

# EMMLi: A maximum likelihood approach to the analysis of modularity

Journal:	Evolution
Manuscript ID	Draft
Manuscript Type:	Brief Communication
Date Submitted by the Author:	n/a
Complete List of Authors:	Goswami, Anjali; University College London, Genetics, Evolution & Environment Finarelli, John; University College Dublin, School of Biology & Environmental Science
ия Саларана и солотории и сол Саларана и солотории и соло	Morphological Evolution, phenotypic integration, trait correlations, mammals, model selection



- 1 EMMLi: A maximum likelihood approach to the analysis of modularity
- 2 Running Header: Maximum likelihood analysis of modularity
- 4 Anjali Goswami<sup>1</sup> and John A. Finarelli<sup>2,3</sup>
- <sup>1</sup>Department of Genetics, Evolution & Environment and Department of Earth Sciences,
- 7 University College London, London, WC1E 6BT, UK
- <sup>8</sup> <sup>2</sup>School of Biology & Environment Science, University College Dublin, Science Centre West,
- 9 Belfield, Dublin 4, Ireland.
- <sup>3</sup>UCD Earth Institute, University of College Dublin, Belfield, Dublin 4, Ireland.
- 11 Emails: a.goswami@ucl.ac.uk; john.finarelli@ucd.ie
- 12
- 13 Keywords: phenotypic integration, trait correlations, mammals, model selection
- 14

#### 15 ABSTRACT

Identification of phenotypic modules, semi-autonomous sets of highly-correlated traits, can be 16 17 accomplished through exploratory (e.g., cluster analysis) or confirmatory approaches (e.g., RV coefficient analysis). While statistically more robust, confirmatory approaches are generally 18 unable to compare across different model structures. For example, RV coefficient analysis finds 19 20 support for both two- and six-module models for the therian mammalian skull. Here, we present a maximum likelihood approach that takes into account model parameterization. We compare 21 22 model log-likelihoods of trait correlation matrices using the finite-sample corrected Akaike Information Criterion, allowing for comparison of hypotheses across different model structures. 23 Simulations varying model complexity and within- and between-module contrast demonstrate 24 that this method correctly identifies model structure and parameters across a wide range of 25 conditions. We further analyzed a dataset of 3-D data, consisting of 61 landmarks from 181 26 macaque (*Macaca fuscata*) skulls, distributed among five age categories, testing 31 models, 27 28 including no modularity among the landmarks, and various partitions of 2, 3, 6, and 8 modules. Our results clearly support a complex six-module model, with separate within- and inter-module 29 30 correlations. Furthermore, this model was selected for all five age categories, demonstrating that 31 this complex pattern of integration in the macaque skull appears early and is highly conserved 32 throughout postnatal ontogeny. Subsampling analyses demonstrate that this method is robust to 33 relatively low sample sizes, as is commonly encountered in rare or extinct taxa. This new 34 approach allows for the direct comparison of models with different parameterizations, providing an important tool for the analysis of modularity across diverse systems. 35

36

### **37 INTRODUCTION**

The related topics of phenotypic integration and modularity, which concern associations among 38 traits and their partitioning into semi-autonomous and highly-correlated subsets, respectively, 39 have received increased attention over the past few decades as a powerful bridge among different 40 scales of evolutionary analysis. Recent years have seen increasing effort to identify and compare 41 phenotypic modularity and integration across taxa, in some cases spanning entire vertebrate 42 'classes' (Goswami 2006b, a; Goswami 2007; Porto et al. 2009; Bell et al. 2011; Bennett and 43 Goswami 2011; Klingenberg and Marugan-Lobon 2013), and even comparing plants and animals 44 (Conner et al. 2014). There has also been a refining of different levels of modularity acting at 45 different scales. The most typically-studied level, termed "variational" (Marquez 2008) or 46 "static" (Klingenberg 2014) modularity, focuses on a single species or population, commonly at 47 a specific ontogenetic stage (e.g., adults). Within this level, analyses focus on identifying drivers 48 of trait integration, whether functional, developmental, genetic, or environmental. Beyond 49 variational modularity, studies have analyzed modularity at the ontogenetic scale (that is, 50 patterns or changes in modularity through ontogeny within a species), and evolutionary 51 modularity (comparative analysis of patterns of modularity across taxa). Coincident with this 52 53 increase in studies of modularity, there has been an explosion in the number of methods 54 proposed to analyze phenotypic modularity and integration, both within and across populations (Klingenberg 2009; Goswami and Polly 2010; Klingenberg 2013; Adams and Felice 2014; 55 56 Bookstein and Mitteroecker 2014; Klingenberg 2014).

Page 4 of 46

58 Analyses of modularity have taken many forms, from entirely exploratory approaches, such as cluster analysis, Euclidean distance matrix analysis, and graphical modelling, to confirmatory 59 approaches, such as partial least squares analysis and the related RV coefficient analysis, 60 integration matrices, and theoretical matrix modelling (reviewed in Klingenberg 2009; Goswami 61 and Polly 2010; Klingenberg 2013, 2014), and there has been a vigorous discussion of the merits, 62 practical considerations, and issues of each approach (Klingenberg 2008; Goswami and Polly 63 2010; Fruciano et al. 2013; Adams and Felice 2014). Not surprisingly, confirmatory methods are 64 generally viewed as more robust, particularly as exploratory methods such as cluster analysis 65 impose hierarchical relationships on traits that may or may not reflect their true biological 66 organization. On the other hand, exploratory approaches have the benefit of not requiring a 67 priori determination of model structure, whereas confirmatory methods depend on a defined 68 69 model structure and are therefore limited to testing pre-selected models. Given the complexity of many biological structures, and the diverse factors that may influence trait relationships 70 (Hallgrimsson et al. 2009), this limitation argues for the continued role of exploratory 71 72 approaches, particularly as studies expand beyond well-established model systems. Recent work has developed relative eigenanalysis for the purpose of comparing two covariance matrices in a 73 more informative manner than do previous methods, such as eigenvalue dispersion or random 74 skewers analysis (Bookstein and Mitteroecker 2014), providing an efficient exploratory approach 75 that can detail the specific ways that high-dimensional covariance matrices differ by identifying 76 the maximal ratios of variance between any two groups. However, this approach does not 77 directly address the problem of describing the pattern of integration for a group, which remains 78 an outstanding issue in this field. 79

81 Another important issue with most current confirmatory approaches is that they are designed to measure support for alternative hypothesized parameter values within a proposed model structure 82 (Wagner 2000). For example, RV coefficient analysis determines the correlations among sets of 83 traits, and then randomizes trait associations to produce an empirical distribution of RV 84 coefficients for the model structure under consideration, testing the hypothesis that the observed 85 RV coefficient is significantly lower than randomized alternatives. But while this methodology 86 can test if a particular model is more structured than random, it does not readily address the 87 question of whether a four-module model describes the pattern of phenotypic integration better 88 than arrangements with three or five modules. The same is true of the recently described 89 Covariance Ratio metric (Adams 2016), which improves upon several statistical issues with RV 90 coefficient analysis, but also can only test one model of modularity against a hypothesis of 91 random associations of traits. Thus far, only one published method allows for comparisons of 92 models with different complexities (Marguez 2008), as demonstrated with a 2-D landmark 93 dataset for rodent mandibles. This method included several innovations that allowed for testing 94 of hundreds of alternative models, including those with overlapping landmarks, but the most 95 relevant is the correction of similarity among the observed and modeled covariance matrices 96 against the number of estimated parameters. This addition facilitates comparison across models 97 with varying structures of different complexity. While this represented an important step in 98 confirmatory tests of modularity, the author noted that a linear correction for the number of 99 estimated parameters may not be appropriate for all test statistics or for more complex 100 approaches (Marquez 2008). Additionally, this method has also never been expanded to 3-D 101 data. 102

Here, we describe a new method for the analysis of phenotypic modularity from trait correlation 104 matrices based on a maximum likelihood approach. We provide a case study applying this 105 approach to a dataset of macaque skulls spanning infant to adult age groups. We use this method 106 107 to compare various models that have been proposed for mammalian skull modularity (including no modularity, a two-module neurocranial/facial hypothesis, and multiple six-module 108 hypotheses; Fig. 1), as well as novel alternative models of varying structure and complexity. 109

