on X_{1A} after setting $X_{1A} = X_{1*} = X_1$ (see Supporting Information for formal proof).**

Notation

We use regular typeface for scalar random variables and boldface for random vectors and matrices. We write $\mathbb{E}(X)$ for the expected value of a random variable X and $\mathbb{E}(X|Y_1, \ldots, Y_n)$ for the expected value of X conditional on the values of $Y_1, \ldots, Y_n$. Similarly, $\mathbb{P}(X) = \mathbb{P}(X = x)$ is the probability that X attains a particular value x and $\mathbb{P}(X|Y_1, \ldots, Y_n)$ is the conditional probability of X given $Y_1, \ldots, Y_n$. The probability distribution of a continuous random variable X is written $d\mathbb{P}(X)$. The variance of X is σ_X^2. Last, we write ∇ for the gradient of a function with respect to its

Table 1. Summary of key variables and notation.

Notation	Meaning
$\mathbb{E}(X)$	Expected value of X
$\mathbb{E}(X\|Y)$	Expected value of X conditional on Y
$\mathbb{P}(X)$	Probability distribution of X
σ_X^2	Variance of X
∇f	Gradient of a function with respect to its arguments: $\nabla f\,(x_1, \ldots, x_n) = [\frac{\partial f}{\partial_{x_1}}, \ldots, \frac{\partial f}{\partial_{x_n}}]^T$
$\boldsymbol{Z}$	Vector of all measured traits, including fitness
$\boldsymbol{X}$	Vector of exogenous traits (of which causes are not modeled explicitly)
$\boldsymbol{Y}$	Vector of endogenous traits (of which causes are modeled explicitly)
$\boldsymbol{U}$	Vector of unmeasured background variables
$\boldsymbol{S}$	Selection differential (vector of covariances between traits and fitness)
$\boldsymbol{P}$	Variance-covariance matrix of phenotypic traits
$\boldsymbol{\beta}_{\mathrm{LR}}$	Linear regression selection gradients: $\boldsymbol{\beta}_{\mathrm{LR}} = \boldsymbol{P}^{-1}\,\boldsymbol{S}$
$\boldsymbol{\beta}_{\mathrm{AP}}$	Average partial selection gradients (average rate that fitness is predicted to change given small changes in each trait, while holding all other measured traits fixed)
$\boldsymbol{G}$	Variance-covariance matrix of genetic values
$\mathrm{ch}(Z)$	Children of a trait Z (i.e., set of measured traits that Z influences directly)
$\mathrm{pa}(Z)$	Parents of a trait Z (i.e., set of measured traits that directly influence Z)
f_Z	Function relating a trait Z to its parents $\mathrm{pa}(Z)$ and background variable U_Z
$\frac{\delta w}{\delta Z}$	Causal derivative of fitness w on a trait Z (rate at which fitness would change due to small changes in Z, holding fixed all nondescendants of Z)
η_Z	Extended selection gradient on Z (average of $\frac{\delta w}{\delta Z}$ across joint distribution of all traits)
$\frac{\delta w}{\delta Z}\|_{\mathcal{H}}$	Path-specific causal derivative of fitness w on a trait Z via paths $\mathcal{H}$ (rate at which fitness would change due to small changes in Z, holding fixed all pathways other than $\mathcal{H}$)
$\eta_Z\|_{\mathcal{H}}$	Path-specific selection gradient on Z via paths $\mathcal{H}$ (average of $\frac{\delta w}{\delta Z}\|_{\mathcal{H}}$ across joint distribution of all traits)

arguments: i.e., $\nabla f\,(x_1, \ldots, x_n) = [\frac{\partial f}{\partial_{x_1}}, \ldots, \frac{\partial f}{\partial_{x_n}}]^T$. Key variables and notation are summarized in Table 1.

Background: The Lande-Arnold Framework

We begin by summarizing Lande and Arnold's (1983) approach to quantifying selection. Let W be absolute fitness (as estimated, for instance, by the number of offspring produced) and let $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^T$ be a vector of traits with covariance matrix $\boldsymbol{P}$. Environmental variables may also be included in $\boldsymbol{Z}$ without fundamentally changing the analysis. The directional *selection differential* $\boldsymbol{S}$ is the difference in mean trait values before and after selection (i.e., within a generation). For example, if fitness is estimated by the number of offspring produced, then the selection differential is the difference in mean trait values between all individuals and the parents of offspring, with the latter weighted by the number of offspring per parent. Writing $w = \frac{W}{\mathbb{E}(W)}$ for relative fitness, the selection differential is given by (Robertson 1966; Henshaw and Zemel 2017; Walsh and Lynch 2018):

$$\boldsymbol{S} = \mathbb{E}\,(w\boldsymbol{Z}) - \mathbb{E}\,(\boldsymbol{Z}) = \mathrm{cov}\,(\boldsymbol{Z}, w)\,. \tag{1}$$

Let us now write $w = \mathbb{E}(w|\boldsymbol{Z}) + \varepsilon_w$, where ε_w represents variation in relative fitness that is not explained by the traits $\boldsymbol{Z}$. By construction, we have $\mathrm{cov}\,(\boldsymbol{Z}, \varepsilon_w) = 0$ and hence:

$$\boldsymbol{S} = \mathrm{cov}\,(\boldsymbol{Z}, \mathbb{E}\,(w|\boldsymbol{Z}))\,. \tag{2}$$

Further, if the trait vector $\boldsymbol{Z}$ is multivariate normal, then by Stein's lemma (Stein 1981; Lande and Arnold 1983; Liu 1994; de Villemereuil et al. 2016; Walsh and Morrissey 2019), we have

$$\boldsymbol{S} = \boldsymbol{P}\,\mathbb{E}\,(\nabla\mathbb{E}\,(w\boldsymbol{Z}))\,. \tag{3a}$$

In particular, the selection differential on the ith trait Z_i can be written as

$$S_i = \sum_{j=1}^{n} \mathrm{cov}\left(Z_i, Z_j\right) \mathbb{E}\left(\frac{\partial}{\partial Z_j}\mathbb{E}\,(w|\boldsymbol{Z})\right) \tag{3b}$$

As a consequence of these useful properties, transformations that bring $\boldsymbol{Z}$ close to multivariate normality are often desirable. Alternatively, measured traits may be modeled as nonlinear functions of underlying "latent" traits that are normally

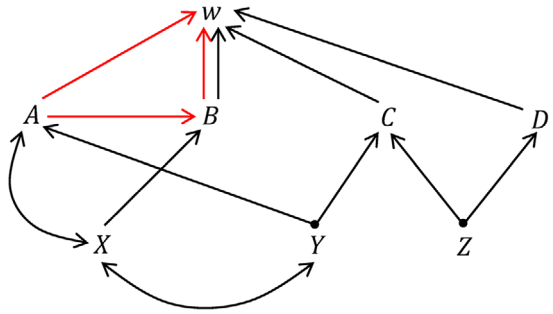

Figure 2. An example causal diagram, showing the potential causal paths (red) and noncausal or "backdoor" paths (black) between a trait *A* and fitness *w*. Background variables are omitted for simplicity, but traits with potential shared influences are joined by double-headed arrows. The causal effect of *A* on fitness can be estimated from observational data by controlling for an appropriate set of covariates. For the pictured graph, such backdoor sets include {*X*, *Y*}, {*X*, *Y*, *C*}, and {*X*, *Y*, *Z*}. Note the following: (i) *B* cannot be included in any backdoor set for *A* because it is a descendant of *A*, (ii) the set {*Y*} is not a backdoor set for *A* because it leaves the backdoor path $A \leftrightarrow X \rightarrow B \rightarrow w$ unblocked, and (iii) the set {*X*, *C*} is not a backdoor set for *A* because the inclusion of *C* unblocks the backdoor path $A \leftarrow Y \rightarrow C \leftarrow Z \rightarrow D \rightarrow w$. This is because, although *A* and *Z* are independent, they may be dependent after conditioning on *C* (see Supporting Information or Pearl 2009 for details).

distributed (for details, see de Villemereuil et al. 2016; de Villemereuil 2018).

