doi:10.1111/evo.13867

Comparing the strength of modular signal, and evaluating alternative modular hypotheses, using covariance ratio effect sizes with morphometric data

Dean C. Adams[1,2] and Michael L. Collyer[3]

[1]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011*

[2]*E-mail: dcadams@iastate.edu*

[3]*Department of Science, Chatham University, Pittsburgh, Pennsylvania 15232*

Received July 2, 2019
Accepted October 9, 2019

The study of modularity is paramount for understanding trends of phenotypic evolution, and for determining the extent to which covariation patterns are conserved across taxa and levels of biological organization. However, biologists currently lack quantitative methods for statistically comparing the strength of modular signal across datasets, and a robust approach for evaluating alternative modular hypotheses for the same dataset. As a solution to these challenges, we propose an effect size measure (Z_{CR}) derived from the covariance ratio, and develop hypothesis-testing procedures for their comparison. Computer simulations demonstrate that Z_{CR} displays appropriate statistical properties and low levels of mis-specification, implying that it correctly identifies modular signal, when present. By contrast, alternative methods based on likelihood (`EMMLi`) and goodness of fit (`MINT`) suffer from high false positive rates and high model mis-specification rates. An empirical example in sigmodontine rodent mandibles is provided to illustrate the utility of Z_{CR} for comparing modular hypotheses. Overall, we find that covariance ratio effect sizes are useful for comparing patterns of modular signal across datasets or for evaluating alternative modular hypotheses for the same dataset. Finally, the statistical philosophy for pairwise model comparisons using effect sizes should accommodate any future analytical developments for characterizing modular signal.

KEY WORDS: **Modularity, morphological evolution, morphometrics, trait covariation.**

Characterizing the extent to which phenotypic traits covary, and deciphering the forces that shape patterns of trait covariation, are perennial topics in evolutionary biology. Empirical research has demonstrated that patterns of covariation are unevenly dispersed across traits, with some variables exhibiting high correlations with one another while other traits are more independent (Olson and Miller 1958; Cheverud 1996; Klingenberg 2008; Goswami and Polly 2010). Indeed, studies of morphological integration and modularity are largely concerned with how these trait correlations are distributed, and the extent to which they are driven by developmental, genetic, or functional linkages among traits (Olson and Miller 1958; Wagner 1984). In particular, modularity describes patterns where trait correlations are high within subsets of variables, termed *modules*, and where trait correlations across modules are comparatively weaker (Cheverud 1982; Goswami 2006; Wagner et al. 2007). Patterns of modularity therefore result in sets of semi-autonomous traits, which have the potential to respond differentially to natural selection, and thus the capacity to promote the evolution of novelty (see Wagner and Altenberg 1996; Wagner et al. 2007; Tokita et al. 2007; Hansen and Houle 2008; Clune et al. 2013).

The past several decades have seen resurging interest in understanding the evolution of modularity, and numerous analytical approaches have been developed for quantifying patterns of modularity in phenotypic datasets (e.g., Magwene 2001; Mitteroecker and Bookstein 2007; Márquez 2008; Klingenberg 2009; Adams

© 2019 The Author(s). *Evolution* © 2019 The Society for the Study of Evolution.

2016; Goswami and Finarelli 2016). Concomitant with these advances is an increasing number of empirical studies that characterize patterns of modularity in distinct phenotypic traits, and across a wide variety of taxa (for recent examples, see Parsons et al. 2012; Parr et al. 2016; Felice and Goswami 2018; Larouche et al. 2018; Bardua et al. 2019). Likewise, evolutionary biologists have striven to decipher whether patterns of modularity are similar among taxa and traits, and across levels of biological organization (Drake and Klingenberg 2010; Renaud et al. 2012; Sanger et al. 2012; Felice and Goswami 2018; Bardua et al. 2019; Marshall et al. 2019). Some studies have investigated whether patterns of modularity are conserved across taxa (e.g., Goswami 2006; Marshall et al. 2019), however, direct quantitative or statistical comparisons of modularity patterns are generally lacking. We assert that a critical aspect of this endeavor should be to determine whether the "strength" of modularity is similar across datasets. However, to our knowledge, no formal statistical procedure has yet been proposed to accomplish this task.

Another key challenge in the study of modularity is identifying which sets of traits represent anatomical modules (Zelditch et al. 1990; Hallgrimsson et al. 2007). Biologically, one expects that traits subjected to common processes may exhibit high correlations with one another, resulting in modular structure (Olson and Miller 1958; Wagner 1984; Cheverud 1996; Goswami and Polly 2010). However, phenotypic traits are influenced by an array of genetic, developmental, and functional forces, not all of which act similarly. Thus, it is reasonable to expect that multiple, competing modular hypotheses may be proposed for the same set of anatomical traits (e.g., Márquez 2008; Goswami and Finarelli 2016; Felice and Goswami 2018; Bardua et al. 2019). This then raises the question: which of the alternative modular hypotheses provides the best description of the observed patterns of trait covariation? Several recent attempts to address this problem (e.g., Márquez 2008; Goswami and Finarelli 2016) have greatly sharpened our notions of modularity, and provided insights as to how phenotypic variation may be expected to evolve. For instance, one attempt (`MINT`; Márquez 2008) uses goodness of fit measures to evaluate alternative modular hypotheses, while another approach (`EMMLi`; Goswami and Finarelli 2016) uses penalized likelihood indices (AICc) to evaluate the fit of alternative modular hypotheses to the observed trait correlations.

Nonetheless, these approaches do not provide a complete analytical toolkit for evaluating alternative modular hypotheses for the full spectrum of phenotypic datasets evolutionary biologists wish to examine. One reason for this is that initial investigations into their performance were relatively limited, and suggested that future improvements or alternatives may be required. For example, simulations revealed that `MINT` tended to identify modular patterns when none existed in the data, suggesting that false positives may be a concern (Márquez 2008: Table 3). Likewise, initial investigations found that `EMMLi` tended to select highly parameterized models (Goswami and Finarelli 2016; Goswami pers. comm.), suggesting that it may display some degree of model misspecification. However, to date, an evaluation of both approaches under a broad set of plausible scenarios has not been conducted, nor has their performance been directly compared (see below). As a consequence, it remains possible that neither approach provides a reliable means of evaluating patterns of modular structure, thereby reducing one's ability to identify evolutionary trends in trait covariation. Therefore, we suggest that a second pressing analytical need in the study of modularity is the development of robust statistical methods for comparing the strength of modular signal across alternative modular hypotheses for the same dataset.

In this article, we develop a single statistical tool that accomplishes both of these tasks. Our approach utilizes a standardized test statistic (an effect size: Z_{CR}) for measuring the degree of modularity between sets of variables, and for comparing these effect sizes statistically. Our procedure is based on the covariance ratio (Adams 2016), and may be used for comparing the strength of modular signal across datasets, as well as for evaluating patterns of modular signal as defined by alternative modular hypotheses for the same dataset. Using computer simulations, we evaluate the statistical properties of tests based on Z_{CR}, and find that they display appropriate type I error rates, high statistical power, and low levels of model mis-specification. Thus, when modular signal is present, Z_{CR} is unlikely to mis-assign that signal to an incorrect modular hypothesis. We then compare these results to those found using two alternative methods for comparing modular hypotheses: `MINT` (Márquez 2008) and `EMMLi` (Goswami and Finarelli 2016). We find that both exhibit poor statistical performance (very high false positive rates, and high model mis-specification rates), making it challenging to arrive at reliable biological inferences when using these approaches. We illustrate the utility of Z_{CR} with an empirical example of mandible shape in sigmodontine rodents, where alternative modular hypotheses, and the strength of modular signal across several species, is compared. Our new method is implemented in R, and is distributed through the package `geomorph` (Adams et al. 2019). Finally, the method can accommodate future analytical developments for characterizing modular signal as they are developed (see Discussion).

Methods

ANALYTICAL DEVELOPMENT

Covariance ratio effect sizes (Z_{CR})

One approach to quantifying the degree of modular signal in morphometric datasets is based on the covariance ratio (Adams 2016). This measure is preferable to alternative estimates, such as the RV coefficient (sensu: Klingenberg 2009), because it is insensitive to the number of variables (p) and the number of

specimens (n) (Adams 2016). To calculate the covariance ratio (CR), one first concatenates the variables for all modules into an $n \times p$ matrix ($\mathbf{Y}$). At a minimum, the variables are represented as mean-centered data, implying that they are residuals from an intercept model. However, one could also perform additional transformations to the data, such as a phylogenetic transformation (e.g., Garland and Ives 2000; Adams 2014; Adams and Collyer 2018a) to account for non-independence among objects as a result of shared evolutionary history (see Adams and Felice 2014 for a related approach measuring morphological integration in a phylogenetic context). Using this concatenated dataset, the covariance ratio for the observed data is found as:

$$CR = \sqrt{\frac{tr(\mathbf{S}_{12}\mathbf{S}_{21})}{\sqrt{tr(\mathbf{S}_{11}^{*}\mathbf{S}_{11}^{*})tr(\mathbf{S}_{22}^{*}\mathbf{S}_{22}^{*})}}} \quad (1)$$

where $\mathbf{S}_{11}^{*}$ and $\mathbf{S}_{22}^{*}$ represent the within-module covariance matrices for each of two modules (with zeroes replacing the diagonal elements), and $\mathbf{S}_{21} = \mathbf{S}_{12}^{T}$ represents the covariation between modules $\mathbf{Y}_1$ and $\mathbf{Y}_2$. For datasets containing more than two modules, the CR coefficient is calculated for each pair of modules, and the average pairwise CR coefficient is used (see Adams 2016). Empirically, the CR coefficient has a lower limit of 0, an expected value of 1 when there is no modularity, and values greater than 1 are also possible if covariances between module variables exceed covariances within modules. Thus, CR values that are closer to 0 describe datasets with a relatively greater degree of modularity, while CR values closer to 1 describe datasets with less modular signal. To evaluate CR statistically, a resampling procedure that randomly assigns variables to modules is used, and in each iteration a CR_{rand} is obtained under the null hypothesis of no modularity. The observed value, CR_{obs}, is then compared to the empirically generated sampling distribution of CR_{rand} values to evaluate significance (for details, see Adams 2016).