110

#### EMMLi: Evaluating Modularity with Maximum Likelihood 111

Model selection approaches using information theory compare likelihood fits across a set of 112 models of varying degree of complexity. In order to estimate likelihoods of models of trait 113 integrations, we first model the expected distribution around a hypothesized value representing 114 the relationship among a set of traits. For the product moment correlation coefficient, and its 115 derivatives including the congruence coefficient and canonical correlation (Goswami and Polly 116 2010), a simple transformation is available in the Fisher r-to-z transformation: 117

118

118  
119 Eq. 1) 
$$\mathbf{z}_r = tanh^{-1}(r) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$$
 (Sokal and Rohlf 1995, pg. 575),

120

where r is the sample correlation coefficient. Here the observed correlation matrix is treated as a 121 122 set of realizations (the values of r) of a hypothesized true correlation coefficient ( $\rho$ ). The

distribution around a hypothesized value of p is approximately normally distributed with 123 124 parameters: 125

126 Eq. 2a) 
$$\mu_{\rho} = \mathbf{z}_{\rho} = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right)$$
 and  $\mathbf{z}_{\rho}$ 

127 Eq. 2b) 
$$\sigma_{\rho}^2 = \left(\frac{1}{\sqrt{n-3}}\right)^2 = \frac{1}{n-3}$$
 (Sokal and Rohlf 1995, page 575),

128

- where n is the sample size used to calculate the correlation coefficient (i.e., the number of 129
- specimens with measured landmarks). The log-likelihood support for a hypothesized value of p, 130
- given an observed value of r, is then: 131

132

131 given an observed value of *r*, is then:  
132  
133 Eq. 3) 
$$LogL \propto -\frac{1}{2}Ln(\sigma_{\rho}^{2}) - \frac{(z_{r}-\mu_{\rho})^{2}}{2\sigma_{\rho}^{2}}$$
 (Edwards 1992).

134

Applying Equation 3 to a matrix of trait correlations, the simplest model structure (no 135

modularity) proposes a single value for the correlation coefficient between all possible trait pairs. 136

- The value that maximizes the summed log-likelihood for all observed correlations in the matrix 137
- would then be the preferred hypothesis, and this log-likelihood would then be the model log-138
- likelihood for the "no modularity" model structure. 139

141 However, given the results of a large number of previous studies (Cheverud 1982, 1989, 1995a,

142 1996; Ackermann and Cheverud 2000; Marroig and Cheverud 2001; Hallgrimsson et al. 2004;

143 Goswami 2006a; Hallgrimsson et al. 2009; Porto et al. 2009; Goswami and Polly 2010;

144 Klingenberg 2013), it is highly likely that a model structure positing a single value of ρ for the

145 entire correlation matrix would not adequately describe trait correlations in a real biological

system. Model structures of varying complexity can be compared using the Akaike Information

147 Criterion (AIC) (Akaike 1973; Burnham and Anderson 2002), assessing the likelihood fit of the

148 models, while controlling for better fit induced by increased model complexity. The finite-

149 sample AIC (AIC<sub>c</sub>) is given by:

150

151 Eq. 4) 
$$AIC_c = -2LogL + 2K + \frac{2K(K+1)}{N-K-1}$$
 (Hurvich and Tsai 1989).

152

In Equation 4, N is the sample size, but in the case of computing AIC<sub>c</sub>, this is the number of 153 between-trait correlations used to calculate the likelihood score. K is the number of estimated 154 parameters, which is the number of distinct, optimal correlations estimated by the model, and an 155 additional parameter for each estimate of the variance around the hypothetical value of p (see: 156 Equation 2b). In the present analysis, this is fixed for all of the examined models within each 157 data set (a single variance was calculated for each data set based on its sample size), and the 158 number of parameters is simply the number of estimated values of p incremented by one for all 159 models. However, this does not need to be the case, as more complex analyses may wish to 160 161 consider whether patterns of modularity are common across multiple data sets which may have

different estimates of variance. In such cases, different variances may be included as estimatedparameters among different models.

164

165	To illustrate the designation of model parameters more clearly, consider a set of landmarks
166	across a mammal cranium (Fig. 2A). Previous study of the mammal skull has proposed six
167	modules for this system (Cheverud 1982; Goswami 2006a). It is possible that that the
168	magnitudes of within-module correlations are effectively the same in all of the modules (Fig. 2B)
169	or that each of these modules has distinct strengths of correlation between landmarks within a
170	given module (Fig. 2C). Furthermore, inter-module correlations could also be distinct for each
171	module-to-module set (Fig. 2E and G), or they could be effectively identical (Fig. 2D and F).
172	These variations then returns four potential model structures with 3, 17, 8 or 22 estimated
173	parameters (the number of estimated $\rho$ 's in each, plus 1 for the estimated variance). Summing the
174	log-likelihoods from Equation 3 for the set of observed correlations within each modeled set for
175	an optimal estimate of $\rho$ , gives the model log-likelihood. These can be compared to one another,
176	to the "no modularity" hypothesis, and to different proposed structures or different groupings of
177	the landmarks within modules using Equation 4. From the model AIC <sub>c</sub> scores, we calculate
178	$\Delta AIC_c$ , the difference between a particular model's $AIC_c$ score and the lowest score observed
179	among the tested models. From this, we calculate the model log-likelihood adjusting for the
180	penalty due to parameterization:

181

182 Eq. 5) *Model LogL*  $\propto -\frac{1}{2}\Delta AIC_c$  (Burnham and Anderson 2002).

Page 10 of 46

183

A set of model posterior probabilities can then be calculated by dividing each model's likelihood by the sum of likelihoods over the set of examined models (N.B. these are likelihoods, and are therefore equal to e<sup>*Model LogL*</sup> (see: Burnham and Anderson 2004)).

187

188 A Note on Sample Size

A value of "n" or sample size appears in both the equations for calculating the variance around an estimated value of  $\rho$  (Equation 2b) and for the calculation of the AIC statistic (Equation 5). We have used upper- and lowercase to distinguish between the two, as *n* for calculation of correlations is based on the number of specimens, whereas, in the case of computing AIC<sub>c</sub>, *N* is the number of between-trait correlations considered in calculating the log-likelihood. For a 61 landmark data matrix, there are 1830 unique between-landmark correlations (i.e., the subdiagonal values of the matrix).

196

### 197 *A note on the use of the Fisher Transformation*

The Fisher r-to-z Transformation converts the bounded correlation coefficient to an unbounded
variable. Comparison of the transformed correlation to a hypothetical population value of ρ
demonstrates that the transformed coefficient is approximately normally distributed about ρ,
making the Fisher Transformation attractive for hypothesis testing. In the case of the correlation
matrix, however, there is a concern about the independence of the sample of correlation

203 coefficients, in that, for example, elements  $r_{12}$  and  $r_{13}$  are not strictly random *iid* draws from a 204 population, but are themselves intercorrelated. However, the Fisher-transformed correlations within a correlation matrix have been shown to be asymptotically, multivariate normal in 205 206 distribution, and robust to the violations of independence (Steiger 1980b; De Leeuw 1983). Specifically, this has been demonstrated for pattern hypotheses within correlation matrices, 207 wherein observed correlation coefficients are tested against a proposed "pattern matrix" (Steiger 208 1980a), and this approach, which is adopted here in the form of the proposed within- and among-209 module correlation estimates, has been applied in a wide range of research questions (Feldman et 210 al. 2007; Wager et al. 2007; LeBel and Gawronski 2009). As such the employing Fisher-211 transformed correlations in a likelihood framework, as proposed here, should prove a reliable 212 approach to evaluating modularity with trait correlation matrices. 213

214

#### 215 SIMULATIONS

Given the above noted concern with respect to independence of the Fisher-transformed 216 correlation coefficients, we evaluated the ability of the maximum likelihood approach as 217 218 implemented in EMMLi to correctly select a known model when choosing among models structures. To do so, we conducted an extensive series of simulations testing a range of model 219 220 structures, contrasting two variables: model complexity (number of parameters) and contrast (difference between within-module and between-module strength of integration). In all cases, 60 221 "landmarks" were simulated as divided into zero, two or six modules, to represent a hypothetical 222 correlation structure that we wish to evaluate. Between-module correlations were set at a mean 223 value of 0.1 for all simulations. Standard deviations for generating correlations were varied from 224

a low value of  $\sigma = 0.01$  to realistic value of  $\sigma = 0.05$  (e.g., Cheverud 1982), encompassing values used in simulations testing other recently described methods for the analysis of modularity (Adams 2016).