Following Lande and Arnold (1983), the vector of directional *selection gradients* $\boldsymbol{\beta}$ on the traits $\boldsymbol{Z}$ is usually defined as the solution to $\boldsymbol{S} = \boldsymbol{P}\boldsymbol{\beta}$. Equivalently, $\boldsymbol{\beta}$ is the vector of partial regression coefficients in an ordinary least-squares linear regression of relative fitness on trait values. In the case where $\boldsymbol{Z}$ is multivariate normal and $\boldsymbol{P}$ is invertible, it follows from equation (3a) that $\boldsymbol{\beta} = \mathbb{E}(\nabla\mathbb{E}(w|\boldsymbol{Z}))$. In other words, the selection gradients equal the average rate that fitness is predicted to change given small changes in each trait, while holding all other measured traits fixed. To avoid ambiguity, we refer to $\boldsymbol{\beta}_{\mathrm{LR}} = \boldsymbol{P}^{-1}\boldsymbol{S}$ as the *linear regression* selection gradients and $\boldsymbol{\beta}_{\mathrm{AP}} = \mathbb{E}(\nabla\mathbb{E}(w|\boldsymbol{Z}))$ as the *average partial* selection gradients (note that our terminology differs from Walsh and Lynch 2018, who call $\boldsymbol{\beta}_{\mathrm{AP}}$ the *Janzen-Stern* gradients).

Causal Model

Following Pearl (2009), we consider causal models consisting of three components: (1) a *causal diagram*, which describes the hypothesized causal relationships among variables in a qualitative way, (2) a model of the *functional relationships* between variables and their causes, and (3) a model of the *probability distributions* of unmeasured "background variables," which generate unexplained variation in measured variables. Most of the mathematical results we rely on below were developed by Judea Pearl and colleagues (Avin et al. 2005; Pearl 2009), but our application of these results to evolutionary biology is largely new.

The causal diagram specifies qualitative assumptions about causal effects in the form of a graph, in which multiple measured traits $\boldsymbol{Z}$ are represented by nodes and causal links between such traits are represented by arrows (e.g., Fig. 1). For the purposes of studying selection, the set $\boldsymbol{Z}$ must always include fitness w and may also include components of fitness like survival, fecundity, or mating success. Single-headed arrows of the form $A \rightarrow B$ represent a direct causal effect of A on B, whereas double-headed arrows $A \leftrightarrow B$ indicate that these traits potentially have shared causal influences that are not accounted for in the causal diagram. We write $\boldsymbol{X}$ for the set of *exogenous* traits, whose causes are not modeled explicitly, and $\boldsymbol{Y}$ for *endogenous traits* with at least one modeled cause. As in the Lande-Arnold framework, environmental variables can be included as "traits" without fundamentally changing the analysis.

Each trait $Z \in \boldsymbol{Z}$ is fully determined by (i) its *parents* $\mathrm{pa}(Z)$, which is a subset of $\boldsymbol{Z}$, and (ii) an unmeasured background variable U_Z, which captures all other causal determinants of Z. We write the functional relationship between a trait and its causal determinants as

$$Z = f_Z\left(\mathrm{pa}\left(Z\right), U_Z\right). \tag{4}$$

We say that a trait Z is a *child* of its parents and we write $\mathrm{ch}(Z)$ for the set of children of Z. An exogenous trait $X \in \boldsymbol{X}$ has no explicitly modeled causes (i.e., $\mathrm{pa}\left(X\right) = \emptyset$) and so for our purposes can be treated as identical to its background variable (i.e., $X = f_X(U_X) = U_X$). The causal diagram $\mathcal{G}$ is formed by drawing a single-headed arrow from each trait to each of its children (e.g., Fig. 1A). In addition, a double-headed arrow is drawn between traits A and B if their background variables U_A and U_B may covary due to shared influences. A *directed path* from A to B is a sequence of traits of the form $(A = P_1, P_2, \ldots, P_n = B)$ such that there is a single-headed arrow $P_i \rightarrow P_{i+1}$ from each trait to the next in the sequence. We say that B is a *descendant* of A if there is a directed path from A to B. This means that A has a (direct or indirect) causal influence on B. These terms obey the everyday genealogical conventions: a variable's descendants consist of its children, its children's children, and so on. We restrict our attention to graphs that are *acyclic*, meaning that no trait is a descendant of itself. Cyclic graphs require more sophisticated data collection and analysis than we consider here and, even in linear cyclic models, key parameters may be inestimable (Gianola and Sorensen 2004). A *causal model* $\mathcal{M}$ consists of a causal diagram, a set of functional relationships f_Z, and a joint

probability distribution of background variables $\mathbb{P}(\boldsymbol{U})$. Note that a given causal diagram $\mathcal{G}$ can support many causal models.

These methods require that the researcher can make plausible assumptions about the qualitative structure of the causal processes that generated their data. These assumptions may be based on prior knowledge of the study species (e.g., due to previous experiments), general principles (e.g., causes precede effects in time), or mathematical relationships (e.g., when fitness is viewed as an "effect" of its components). Most importantly, the researcher must identify the potential presence and direction of causal effects between measured traits (i.e., the structure of the causal diagram). Including a potential causal effect in a model does not mean that the effect must be nonzero; on the other hand, the absence of an effect (i.e., a missing arrow) is a substantive assumption. Nonetheless, including unnecessary arrows in a model can introduce additional noise in estimates of causal effects, so researchers should only include an arrow if they believe the effect it represents is plausible. As in any selection analysis, our causal interpretations hold only in the absence of certain confounding variables. Confounding occurs when the relationship between two measured traits is misestimated because both traits are influenced by an unmeasured third variable. The nature and strictness of no-confounding assumptions depend on the result in question, so we outline them as necessary below. For a general review of how departures from causal assumptions can complicate or invalidate causal inference, we refer the reader to Antonakis et al. (2010).

Causal Derivatives and Extended Selection Gradients

Suppose that, by some external intervention, an individual's trait value Z is changed by a small amount, while holding fixed all traits and background variables that are not causally influenced by Z (i.e., all nondescendants of Z). This intervention induces cascading effects on the descendants of Z, which are described by the deterministic relationships of equation (4). Let us write $\pi(Z, w)$ for the set of directed paths from Z to fitness w. For any such path P, we denote the variables along the path in order as $(Z = P_0\,, P_1, \ldots,\ P_{m(P)} = w)$, where $m(P)$ is the length of the path. In the limit of infinitesimally small interventions, we can consider the instantaneous rate of fitness change due to changes in Z. We refer to this rate as the *causal derivative* of w on Z, denoted $\frac{\delta w}{\delta Z}$ (this notation indicates a special kind of partial derivative, where only nondescendants of Z are held fixed). If all the variables P_i are continuous, we can apply the chain rule of calculus to express the causal derivative as

$$\frac{\delta w}{\delta Z} = \sum_{P \in \pi(Z,w)} \prod_{j=1}^{m(P)} \frac{\partial P_j}{\partial P_{j-1}}. \tag{5}$$

The partial derivatives on the right are understood as $\frac{\partial P_j}{\partial P_{j-1}} = \frac{\partial}{\partial P_{j-1}} f_{P_j}(\mathrm{pa}(P_j), U_{P_j})$. Hypothetical interventions, as considered here, are a standard tool of modern causal theory (e.g., as implemented in Judea Pearl's do-calculus: Pearl 2009). Surprisingly, however, we could find no precedent for the concept of a causal derivative, perhaps because the mathematical literature on causality tends to focus on discrete interventions. For readers familiar with Pearl's approach to causal modeling (Pearl 2009, 2018), we provide an alternative definition of $\frac{\delta w}{\delta Z}$ in the Supporting Information (see also eq. 15). This alternative definition can also be applied when the descendants of Z are not all continuous.

The *extended selection gradient* η_Z is the mean causal derivative of Z, taken over the joint probability distribution of all traits (Morrissey 2014b, 2015):

$$\eta_Z = \mathbb{E}\left(\frac{\delta w}{\delta Z}\right). \tag{6}$$

The extended selection gradient can be thought of as the average causal effect of small changes in a focal trait on fitness. Morrissey (2015) discusses extended selection gradients on both measured traits η_Z and background variables η_{U_Z} (note that these two gradients coincide if $\frac{\partial Z}{\partial U_Z} = 1$). Here, we mainly emphasize selection on measured traits. For normally distributed traits Z, the portion of the selection differential arising from the causal effect of Z on fitness is given by (proof in Supporting Information):

$$S_{\mathrm{causal},Z} = \sigma_Z^2\, \eta_Z. \tag{7}$$

This is the portion of the change in the mean of Z within a generation that arises via causal effects of Z on fitness. Note here that only Z is required to be normal; all other traits may be nonnormal.