The permutation procedure above is sufficient to statistically determine whether the degree of modular signal is greater than expected by chance for a single dataset, but evaluating modular signals across datasets requires a standardized effect size to ensure statistical comparability. To accomplish this, we make use of prior theoretical developments for effect sizes obtained from statistics from multivariate data and their empirical sampling distributions (Collyer et al. 2015; Adams and Collyer 2016). Previous studies used residual randomization in permutation procedures (RRPP) to generate sampling distributions. The important insight from Collyer et al. (2015) was that a permutation-based effect size may be obtained from any observed test statistic, found in relation to the mean and SD of its empirical sampling distribution obtained from RRPP. Although the resampling procedure for generating sampling distributions of the CR statistic is different than RRPP, the concept of using the random outcomes to evaluate the size of the observed effect is the same. Thus, for the case of the covariance ratio, this effect size is obtained as:

$$Z_{CR} = \frac{CR_{obs} - \hat{\mu}_r}{\hat{\sigma}_r} \quad (2)$$

where CR_{obs} is the observed covariance ratio for the dataset, $\hat{\mu}_r$ is the expected value of CR under the null hypothesis of no modularity (found as the mean of the empirical sampling distribution), and $\hat{\sigma}_r$ is the SE of the mean, found as the SD of the empirical sampling distribution. Note that the calculation of Z_{CR} is the same regardless of how many modules are represented, because CR_{obs} in equation (2) represents the average CR value obtained across all pairs of modules (see Adams 2016). One should recognize that more negative values of Z_{CR} represent greater modular signal, because when modular signal is present, CR_{obs} will be less than $\hat{\mu}_r$.

Importantly, it can be shown that Z_{CR} is normally distributed, and thus represents a valid standardized effect size. To demonstrate this property, we performed a simulation experiment, where 500 multivariate datasets were generated at a given n and p, and with all variates drawn from a normal distribution: $\mathcal{N}(0, 1)$. Next, variables were randomly assigned to modules, and both the CR and Z_{CR} were obtained from each dataset, with Z_{CR} obtained using 999 permutations. The mean and SD of Z_{CR} across the 500 datasets were then calculated, and the sampling experiment was repeated across differing levels of n and p. As is clear from Figure 1, under the null hypothesis of no modularity (i.e., a random association of variables to modules), the covariance ratio effect size exhibits a constant expected value near zero, and constant variance (inferred from the constant confidence interval) across the entire spectrum of sample sizes (n) and variable number (p). Furthermore, the distribution of Z_{CR} values from the 500 datasets for any set of simulation conditions was also found to be normally distributed (Results not shown). Thus, Z_{CR} represents a valid standardized effect size that characterizes the strength of modular signal in morphometric datasets.

A two-sample Z-score for comparing modular signals

Once the degree of modular signal has been characterized, one may wish to evaluate hypotheses that compare multiple effect sizes. For hypotheses of modularity, this is particularly useful in two biological contexts. First, it may be of interest to determine whether the "strength" of modular signal is greater in one dataset as compared to another. To date, no explicit analytical approach has been proposed to statistically compare the strength of modular signal across datasets, though some studies have evaluated whether general patterns of modularity are conserved across taxa (e.g., Goswami 2006; Sanger et al. 2012; Marshall et al. 2019). Second, if several alternative modular hypotheses have been

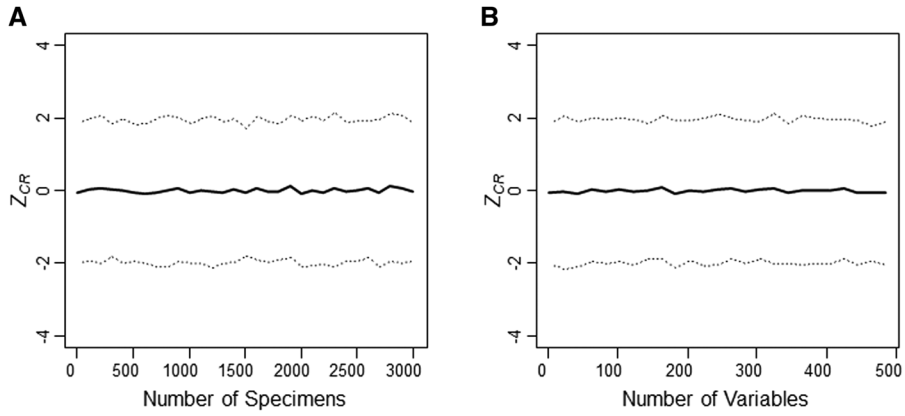


Figure 1. Evaluation of covariance ratio effect sizes (Z_{CR}) under the hypothesis of no modularity (i.e., a random association of variables to modules). Mean and 95% confidence intervals of Z_{CR} obtained from (A) 500 datasets simulated across a range of sample sizes, and from (B) 500 datasets simulated across a range of variable number.

proposed for the same structure (based on developmental, genetic, or functional grounds), it may be of interest to characterize the degree of modular signal under each of these hypotheses and determine which provides the best description of the observed patterns of morphological covariation (see Márquez 2008; Goswami and Finarelli 2016; Felice and Goswami 2018).

As an analytical solution to both of these tasks, we propose a two-sample test, which calculates the effect size of the difference between pairs of modular effect sizes, Z_{CR}. The approach is analogous to that used to compare patterns of morphological integration across datasets (Adams and Collyer 2016). In this case, the paired effect size is found as:

$$\hat{Z}_{12} = \frac{|(CR_1 - \hat{\mu}_1) - (CR_2 - \hat{\mu}_2)|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} \quad (3)$$

where CR_1, CR_2, $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1$, and $\hat{\sigma}_2$ are as defined above for equation (2). Note that this two-sample effect size is simply the difference between the numerators of Z_{CR_1} and Z_{CR_2}, standardized by the pooled within-sample SD. The probability of $\hat{Z}_{12}$ may then be estimated from a standard normal distribution. Typically, this is treated as a two-tailed test, though directional (one-tailed) tests may be specified should the empirical situation require it. Finally, confidence intervals for $\hat{Z}_{12}$ are found as:

$$(1-\alpha)100\%CI = |(CR_1 - \hat{\mu}_1) - (CR_2 - \hat{\mu}_2)| \pm z_{\alpha/2}\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}. \quad (4)$$

where $z_{\alpha/2}$ is the quantile from a standard normal distribution corresponding to the two-tailed probability for the level of significance, α.

The two-sample effect size described here is appropriate for comparing the relative strength of modular signal across datasets, as well as for comparing modular signal as defined by alternative modular hypotheses for the same dataset. For the latter application, it is often of interest to include the null hypothesis of no modularity within the set of modular hypotheses (see e.g., Márquez 2008; Goswami and Finarelli 2016). For this case, all variables are considered to belong to a single module, so $Z_{CR} = 0.0$ for this hypothesis. The reason is that, under the hypothesis that all traits belong to a single module, each permutation iteration results in the same outcome, as each permutation yields the same covariance matrix (with rows and their corresponding columns permuted). Thus, all computations of CR result in identical values; meaning that the numerator of Z_{CR} (equation 2) is exactly 0, as $CR_{obs} = \hat{\mu}$. Likewise, because all CR_r are identical under this hypothesis, $\hat{\sigma}_r = 0$. Thus, the null hypothesis that all variables belong to a single module may be included in the analysis simply by substituting these values as $\hat{\mu}_2$ and $\hat{\sigma}_2$ in equation (3) above to compare Z_{null} to the set of alternative modular hypotheses.