228

Simulating datasets without any modular structure allowed for assessment of Type I error rates. 100 permutations each were run with the mean correlations among all traits simulated as r =0.15, 0.3, 0.5, 0.7, or 0.9, with  $\sigma = 0.01$  or 0.05, for a total of 1000 simulations. In these cases, the correct model would be equivalent in structure to model 1 (K=2) in Table 1.

233

For the two and six module structures, both simple and complex models were tested. The simple models involved two or six modules which all had the same within-module correlations, set to five mean values ranging from r = 0.15 in the lowest contrast model to r = 0.9 in the highest contrast model (i.e., mean within-module r = 0.15, 0.3, 0.5, 0.7, and 0.9 were all simulated).

238

For the complex models, all two or six modules had different within-module correlations. In the high contrast, complex two-module model, these values were set to mean within-module r = 0.7and 0.9; in the mix contrast model, mean within-module r = 0.3 and 0.8; and in the low contrast case, mean within-module r = 0.15 and 0.3. In the high contrast, complex six-module model, mean within-module r = 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95; in the mix contrast case, mean withinmodule r = 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8; and in the low contrast case, mean within-module r =0.15, 0.2, 0.25, 0.3, 0.35, and 0.4. For the simple two-module structure, the correct model would

246	be equivalent in structure to model 2 (K=3) in Table 1, and the complex structure would be
247	equivalent to model 3 (K=4). For the simple six-module structure, the correct model would be
248	equivalent in structure to model 4 or 8 (K=3) in Table 1, and the complex structure would be
249	equivalent to model 5 or 9 (K=8). 100 permutations each of these 16 models were run, using
250	each of the standard deviation levels, resulting in 3200 total simulations of these modular
251	structures.
252	
253	
254	CASE STUDY: MAXIMUM LIKELIHOOD ANALYSIS OF MACAQUE CRANIAL
255	MODULARITY
255 256	MODULARITY Materials
255 256 257	MODULARITY Materials We use a data set of 3-D coordinates for 61 landmarks taken on the cranium of Japanese
255 256 257 258	MODULARITY         Materials         We use a data set of 3-D coordinates for 61 landmarks taken on the cranium of Japanese         macaque (Macaca fuscata) from the Primate Research Institute at Inuyama, Japan, previously
255 256 257 258 259	MODULARITY         Materials         We use a data set of 3-D coordinates for 61 landmarks taken on the cranium of Japanese         macaque (Macaca fuscata) from the Primate Research Institute at Inuyama, Japan, previously         described in (Goswami and Polly 2010) (see Supporting Information). Individuals were divided
255 256 257 258 259 260	MODULARITY         Materials         We use a data set of 3-D coordinates for 61 landmarks taken on the cranium of Japanese         macaque (Macaca fuscata) from the Primate Research Institute at Inuyama, Japan, previously         described in (Goswami and Polly 2010) (see Supporting Information). Individuals were divided         into five datasets representing four age classes: infants with deciduous dentition only (n = 42),
255 256 257 258 259 260 261	MODULARITY <i>Materials</i> We use a data set of 3-D coordinates for 61 landmarks taken on the cranium of Japanese macaque ( <i>Macaca fuscata</i> ) from the Primate Research Institute at Inuyama, Japan, previously described in (Goswami and Polly 2010) (see Supporting Information). Individuals were divided into five datasets representing four age classes: infants with deciduous dentition only (n = 42), juveniles with M1 erupted (n = 42), sub-adult with M2 erupted (n = 48), and adults with the
255 256 257 258 259 260 261 262	MODULARITYMaterialsWe use a data set of 3-D coordinates for 61 landmarks taken on the cranium of Japanesemacaque (Macaca fuscata) from the Primate Research Institute at Inuyama, Japan, previouslydescribed in (Goswami and Polly 2010) (see Supporting Information). Individuals were dividedinto five datasets representing four age classes: infants with deciduous dentition only (n = 42),juveniles with M1 erupted (n = 42), sub-adult with M2 erupted (n = 48), and adults with theentire adult dentition, further divided into male and female data partitions (n <sub>m</sub> = 25, n <sub>f</sub> = 24). See

265 The landmark data were superimposed with Generalized Procrustes superimposition to remove 266 the effects of rotation, translation and size (scaling all specimens to unit centroid size). All five datasets were analyzed separately. We calculated vector congruence coefficient correlation 267 268 matrices, producing 61x61 element matrices. This vector-based approach allows for simultaneous analysis of all three coordinates representing a single landmark (Goswami 2006a; 269 Goswami and Polly 2010). There has been some debate about the use of vector-based versus 270 coordinate-based correlations in studies of phenotypic integration and modularity (Klingenberg 271 2008; Goswami and Polly 2010; Klingenberg 2013). Here, we use the vector-based matrices, as 272 we feel these better reflects biological relationships, treating each landmark as a single unit of 273 information. However, we also include an example using the correlation matrix for individual 274 coordinates for the M1-erupted data set (see Supporting Information). This is a 183x183 matrix 275 (x-, y- and z-coordinates for each of 61 landmarks). Allometric effects and asymmetric variation 276 have not been removed from the example dataset, for comparability with previously published 277 analyses of macaque skull modularity (Cheverud 1982; Goswami and Polly 2010), although, as 278 279 with selection of metric of trait correlation, the model presented here is applicable to datasets that do remove, or focus entirely on, those aspects of shape. 280

281

#### 282 Models

We investigated 31 model structures within several broad hypotheses of cranial modularity. The first, and simplest, model structure is that there are no distinct modules within the cranium, and that the cranium can be analyzed as a single entity. Further, more complex, models of modularity consist of a two-module (neurocranial vs. facial) structure (Drake and Klingenberg 2010), two 287 six-module structures (primate-specific (Cheverud 1995b) and general mammalian (Goswami 2006a)), and an eight-module structure combining the two six-module models (see: Table S1). 288 We investigated further refinements for both configurations of the six-module structure: first, 289 290 leaving some landmarks "unintegrated", i.e., outside of any module, based on a monotreme model of integration (Goswami 2006a), resulting in 3-module + "unintegrated" models; and, 291 second, considering a tissue-origin model (Goswami 2006a), in which landmarks were grouped 292 293 based on their derivation from neural crest, mesodermal, or mixed germ-layer derived bone (see: Table S1). 294

295

As detailed above, each hypothesized model structure may have many potential 296 297 parametrizations, depending on whether within-module or across-module correlations are 298 modeled as being the same for all cases (e.g., a single high hypothesized correlation within modules and a single, across-module correlation), or all module cases are considered unique, or 299 300 some mixture of these extremes. For example, the 2-module neurocranial/facial model structure 301 comprises Models 2 and 3 (Table 1), with the difference being the number of proposed withinmodule estimates. Models with increasing numbers of modules have correspondingly greater 302 303 complexity in their potential parameterizations. As described above, the six-module model has four different parameterizations examined here (Fig. 2). In the simplest model (Model 4, Fig. 304 2D), there is a single within-module estimate and a single across-module estimate. Other models 305 propose six freely-varying within-module estimates with a constant across-module estimate 306 (Model 5, Fig. 2F), fifteen freely-varying across-module estimates with a single within-module 307 estimate (Model 6, Fig 2E) and a completely varying model with six within-module estimates 308 and 15 across-module estimates (Model 7, Fig. 2G). All model structures that were explored and 309

Page 16 of 46

their corresponding parameterizations are given in Table 1. The R code used in this analysis and

example data files are provided in the online supporting information for this article and are

312 available for download from: http://www.goswamilab.com/#!software/c1cxq.