Our focus on continuous traits and on interventions of small effect is partly for mathematical convenience. In particular, this focus will allow us to separate extended selection gradients cleanly into components arising via distinct causal pathways (eq. 16 below). Nonetheless, many of our results can be recast in forms applicable to discrete variables, although we do not present these additional results here.

Estimating Extended Selection Gradients Using the Backdoor Criterion

The causal derivative $\frac{\delta w}{\delta Z}$ potentially depends on the values of the unmeasured background variables $\boldsymbol{U}$. As a consequence, the value of $\frac{\delta w}{\delta Z}$ is generally unknowable for any given observation. On the other hand, the extended selection gradient η_Z can often be estimated from observational data if the effects of confounding variables can be removed statistically. If we knew all of the

functional relationships f_Z among traits, as well as the joint probability distributions of the background variables $\mathbb{P}(\boldsymbol{U})$, then we could estimate the extended selection gradients on each trait by applying equations (5) and (6) directly. In practice, such information is generally not available, and more nuanced approaches are required (for gentle introductions, see Pearl 2018; Rohrer 2018; a more thorough treatment is given by Pearl 2009).

How can one estimate the causal effects of a trait without having measured all possible variables? A familiar solution to biologists is to control for a set of variables that intercepts noncausal paths between the focal trait Z and fitness. If we want to isolate the causal effect of a trait on fitness, it is important to choose this set carefully. For instance, we should not control for a descendant D of Z, as this will remove any indirect causal effects of the form $Z \to D \to w$, where each arrow indicates a directed path. Our set of covariates should also block *all* noncausal pathways between Z and fitness and not inadvertently open any new noncausal paths. The "backdoor criterion" is a formal method for ascertaining whether controlling for a particular set of covariates will isolate the causal effect of one variable on another (Pearl 2009, 2018). We call $\boldsymbol{B}$ a *backdoor set* with respect to selection on a trait Z if (i) $\boldsymbol{B}$ contains no descendants of Z and (ii) $\boldsymbol{B}$ blocks all backdoor (i.e., noncausal) paths from Z to fitness (Fig. 2: see Appendix for details). If $\boldsymbol{B}$ is a backdoor set for Z, then the extended selection gradient on Z can be expressed as

$$\eta_Z \;=\; \mathbb{E}\left(\frac{\partial}{\partial Z}\mathbb{E}\,(w|Z,\boldsymbol{B})\right). \tag{8}$$

In other words, η_Z can be calculated as an average partial selection gradient on Z with the traits $\boldsymbol{B}$ as covariates. If, in addition, Z is normally distributed, then equations (7) and (8) allow us to calculate how much of the selection differential arises from the causal effect of Z on fitness.

A special case arises when we are only interested in selection on the exogenous traits $\boldsymbol{X}$ (i.e., traits whose causes are not modeled explicitly). Suppose here that the set of background variables $\boldsymbol{U}_Y$ of the endogenous traits is independent of $\boldsymbol{X}$. For any focal exogenous trait, the remaining exogenous traits then constitute a backdoor set. We can consequently write the vector of extended selection gradients $\boldsymbol{\eta}_X$ on the exogenous traits as

$$\boldsymbol{\eta}_X \;=\; \mathbb{E}\,(\nabla\mathbb{E}\,(w|\boldsymbol{X}))\,. \tag{9}$$

In other words, the extended selection gradients $\boldsymbol{\eta}_X$ equal the average partial selection gradients $\boldsymbol{\beta}_{\mathrm{AP}}$ on $\boldsymbol{X}$ (it is important that the endogenous traits are excluded when calculating these gradients). If the exogenous traits are jointly multivariate normal, then the linear selection differentials on $\boldsymbol{X}$ can be expressed as $\mathbf{S_X} = \mathbf{P_X}\boldsymbol{\eta}_{\mathbf{X}}$, where $\boldsymbol{P}_X$ is the phenotypic variance-covariance matrix of $\boldsymbol{X}$ (cf. eq. 3a).

Decomposing Extended Selection Gradients Along Causal Pathways

So far, we have considered selection on a trait Z arising via all causal paths from Z to w. Suppose now that we only wish to consider selection via some subset of these pathways. For instance, we might want to consider all pathways passing through the variable(s) representing an individual's mating success; this would provide a measure of (premating) sexual selection (Henshaw et al. 2018). Let us write $\mathcal{H}$ for the subgraph of $\mathcal{G}$ consisting of the union of our paths of interest. For example, $\mathcal{H}$ might consist of a single path, or several paths representing similar types of selection. Analogously to above, we write $\pi_{\mathcal{H}}(Z, w)$ for the set of directed paths from Z to w that lie entirely in $\mathcal{H}$. We define the *path-specific causal derivative* of w on Z via paths in $\mathcal{H}$ as the rate at which fitness changes due to changes in Z, while holding all paths outside of $\mathcal{H}$ fixed. If all variables in $\pi_{\mathcal{H}}(Z, w)$ are continuous, we can express the path-specific causal derivative as

$$\frac{\delta w}{\delta Z}|_{\mathcal{H}} = \sum_{P\in\pi_{\mathcal{H}}(Z,w)}\prod_{j=1}^{m(P)}\frac{\partial P_j}{\partial P_{j-1}}. \tag{10}$$

Similarly, the *path-specific selection gradient* on Z is the average path-specific causal derivative:

$$\eta_Z\,|_{\mathcal{H}} \;=\; \mathbb{E}\left(\frac{\delta w}{\delta Z}|_{\mathcal{H}}\right). \tag{11}$$

We also provide a more general definition of $\eta_Z|_{\mathcal{H}}$ that does not require the descendants of Z to be continuous (see Supporting Information and eq. 16). Note that the extended selection gradient is simply the path-specific selection gradient over the whole graph $\mathcal{G}$ (i.e., $\eta_Z = \eta_Z|_{\mathcal{G}}$). Estimating path-specific selection gradients often requires stronger assumptions than estimating extended selection gradients. As we will see in the next section, even given strong assumptions that exclude certain types of confounding, only some path-specific selection gradients can be estimated from observational data alone.

Identifiability of Path-Specific Selection Gradients

We now consider which path-specific selection gradients are *identifiable*. This means that they can be estimated from observational data, assuming that the causal diagram is correct, without any prior knowledge of the functional relationships between causes and their effects (i.e., we know nothing about the shape of these functions, such as whether they are linear or additive). We assume that the background variables $\boldsymbol{U}_Y$ of the endogenous traits are mutually independent and that $\boldsymbol{U}_Y$ is independent of the exogenous traits $\boldsymbol{X}$. Given these assumptions, there is a formal criterion that allows one to decide if any given path-specific

selection gradient is identifiable (Avin et al. 2005; proof in Supporting Information). We first present the criterion and then consider its interpretation:

(**First-edge criterion**) Suppose C is a child of a trait Z. Let $\mathcal{H}(C)$ be the subgraph of $\mathcal{G}$ consisting of all directed paths from Z to w whose first edge is $Z \to C$. Then the path-specific selection gradient $\eta_Z|_{\mathcal{H}(C)}$ is identifiable (we will use the shorthand $\eta_Z|_C = \eta_Z|_{\mathcal{H}(C)}$ and refer to the *C-specific selection gradient* on Z). Further, if $\mathcal{H}$ is any subgraph for which the path-specific selection gradient $\eta_Z|_{\mathcal{H}}$ is identifiable, then $\mathcal{H}$ can be expressed as a union of the form $\mathcal{H} = \bigcup_{C \in \mathbf{C}} \mathcal{H}(C)$, where $\boldsymbol{C}$ is a subset of the children of Z.

The first-edge criterion might seem obscure initially, but its essence is very simple (Fig. 1A). If there are multiple causal pathways from Z to w that begin with the edge $Z \to C$, then these pathways cannot be teased apart using observational data alone (at least not in the absence of further assumptions; see below). In this sense, the C-specific selection gradients are *minimal* identifiable components of the extended selection gradient. Moreover, they are exhaustive, in the sense that any identifiable path-specific gradient can be built up out of one or more C-specific gradients. In fact, there is a one-to-one correspondence between identifiable path-specific selection gradients and subsets of the children of Z.