STATISTICAL PERFORMANCE OF $\hat{Z}_{12}$

To determine whether the two-sample test statistic ($\hat{Z}_{12}$) displays appropriate statistical properties, we performed a series of computer simulations. For each simulation, we first selected the number of specimens: $n = (50, 100, 250)$, the number of variables: $p = (16, 32, 64)$, and the number of modules into which the p-variables would be equally partitioned: $M = (2, 4, 8)$. Next, an initial $p \times p$ covariance matrix was constructed ($\mathbf{S}_{in}$), such that it contained a known difference between the covariation within modules (S_W) and the covariation between modules (S_B). To accomplish this, all between-module covariance elements (i.e., those elements describing the covariance between variables assigned to different modules) were drawn from a normal distribution with an average value of 0.6: $\mathcal{N}(\mu = 0.6, \sigma = 0.01)$, while the within-module covariance elements were

drawn from a normal distribution: $\mathcal{N}(\mu = S_W, \sigma = 0.01)$, with an average value equal to, or exceeding, the between-module covariance, depending upon simulation conditions (S_W = 0.60, 0.625, 0.650, 0, 675, 0.70; additional simulations with differing covariance patterns are described in the Supporting Information). The elements of $\mathbf{S}_{in}$ were then adjusted as needed using the `nearPD` function in the `Matrix` package, so that the resulting matrix conformed to the properties of a valid covariance matrix (i.e., symmetric, positive definite). Finally, for each simulation condition (i.e., each combination of n, p, M, and S_W), 500 multivariate datasets were generated using the desired input covariance matrix ($\mathbf{S}_{in}$) by drawing a set of variates from a normal distribution: $\mathcal{N}(0, 1)$, and post-multiplying them by the Cholesky decomposition of $\mathbf{S}_{in}$. This resulted in random normal data following: $\mathcal{N}(0, \mathbf{S}_{in})$, from which CR, and Z_{CR} were obtained (using 999 random permutations, in addition to the observed case). Importantly, this simulation procedure enabled the evaluation of both the type I error and statistical power of $\hat{Z}_{12}$, depending upon simulation conditions. For instance, simulations with input covariance matrices displaying equal within- and between-module covariation ($S_W \approx S_B$) represented situations where there was no modular signal, and thus evaluated type I error. Alternatively, simulations with input covariance matrices containing differing covariation levels ($S_W > S_B$) represented situations where there were increasing amounts of modular signal, and thus evaluated statistical power.

Because Z_{CR} may be used to evaluate two distinct biological scenarios (comparing the strength of modular signal across datasets, and evaluating alternative modular hypotheses for the same dataset), we conducted two parallel sets of simulations to evaluate its performance. For the first set of simulations (comparing Z_{CR} across datasets), we generated two datasets for each simulation using the procedure above, where the first dataset was simulated under $\mathbf{S}_{in}$ containing no modularity (i.e., $S_W \approx S_B$), while the second dataset was simulated under $\mathbf{S}_{in}$ containing some degree of modular signal (i.e., $S_W > S_B$). For these simulations, a two-module alternative was used ($M = 2$), although additional simulations using $M = 4$ yielded similar findings (Results not shown). Then, for each pair of simulated datasets, we evaluated the difference in strength of modular signal using $\hat{Z}_{12}$ as described above. The proportion of datasets (out of 500) that differed significantly from the null model (no modular signal) was treated as the type I error or statistical power, depending on initial levels of S_W for the second dataset. For all considerations, an $\alpha = 0.05$ level of significance was used.

The second set of simulations was designed to evaluate the ability of $\hat{Z}_{12}$ to distinguish among alternative modular hypotheses for the same dataset. In this case, datasets were simulated to contain one, two, four, or eight modular partitions, under the set of input conditions described above (i.e,. for a given n, p, and S_W). Next, alternative hypotheses partitioning variables into modules were developed ($M = 1, 2, 4, 8$), and Z_{CR} values were calculated for each dataset under each of the alternative modular hypotheses. Then, for each dataset, the most negative Z_{CR} was identified, as this represented the hypothesis representing the strongest modular signal. This was then compared with the value of M used in that simulation to determine whether the approach correctly identified the modular hypothesis that generated the data. All hypotheses were statistically compared using our test statistic, $\hat{Z}_{12}$. The proportion of datasets (out of 500) that differed significantly from the null model (no modular signal), and where the correct (input) modular hypothesis was identified, was treated as statistical power of the test (i.e., when $S_W > S_B$). Likewise, type I error was calculated as the proportion of datasets that differed significantly from the null model (no modular signal) when there was no input modular signal (i.e., when $S_W \approx S_B$). Finally, because it was known which modular hypothesis generated the data, the proportion of datasets assigned to an incorrect modular hypothesis that differed significantly from the null model was also determined, and treated as an estimate of model mis-specification. Note that in some cases power and model misspecification may not sum to 1.0, because there may be instances where an incorrect modular hypothesis was identified as the best hypothesis, but where this model did not differ statistically from the null. For all considerations, an $\alpha = 0.05$ level of significance was used.

For comparing the performance of $\hat{Z}_{12}$ to previous methods, we subjected each dataset from the second set of simulations to both the `MINT` (Márquez 2008) and `EMMLi` (Goswami and Finarelli 2016) procedures. `MINT` compares the observed phenotypic covariance matrix to a set of covariance matrices that describe alternative modular hypotheses using a goodness of fit test. A γ^* statistic measures the degree of fit, and is used to rank models, with lower γ^* values providing stronger support for that model. A jackknife procedure may also be used to obtain confidence intervals for statistical evaluation (see details in Márquez 2008). Using `MINT`, the best modular hypothesis was identified for each simulated dataset above. Then, when modular signal was present (i.e., $S_W > S_B$), the proportion of datasets where the correct model was ranked as the best model was determined. This proportion (true positives) was treated as an estimate of power. Likewise, the proportion of datasets where an incorrect model was identified as the best model described levels of model misspecification. Finally, for datasets containing no modular structure (i.e., $S_W \approx S_B$), the proportion of datasets where the best model was distinct from the null model was treated as the false positive rate, which was analogous to a type I error rate. For these datasets, a jackknife procedure was also used to generate 95% confidence intervals (sensu Márquez 2008) to determine whether the best hypothesis differed from the (correct) null hypothesis.

A second procedure, `EMMLi`, describes the fit of the data to multiple modular hypotheses by estimating the likelihood of the observed trait correlation matrix relative to a particular modular hypothesis. Penalized likelihood indices (AICc) are then obtained for each model based on the number of parameters the model requires, and models are ranked using this index to determine which model exhibits the highest support (for details see Goswami and Finarelli 2016). Using `EMMLi`, we identified the best modular hypothesis for each simulated dataset above. When modular signal was present (i.e., $S_W > S_B$), the proportion of datasets where the correct model was ranked as the best model was determined, and treated as an estimate of power (i.e., true positives). Likewise, the proportion of datasets where an incorrect model was identified as the best model described levels of model misspecification. Finally, for datasets containing no modular structure (i.e., $S_W \approx S_B$), the proportion of datasets where the best model was distinct from the null model was treated as the false positive rate, which was analogous to a type I error rate.

Finally, an increasing number of empirical studies investigate patterns of modularity under the "high p: small N" scenario ($p>>N$), where the number of variables greatly exceeds the number of specimens (e.g., Parr et al. 2016; Felice and Goswami 2018; Bardua et al. 2019; Goswami et al. 2019). Such situations pose particular problems for parametric statistical approaches, as the likelihood of the model, and many summary test measures based on them, cannot be computed, as the trait covariance matrix is singular (see Adams 2014; Adams and Collyer 2018b). Thus, with respect to investigations of modularity, the potential for model mis-specification may be high. To evaluate the extent to which high-dimensional data (i.e., $p>>N$) affects the performance of modularity methods, we performed additional computer simulations where the number of variables, $p = (180, 360)$, greatly exceeded the number of specimens, $N = (50, 100)$. Datasets were generated using the identical procedure as described above, and were then subjected to statistical evaluation using several modularity procedures (see Supporting Information).

All simulations were performed in R 3.6.0 (R Core Team 2019) using the packages `geomorph` (Adams and Otárola-Castillo 2013; Adams et al. 2019), `EMMLi` (Goswami and Finarelli 2016), and `evolqg` (Melo et al. 2016).

EMPIRICAL EXAMPLE

To illustrate the use of $\hat{Z}_{12}$, we conducted two analyses of modularity patterns in mandible shape from sigmodontine rodents, using the dataset of Márquez (2008). The rodent mandible has long served as a model system for understanding the development and evolution of complex morphological structures (Atchley and Hall 1991; Cheverud et al. 1997; Klingenberg et al. 2003). Developmentally, the mandible forms from six condensations, arguably the developmental modules of the mandible (sensu Hall 2003), which can be mapped onto the form of the adult mandible (Fig. 6A). Much is known about mandibular development, including the developmental origins of the cells that give rise to the adult mandible, the extracellular signaling pathways that regulate the migration, proliferation, and differentiation of those cells, as well as the interactions between the developing skeleton and other tissues, including teeth and muscles. Those various interactions potentially integrate modules due to epigenetic or pleiotropic effects (sensu Atchley and Hall 1991). The most complex hypothesis of mandibular modularity proposed by Márquez (2008) is that the condensations themselves are developmental modules; the landmarks are thus partitioned as shown in Figure 6A so that each subset is within the region formed by one condensation. Note, however, that partition 4 spans a region derived from two condensations.

Our dataset (from Márquez 2008) consisted of morphometric shape data for mandibles from 546 adult individuals across nine species of rodents. Mandible shape was quantified using 18 landmarks and 51 semilandmarks (Fig. 5A). Following the procedures in Márquez (2008), we superimposed all specimens using a Generalized Procrustes Analysis, where the positions of semilandmarks were adjusted using the Procrustes distance criterion. Shape allometry was then removed via regression, and shape residuals were subsequently used for all analyses below (see details in Márquez 2008). Because shape variables were obtained from individuals across several species, we removed species-specific shape differences as well.