313

314 Subsampling analysis

While analyses of integration are often performed on model systems with the ability to sample 315 large numbers of individuals, questions about the evolution of integration can require the 316 317 incorporation of fossil or rare taxa (Goswami et al. 2015) for which sample sizes are constrained. To evaluate potential sensitivity of this method to small sample sizes, we conducted a 318 subsampling analysis of the best sampled dataset (subadult *Macaca*, 48 specimens), producing 319 50 random subsets each of 25 specimens, 15 specimens, and 10 specimens. Each subset was 320 subjected to generalized Procrustes analysis prior to calculation of vector congruence coefficient 321 correlation matrices, producing 61x61 element matrices and analyzed in EMMLi. 322

323

#### 324 **RESULTS**

#### 325 *Simulations*

When a low standard deviation ( $\sigma = 0.01$ ) around the simulated correlation values was used, the correct model structure was identified as the best fit model in 100% of cases for all no-module, two-module, and six-module structures (Fig. 3A). Reconstructed  $\rho$  values were consistently within 0.01 of the simulated values. For the simulations of a no-modularity data set, posterior 330 probabilities were generally low, ~0.24, even for the best fit model. All posterior probabilities 331 for the correct model were greater than 0.5 for the simulations in which there was a modular 332 structure to the data. In all cases, estimated  $\rho$  values exactly matched those used to generate the 333 simulated datasets.

334

When a higher standard deviation of 0.05 was used, the correct model was identified in most 335 cases, although accuracy decreased at the highest levels of mean correlations for simple 336 structures (Fig. 3B). The correct model was selected with high (>0.90) posterior probability in 337 338 100% of cases for the simple six-module model with within-module correlations ranging from 0.15 to 0.70. It was also correct, with 100% posterior probability, in all cases for the complex 339 340 six-module structure, using either high, mixed, or low correlations. When all within-module correlations were set to 0.90, the correct model was selected in 23/100 runs, and receives a 341 posterior probability > 0.05 in 36/100 runs, with a different parameterization of the same model 342 structure (six modules, K=8) selected in all remaining cases. For the two-module model, the 343 correct model was selected in 100% of cases for within-module correlations of 0.15, 0.30, and 344 0.50. The correct model is selected in 84/100 cases when the within-module correlation is 0.7, 345 346 and receives a posterior probability > 0.05 in 100% of cases. In the remaining 16 runs, the closely related, more parameterized two-model model (K=3) was selected as the best fit model. 347 When within-module correlations are centered around 0.90, an unrelated model was selected in 348 the majority of cases. The correct model was selected in 100% of cases with the complex two-349 module model using low or mixed correlations. When only the highest correlations (0.70 and 350 0.90) were used to simulate a complex two-module structure, the correct model was selected in 351 77/100 cases and had a posterior probability > 0.05 in 83/100 cases. 352

353

The strongest effects of high correlations and higher standard deviation were observed in cases 354 355 of no modularity in the simulated structure (Fig. 3B). The correct model was selected in 100% of cases when the overall correlation was 0.15 or 0.30. When the overall correlation was 0.50, 356 the correct model was selected as the best fit model in 98/100 runs and had a posterior 357 probability > 0.05 in all runs. With overall correlations of 0.70, the correct model was selected 358 as the best fit model in 53/100 cases and had a posterior probability > 0.05 in 95 cases. In the 359 cases where the wrong model was selected, the posterior probability was < 0.50 in all but five 360 cases, although, as noted above, posterior probabilities are generally low ( $\sim 0.2$ ) for models of no 361 modularity, even when the correct model was selected. When the overall correlation was 362 extremely high, 0.90, the wrong model was selected with posterior probability > 0.50 in all runs. 363 Even in cases where the wrong model was supported, estimated  $\rho$  values were within 0.03 of the 364 values used to simulate each dataset. 365

366

367 *Case study* 

For all five data sets, the optimal model selected by AIC<sub>c</sub> was Model 7 (Fig. 1C), with over 99% of the posterior probability centered on this model for each data set, with the remaining model posterior probabilities were effectively zero for all other models considered (Tables 2, S2-S5). Additionally, the 183x183 raw coordinate data the juvenile (M1 erupted) data set (Table S6) also returned Model 7 as the unambiguously best-supported model. Model 7 can thus be considered the single optimal model describing the pattern of cranial integration in the macaque data set (Edwards 1992; Royall 1997; Burnham and Anderson 2002). 375

376	Model 7 is based on Cheverud's primate-specific six-module structure (Cheverud 1982),
377	proposing distinct within-module $\rho$ 's for all six modules, as well as separate $\rho$ 's for all possible
378	across-module comparisons (total of 22 estimated parameters). Model 16, for the adult female
379	data set only, had a posterior probability of ~ 0.001 (Table S2). This model is a variant of Model
380	7, in which the oral, nasal, and occipital modules are maintained, but all other landmarks are
381	treated as unintegrated, which is broadly similar to the pattern of modularity displayed by
382	monotremes (Goswami 2006a). All other model structures, including those that proposed no
383	modularity, a neurocranial/facial module structure, more than six cranial modules, or non-
384	primate specific module structures, received no support.

385

Estimated values for p were similar for each of the 21 model parameters across the four data sets
(Table 3), with very strongly integrated anterior modules (Modules 1 and 2, corresponding to the
anterior dentition and nasal/facial bones) and a moderately integrated occipital region (Module
6). Other modules, corresponding to the basicranium, neurocranium, and palatal/molar region
were less well integrated, as were inter-module correlations. This is in approximate agreement
with previous analyses of integration patterns in mammalian crania (Goswami 2006a).

392

393 Subsampling analysis

For the subsampling analyses, the unambiguously best supported model (posterior probabilities >

395 0.95) was the same as for the full dataset (Model 7) 100% of the time, for the rarefaction to 25

396 specimens. With 15 specimens, the same model was selected in 48/50 analyses. In the two cases 397 of mismatch, Model 7 was one of three top models (posterior probability > 0.05), sharing support with alternative parameterizations of the same Cheverud six-module structure. Subsampling to 398 399 10 specimens recovered Model 7 in 36/50 of runs. In three of the remaining runs in which it wasn't the best fit model, it was selected as one of the top models (>0.05 posterior probability), 400 in all cases along with alternative parameterizations of the Cheverud six-module structure. For 401 11 runs, Model 7 had a posterior probability less than 0.05. Thus, even at n=10, this method was 402 successful at identifying the correct model as having a significant posterior probability 78% of 403 the time. Moreoever, of the 14 cases where Model 7 was not the top model, the best supported 404 model was a variation on the Cheverud model in 12 cases. In only 2 of the 50 runs was the top 405 model unrelated to Model 7; thus, a relevant model structure, if not the correct parameterization, 406 was recovered in 96% of cases at n=10. 407

408

Reconstructed  $\rho$  values were consistently very similar to those of the full dataset (Table 4), even at n =10, with mean deviations from  $\rho$  values for the full dataset of 0.020 for n = 25, to 0.037 for n = 15, and 0.062 for n = 10. Standard deviations of reconstructed  $\rho$  values were similarly low, but unsurprisingly increasing with decreasing sample sizes: 0.023 for n = 25, 0.036 for n = 15, and 0.042 for n = 10. Thus, these further analyses provide strong support that this method is remarkably robust to quite low sample sizes.

415

417

#### 418 **DISCUSSION**

Extensive simulations varying model complexity, magnitude of mean within-module correlation, 419 and standard deviation of correlations demonstrates that this method is robust under biologically 420 realistic conditions. It performs exceedingly well (perfectly, in fact), when correlations are 421 tightly grouped around hypothetical values of  $\rho$  (low standard deviation simulations), regardless 422 of whether the simulated structure is highly modular or entirely lacks any modular structure. 423 424 With increased dispersion around the p values (higher standard deviations), this method is robust 425 under most conditions, but struggles with highly integrated structures, specifically those that combine two biologically unlikely situations: 1) complete lack of modularity and 2) uniformly 426 427 and, in most cases, unrealistically high correlations. Only in the case of very high within-module 428 correlations (mainly  $\rho = 0.90$ , but also involving  $\rho = 0.70$  in the no-modularity model and in the 429 high-correlation complex two-module model) does the method return incorrect model structures 430 with high posterior probability. Observing such high correlations, uniformly across all modules 431 or an entire structure is unusual. Previous studies (Conner et al. 2014) have shown that vertebrates, plants, and hemimetabolous insects display mean phenotypic correlations among 432 433 linear traits ranging from 0.35 to 0.5, although mean correlations among linear traits in holometabolous insects may be much higher ( $\sim 0.84$ ). In the case study presented here, only a 434 single module (Module 2) shows mean within-module correlations above 0.7 (Table 3), while all 435 other modules are in the moderate to low range of within-module correlations used in these 436 simulations. Our simulations also show that this method is extremely robust in identifying 437 complex models of modularity in which some modules have high within-module correlations and 438 others have moderate or low within-module correlations. Thus, outside of the unusual conditions 439

Page 22 of 46

noted above, our method proves to work with high efficacy, and the few cases of "failure" in
conditions typically encountered in most biological systems involved selection of a differently
parameterized version of the same model structure.