The extended selection gradient on a trait Z can now be rewritten as the sum of path-specific selection gradients via each of its children:

$$\eta_Z = \sum_{C \in \mathrm{ch}(Z)} \eta_Z|_C. \tag{12}$$

If Z is normally distributed, then the component of the selection differential on Z that arises via the effect of Z on its child C can be expressed as $\sigma_Z^2 \eta_Z|_C$. This represents the portion of the change in the mean of Z within a generation that arises due to the direct effect $Z \to C$.

Three subtleties are worth noting. First, the first-edge criterion applies in a very general setting, where nothing is assumed about probability distributions or the functional relationships among traits, other than the correctness of the causal diagram (which includes the independence assumptions on $\boldsymbol{U_Y}$). Further pathways may become identifiable if additional assumptions about the data-generating process can be made. In particular, if the effects of background variables can be assumed to be additive, then all path-specific selection gradients are identifiable (see below). This is a standard assumption in linear path analyses. Second, the C-specific selection gradient on a trait Z may not account for all pathways that pass from Z via C to w. Rather, it represents exactly those pathways that begin with the edge $Z \to C$ (see, e.g., the C-specific selection gradient on X_1 in Fig. 1A, which does not include the pathway $X_1 \to B \to C \to w$). In other words, the C-specific selection gradient represents selection arising from the *direct effect* of Z on C. Third, some downstream edges may be included in path-specific selection gradients via more than one child (e.g., the edge $C \to w$ in Fig. 1A), consistent with the fact that multiple mechanisms of selection may share a causal link.

Estimating Path-Specific Selection Gradients

So far, we have demonstrated the conditions under which path-specific selection gradients are identifiable, assuming that nothing is known about the functional relationships between causes and their effects. Fortunately, empirically interesting path-specific selection gradients should be estimable in a range of practical settings, requiring similar types of assumptions to those that practitioners of more elementary types of path analyses are regularly willing to make. We now provide an explicit way to estimate path-specific gradients whenever they are identifiable. We assume, as above, that the background variables $\boldsymbol{U}_Y$ of the endogenous traits are mutually independent and that $\boldsymbol{U}_Y$ is independent of the exogenous traits $\boldsymbol{X}$. It follows that the joint distribution of the traits $\boldsymbol{Z}$ equals the distribution of the exogenous traits $\boldsymbol{X}$ multiplied by the product of the conditional distributions of each endogenous trait Y given its parents (Pearl 2009):

$$\mathbb{P}(\mathbf{Z}) = \mathbb{P}(\mathbf{X}) \prod_{Y \in \mathbf{Y}} \mathbb{P}\left(Y|\mathrm{pa}(Y)\right). \tag{13}$$

We say that $\mathcal{G}$ is *Markovian* conditional on $\boldsymbol{X}$.

Suppose now that we have estimated the distribution of each endogenous trait Y, conditional on its parents $\mathrm{pa}(Y)$ (i.e., the distribution $\mathbb{P}(Y|\mathrm{pa}(Y))$). Such estimates can be obtained using a variety of regression-based methods that predict both the conditional expectation of a trait given its predictors, as well as the distribution of residuals (Morrissey 2014a). These include linear or generalized linear models, as well as more flexible frameworks such as generalized additive models and multidimensional smoothing (O'Hara 2009; Wood 2017). Given such estimates for each endogenous trait Y, we can write expected fitness as

$$\mathbb{E}(w) = \int w \; d\mathbb{P}(\boldsymbol{X}) \prod_{Y \in \boldsymbol{Y}} d\mathbb{P}\left(Y|\mathrm{pa}(Y)\right). \tag{14}$$

In other words, we integrate realized fitness w over the joint probability distribution of the exogenous traits $d\mathbb{P}(\boldsymbol{X})$ and the conditional probability distributions $d\mathbb{P}(Y\,\mathrm{pa}(Y))$ of each endogenous variable given its parents. Now, the extended selection gradient η_Z on a continuous trait Z is fully determined by the trait's effects on its children $\mathrm{ch}(Z)$. Analogously to equation (14), we can write the extended selection gradient as (proof in Supporting

Information):

$$\eta_Z = \mathbb{E}\left(\frac{\delta w}{\delta Z}\right) = \int w \; d\mathbb{P}(\boldsymbol{X}) \underbrace{\left(\frac{\partial}{\partial Z}\prod_{C\in \mathrm{ch(Z)}} d\mathbb{P}\left(C|\mathrm{pa}\left(C\right)\right)\right)}_{\text{children of } Z} \underbrace{\left(\prod_{Y\in \boldsymbol{Y}\setminus \mathrm{ch(Z)}} d\mathbb{P}\left(Y|\mathrm{pa}\left(Y\right)\right)\right)}_{\text{nonchildren of } Z}. \quad (15)$$

Note that the product term from equation (14) has been split up into two factors, representing endogenous traits that are children and nonchildren of Z, respectively. The partial derivative in the first of these factors quantifies how small changes in Z affect the distributions of its children $d\mathbb{P}(C|\mathrm{pa}(C))$. Unlike equation (5), this expression still applies if the descendants of Z are not all continuous.

Equation (15) tracks how changes in Z affect *all* of its children. In contrast, the path-specific selection gradients relate to changes in some subset of these children. One way to conceptualize this is via a "twin-network" model (Fig. 1B), where the focal trait Z can take on one value for the purposes of influencing some children, but another value for influencing its remaining children (Avin et al. 2005; Pearl 2009; details in Supporting Information). This construction allows one to control for (i.e., hold fixed) all nonfocal pathways when quantifying path-specific selection gradients. Equivalently, one can decompose the extended selection gradient into its child-specific components by applying the product rule of calculus to the partial derivative in equation (15). For any child C, the C-specific selection gradient equals:

$$\eta_Z|_C = \int w \; d\mathbb{P}(\mathbf{X}) \left(\frac{\partial}{\partial Z} d\mathbb{P}\left(C|\mathrm{pa}\left(C\right)\right)\right) \left(\prod_{Y\in \mathbf{Y}\setminus C} d\mathbb{P}\left(Y|\mathrm{pa}\left(Y\right)\right)\right). \quad (16)$$

Note that the right-hand side of this equation depends only on observable quantities and can consequently be estimated from observational data (see Appendix for a worked example). Unlike equation (10), this expression still applies when the descendants of Z are not all continuous. The joint distribution of $\boldsymbol{X}$ can be estimated as its empirical (i.e., sample) distribution or by parametric means. Note that our focus on infinitesimal interventions allows a clean separation of the extended selection gradient into path-specific gradients, each of which corresponds to the trait's effects via one of its children. For discrete interventions, it is more challenging to cleanly separate a variable's effects arising via multiple causal pathways.

Decomposing Path-Specific Coefficients into Multiplicative Components, Assuming that the Effects of Background Variables are Additive

The above analysis is very general, in the sense that the functional relationships f_Z and the distribution of $\boldsymbol{U}$ are arbitrary (aside from the independence assumptions encoded in the causal diagram). We now make an additional assumption that, when met, allows us to further decompose causal effects. Suppose that the effects of background variables are additive, in the sense that for each trait Z there exist functions g_Z and h_Z such that

$$f_Z\left(\mathrm{pa}\left(Z\right), U_Z\right) = g_Z\left(\mathrm{pa}\left(Z\right)\right) + h_Z\left(U_Z\right). \quad (17)$$

This assumption is standard in classical path analysis, which assumes that residuals are additive. If, in addition, U_Z is independent of $\mathrm{pa}(Z)$, then the partial derivative of Z with respect to any of its parents $\mathrm{pa}(Z)_i$ is given by

$$\frac{\partial Z}{\partial \mathrm{pa}(Z)_i} = \frac{\partial}{\partial \mathrm{pa}(Z)_i}\mathbb{E}\left(Z|\mathrm{pa}\left(Z\right)\right). \quad (18)$$

Note that this expression is independent of U_Z and thus can be estimated from observational data (i.e., it is identifiable). Consequently, all path-specific derivatives and effects are identifiable in this case. From equation (10), the path-specific causal derivative of w on Z via the single directed path P is then given by

$$\frac{\delta w}{\delta Z}|_P = \prod_{j=1}^{m(P)} \frac{\partial}{\partial P_{j-1}}\mathbb{E}\left(P_j|\mathrm{pa}\left(P_j\right)\right). \quad (19)$$

In other words, the causal derivative is equal to the product of partial effects corresponding to each edge along a chain of sequential causal effects. This is a restatement of Wright's rules that relies on the local linearity of the functions f_{P_i}. Note that this expression does not explicitly depend on the background variables $\boldsymbol{U}$, but rather only on the observed traits $\boldsymbol{Z}$. Using the Markovian property from equation (13), the path-specific selection gradient on Z via the path P is then

$$\eta_Z|_P = \mathbb{E}_{\boldsymbol{Z}}\left(\prod_{j=1}^{m(P)} \frac{\partial}{\partial P_{j-1}}\mathbb{E}\left(P_j|\mathrm{pa}\left(P_j\right)\right)\right) = \int\left(\prod_{j=1}^{m(P)} \frac{\partial}{\partial P_{j-1}}\mathbb{E}\left(P_j|\mathrm{pa}\left(P_j\right)\right)\right) d\mathbb{P}(\boldsymbol{X})\prod_{Y\in \boldsymbol{Y}} d\mathbb{P}(Y|\mathrm{pa}\left(Y\right)). \quad (20)$$

This formula allows us to calculate path-specific selection gradients via *any* causal pathway, under the assumptions that (i)

background variables $\boldsymbol{U_Y}$ of endogenous traits are both mutually independent and independent of $\boldsymbol{X}$, and (ii) the effects of background variables are additive.