Using the aligned shape data, we conducted two separate modularity comparisons. First, we compared several alternative modularity hypotheses to determine which provided the best description of shape covariation patterns. Several hypotheses of modularity were formulated by combining these partitions into modules (ranging from two to six) that predict within-module covariances based on their developmental, genetic, and functional connections (see Table 1). These included a single-module model, several two-module models, a three-module model, and a six-module model (Table 1; for further details, see Márquez 2008). The Z_{CR} was then calculated for each modularity hypothesis, and these were compared in pairwise fashion using $\hat{Z}_{12}$ as described above. Importantly, because these analyses were performed on shape variables whose species-specific shape differences were removed, the above modularity analyses were equivalent to those performed on the pooled within-group covariance matrix. Second, once the optimal modularity model was identified for the entire dataset, we estimated CR and Z_{CR} using that model for each of the nine species separately, and compared the strength of modular signal among species using $\hat{Z}_{12}$ to determine whether some species displayed a greater degree of modularity than did others. All analyses were performed in R 3.6.0 (R Core Team 2019)

Table 1. Description of modular hypotheses examined in the empirical example. Each modularity hypothesis describes which developmental unit (Fig. 5A) is assigned to which module. Hypotheses define putative modules based on developmental and functional considerations (for details see Márquez 2008).

Model	Modules	Description
H-1	[1,2,3,4,5,6]	All developmental units belong to a single module
H-2A	[1,2][3,4,5,6]	Separate tooth and muscle bearing modules (Cheverud et al. 1997)
H-2B	[1,2,4_{part}][3,4_{part},5,6]	Alternative tooth and muscle bearing modules (Klingenberg et al. 2003)
H-2C	[1,2,5][3,4,6]	Two modules: rotation at TMJ on occlusion patterns (Björk 1969)
H-3	[1][2][3,4,5,6]	Three modules: two dental, and the posterior region (Márquez 2008)
H-6	[1][2][3][4][5][6]	Each developmental unit is a separate module (Hall 2003)

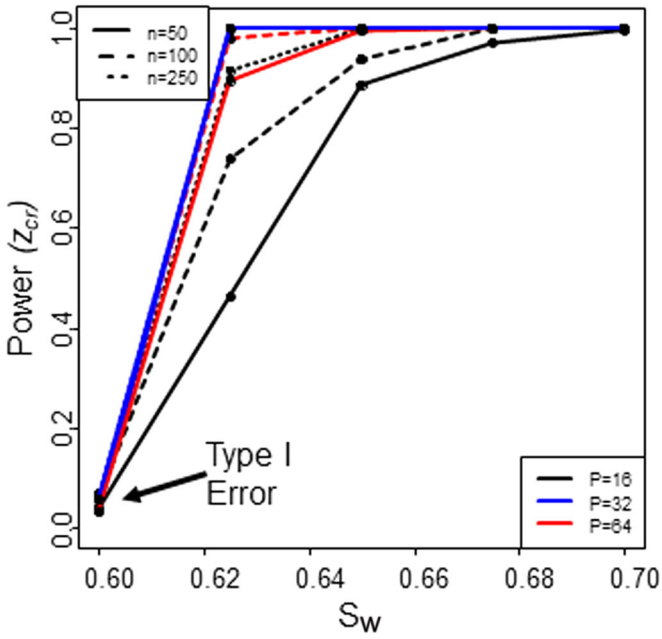


Figure 2. Simulation results evaluating the type I error and statistical power of hypothesis testing procedures for comparing the strength of modular signal across datasets using Z_{CR}. Data were simulated using differing numbers of variables *(p)* and at different samples sizes (*n*). Between-module covaration was: $S_B = 0.6$, and within-module covariation is shown along the *X*-axis (see text for details).

using `geomorph` (Adams and Otárola-Castillo 2013; Adams et al. 2019).

Results

SIMULATION RESULTS

Using simulations, we found that comparisons of effect sizes across datasets using $\hat{Z}_{12}$ displayed appropriate type I error rates (~5%), which remained consistent across simulation conditions. Additionally, statistical power was maintained as the degree of modularity increased, and quickly attained very high power, even when the degree of modular signal was relatively modest (Fig. 2). Predictably, statistical power increased with both increasing sample size (n) and an increase in the number of variables (p). Patterns remained robust, and were even stronger, when a four-module hypothesis was used in the simulations (Results not shown). These results confirm that statistical comparisons of Z_{CR} across datasets can be used to reliably identify differences in the strength of modular signal when it is present.

Likewise, when comparing effect sizes obtained from alternative modular hypotheses, we found that tests based on Z_{CR} also displayed appropriate type I error rates (~5%), which remained consistent across simulation conditions. Statistical power increased rapidly as the degree of modularity increased (Fig. 3), and power increased with the number of variables (p). In the case of high-dimensional data (where $p >> N$), this increase in power was retained (Supplemental Material: Fig. S1), demonstrating that tests based on Z_{CR} exhibited increasing statistical power with increasing trait dimensionality, a pattern observed in other permutation-based statistical procedures for high-dimensional data (see Adams 2014, 2016; Adams and Collyer 2018a). Notably, this high power was observed regardless of whether the input modular signal was generated from a two-module (Fig. 3A), a four-module (Fig. 3C), or an eight-module (Fig. 3E) model. Finally, results based on Z_{CR} were unaffected by differing levels of covariation among pairs of modules, implying that the approach was generally robust to such variation (see Supplemental Material).

As with the previous simulations, $\hat{Z}_{12}$ quickly attained very high power, even when the degree of modular signal was relatively modest. And as before, statistical power increased with both increasing sample size (n) and an increase in the number of variables (p). The modular hypothesis identified as displaying the strongest modular signal corresponded with the input model in nearly all cases. Finally, in the few instances where an alternative model was incorrectly identified as the best model (i.e., displayed the lowest $\hat{Z}_{12}$), this model did not provide a significantly better fit than did the correct input model. As a consequence, $\hat{Z}_{12}$ displayed very low levels of model misspecification (Fig. 3B, 3D, 3F), implying that when modular signal was present, the method exhibited a high propensity for identifying the correct modular

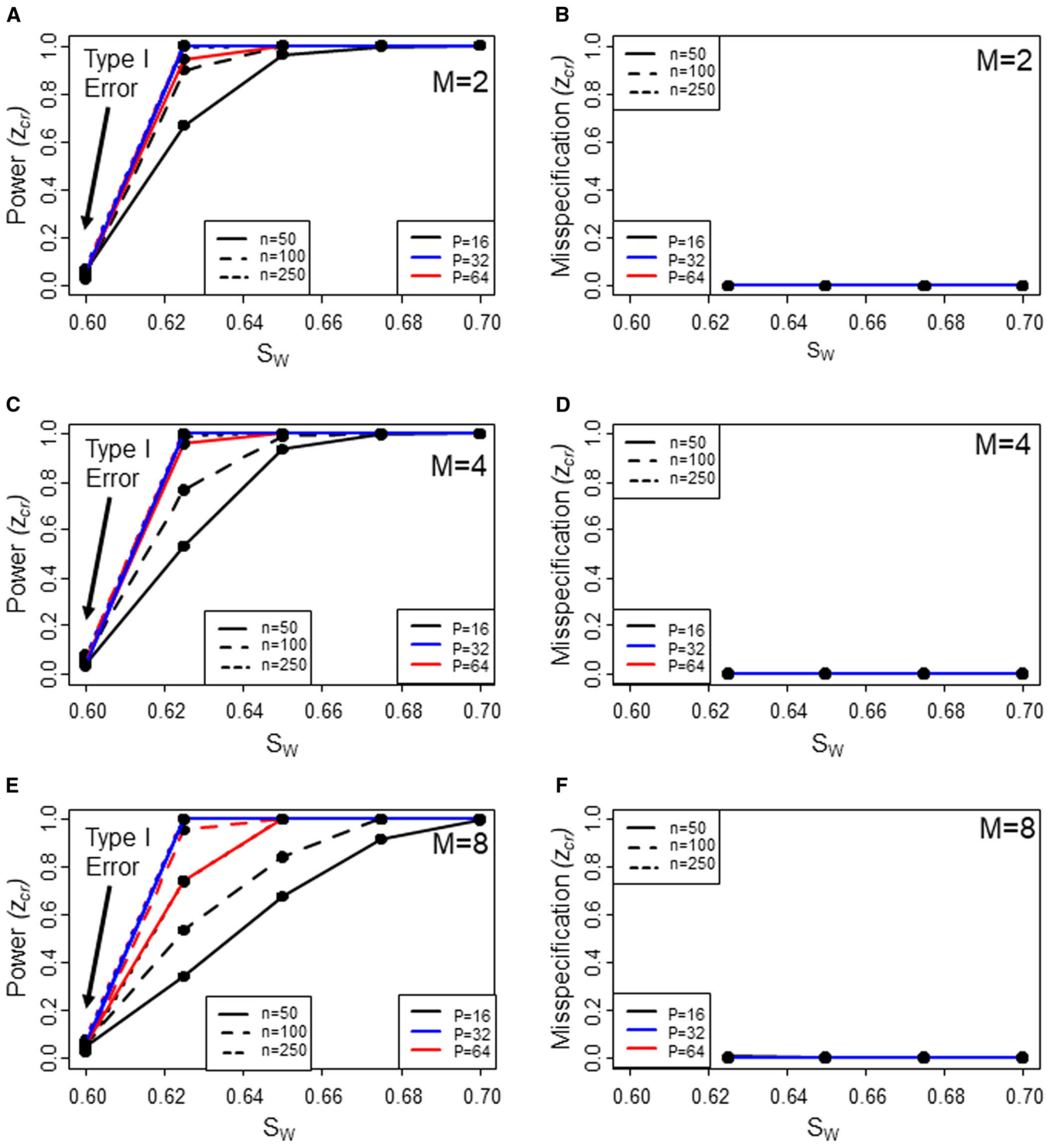


Figure 3. Results of simulations evaluating the type I error, statistical power, and model mis-specification rates of hypothesis testing procedures for comparing alternative modular hypotheses for the same dataset using Z_{CR}. Data were simulated using differing numbers of variables *(p)* and at different samples sizes (*n*). Between-module covaration was: $S_B = 0.6$, and within-module covariation is shown along the *X*-axis. Panels on the left display power curves and panels on the right display model mis-specification rates for differing input conditions: (A) and (B) two-module input signal, (C) and (D) four-module input signal, and (E) and (F) eight-module input signal. Note that when $S_W = S_B = 0.6$, simulations contain no modular signal (see text for details).