443

We further note that no other method currently available for confirmatory analysis of modularity 444 directly compares models of modularity against a model of total integration (e.g., Marquez 2008; 445 Klingenberg 2009; Adams 2016). For example, in the description of the covariance ratio metric, 446 447 the author provided the important cautionary note that covariance ratio analysis be used only for 448 evaluating patterns of modularity and suggested that Partial Least Squares analysis (Rohlf and 449 Corti 2000; Adams and Felice 2014) be used to evaluate hypotheses of integration (Adams 450 2016). EMMLi thus provides unprecedented ability to evaluate models of total integration as well as models of modularity, but struggles with correctly identifying the lack of modularity 451 when both standard deviations of correlations and mean correlations are high. For this reason, 452 453 we urge caution in interpreting results if the returned posterior probabilities of the best fit models 454 are low (< 0.50), if reconstructed correlations are exceptionally high (uniformly > 0.70), or if multiple unrelated models are returned with posterior probability > 0.05, particularly if standard 455 456 deviations of within-module correlations are high. Under those circumstances, we follow Adams (2016) in suggesting that it may prove useful to employ Partial Least Squares analysis to 457 evaluate the support for a highly integrated structure. We further advise users to consider and 458 report all models with posterior probabilities greater than 0.05. 459

460

461 With regard to the macaque case study, for all five data sets, greater than 99% of the posterior probability distribution was explained by Model 7, the most parameterized version of 462 Cheverud's model of six cranial modules. This result indicates very strong support for this model 463 of cranial modularity in macaques. Cheverud's (1982) model structure was based on analysis of 464 correlations among inter-landmark distances (length measurements) from a dataset of 462 rhesus 465 macaques (Macaca mulatta). Cheverud (1982) identified support for this model by calculating an 466 agreement statistic between the hypothesized F-sets and empirical P-sets, the latter derived by 467 cluster analysis of inter-landmark distances in principal component space. This model structure 468 has subsequently tested using theoretical matrix correlation analysis and RV coefficient analysis, 469 with the present Japanese macaque dataset (*M. fuscata*) (Goswami and Polly 2010). However, 470 that study also tested two alternative models: the two-module facial/neurocranial model (Models 471 472 2-3 in Table 1), and an alternative six-module structure (the "Goswami" models, Models 8-11 in Table 1), based on general patterns of integration among therian mammals (Goswami 2006a). In 473 that study, model selection was not directly possible, as RV coefficient analysis makes no 474 475 specific hypothesis regarding model parameterization beyond the total number of modules and theoretical matrix correlation analysis simply compares the correspondence between two 476 matrices, usually with a permutation test to assess support. All three model structures were 477 supported at p < 0.01 using theoretical matrix correlation analysis with Mantel's test, although it 478 should be noted that Cheverud's model showed the highest correlations with the empirical data. 479 In the RV coefficient analyses, the two-module model was supported in three of the five datasets 480 (p < 0.05), the Goswami model was supported in two of five datasets, and the Cheverud model 481 supported in three of the five datasets, and, where supported, the Cheverud model received the 482

strongest support (p < 0.001). However, it was not supported for either adult dataset, whereas</li>
both the two-module and the Goswami models received support for the adult male dataset.

485

486	The Goswami and Polly (2010) analysis highlighted an important issue with the existing range of
487	confirmatory approaches to analyzing modularity: the lack of a clear way to compare among
488	models across proposing fundamentally different structures of modularity/integration. One can
489	compare the Cheverud six-module model to the Goswami six-module model with RV coefficient
490	analysis, as they both are based on six cranial modules, yet neither can be meaningfully
491	compared to the two-module neurocranial/facial model (Fig. 1). Moreover, there are a range of
492	possibilities, from unintegrated traits within a partially modular structure, to entirely different
493	modular structures that are biologically interesting and potentially informative, but which are
494	impossible to approach with the existing methods.

495

The results presented demonstrate the unambiguous support for Cheverud's structure of 496 phenotypic modularity for the macaque cranium, with distinct within- and among-model 497 correlation values. Here, we used maximum likelihood analysis of congruence coefficients 498 derived from multidimensional vector variables, as well as the more standard individual 499 coordinate correlations for one dataset. We focused on trait correlation matrices, rather than 500 variance-covariance matrices, in this method, as the relationships among traits, and not their 501 individual variances, are the primary concern in studies of phenotypic integration and modularity 502 (Olson and Miller 1951; Olson and Miller 1958; Pavlicev et al. 2009; Goswami and Polly 2010; 503 Conner et al. 2014). Benefits of the model selection approach employed here include: 1) ability 504

to directly compare models of different complexities (such as two- and six-module models) or models of similar complexity which do not constitute nested subsets of one another (such as the Cheverud (1982) and Goswami (2006a) six-module models), 2) increased precision in model description, in terms of varying numbers of within- and between-module values for  $\rho$ ; and 3) expansion to mixed models, in which a structure can include both modules and unintegrated traits (e.g., models 20-31 in Table 1).

511

As noted above, there is an existing method to compare competing models of variational modularity using subspace analysis (Marquez 2008). As with the maximum likelihood approach described here, subspace analysis is a remarkably flexible approach that accurately reflects the complexity of biological systems and is capable of comparing hundreds of models (and indeed performs better with more models).

517

Both subspace analysis and EMMLI can test multiple variations on a basic model structure, 518 allow for combined or overlapping modules, and conduct direct comparison of models with 519 similar or different parametrizations. In contrast to maximum likelihood analysis as implemented 520 521 in EMMLi, subspace analysis creates a specific hypothetical covariance matrix for each matrix that fixes between-module covariances at zero. This is rarely the case in biological systems, 522 particularly in proximal modules, and therefore oversimplifies the apparent hierarchical pattern 523 of modularity in systems such as the cranium. The maximum likelihood-based approach 524 525 described here could be considered preferable because it does not assign an *a priori* value to between-module correlations, and by returning all estimated p values for the best supported 526

Page 26 of 46

527 model(s), allows for direct assessment of every within- and between-module correlation, which 528 can inform on alternative model structures to test (for example, if two modules show a between-529 module  $\rho$  that is equal or similar to their respective within-module  $\rho$  values, one could add an 530 additional model that unites those modules into a single grouping).

531

The two methods also differ on the method of model selection. As a measure of goodness of fit 532 between the observed and model covariance matrices, subspace analysis as implemented in 533 534 MINT (Marquez 2008) uses  $\gamma$ , and corrects for differences in the parametrizations of each model by regressing  $\gamma$  against the number of zero elements in each model, generating  $\gamma^*$ , with 535 significance evaluated against expectations from random covariance matrices. In order to 536 537 strengthen the evaluation of model rank, a jackknifing approach was used, with model support 538 reflecting how often a model ranked first in the jackknifed samples. The method described here 539 does not require fixing any values, but instead provides an overall model structure and searches 540 for values of p that return the maximum likelihood for that structure. The complexity of the model, and correction for the goodness of fit or model selection, is a function of the number of 541 independent estimates of  $\rho$ , rather than the number of zero elements in the model. 542

543

Because subspace analysis as implemented in MINT has never been developed for 3-D data, we did not conduct a direct comparison of these two methods. Qualitative comparison of the simulations of subspace analysis (Marquez 2008) and those described here suggest that the maximum likelihood approach is more robust to sample size, number of models, model complexity, and magnitude of integration, as well as being available for use with any

morphometric dataset. Nonetheless, subspace analysis represented a major improvement on
existing methods, and there are numerous interesting aspects to subspace analysis as
implemented in MINT, such as the heuristic modeling of additional hypotheses of modularity
and the construction of consensus models, both of which could be developed as exploratory tools
within a likelihood framework.

554

In addition to the possibility of incorporating aspects of the Marquez (2008) method, which was 555 developed for the same purpose as the maximum likelihood method described here, there is also 556 557 vast potential for combining with methods developed for different goals. For example, the Reimmanian spaces for covariance matrices and the distances therein provide a framework for 558 559 comparing the relative likelihood of one covariance matrix to that of another (Bookstein and 560 Mitteroecker 2014) and could be combined with the method we describe here. In whatever 561 combination, all of these methods are beginning to fill an important need for approaches that are 562 more flexible to the biological reality of complex anatomy.