In classical path-analytic models, each endogenous trait is a linear function of its parents. We can then write $\frac{\partial}{\partial P_{j-1}}\mathbb{E}\,(P_j|\mathrm{pa}(P_j)) = \beta_{P_jP_{j-1}}$, where the *path coefficient* $\beta_{P_jP_{j-1}}$ is a fixed constant that does not depend on the values of the parent variables $\mathrm{pa}(P_j)$. In this case, we can rewrite equation (20) as

$$\eta_Z|_P = \prod_{j=1}^{m(P)} \beta_{P_jP_{j-1}}. \tag{21}$$

Thus, we recover the well-known rule for linear models that the P-specific effect is simply the product of path coefficients along the path P (Wright 1921; Loehlin 2004; Pearl 2009; Zhang and Bareinboim 2018). The extended selection gradient is the sum of such products over every directed path P joining Z to w. For nonlinear relationships, this simple decomposition does not hold, as was recognised already by Wright (1921).

Decomposing the Predicted Evolutionary Response to Selection Between Generations Across Causal Pathways

As well as inducing a change in phenotype, selection may also change the distribution of alleles in a population within a single generation. The phenotypic effects of transmitting such changes between generations are known as the *evolutionary response to selection*. Here, we show how to decompose the predicted evolutionary response to selection along causal pathways. Let us suppose that each background variable U_Z can be modeled as a function of a *genetic value* A_Z and a *residual value* R_Z. In other words, $U_Z = f_{U_Z}(A_Z, R_Z)$ for some function f_{U_Z}. A common model is $U_Z = A_Z + R_Z$, where A_Z is an additive genetic effect (i.e., a breeding value) and R_Z incorporates both environmental and nonadditive genetic effects (Lande 1979; Morrissey 2015; Walsh and Lynch 2018). We write $\boldsymbol{A}$ and $\boldsymbol{R}$ for the vectors of genetic and residual values for each trait. Note that these values determine the *background variables* rather than the measured traits (in the language of Morrissey 2015, they determine the "exogenous inputs" to the system).

The evolutionary response to selection is generally predicted via the change in mean additive genetic values due to selection within a generation (i.e., the difference in $\boldsymbol{A}$ between all individuals and the parents of offspring, with the latter weighted by the number of offspring per parent). This change is given by the *genetic selection differentials* $\boldsymbol{S_A} = \mathrm{cov}(\boldsymbol{A}, w)$ (Robertson 1966; Walsh and Lynch 2018). If the additive genetic values $\boldsymbol{A}$ are multivariate normal, then Stein's lemma tells us that (Walsh and Morrissey 2019)

$$\boldsymbol{S_A} = \boldsymbol{G}\boldsymbol{\beta_A}. \tag{22}$$

Here, $\boldsymbol{G} = \mathrm{cov}(\boldsymbol{A})$ is the covariance matrix of genetic effects and $\boldsymbol{\beta_A} = \mathbb{E}(\nabla\mathbb{E}(w|\boldsymbol{A}))$ is a vector of average partial selection gradients on the genetic values $\boldsymbol{A}$. Suppose further that (i) $U_Z = A_Z + R_Z$ for all traits Z (i.e., background variables equal the sum of genetic and residual values), and (ii) $\boldsymbol{A}$ is independent of $\boldsymbol{R}$ (Geyer and Shaw 2008; Queller 2017). We then have $\boldsymbol{\beta_A} = \boldsymbol{\eta_U}$, where $\boldsymbol{\eta_U} = \mathbb{E}(\nabla\mathbb{E}(w|\boldsymbol{U}))$ is the vector of extended selection gradients on the background variables (proof in Supporting Information). This allows us to write (Morrissey 2015):

$$\boldsymbol{S_A} = \boldsymbol{G}\boldsymbol{\eta_U} \tag{23}$$

This equation describes the within-generation genetic change in background variables as a result of selection. The actual phenotypic change between generations depends on this genetic change, along with any environmental change and the details of inheritance, where only some of the latter will be captured by the causal model (Morrissey 2015; Walsh and Lynch 2018). Note that this equation holds under arbitrary dependencies among the background variables $\boldsymbol{U}$.

The above equations depend explicitly on the unmeasured background variables $\boldsymbol{U}$. To apply them, we must model the functional relationships f_Z between unmeasured background variables and measured traits, as well as the distribution of background variables given $\mathrm{pa}(Z)$. In the particular case where a background variable U_Z is assumed to affect the trait Z additively (i.e., $Z = g_Z(\mathrm{pa}(Z)) + h_Z(U_Z)$, as above), we can transform U_Z such that h_Z is the identity function and $\frac{\partial Z}{\partial U_Z} = 1$. We then have $\eta_Z = \eta_{U_Z}$. The component of the predicted evolutionary response to selection on Z that is due to the causal effects of Z on fitness can then be written as

$$S_{\mathrm{causal},A_Z} = \sigma^2_{A_Z}\eta_Z. \tag{24}$$

This represents the portion of the change in genetic values for Z within a generation that arises due to the causal effect of Z on fitness, rather than by spurious mechanisms. Moreover, using equation (12), we can further decompose this quantity into contributions arising from the causal effects of Z on fitness via each of its children:

$$S_{\mathrm{causal},A_Z} = \sigma^2_{A_Z}\sum_{C\in\mathrm{ch}(Z)}\eta_Z|_C. \tag{25}$$

Empirical Applications

In a companion article, we apply the above method to study selection on body mass and early pregnancy in Soay sheep (*Ovis aries*) (Janeiro et al. pre-print). We also provide a worked example in the Appendix to this article, where we analyze simulated data inspired by the life history of the marsupial genus *Antechinus*.

Discussion

We have presented general principles for decomposing selection into components arising via distinct causal pathways. In particular, our approach allows researchers to calculate selection *via* intermediate traits such as fitness components, which leads naturally to operational definitions of concepts like sexual selection (i.e., selection via mating or fertilization success), fecundity selection, and longevity selection. Such nuanced causal breakdowns are not provided by classical approaches based on selection differentials and selection gradients alone. The framework we present is very general, in that it accommodates arbitrary trait-fitness relationships and arbitrary distributions of variables. It is not necessary that researchers know the shapes of such relationships or distributions in advance. Rather, each endogenous variable in the causal diagram can be modeled using any appropriate regression technique, including semiparametric methods that require few statistical assumptions (Schluter 1988; Morrissey and Sakrejda 2013; Morrissey 2014a; Lefcheck 2016; Wood 2017). Our approach is consequently much more flexible than classical path analyses, which assume that the relationships between effects and their causes are linear and additive (Wright 1921, 1934; Arnold and Duvall 1994; Frank 1997). The framework presented here complements that of Morrissey (2015) by estimating not only the extended selection gradients, but also the path-specific selection gradients via causally intermediate traits. It can also be used to obtain path-specific decompositions of the predicted evolutionary response to selection, thus integrating smoothly with evolutionary quantitative genetics.