signal, and a low probability of mis-assigning the signal to an incorrect modular hypothesis. Overall, these results demonstrate that methods based on covariance ratio effect sizes (Z_{CR}) can identify modular signal when it is present, have high accuracy, and exhibit very low rates of model misspecification.

In stark contrast with $\hat{Z}_{12}$, we found that MINT displayed extremely high false positive rates that approached 100%, implying that when no modular signal was present, the method exhibited a propensity to identify patterns that do not exist in the data (Fig. 4A, 4C, 4E). Additionally, when modular signal was present, MINT displayed low power and high levels of model misspecification; even when modularity was observed the method tended often to mis-assign the signal to an alternative modular hypothesis (Fig. 4B, 4D, 4F). Part of the reason for this tendency is that MINT considers all possible combinations of modules across all input models as possible modular hypotheses to be examined, whether or not they are biologically plausible. While this 'unsupervised' combinatoric exploration has previously been considered a strength of the approach, our results reveal that instead they contribute to model misspecification, as biologically implausible modular hypotheses may be selected by MINT as representing the best hypothesis. Thus at present, it appears that reliance on prior biological knowledge to generate putative modular hypotheses is preferred. Overall we find that MINT displays poor statistical properties (Fig. 4), with very high false positive rates and high rates of model misspecification. For the purposes of comparing modular signal across hypotheses, we conclude that methods based on covariance ratio effect sizes (Z_{CR}) are preferred.

Likewise, we found that EMMLi displayed very high false positive rates that ranged from 40% - 90% when no modular signal was present, and exhibited variable power, regardless of simulation condition (Fig. 5A, 5C, 5E). Even when considering whether the null model was within $\Delta AIC < 4.0$ units of the best model, the false positive rates of EMMLi were still extremely high (20% - 80%). We also found that the method displayed high rates of model misspecification (Fig. 5B, 5D, 5F), implying that when modular signal was present, EMMLi frequently mis-assigns that signal to an incorrect modular hypothesis. EMMLi's performance did improve slightly as the number of modules increased, suggesting that the method was sensitive to how many data partitions were present. This finding is in accord with previous studies showing that EMMLi tends to prefer more complex models over simpler ones, and can possibly identify complex models when they are present (Goswami and Finarelli 2016; Goswami pers. comm.). Additionally, when high-dimensional datasets ($p>>N$) were examined, we found that EMMLi still displayed extremely high false positive rates and variable mis-specification rates, implying that under both $p < N$ and $p > N$ scenarios, these analytical challenges persist (see Supporting Information).

One possible reason for the very poor performance of EMMLi may be the fact that its calculation of AICc—based on the univariate method proposed by Hurvich and Tsai (1989), used by Goswami and Finarelli (2016)—does not consider the number of variables (p), but only the number of modules. For instance, the parameter penalty for a four-module hypothesis with separate trait correlations within each module but the same correlation structure between modules is $K = 6$, regardless of whether the dataset contains 16 or 160 variables. Such formulations are akin to a univariate view of AIC, whereas for multivariate data, minimally a multivariate parameter penalty should be utilized (see Bedrick and Tsai 1994). Ideally, the log-likelihood for EMMLi would also be based on a single matrix covariance statistic and the parameter penalty for AICc would consider both sample size and data dimensionality (Bedrick and Tsai 1994) rather than using log-likelihood summation over disparate numbers of variable correlations and parameter penalties based only on the number of modules (thus not providing enough penalty for multivariate data, and inherently favoring more complex modular hypotheses). It is possible that if EMMLi were to be updated to better evaluate log-likelihoods and the number of parameters—including covariance matrix estimation, related to the number of variables—the elevated false positive rates we observed would be mitigated. However, such formulations must also be capable of accommodating high-dimensional cases when $p>>N$, a scenario where likelihoods are typically unable to be computed due to covariance matrix singularity (see Adams 2014; Adams and Collyer 2018b). Overall our findings revealed that while EMMLi may be capable of detecting complex modular signal in datasets with very large sample sizes, a large number of variables, and very strong modular signal, such instances cannot be distinguished from false positives, because of the method's extremely high false positive rate. Therefore, we find that the EMMLi approach (based on the current philosophy of summing univariate log-likelihoods) displays poor statistical properties (Fig. 5), with very high false positive rates and high rates of model mis-specification. For the purposes of comparing modular signal across hypotheses, we conclude that methods based on covariance ratio effect sizes (Z_{CR}) are preferred.

EMPIRICAL EXAMPLE

Using Z_{CR}, we found that the model displaying the greatest degree of modular signal was the six-module model, a hypothesis where each of the six developmental units represented its own anatomical module (Fig. 6A). We recognize that this was the most complex modular hypothesis considered in this example. However, this result does not imply that Z_{CR} favors overparameterized models, because our simulations above demonstrated that Z_{CR} rejects false, more complex models, and correctly identifies simpler models when they generated the data (see Fig. 3). For