563

These benefits are important, as many studies of phenotypic modularity to date have either assumed a hypothesized set of modules without explicitly testing its validity for the taxon of interest (e.g., applying the Cheverud model to other mammals, as in Marroig et al. 2009; Porto et al. 2009), or have tested a single model in the absence of comparison to other potential models, regardless of the support for that one model (e.g., Klingenberg and Marugan-Lobon 2013). Ongoing analyses of other groups suggest that the Cheverud model does not adequately describe all mammalian taxa. For example, EMMLi analysis of a 55 landmark data set for the red

571 fox, *Vulpes vulpes* (Table S7) recovered the 22-parameter version of the Goswami six-module 572 model as the unambiguous best fit model (for details of dataset, see Goswami 2006b). This result is perhaps unsurprising, as that model was initially based on cluster analyses of a comparative 573 574 dataset that included a large sample of carnivorans (Goswami 2006a). However, it underscores the flexibility of the model selection approach advocated here, in that many different proposed 575 model structures can be simultaneously compared. The approach implemented in EMMLi, and 576 its many possible future extensions, provides the ability to directly compare diverse hypotheses 577 on the evolution of modularity and integration, which will become increasingly crucial as we 578 drift further from well-established model systems. Further work along these lines will be crucial 579 to identifying where shifts in modularity occur in the tree of life, and what the consequences of 580 those shifts may be for the morphological evolution. 581

582

With respect to cranial modularity in macaques, the results from maximum likelihood analyses 583 584 as implemented in EMMLi underscore two important biological points: 1) the model of two cranial modules based on a neurocranial and a facial module is not supported when compared 585 with more complex six-module hypotheses, and 2) the 8-module structure, although biologically 586 587 plausible, is not supported. This implies that while a functional model of a facial (masticatory) vs. neurocranial organization of the skull is too simplistic to describe phenotypic integration, 588 there is also likely an upper limit to the complexity of cranial integration in the macaque system. 589 590 In addition, because Model 7 is highly-supported in the infant, juvenile, and subadult data sets in addition to the two adult data sets, this pattern of morphological integration appears to be 591 established very early in postnatal ontogeny in *Macaca*. This consistency through ontogeny 592 confirms the previous analyses of this dataset (Goswami and Polly 2010), which suggested that, 593

although relative level of integration decreases through ontogeny, the overall pattern isconserved from infancy to adulthood.

596

597 CONCLUSIONS

598 The study of phenotypic modularity has seen rapid growth in recent years. New empirical studies are expanding the topic beyond model systems through development (Young 1959; Zelditch 599 1988; Hallgrimsson et al. 2004; Zelditch et al. 2006; Goswami et al. 2009; Hallgrimsson et al. 600 601 2009; Zelditch et al. 2009; Sears et al. 2012), across the tree of life (Armbruster et al. 2004; Young and Hallgrimsson 2005; Goswami 2006b, a; Goswami 2007; Bell et al. 2011; Bennett and 602 Goswami 2011; Armbruster et al. 2014; Conner et al. 2014; Goswami et al. 2014), and even into 603 the distant past (Goswami 2006a; Bell et al. 2011; Gerber and Hopkins 2011; Webster and 604 Zelditch 2011a, b; Maxwell and Dececchi 2012; Meloro and Slater 2012; Gerber 2013; Goswami 605 et al. 2015). Alongside this extension of taxonomic and temporal sampling, there has been an 606 expansion of analytical tools for the evaluation of modularity and integration. Confirmatory 607 approaches, in particular, have received much attention in recent years, with RV coefficient 608 analysis in particular being heavily applied to the analysis of modularity. However, these 609 approaches by and large are limited to the direct comparison of models with similar complexities 610 and do not allow for mixed models, where some traits are highly integrated and others are not. 611 The issues caused by these weaknesses in the existing approaches will become increasing 612 problematic as workers diverge from well-studied models into new systems without well-613 established a priori hypotheses of trait relationships. 614

616 Here, we have presented a maximum likelihood and model selection approach to the evaluation 617 of modularity, which can directly compare highly complex hypotheses of trait relationships, including comparisons of nested and non-nested models. We demonstrate this approach using 618 619 multidimensional vector correlation matrices for a large dataset of macaque crania, confirming the results of previous analyses, but allowing, for the first time, robust discrimination of 620 alternative models. Our results support a highly parameterized model of six cranial modules, 621 with distinct levels of integration within modules, as well as between pairs of modules. This 622 method is applicable to any metric of trait relationship, given the availability of an appropriate 623 transformation, has appropriate Type I error rates, is robust to low sample sizes, and should be 624 incorporated into the existing toolbox for the study of phenotypic modularity in diverse systems. 625

626

#### 627 ACKNOWLEDGEMENTS

We thank Prabu Sivasubramaniam and Tim Lucas for development of the R code for EMMLi. 628 Data were gathered at the Primate Research Institute in Inuvama, Japan, with funding provided 629 by the Japanese Society for the Promotion of Science HOPE grant to AG. We thank M. Takai for 630 631 his role in encouraging and facilitating the research trip to PRI. We thank D. Adams, P.D. Polly, E. Sherratt, C. Klingenberg, UCL's ADaPTiVE lab group, and the NHM-UCL-IC Palaeobiology 632 journal club for relevant discussions and comments on this work. The development of this 633 method was supported by a Leverhulme Trust research grant to AG (RPG 2013-124) and a 634 European Research Council grant (ERC-STG-2014-637171) to AG. 635

#### 637 REFERENCES

- Ackermann, R. R. and J. M. Cheverud. 2000. Phenotypic covariance structure in tamarins (genus
   Saguinus): a comparison of variation patterns using matrix correlation and common
   principal components analysis. American Journal of Physical Anthropology 111:489-501.
- Adams, D. C. 2016. Evaluating modularity in morphometric data: challenges with the RV

642 coefficient and a new test measure. Methods Ecol. Evol. in press.

- Adams, D. C. and R. N. Felice. 2014. Assessing trait covariation and morphological integration
   on phylogenies using evolutionary covariance matrices. PLoS ONE 9:e94335.
- Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pp.
   267-281 *in* B. N. Petrov, and F. Csaki, eds. Second International Symposium on
   Information Theory. Akademiai Kiado, Budapest.
- Armbruster, W. S., C. Pelabon, G. H. Bolstad, and T. F. Hansen. 2014. Integrated phenotypes:
   understanding trait covariation in plants and animals. Phil Trans Roy Soc Lon B
   369:20130245.
- Armbruster, W. S., C. Pélabon, T. F. Hansen, and C. P. H. Mulder. 2004. Floral integration,
   modularity, and accuracy: distinguishing complex adaptations from genetic constraints.
   Pp. 23-49 *in* M. Pigliucci, and K. Preston, eds. Phenotypic integration. Oxford University
   Press, Oxford.
- Bell, E., B. Andres, and A. Goswami. 2011. Limb integration and dissociation in flying
   vertebrates: a comparison of pterosaurs, birds, and bats. J. Evol. Biol. 24:286-2599.
- Bennett, C. V. and A. Goswami. 2011. Does reproductive strategy drive limb integration in
   marsupials and monotremes? Mammalian Biology 76:79-83.
- Bookstein, F. L. and P. Mitteroecker. 2014. Comparing covariance matrices by relative
   eigenanalysis, with applications to organismal biology. Evol. Biol. 41:336-350.
- Burnham, K. P. and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A
   Practical Information-Theoretic Approach. Springer, New York.
- Burnham, K. P. and D. R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC
   in model selection. Sociological Methods and Research 33:261-304.
- 665 Cheverud, J. M. 1982. Phenotypic, Genetic, and Environmental Morphological Integration in the
   666 Cranium. Evolution 36:499-516.
- Cheverud, J. M. 1989. A comparative analysis of morphological variation patterns in the
   Papionines. Evolution 43:1737-1747.
- Cheverud, J. M. 1995a. Morphological integration in the saddle-back tamarin (*Saguinus fuscicollis*) cranium. Am. Nat. 145:63-89.
- 671 Cheverud, J. M. 1995b. Morphological integration in the saddle-back tamarin (Saguinus fuscicollis) cranium. Am. Nat. 145:63-89.
- 673 Cheverud, J. M. 1996. Developmental integration and the evolution of pleiotropy. Am. Zool.
   674 36:44-50.
- Conner, J. K., I. A. Cooper, R. J. L. Rosa, S. G. Perez, and A. M. Royer. 2014. Patterns of
   phenotypic correlations among morphological traits in plants and animals. Phil Trans Roy
   Soc Lon B 369:20130246.
- De Leeuw, J. 1983. Models and methods for the analysis of correlation coefficients. Journal of
   Econometrics 22:113-137.
- 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.