Very little causal knowledge can be gained from observational data alone (Rohrer 2018). However, a few judicious assumptions about the qualitative nature of causal pathways often suffice to permit extensive quantitative inferences regarding their strengths (Pearl 2009). Our method requires that the researcher sketch a plausible causal diagram outlining the qualitative structure of the causal processes that generated their data, including the potential presence (or assumed absence) and direction of causal effects among measured traits. In many cases, we believe such models will be naturally suggested by a species' life cycle and ecology. For instance, fitness may be determined by life history components such as survival, fecundity, and so on, which in turn are influenced by morphological and behavioral traits (Arnold and Duvall 1994). The causal relationships among traits that are more distantly linked to fitness will often be less clear. As with approaches based on selection gradients, however, our framework allows the researcher to remain agnostic about the causal determinants of exogenous traits. Although methods exist to compare the plausibility of competing causal hypotheses, at least in some circumstances, these methods are not the focus of the current work (for details, see Pearl 2009; Kline 2016; Shipley 2016).

Like all phenotypic analyses of selection, our results hold only in the absence of certain types of confounding variables (i.e., unmeasured variables that causally influence two or more measured variables). The problem of "hidden variables" is well-known from regression-based studies of selection (e.g., Mitchell-Olds and Shaw 1987; Rausher 1992; Morrissey et al. 2010; Walker 2014; Reed et al. 2016; Walsh and Lynch 2018) and is fundamentally no different here. Analyses will be improved by measuring and including any obvious confounders, to the extent that this is feasible. Although some relevant variables will inevitably be missed, the framework presented here potentially provides a more nuanced understanding of selection than regression-based approaches, which also assume no hidden variables. Ultimately, information from a variety of sources, including manipulative experiments, should be integrated together to unambiguously resolve causal patterns.

A prime application of the theory here will be to analyzing sexual selection. Sexual selection is usually defined as selection arising via competition for mating or fertilization opportunities (Andersson 1994; Shuker 2010). It can be further partitioned into *premating* sexual selection, which acts via the number and quality of mates, and *postmating* sexual selection, which acts via fertilization success after mating (e.g., paternity under sperm competition: Parker and Pizzari 2010). We currently lack general methods to partition natural selection into sexual and nonsexual components, and to further disentangle sexual selection into selection via mate number, mate quality, and fertilization success (Fitzpatrick 2015; Henshaw et al. 2018). Current approaches are limited to quantifying multiple components of sexual selection *en bloc* (Henshaw et al. 2018) or are constrained by inappropriate assumptions of linearity and additivity (Arnold and Duvall 1994). Decomposing the variance in fitness into pre- and postmating components may be heuristically useful (Arnold and Wade 1984; Rose et al. 2013; Janicke et al. 2015; Evans and Garcia-Gonzalez 2016; Marie-Orleach et al. 2016), but such decompositions are inexact and do not indicate the extent to which fitness variance is due to selection versus environmental effects (cf. Klug et al. 2010). The method outlined here would allow selection to be partitioned along these constituent pathways, a goal that we will pursue in future work.

Although we have presented our framework in the context of selection theory, it applies naturally to many problems where the causal relationships among continuous variables can be represented via directed acyclic graphs (i.e., graphs with no causal loops). Assuming there are no unmeasured confounders or other sources of bias, the total effect of a continuous variable X on a variable Y can be decomposed into path-specific effects via each child of X, while controlling for other measured variables (eq. 16). For a normally distributed variable X, the covariance between X and Y can then be separated into causal and spurious components, and the causal component can be decomposed into effects arising via the direct effects of X on each of its children. Like the causal theory that it draws from (Avin et al. 2005; Pearl 2009, 2018), our approach is consequently of broad applicability.

AUTHOR CONTRIBUTIONS

JMH and AGJ conceived of the study; JMH and MBM worked out the mathematical details; JMH wrote the paper with input from MBM and AGJ; all authors approved the final version.

ACKNOWLEDGMENTS

We are grateful to K. Fritzsche, M. J. Janeiro, L. Kruuk, and L. Fromhage for their comments on the manuscript and to M. Jennions for many helpful discussions. E. Kisdi, J. Grace, and an anonymous reviewer provided constructive feedback that greatly improved the manuscript. MBM is supported by a University Research Fellowship from the Royal Society (London).

Open access funding enabled and organized by Projekt DEAL.

DATA ARCHIVING

The simulated dataset and R code for the worked example are available on Dryad (http://doi.org/10.5061/dryad.j0zpc86c8).

LITERATURE CITED

Andersson, M. 1994. Sexual selection. Princeton Univ. Press, Princeton, NJ.

Antonakis, J., S. Bendahan, P. Jacquart, and R. Lalive. 2010. On making causal claims: a review and recommendations. Leadersh. Q. 21:1086–1120.

Arnold, S. J., and D. Duvall. 1994. Animal mating systems: a synthesis based on selection theory. Am. Nat. 143:317–348.

Arnold, S. J., and M. J. Wade. 1984. On the measurement of natural and sexual selection: theory. Evolution 38:709–719.

Avin, C., I. Shpitser, and J. Pearl. 2005. Identifiability of path-specific effects. Pp. 357–363 *in* Proceedings of International Joint Conference on Artificial Intelligence. Edinburgh, U.K.

Conner, J. K., S. Rush, and P. Jennetten. 1996. Measurements of natural selection on floral traits in wild radish (*Raphanus raphanistrum*). I. Selection through lifetime female fitness. Evolution 50:1127–1136.

Crespi, B. J., and F. L. Bookstein. 1989. A path-analytic model for the measurement of selection on morphology. Evolution 43:18–28.

de Villemereuil, P. 2018. Quantitative genetic methods depending on the nature of the phenotypic trait. Ann. N.Y. Acad. Sci. 1422:29–47.

de Villemereuil, P., H. Schielzeth, S. Nakagawa, and M. B. Morrissey. 2016. General methods for evolutionary quantitative genetic inference from generalized mixed models. Genetics 204:1281–1294.

Evans, J. P., and F. Garcia-Gonzalez. 2016. The total opportunity for sexual selection and the integration of pre- and post-mating episodes of sexual selection in a complex world. J. Evol. Biol. 29:2338–2361.

Fisher, D. O., M. C. Double, S. P. Blomberg, M. D. Jennions, and A. Cockburn. 2006. Post-mating sexual selection increases lifetime fitness of polyandrous females in the wild. Nature 444:89–92.

Fitzpatrick, C. L. 2015. Expanding sexual selection gradients; a synthetic refinement of sexual selection theory. Ethology 121:207–217.

Frank, S. A. 1997. The Price equation, Fisher's fundamental theorem, kin selection, and causal analysis. Evolution 51:1712–1729.

Fromhage, L., and M. D. Jennions. 2019. The strategic reference gene: an organismal theory of inclusive fitness. Proc. R. Soc. B 286: 20190459.

Geyer, C. J., and R. G. Shaw. 2008. Commentary on Lande-Arnold analysis. School of Statistics, University of Minnesota, Minneapolis, MN.

Gianola, D., and D. Sorensen. 2004. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics 167:1407–1424.

Godfrey-Smith, P. 2007. Conditions for evolution by natural selection. J. Phil. 104:489–516.

Grace, J. B., and K. M. Irvine. 2020. Scientist's guide to developing explanatory statistical models using causal analysis principles. Ecology 101:e02962.

Grace, J. B., D. R. Schoolmaster, G. R. Guntenspergen, A. M. Little, B. R. Mitchell, K. M. Miller, and E. W. Schweiger. 2012. Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere 3:art73.

Henshaw, J. M., and Y. Zemel. 2017. A unified measure of linear and nonlinear selection on quantitative traits. Methods Ecol. Evol. 8:604–614.

Henshaw, J. M., A. T. Kahn, and K. Fritzsche. 2016. A rigorous comparison of sexual selection indexes via simulations of diverse mating systems. Proc. Natl. Acad. Sci. USA 113:E300–E308.

Henshaw, J. M., M. D. Jennions, and L. E. B. Kruuk. 2018. How to quantify (the response to) sexual selection on traits. Evolution 72:1904–1917.

Janeiro, M. J., J. M. Henshaw, J. M. Pemberton, J. G. Pilkington, and M. B. Morrissey (pre-print). Selection of lamb size and early pregnancy in Soay sheep (Ovies aries). bioRxiv. https://doi.org/10.1101/2020.09.16.299685

Janicke, T., P. David, and E. Chapuis. 2015. Environment-dependent sexual selection: Bateman's parameters under varying levels of food availability. Am. Nat. 185:756–768.

Kingsolver, J. G., and D. W. Schemske. 1991. Path analyses of selection. Trends Ecol. Evol. 6:276–280.