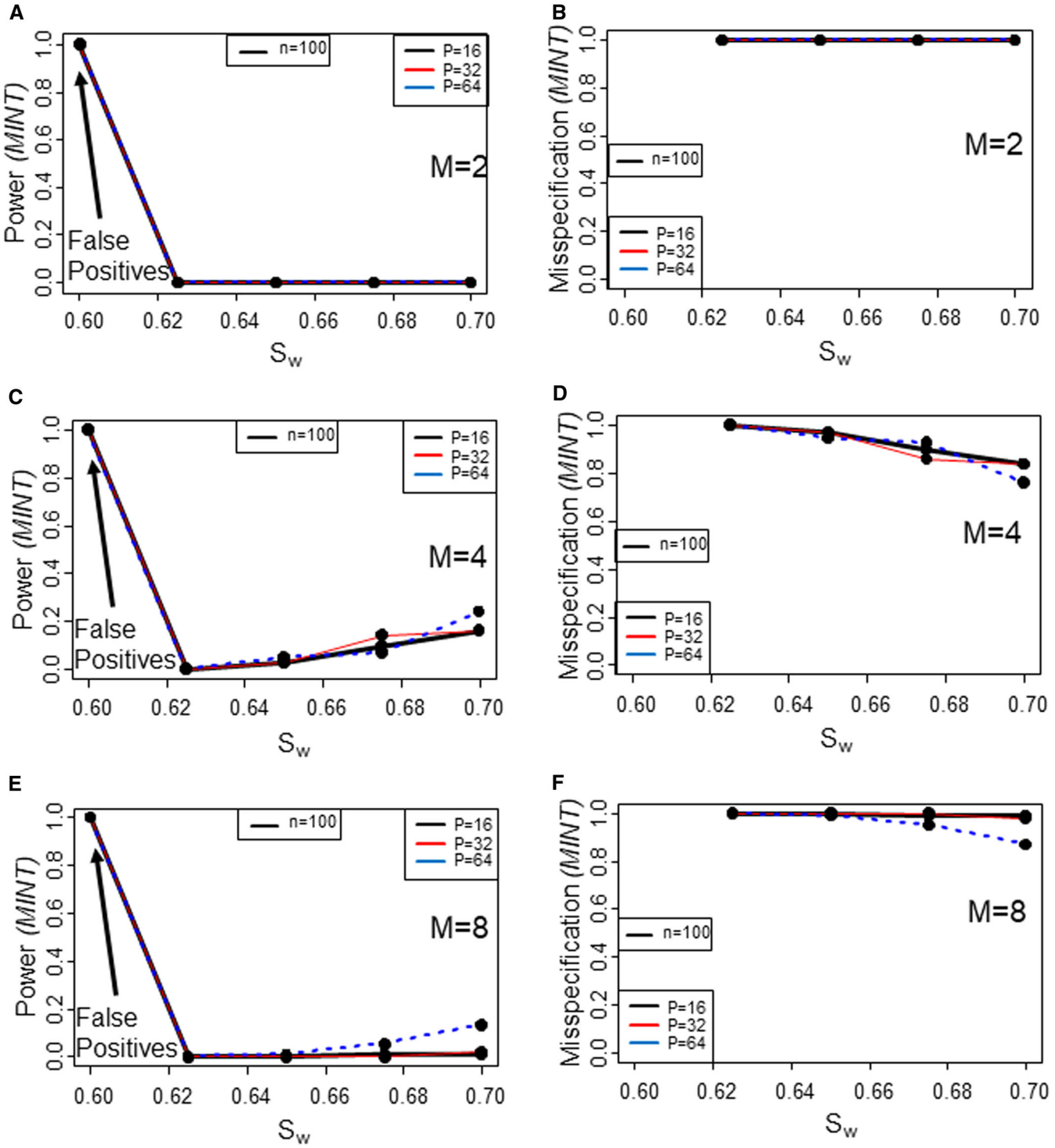


Figure 4. Results of simulations evaluating the false positive rates, statistical power, and model misspecification rates of hypothesis testing procedures for comparing alternative modular hypotheses for the same dataset using using MINT. Data were simulated using differing numbers of variables *(p)* and at different samples sizes (*n*). Within-module covariation S_W and between-module covariation S_B are defined as described in the text. For datasets containing no modular structure, the proportion of datasets where the best model was distinct from the null model was treated as the false positive rate, which was analogous to a type I error rate. For datasets containing modular signal, the proportion of datasets where the correct model was ranked as the best model (i.e., true positives) was treated as an estimate of power. Panels on the left display power curves and panels on the right display model misspecification rates for differing input conditions: (A) and (B) two-module input signal, (C) and (D) four-module input signal, and (E) and (F) eight-module input signal. Note that when $S_W = S_B = 0.6$, simulations contain no modular signal (see text for details).

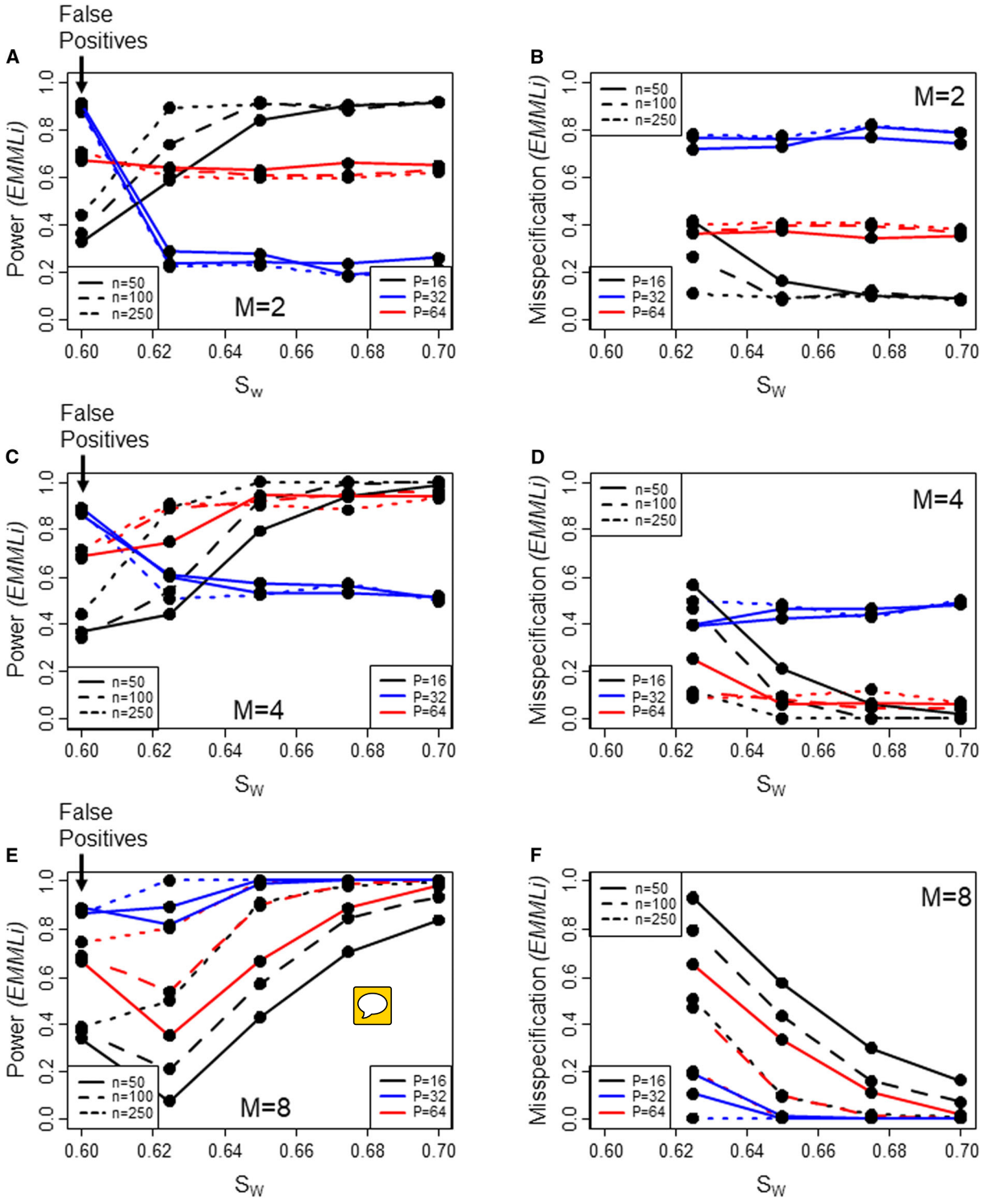


Figure 5. Results of simulations evaluating the false positive rates, statistical power, and model misspecification rates of hypothesis testing procedures for comparing alternative modular hypotheses for the same dataset using using `EMMLi`. Data were simulated using differing numbers of variables *(p)* and at different samples sizes (*n*). Within-module covariation S_W and between-module covariation S_B are defined as described in the text. For datasets containing no modular structure, the proportion of datasets where the best model was distinct from the null model was treated as the false positive rate, which was analogous to a type I error rate. For datasets containing modular signal, the proportion of datasets where the correct model was ranked as the best model (i.e., true positives) was treated as an estimate of power. Panels on the left display power curves and panels on the right display model mis-specification rates for differing input conditions: (A) and (B) two-module input signal, (C) and (D) four-module input signal, and (E) and (F) eight-module input signal. Note that when $S_W = S_B = 0.6$, simulations contain no modular signal (see text for details).