682	Edwards, A. W. F. 1992. Likelihood: Expanded Edition. The Johns Hopkins University Press,
683	Baltimore.
684	Feldman, G., A. Hayes, S. Kumar, J. Greeson, and JP. Laurenceau. 2007. Mindfulness and
685	emotion regulation: The development and initial validation of the Cognitive and
686	Affective Mindfulness Scale-Revised (CAMS-R). Journal of Psychopathology and
687	Benavioral Assessment $29:1//-190$ .
688	Fructano, C., P. Franchini, and A. Meyer. 2013. Resampling-based approaches to study variation
689	In morphological modularity. PLoS ONE 8.009570.
690	membral and integration and membral acidal dimension. DL as ONE 9:6(2012)
691	Carbon S and M L Honking 2011 Magaia betarashrany and evolutionary modularity the
692	trilabite games Zacanthomais as a case study. Evolution 65:2241, 2252
693	Consumi A 2006 Cranial modularity shifts during mammalian evolution Am Nat 168:270
694 695	280.
696	Goswami, A. 2006b. Morphological integration in the carnivoran skull. Evolution 60:169-183.
697	Goswami, A. 2007. Phylogeny, diet, and cranial integration in australodelphian marsupials.
698	PLoS One 2:e995.
699	Goswami, A., W. J. Binder, J. A. Meachen, and F. R. O'Keefe. 2015. The fossil record of
700	phenotypic integration and modularity: a deep-time perspective on developmental and
701	evolutionary dynamics. Proc. Natl. Acad. Sci. U. S. A. 112:4891-4896.
702	Goswami, A. and P. D. Polly. 2010. Methods for studying morphological integration and
703	modularity. Pp. 213-243 in J. Alroy, and E. G. Hunt, eds. Quantitative Methods in
704	Paleobiology. Paleontological Society Special Publications.
705	Goswami, A., J. B. Smaers, C. Soligo, and P. D. Polly. 2014. The macroevolutionary
706	consequences of phenotypic integration: from development to deep time. Phil Trans Roy
707	Soc Lon B 369:20130254.
708	Goswami, A., V. Weisbecker, and M. R. Sanchez-Villagra. 2009. Developmental Modularity and
709	the Marsupial-Placental Dichotomy. J. Exp. Zool. Part B 312B:186-195.
710	Hallgrimsson, B., H. Jamniczky, N. M. Young, C. Rolian, T. E. Parsons, J. C. Boughner, and R.
711	S. Marcucio. 2009. Deciphering the palimpsest: studying the relationship between
712	morphological integration and phenotypic covariation. Evol. Biol. 36:355-376.
713	Hallgrimsson, B., K. Willmore, C. Dorval, and D. M. L. Cooper. 2004. Craniofacial variability
714	and modularity in macaques and mice. J. Exp. Zool. Part B 302B:207-225.
715	Hurvich, C. M. and CL. Tsai. 1989. Regression and time series model selection in small
716	samples. Biometrika 76:297-307.
717	Klingenberg, C. P. 2008. Morphological integration and developmental modularity. Annual
718	Review of Ecology, Evolution, and Systematics 39:115-132.
719	Klingenberg, C. P. 2009. Morphometric integration and modularity in configurations of
720	landmarks: tools for evaluating a prior hypotheses. Evol. Dev. 11:405-421.
721	Klingenberg, C. P. 2013. Cranial integration and modularity: insights into evolution and
722	development from morphometric data. Hystrix 24:43-58.
723	Klingenberg, C. P. 2014. Studying morphological integration and modularity at multiple levels:
724	concepts and analysis. Phil Trans Roy Soc Lon B 369:in press.
725	Klingenberg, C. P. and J. Marugan-Lobon. 2013. Evolutionary covariation in geometric
726	morphometric data: analyzing integration, modularity and allometry in a phylogenetic
727	context. Syst. Biol. 62:591-610.

- LeBel, E. P. and B. Gawronski. 2009. How to find what's in a name: Scrutinizing the optimality
  of five scoring algorithms for the name-letter task. European Journal of Personality
  23:85-106.
- Marquez, E. J. 2008. A statistical framework for testing modularity in multidimensional data.
   Evolution 62:2688-2708.
- Marroig, G. and J. M. Cheverud. 2001. A comparison of phenotypic variation and covariation
   patterns and the role of phylogeny, ecology, and ontogeny during cranial evolution of
   New World monkeys. Evolution 55:2576-2600.
- Marroig, G., L. Shirai, A. Porto, F. B. de Oliveira, and V. De Conto. 2009. The evolution of
   modularity in the mammalian skull II: evolutionary consequences. Evol. Biol. 36:136 148.
- Maxwell, E. E. and T. A. Dececchi. 2012. Ontogenetic and stratigraphic influence on observed
   phenotypic integration in the limb skeleton of a fossil tetrapod. Paleobiology 39:123-134.
- Meloro, C. and G. J. Slater. 2012. Covariation in the skull modules of cats: the challenge of
   growing saber-like canines. J. Vert. Paleontol. 32:677-685.
- Olson, E. C. and R. L. Miller. 1951. A mathematical model applied to the evolution of species.
   Evolution 5:325-338.
- Olson, E. C. and R. L. Miller. 1958. Morphological Integration. University of Chicago Press,
   Chicago
- Pavlicev, M., J. M. Cheverud, and G. P. Wagner. 2009. Measuring morphological integration
   using eigenvalue variance. Evol. Biol. 36:157-170.
- Porto, A., F. B. de Oliveira, L. Shirai, V. De Conto, and G. Marroig. 2009. The evolution of
   modularity in the mammalian skull I: morphological integration patterns and magnitudes.
   Evol. Biol. 36:118-135.
- Rohlf, F. J. and M. Corti. 2000. Use of two-block partial least-squares to study covariation in
   shape. Syst. Biol. 49:740-753.
- Royall, R. M. 1997. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall, New
   York.
- Sears, K. E., C. Doroba, X. Cao, D. Xie, and S. Zhong. 2012. Molecluar determinants of
  marsupial integration and constraint *in* R. J. Asher, and J. Mueller, eds. From clone to
  bone: the synergy of morphological and molecular tools in palaeobiology. Cambridge
  University Press, Cambridge.
- 760 Sokal, R. R. and F. J. Rohlf. 1995. Biometry. W. H. Freeman, New York.
- Steiger, J. H. 1980a. Testing pattern hypotheses on correlation matricies: alternative statistics and
   some empirical results. Multivariate Behavioral Research 15:335-352.
- Steiger, J. H. 1980b. Tests for comparing elements of a correlation matrix. Psychological
   Bulletin 87:245-251.
- Wager, T. D., D. J. Scott, and J.-K. Zubieta. 2007. Placebo effects on human μ-opioid activity
   during pain. Proceedings of the National Academy of Sciences 104:11056-11061.
- Wagner, P. J. 2000. Likelihood tests of hypothesized durations: determining and accommodating
   biasing factors. Paleobiology 26:431-449.
- Webster, M. and M. L. Zelditch. 2011a. Evolutionary lability of integration in Cambrian
   ptychopariod trilobites. Evol. Biol. 38:144-162.
- Webster, M. and M. L. Zelditch. 2011b. Modularity of a Cambrian ptychoparioid trilobite
   cranidium. Evol. Dev. 13:96-109.