Kline, R. B. 2016. Principles and practice of structural equation modeling. 4th ed. Guilford Press, New York.

Klug, H., J. Heuschele, M. D. Jennions, and H. Kokko. 2010. The mismeasurement of sexual selection. J. Evol. Biol. 23:447–462.

Lande, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

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

Latta, R. G., and C. McCain. 2009. Path analysis of natural selection via survival and fecundity across contrasting environments in *Avena barbata*. J. Evol. Biol. 22:2458–2469.

Lefcheck, J. S. 2016. piecewiseSEM: piecewise structural equation modelling in R for ecology, evolution, and systematics. Methods Ecol. Evol. 7:573–579.

Liu, J. S. 1994. Siegel's formula via Stein's identities. Stat. Probab. Lett. 21:247–251.

Loehlin, J. C. 2004. Latent variable models: an introduction to factor, path, and structural equation analysis. 4th ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Marie-Orleach, L., T. Janicke, D. B. Vizoso, P. David, and L. Schärer. 2016. Quantifying episodes of sexual selection: insights from a transparent worm with fluorescent sperm. Evolution 70:314–328.

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

Morrissey, M. B. 2014a. In search of the best methods for multivariate selection analysis. Methods Ecol. Evol. 5:1095–1109.

———. 2014b. Selection and evolution of causally covarying traits. Evolution 68:1748–1761.

———. 2015. Evolutionary quantitative genetics of nonlinear developmental systems. Evolution 69:2050–2066.

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

Morrissey, M. B., L. E. B. Kruuk, and A. J. Wilson. 2010. The danger of applying the breeder's equation in observational studies of natural populations. J. Evol. Biol. 23:2277–2288.

O'Hara, R. B. 2009. How to make models add up—a primer on GLMMs. Ann. Zool. Fenn. 46:124–137.

Okasha, S. 2016. The relation between kin and multilevel selection: an approach using causal graphs. Brit. J. Phil. Sci. 67:435–470.

Okasha, S., and J. Otsuka. 2020. The Price equation and the causal analysis of evolutionary change. Phil. Trans. R. Soc. B 375:20190365.

Otsuka, J. 2016. Causal foundations of evolutionary genetics. Brit. J. Phil. Sci. 67:247–269.

Parker, G. A., and T. Pizzari. 2010. Sperm competition and ejaculate economics. Biol. Rev. 85:897–934.

Pearl, J. 2009. Causality. 2nd ed. Cambridge Univ. Press, Cambridge, U.K.

———. 2018. The book of why. Basic Books, New York.

Queller, D. C. 2017. Fundamental theorems of evolution. Am. Nat. 189:345–353.

———. 2020. The gene's eye view, the Gouldian knot, Fisherian swords and the causes of selection. Phil. Trans. R. Soc. B 375:20190354.

Rausher, M. D. 1992. The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. Evolution 46:616–626.

Reed, T. E., P. Gienapp, and M. E. Visser. 2016. Testing for biases in selection on avian reproductive traits and partitioning direct and indirect selection using quantitative genetic models. Evolution 70:2211–2225.

Robertson, A. 1966. A mathematical model of the culling process in dairy cattle. Anim. Prod. 8:95–108.

Rohrer, J. M. 2018. Thinking clearly about correlations and causation: graphical causal models for observational data. Adv. Methods Pract. Psychol. Sci. 1:27–42.

Rose, E., K. A. Paczolt, and A. G. Jones. 2013. The contributions of premating and postmating selection episodes to total selection in sex-role-reversed gulf pipefish. Am. Nat. 182:410–420.

Scheiner, S. M., R. J. Mitchell, and H. S. Callahan. 2000. Using path analysis to measure natural selection. J. Evol. Biol. 13:423–433.

Schluter, D. 1988. Estimating the form of natural selection on a quantitative trait. Evolution 42:849–861.

Shipley, B. 2016. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R. 2nd ed. Cambridge Univ. Press, Cambridge, U.K.

Shuker, D. M. 2010. Sexual selection: endless forms or tangled bank? Anim. Behav. 79:e11–e17.

Sober, E. 1984. The nature of selection: evolutionary theory in philosophical focus. Univ. of Chicago Press, Chicago.

Stein, C. M. 1981. Estimation of the mean of a multivariate normal distribution. Ann. Stat. 9:1135–1151.

Uller, T., and K. N. Laland, eds. 2019. Evolutionary causation: biological and philosophical reflections. MIT Press, Cambridge, MA.

Wade, M. J., and S. Kalisz. 1990. The causes of natural selection. Evolution 44:1947–1955.

Walker, J. A. 2014. The effect of unmeasured confounders on the ability to estimate a true performance or selection gradient (and other partial regression coefficients). Evolution 68:2128–2136.

Walsh, B., and M. Lynch. 2018. Evolution and selection of quantitative traits. Oxford Univ. Press, Oxford, U.K.

Walsh, B., and M. B. Morrissey. 2019. Evolutionary quantitative genetics. Pp. 421–430 *in* D. Balding, I. Moltke, and J. Marioni, eds. Handbook of statistical genomics. Wiley, Hoboken, NJ.

Wood, S. N. 2017. Generalized additive models: an introduction with R. 2nd ed. CRC Press, Boca Raton, FL.

Wright, S. 1921. Correlation and causation. J. Agr. Res. 20:557–585.

———. 1934. The method of path coefficients. Ann. Math. Stat. 5:161–215.

Zhang, J., and E. Bareinboim. 2018. Non-parametric path analysis in structural causal models. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence.

Associate Editor: E. Kisdi
Handling Editor: D. W. Hall

Appendix

The Backdoor Criterion

Here, we briefly introduce the "backdoor criterion," which is a formal method for ascertaining whether controlling for a particular set of covariates is sufficient to isolate the effect of one variable on another (see Pearl 2009 for a more extensive treatment). We first require some further terminology. A *path* from a variable X to another variable Y is sequence of traits $(X = P_1, P_2, \ldots, P_n = Y)$ such that there is an arrow between each adjacent pair of traits P_i and P_{i+1} in the sequence. The arrows may point in either direction or be double headed (in contrast to a "directed path," where all arrows are single headed and point in the same direction). A *backdoor path* from X to Y is a path that begins with an arrow pointing into X (i.e., a path of the form $X \leftarrow \ldots Y$ or $X \leftrightarrow \ldots Y$). Backdoor paths may generate spurious covariation between traits and must consequently be controlled for when estimating causal effects. We say that a set of traits $\boldsymbol{B}$ *blocks* a backdoor path P if and only if:

1. P contains a *chain* of the form $P_{i-1} \rightarrow P_i \rightarrow P_{i+1}$ or $P_{i-1} \leftarrow P_i \leftarrow P_{i+1}$ or a *fork* of the form $P_{i-1} \leftarrow P_i \rightarrow P_{i+1}$ such that the middle node P_i is in $\boldsymbol{B}$, or
2. P contains a *collider* $P_{i-1} \rightarrow P_i \leftarrow P_{i+1}$ such that neither the middle node P_i nor any of its descendants is in $\boldsymbol{B}$.

Graphs with double-headed arrows should be evaluated by first replacing edges of the form $R \leftrightarrow S$ with $R \leftarrow U_{RS} \rightarrow S$,

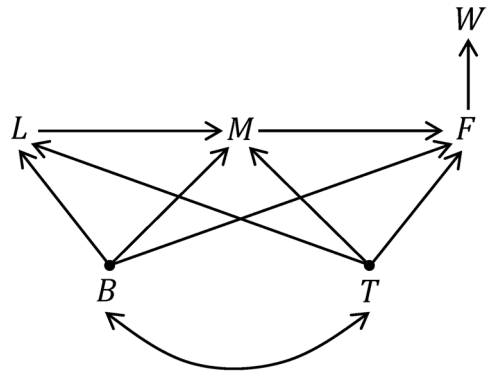

Figure A1. Causal diagram for the worked example on antechinus females. An individual's body size *B* and her last date of torpor *T* influence her absolute fitness *W* via their effects on her survival to breed *L*, and her number of matings *M* and total fecundity *F* if she does survive. In addition, fecundity may depend on the (nonzero) number of matings. The covariance between body size and the last date of torpor is not analyzed causally. Unmeasured background variables are omitted.

where U_{RS} is an unmeasured common cause. Informally, if $\boldsymbol{B}$ blocks a backdoor path P, then controlling for $\boldsymbol{B}$ should remove spurious effects arising via P (see Fig. 2 in the main text for examples).