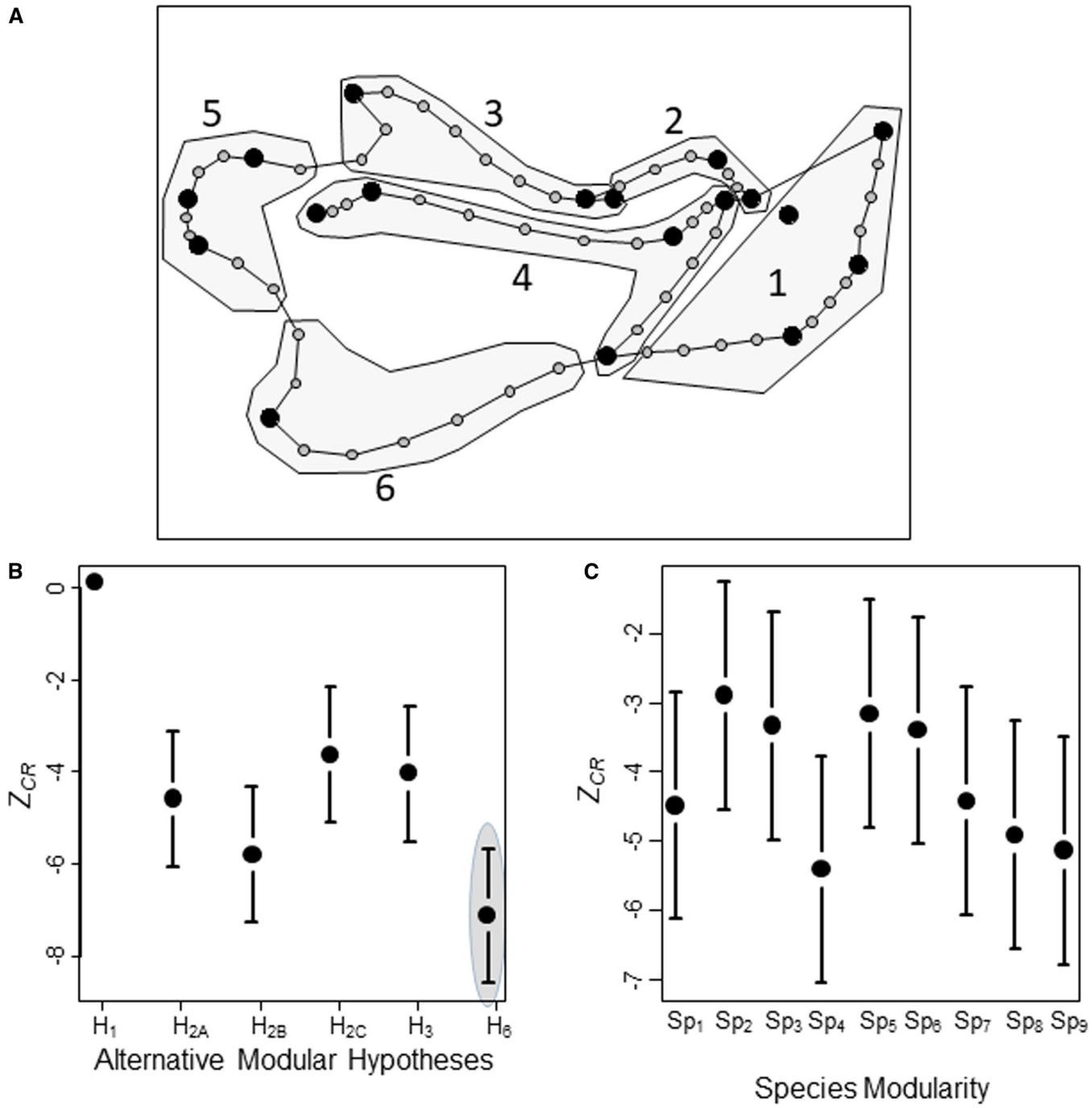


Figure 6. (A) Landmarks (black) and semilandmarks (gray) used in the empirical example. The six developmental units (sensu Hall 2003) are shown. Combinations of these six anatomical regions comprise the modules of the various alternative hypotheses. (B) Effect sizes (Z_{CR}) and their 95% confidence intervals for six alternative modular hypotheses commonly used in the study of modularity of the mouse mandible. Hypothesis labels follow the number of modules they comprise, and are described in Table 1. The optimal hypothesis is shaded, and represents the case where all six developmental units are considered a separate module. (C) Effect sizes (Z_{CR}) and their 95% confidence intervals for each of nine species of sigmodontine rodent. Species are: Sp_1 = *Holochilus chacarius*; Sp_2 = *Melanomys caliginosus*; Sp_3 = *Microryzomys minutus*; Sp_4 = *Nectomys squamipes*; Sp_5 = *Oligoryzomys nigripes*; Sp_6 = *Oryzomys couesi*; Sp_7 = *Oryzomys palustris*; Sp_8 = *Oryzomys xantheolus*; and Sp_9 = *Sigmodontomys alfari* (see text for details).

Table 2. **Pairwise modularity scores as described using the covariance ratio (*CR*), for the six-module hypothesis describing patterns of modularity in sigmodontine rodents. This represents the optimal modularity hypothesis as identified by Z_{CR}.**

	M1	M2	M3	M4	M5	M6
M1	–					
M2	0.504	–				
M3	0.434	0.465	–			
M4	0.649	0.364	0.390	–		
M5	0.461	0.232	0.513	0.471	–	
M6	0.449	0.275	0.390	0.418	0.571	–

the empirical example, overall patterns of modularity were quite strong, with an average modularity score considerably less than 1.0 ($CR_{\mu} = 0.504$), and all pairwise *CR* values ranging between 0.23 and 0.65 (Table 2). Thus, this hypothesis revealed strong modular signal across all six modules, rather than any subset. This hypothesis provided a significantly better fit to patterns of shape covariation as compared with most other models examined, including several commonly used two-module hypotheses (Fig. 6B). Interestingly, the three-module hypothesis found to represent a consensus hypothesis for this dataset (Márquez 2008) was not as well supported when using Z_{CR}. Finally, when modularity patterns were compared across species, we found that no species displayed greater levels of modularity than any other (Fig. 6C). Thus the strength of modular signal was consistent across species in this dataset.

Discussion

A major goal in evolutionary biology is to determine whether different taxa or traits display similar degrees of modularity in their phenotypic attributes. However, quantitative comparisons of the strength of modular signal across datasets, and comparisons of alternative modular hypotheses for the same dataset, have been limited by a lack of statistical tools to directly facilitate these comparisons. In this article, we described an effect size measure (Z_{CR}) derived for this purpose, and proposed a two-sample test statistic that provides an analytical solution to both of these challenges. We demonstrated that Z_{CR} is a valid standardized effect size (i.e., that it exhibits a constant expected value and sample variance under the null hypothesis; Fig. 1). Through computer simulations, we explored the method's statistical properties, and found that the approach displays appropriate type I error rates (5%), exhibits high statistical power (even for modest effects), and low levels of model misspecification, both when comparing effect sizes across datasets, and when evaluating patterns across alternative modular hypotheses. These properties remained consistent even for high-dimensional data (when the number of variables exceeds the number of specimens: $p>>N$), implying that tests based on Z_{CR} may be applied even in these circumstances. Thus, when modular signal is present in morphometric datasets, the method is capable of detecting it, and displays a low chance of mis-assigning that signal to an incorrect modular hypothesis.

On the other hand, our study revealed that several alternative approaches for evaluating modular signal (`MINT` and `EMMLi`) display poor statistical performance as currently implemented, which limits their utility. For instance, our broad set of simulations confirmed that `MINT` identifies modular patterns when none exist in the data, revealing very high false positive rates, a pattern that confirms earlier suspicions that false positives may be a concern (Márquez 2008: Table 3). We further identified that model misspecification, even when modular signal was present, is also an issue with this approach. Likewise, our simulations revealed that when no modular signal was present, `EMMLi` also suffers from extreme false positive rates, and in cases with modular signal displays a propensity for selecting more highly parameterized models. Indeed this latter finding had been previously identified by other authors (e.g., Goswami and Finarelli 2016; Goswami pers. comm.), which has necessitated a coupling of modularity analyses based on `EMMLi` with other approaches based on the CR coefficient to corroborate biological inferences (see e.g., Felice and Goswami 2018; Bardua et al. 2019). Taken together, our results revealed that both `MINT` and `EMMLi` exhibit poor statistical performance (very high false positive rates, and high model misspecification rates), making it challenging to arrive at reliable biological inferences when using these approaches.

When viewing our simulation results across approaches, our results therefore show that of the present alternatives, statistical tests based on Z_{CR} provide the most appropriate and rigorous means of evaluating whether the degree of modularity differs among datasets, and simultaneously, can be used to accurately distinguish among alternative modular hypotheses for the same dataset. In addition, we note the flexibility of the Z_{CR} approach to evaluate and compare patterns of modularity, even when those patterns differ across datasets. Specifically, because Z_{CR} is a standardized effect size, the strength of modular signal may be summarized via Z_{CR} and compared across datasets, even when the underlying modular hypothesis is not the same (e.g., covariation in species X is characterized by two modules, while covariation in species Y is characterized by three modules). To our knowledge, no other approach is capable of making this empirical comparison of the strength of modular signal across datasets in this manner. Furthermore, we note that log-likelihoods could also, in theory, be calculated from the Z_{CR} values and their distributional moments (Edwards 1994). In this case, likelihood ratio tests with inference from χ^2 distributions performed on these statistics will yield *P*-values similar to those obtained from the *Z*-tests introduced here. Likewise, one could use the log-likelihoods for information

criteria (like AIC) to compare modular hypotheses as an alternative means of comparing alternative modular hypotheses. Such implementations would provide a complement to the Z-tests introduced here.

The results presented here demonstrate that statistical tests of the strength of modularity using covariance ratio effect sizes (Z_{CR}) represent a substantial advance over currently available approaches. Nevertheless, we recognize that additional extensions and enhancements to the quantitative study of modularity are likely, particularly as morphometrics is a vibrant and active field (Mitteroecker and Gunz 2009; Adams et al. 2013). In particular, we acknowledge that there are ongoing discussions as to how morphometric data should be standardized, and what analytical pipeline should be used prior to the assessment of modular structure in morphometric datasets (Cardini 2019; but see Goswami et al. 2019). However, we note that such considerations do not fundamentally alter the effect size approach developed here: rather, they affect how one first processes the data prior to the quantification of modular signal via the *CR* and its effect size, Z_{CR}. Likewise, it is conceivable that future methods may be developed that characterize modularity while accounting for the spatial proximity among landmarks, or some other factor. To the extent that these enhancements alter estimation of the covariance ratio (or its empirical sampling distribution), the approach developed here likewise remains unaffected by those methodological enhancements.

In fact, even the development of an entirely new statistic that characterizes patterns of modularity in a different manner, or that envisages new ways of proposing modularity hypotheses for quantification, does not impose an impediment to use of comparisons of effect sizes via $\hat{Z}_{12}$, so long as those new modularity measures are accompanied by empirical sampling distributions obtained through resampling procedures. The reason is that the hypothesis testing framework proposed here simply represents a methodological extension of a two-sample Z-test, but where the expected values and their standard errors are found not through theoretical distributions, but from empirically derived sampling distributions, from resampling procedures (Collyer et al. 2015; see Adams and Collyer 2016 for a related discussion). Indeed, the biological question of "what is a module?," has long vexed evolutionary biologists, and remains a critical challenge in the study of modularity and the evolution of patterns of trait covariation (see Zelditch et al. 1990; Hallgrimsson et al. 2007). Thus, as evolutionary biologists' notions of what constitutes a module continue to be refined, it is to be expected that methods for characterizing such patterns will follow suit. It is important to note however, that while new analytical methods may be developed for the characterization of modular signal, the statistical framework proposed herein for comparing modular signals across datasets, or between modular partitions of the same structure, remains fundamentally unaltered.