Worked Example

Here, we illustrate our method via a hypothetical example that was inspired by the genus *Antechinus* (the simulated dataset and R code for this example are available on Dryad:http://doi.org/10.5061/dryad.j0zpc86c8). Antechinus are small carnivorous marsupials that use torpor to reduce energy consumption from late summer to early winter (Fisher et al. 2006). They reproduce once per year in late winter or early spring, following which most individuals die. We suppose that researchers tracked female antechinus from mid-summer to the end of the breeding season. They measured the following variables: body size B at the beginning of the study period; the last date T at which individuals used torpor; survival L to the beginning of the breeding season; and the number of matings M and the fecundity F of surviving females. Fecundity was used as an estimate of absolute fitness, under the assumption that very few individuals would survive to breed again. The researchers constructed a path model of selection on body size and the date of last torpor (Fig. A1). They supposed that both these traits might directly affect survival, mating success and fecundity. Because only surviving individuals can mate, we necessarily have $M = 0$ when $L = 0$. Similarly, only mated individuals can produce offspring, so $F = 0$ when $L = 0$. In addition, the researchers allowed for a direct effect of the (nonzero) number of mates on fecundity.

We begin by estimating the distribution of each endogenous variable, conditional on its parents, using standard regression techniques. Given that the fitting of regression models is not the focus of the current work, we ask that readers accept that the models we have chosen are appropriate for the dataset. Of course, the selection of suitable models is a crucial step in any empirical application of our method. We modeled all three intermediate variables (i.e., survival, fecundity, and the number of matings) using generalized additive models (Wood 2017). We explicitly modeled mating success only for surviving individuals, and fecundity only for those individuals that mated at least once, because the effects of nonsurvival or zero mating are certainly not additive with respect to other variables. To allow for nonlinear relationships, the effects of all predictors (i.e., the parents of each intermediate variable) were fitted as smooth functions using thin plate regression splines (R package "mgcv": Wood 2017). Survival was modeled as a binomial variable with a logit link function. Fecundity and the number of matings were modeled as Poisson variables with log link functions. Absolute fitness was estimated by F.

From these regression models, we estimated the extended selection gradients for each trait, as well as the path-specific selection gradients via survival, fecundity, and the number of matings. We write $\hat{W}(B, T)$ for the expected value of absolute fitness given any pair of (B, T). Note that each of these traits constitutes a backdoor set for the other trait. The extended selection gradient on B (eq. 8 in the main text) can then be estimated as

$$\eta_B = \frac{1}{\mathbb{E}(W)} \lim_{\varepsilon \to 0} \mathbb{E}_{B,T} \left(\frac{\hat{W}(B + \varepsilon, T) - \hat{W}(B, T)}{\varepsilon} \right). \quad \text{(A1)}$$

In practice, this function can be closely approximated by setting ε to a small value. The factor $\frac{1}{\mathbb{E}(W)}$ corresponds to a transformation from absolute to relative fitness, assuming that population mean fitness is relatively insensitive to changes in an individual's trait values (i.e., the population is fairly large). For fixed inputs, expected fitness $\hat{W}(B, T)$ was estimated as

$$\hat{W}(B, T) = \mathbb{P}(L = 1|B, T) \sum_{M \geq 1} \mathbb{P}(M|B, T, L = 1) \mathbb{E}(F|M, B, T, L = 1). \quad \text{(A2)}$$

Each component of this equation was calculated from the fitted generalized additive models described above. We then averaged over the empirical distribution of (B, T) to obtain η_B.

For the path-specific selection gradients, rather than working with equation (16) in the main text directly, it is more convenient to rely on the underlying twin-network model (Fig. 2B in the main text; see Supporting Information for the formal model). Let us write $\hat{W}_X(B_X, B_*, T)$ for the expected fitness when B has the value B_X for the purposes of influencing X, but the value B_*

Table A1. Breakdown of selection on body size (B) and the last day of torpor (T, in days) into contributions from each modeled causal pathway in a simulated dataset of antechinus females. Raw values are based on unstandardized traits, whereas standard values are calculated using traits that have been standardized to have variances of one.

	Measure	Formula	Raw value	Standard value
	Variance in body size	σ_B^2	25.00	1
	Variance in torpor emergence date	σ_T^2	17.20	1
	Covariance between body size and torpor emergence date	$\mathrm{cov}(B, T)$	–5.13	–0.247
Selection on body size, B	Path-specific selection gradients via L, M, F	$\eta_B\|_L \eta_B\|_M \eta_B\|_F$	0.010 0.003 0.008	0.052 0.017 0.038
	Extended selection gradient	η_B	0.021	0.107
	Linear regression selection gradient	$\beta_{LR,B}$	0.021	0.107
	Causal components of selection differential arising via effects of B on L, M, F	$\sigma_B^2\eta_B\|_L \sigma_B^2\eta_B\|_M \sigma_B^2\eta_B\|_F$	0.260 0.083 0.192	0.052 0.017 0.038
	Causal component of selection differential	$\sigma_B^2\eta_B$	0.535	0.107
	Direct selection (Lande and Arnold 1983)	$\sigma_B^2\beta_{\mathrm{LR},B}$	0.534	0.107
	Selection differential	$S_B = \mathrm{cov}(B, w)$	0.455	0.091
Selection on last torpor date, T	Path-specific selection gradients via L, M, F	$\eta_T\|_L \eta_T\|_M \eta_T\|_F$	0.016 0.001 –0.002	0.066 0.002 –0.006
	Extended selection gradient	η_T	0.015	0.062
	Linear regression selection gradient	$\beta_{LR,T}$	0.015	0.064
	Causal components of selection differential arising via effects of T on L, M, F	$\sigma_T^2\eta_T\|_L \sigma_T^2\eta_T\|_M \sigma_T^2\eta_T\|_F$	0.273 0.009 –0.027	0.066 0.002 –0.006
	Causal component of selection differential	$\sigma_T^2\eta_T$	0.256	0.062
	Direct selection (Lande and Arnold 1983)	$\sigma_T^2\beta_{\mathrm{LR},T}$	0.264	0.064
	Selection differential	$S_T = \mathrm{cov}(T, w)$	0.155	0.037

for influencing all other children of B. The X-specific selection gradient can then be estimated as

$$\eta_B|_X = \frac{1}{\mathbb{E}(W)} \lim_{\varepsilon \to 0} \mathbb{E}_{B,T}\left(\frac{\hat{W}_X(B+\varepsilon, B, T) - \hat{W}_X(B, B, T)}{\varepsilon}\right). \quad \text{(A3)}$$

The quantities $\hat{W}_X(B_X, B_*, T)$ are straightforward to estimate from our regression models. For instance, to calculate the survival-specific selection gradient on body size, we used

$$\hat{W}_L(B_L, B_*, T) = \mathbb{P}(L = 1|B_L, T) \sum_{M \geq 1} \mathbb{P}(M|B_*, T, L = 1) \mathbb{E}(F|M, B_*, T, L = 1). \quad \text{(A4)}$$

We then averaged over the empirical distribution of (B, T) to obtain $\eta_B|_L$. Last, noting that body size is approximately normal in distribution, we estimated the component of the selection differential arising due to the causal effect of body size on fitness via survival as $\sigma_B^2\eta_B|_L$ (Table A1). We used analogous calculations to estimate path-specific selection on body size and torpor date via survival, mating success, and fecundity.

Results

Our analysis shows that there is weak overall selection on both body size and the date of last torpor (selection differentials in Table A1). Selection led to an increase of almost half a gram in body size ($S_B = 0.455$), and an extension of torpor by around one sixth of a day ($S_T = 0.155$). The joint distribution of body size and the

date of last torpor was approximately bivariate normal. Accordingly, our estimates of the extended selection gradients on these traits were very close to the linear regression selection gradients $\boldsymbol{\beta}_{LR}$. For both traits, causal selection for larger trait values (i.e., for larger body size or later emergence from torpor) was offset by spurious selection for smaller trait values, which arose due to the negative covariance between body size and the date of last torpor.

Selection on body size occurred mainly via its effect on survival and fecundity, with selection via mating success contributing only modestly to overall selection. Selection on the date of last torpor occurred almost exclusively via its effect on survival, with individuals that emerged late from torpor being more likely to survive until the beginning of the breeding season. Surviving individuals had similar mating success and fecundity, regardless of when they emerged from torpor.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material