Thus, we conclude that the comparison of effect sizes derived from quantitative measures of modularity, combined with empirical sampling distributions of the strength of their signal, provides the best current means of comparing modular signal across datasets, and between alternative modular hypotheses for the same data.

AUTHOR CONTRIBUTIONS

D.C.A. and M.L.C. contributed equally to all parts of this manuscript.

ACKNOWLEDGMENTS

We thank A. Kaliontzopoulou for comments on drafts of this manuscript. E. Márquez kindly provided the data for the empirical example. This work was sponsored in part by National Science Foundation Grants DEB-1556379 and DBI-1902511 (to DCA) and DEB-1737895 and DBI-1902694 (to MLC).

DATA ARCHIVING

The doi for our data is 10.5061/dryad.h18931zfs20.

LITERATURE CITED

Adams, D. C. 2014. A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. Evolution 68:2675–2688.

Adams, D. C. 2016. Evaluating modularity in morphometric data: challenges with the rv coefficient and a new test measure. Methods Ecol. Evol. 7:565–572.

Adams, D. C., and M. Collyer. 2018a. Phylogenetic anova: group-clade aggregation, biological challenges, and a refined permutation procedure. Evolution 72:1204–1215.

Adams, D. C., and M. L. Collyer. 2018b. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. Syst. Biol. 67:14–31.

Adams, D. C., and M. L. Collyer. 2016. On the comparison of the strength of morphological integration across morphometric datasets. Evolution 70:2623–2631.

Adams, D. C., and R. Felice. 2014. Assessing trait covariation and morphological integration on phylogenies using evolutionary covariance matrices. PLoS ONE 9:e94335.

Adams, D., M. Collyer, and A. Kaliontzopoulou. 2019. Geomorph: software for geometric morphometric analyses. R package version 3.1.1. R Foundation for Statistical Computing, Vienna, Austria.

Adams, D. C., and E. Otárola-Castillo. 2013. Geomorph: an r package for the collection and analysis of geometric morphometric shape data. Methods Ecol. Evol. 4:393–399.

Adams, D. C., F. J. Rohlf, and D. E. Slice. 2013. A field comes of age: geometric morphometrics in the 21st century. Hystrix 24:7–14.

Atchley, W. R., and B. K. Hall. 1991. A model for development and evolution of complex morphological structures. Biol. Rev. 66:101–157.

Bardua, C., M. Wilkinson, D. J. Gower, E. Sherratt, and A. Goswami. 2019. Morphological evolution and modularity of the caecilian skull. BMC Evol. Biol. 19:30.

Bedrick, E. J., and C. Tsai. 1994. Model selection for multivariate regression in small samples. Biometrics 50:226–231.

Cardini, A. 2019. Integration and modularity in procrustes shape data: is there a risk of spurious results? Evol. Biol. 46:90–105.

Cheverud, J. M. 1996. Developmental integration and the evolution of pleiotropy. Am. Zool. 36:44–50.

Cheverud, J. M. 1982. Phenotypic, genetic, and environmental morphological integration in the cranium. Evolution 36:499–516.

Cheverud, J. M., E. J. Routman, and D. J. Irschick. 1997. Pleiotropic effects of individual gene loci on mandibular morphology. Evolution 51:2006–2016.

Clune, J., J. B. Mouret, and H. Lipson. 2013. The evolutionary origins of modularity. Proc. R. Soc. B Biol. Sci. 280:20122863.

Collyer, M. L., D. J. Sekora, and D. C. Adams. 2015. A method for analysis of phenotypic change for phenotypes described by high-dimensional data. Heredity 115:357–365.

Drake, A. G., and C. P. Klingenberg. 2010. Large scale diversification of skull shape in domestic dogs: disparity and modularity. Am. Nat. 175:289–301.

Edwards, A. W. 1994. Likelihood. Expanded edition. Johns Hopkins Univ. Press, Baltimore, MA.

Felice, R. N., and A. Goswami. 2018. Developmental origins of mosaic evolution in the avian cranium. Proc. Natl. Acad. Sci. U.S.A. 115:555–560.

Garland, T. J., and A. R. Ives. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346–364.

Goswami, A. 2006. Cranial modularity shifts during mammalian evolution. Am. Nat. 168:270–280.

Goswami, A., and J. A. Finarelli. 2016. `EMMLi`: a maximum likelihood approach to the analysis of modularity. Evolution 70:1622–1637.

Goswami, A., and P. D. Polly. 2010. Methods for studying morphological integration and modularity. Pp. 213–243 *in* J. Alroy and G. Hunt, eds. Quantitative methods in paleobiology. Paleontological Society, Boulder, CO.

Goswami, A., A. Watanabe, R. N. Felice, C. Bardua, and D. Fabre A. -C. Polly. 2019. High-density morphometric analysis of shape and integration: The good, the bad, and the not-really-a-problem. Integr. Comp. Biol. 59:669–683.

Hall, B. K. 2003. Unlocking the black box between genotype and phenotype: Cell condensations as morphogenetic (modular) units. Biol. Philos. 18:219–247.

Hallgrimsson, B., D. E. Lieberman, N. M. Young, T. Parsons, and S. Wat. 2007. Evolution of covariance in the mammalian skull. Pp. 164–190 *in* G. R. Bock and J. A. Goode, eds. Tinkering: the microevolution of development. John Wiley & Sons, Ltd., Hoboken, NJ.

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

Hurvich, C. M., and C. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Klingenberg, C. P. 2008. Morphological integration and developmental modularity. Annu. Rev. Ecol. Evol. Syst. 39:115–132.

Klingenberg, C. P. 2009. Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. Evol. Dev. 11:405–421.

Klingenberg, C. P., K. Mebus, and J. C. Auffray. 2003. Developmental integration in a complex morphological structure: How distinct are the modules in the mouse mandible? Evol. Dev. 5:522–531.

Larouche, O., M. L. Zelditch, and R. Cloutier. 2018. Modularity promotes morphological divergence in ray-finned fishes. Sci. Rep. 8:7278.

Magwene, P. M. 2001. New tools for studying integration and modularity. Evolution 55:1734–1745.

Marshall, A. F., C. Bardua, D. J. Gower, M. Wilkinson, E. Sherratt, and A. Goswami. 2019. High-density three-dimensional morphometric analyses support conserved static (intraspecific) modularity in caecilian (amphibia: Gymnophiona) crania. Biol. J. Linn. Soc. 126:721–742.

J Márquez, E. 2008. A statistical framework for testing modularity in multidimensional data. Evolution 62:2688–2708.

Melo, D., G. Garcia, A. Hubbe, A. P. Assis, and G. Marroig. 2016. EvolQG - an r package for evolutionary quantitative genetics. F1000Research 4:1–9.

Mitteroecker, P., and F. L. Bookstein. 2007. The conceptual and statistical relationship between modularity and morphological integration. Syst. Biol. 56:818–836.

Mitteroecker, P., and P. Gunz. 2009. Advances in geometric morphometrics. Evol. Biol. 36:235–247.

Olson, E. C., and R. L. Miller. 1958. Morphological integration. Univ. of Chicago Press, Chicago.

Parr, W. C. H., L. A. B. Wilson, S. Wroe, N. J. Colman, M. S. Crowther, and M. Letnic. 2016. Cranial shape and the modularity of hybridization in dingoes and dogs: hybridization does not spell the end for native morphology. Evol. Biol. 43:171–187.

Parsons, K. J., E. Márquez, and R. C. Albertson. 2012. Constraint and opportunity: The genetic basis and evolution of modularity in the cichlid mandible. Am. Nat. 179:64–78.

R Core Team. 2019. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Renaud, S., P. Alibert, and J. C. Auffray. 2012. Modularity as a source of new morphological variation in the mandible of hybrid mice. BMC Evol. Biol. 12:1–16.

Sanger, T. J., D. L. Mahler, A. Abzhanov, and J. B. Losos. 2012. Roles for modularity and constraint in the evolution of cranial diversity among *Anolis* lizards. Evolution 66:1525–1542.

Tokita, M., T. Kiyoshi, and K. N. Armstrong. 2007. Evolution of craniofacial novelty in parrots through developmental modularity and heterochrony. Evol. Dev. 9:590–601.

Wagner, G. P. 1984. Coevolution of functionally constrained characters: Prerequisites of adaptive versatility. BioSystems 17:51–55.

Wagner, G. P., and L. Altenberg. 1996. Complex adaptations and the evolution of evolvability. Evolution 50:967–976.

Wagner, G. P., M. Pavlicev, and J. M. Cheverud. 2007. The road to modularity. Nat. Rev. Genet. 8:921–931.

Zelditch, M. L., D. O. Straney, D. L. Swiderski, and A. C. Carmichael. 1990. Variation in developmental constraints in *Sigmodon*. Evolution 44:1738–1747.

Associate Editor: G. Slater
Handling Editor: M. R. Servedio

Supporting Information

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

Figure S1. Results of additional simulations evaluating the type I error and statistical power of Z_{CR} for data simulated using $p = (180, 360)$ variables and $n = (50, 100)$.

Figure S2. Results of additional simulations evaluating the type I error and statistical power of `EMMLi` for data simulated using $p = (180, 360)$ variables and $n = (50, 100)$.

Figure S3. Results of additional simulations evaluating the statistical power and model misspecification rates of Z_{CR} for data simulated under increasingly uneven levels of covariation between pairs of modules.