- 773 Young, N. M. and B. Hallgrimsson. 2005. Serial homology and the evolution of mammalian limb covariation structure. Evolution 59:2691-2704. 774
- Young, R. W. 1959. The influence of cranial contents on postnatal growth of the skull in the rat. 775 776 American Journal of Anatomy 105:383-415.
- Zelditch, M. L. 1988. Ontogenetic variation in patterns of phenotypic integration in the 777 778 laboratory rat. Evolution 42:28-41.
- 779 Zelditch, M. L., J. G. Mezey, H. D. Sheets, B. L. Lundrigan, and J. T. Garland. 2006.
- Developmental regulation of skull morphology II: Ontogenetic dynamics of covariance. 780 Evol. Biol. 8:46-60. 781
- 782 Zelditch, M. L., A. R. Wood, and D. L. Swiderski. 2009. Building developmental integration into ΔL tion-1. functional systems: function-induced integration of mandibular shape. Evol. Biol. 36:71-783 87. 784
- 785

#### 787 FIGURE CAPTIONS

Figure 1. Schematic depiction of three alternative partitions of the macaque cranium. A) No
modularity, with similar levels of correlation among all landmarks. B) Two modules,
corresponding to facial and neurocranial regions. C) Six modules, corresponding approximately
to Cheverud's model (1982). Colored circles indicate module associations. Solid lines indicate
within-module correlations. Dotted lines indicate between-module correlations.

793

794 Figure 2. Schematic depiction of the four alternative parameterizations of a single six-module model structure. A) Basic structure of landmark associations in six modules, indicated by 795 colours. The six modules may have either similar (B) or different (C) magnitudes of within-796 module correlations. The intermodule correlations may also be similar (D and F) or different (E 797 and G) among all pairs of modules. Each distinct estimated value of p is counted as a parameter, 798 along with one additional parameter for estimated variance. Solid lines indicate within-module 799 correlations. Dashed lines indicate between-module correlations. Line colours indicate similar or 800 different estimated values for  $\rho$  (e.g., in B, the black lines indicate that all of the six modules 801 have the same estimated within-module correlation). 802

803

Figure 3. Results of simulations demonstrating accuracy in model selection for different model structures (no modularity, two modules, or six modules), complexity (similar or different withinmodule correlations), and magnitudes of within-module correlations, modelled with varying standard deviations of A)  $\sigma = 0.01$  or B)  $\sigma = 0.05$ . Stacked bars show percentage of simulations

identifying: the correct model (green), an alternative parameterization of the same model 808 structure, i.e., a related model, with posterior probability < 0.50 (dark blue), a related model with 809 posterior probability > 0.50 (light blue), an unrelated model with posterior probability < 0.50810 811 (pink), or an unrelated model with posterior probability > 0.50 (red). Simulated mean withinmodule correlations, or all correlations for no modularity models, are indicated on the x-axis. 812 100 simulations were run for each model, resulting in a total of 4200 simulations. Results show 813 814 that this method is highly accurate at identifying the correct model structure, except where higher ι extrei. standard deviations are combined with extremely high correlations and simple model structures 815 816 (no modularity, in particular).

## TABLES

Table 1: Model descriptions and parameterizations for the 31 model structures explored in this study. Base models structures follow the allocation of landmark variables in Table S1. Model parameters are a sum of the number of estimated correlations within modules and across modules, plus one (for the estimate of the variance of the population correlation).

Model ID		# Modules	Model description	# Parameters
1		0	$1 \rho$ for all correlations	2
2		2	1 within module $\rho$ for both modules, 1 between-module $\rho$	3
3	иала собъекти и собъекти Собъекти и собъекти и с	2	2 within-module $\rho$ 's and 1 between-module $\rho$	4
4		6	1 within-module $\rho$ and 1 between-module $\rho$	3
5		6	Separate within-module $\rho$ 's and 1 between-module $\rho$	8
6		6	1 within-module $\rho$ and separate between-module $\rho$ 's	17
7	햜탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒탒	6	Separate within-module $\rho$ 's and separate between-module $\rho$ 's	22
8	Goswami model	6	1 within-module $\rho$ and 1 between-module $\rho$	3
9	Goswami model	6	Separate within-module $\rho$ 's and 1 between-module $\rho$	8
10	Goswami model	6	1 within-module $\rho$ and separate between-module $\rho$ 's	17
11	Goswami model	6	Separate within-module $\rho$ 's and separate between-module $\rho$ 's	22
12	Cheverud/Goswami	8	1 within-module $\rho$ and 1 between-module $\rho$	3

1		1 1
0000	hinod	modal
сонн	Diffect	moder
••••	U III U U	1110 401

13	Cheverud/Goswami 8 Separate within-module ρ's and 1 betwe		Separate within-module $\rho$ 's and 1 between-module $\rho$	10
14	Cheverud/Goswami	8	1 within-module $\rho$ and separate between-module $\rho$ 's	30
15	Cheverud/Goswami combined model	8	Separate within-module $\rho$ 's and separate between-module $\rho$ 's	37
16	Tissue Origin model	3	1 within-module $\rho$ and 1 between-module $\rho$	3
17	Tissue Origin model	3	1 within-module $\rho$ and separate between-module $\rho$ 's	5
18	Tissue Origin model	and and a second and	Separate within-module $\rho$ and 1 between-module $\rho$ 's	5
19	Tissue Origin model	and and a second and	Separate within-module $\rho$ and separate between-module $\rho$ 's	7
20	Cheverud-based	and and a second and	1 within-module $\rho$ (for modules 1, 2, and 6 only), 1 pooled	2
20	"monotreme" model	and and a second and	between-module and unintegrated $\rho$	3
21	Cheverud-based	and and an and an and an and an analysis and an	1 within-module $\rho$ (for modules 1, 2, and 6 only), 1 between-	1
21	"monotreme" model	and and an an	module $\rho$ , and 1 unintegrated $\rho$	4
22	Cheverud-based	and and an and an and an and an analysis and an	Separate within-module $\rho$ 's (for modules 1, 2, and 6 only), 1	-
22	"monotreme" model	and and an an	pooled between-module and unintegrated $\rho$	5
23	Cheverud-based	and and a second and	Separate within-module $\rho$ 's (for modules 1, 2, and 6 only), 1	6

	"monotreme" model		between-module $\rho$ , and 1 unintegrated $\rho$		
24	Cheverud-based	2	1 within-module $\rho$ (for modules 1, 2, and 6 only), separate		
24	"monotreme" model	3	between-module $\rho$ 's, and 1 unintegrated $\rho$	6	
25	Cheverud-based	2	Separate within-module $\rho$ 's (for modules 1, 2, and 6 only),	0	
25	"monotreme" model	3	separate between-module p's, and 1 unintegrated $\rho$	8	
26	Goswami-based		1 within-module $\rho$ (for modules 1, 2, and 6 only), 1 pooled	2	
26	"monotreme" model	3	between-module and unintegrated $\rho$	3	
27	Goswami-based		1 within-module $\rho$ (for modules 1, 2, and 6 only), 1 between-	4	
21	"monotreme" model	3	module $\rho$ , and 1 unintegrated $\rho$	4	
20	Goswami-based	2	Separate within-module $\rho$ 's (for modules 1, 2, and 6 only), 1	-	
28	"monotreme" model	3	pooled between-module and unintegrated $\rho$	5	
20	Goswami-based	2	Separate within-module $\rho$ 's (for modules 1, 2, and 6 only), 1	ſ	
29	"monotreme" model	3	between-module $\rho$ , and 1 unintegrated $\rho$		
20	Goswami-based	2	1 within-module $\rho$ (for modules 1, 2, and 6 only), separate	ć	
30	"monotreme" model	3	between-module $\rho$ 's, and 1 unintegrated $\rho$	6	
21	Goswami-based	2	Separate within-module $\rho$ 's (for modules 1, 2, and 6 only),	-	
31	"monotreme" model	3	separate between-module p's, and 1 unintegrated $\rho$	8	

Table 2: Results for the Sub-Adult (M2 erupted) data set (n=48) using congruence coefficients. Model parameters, raw loglikelihood fits for each tested model, AIC<sub>c</sub> and  $\Delta$ AIC<sub>c</sub> scores are provided. Model log-likelihoods and the model posterior probability are also shown. Sample size used to calculate AIC<sub>c</sub> was 1830. See methods for details. Model ID's correspond to the numbering in Table 1. The optimal model in the set of evaluated models is highlighted in bold italics.

	K	LogL		ΔAIC <sub>c</sub>	Model LogL	Model Post. Prob
and and an an	2	2078.86		916.21	1.11E-199	1.11E-199
2	3	2134.49		806.96	5.89E-176	5.89E-176
3	4	2147.54		782.88	1.00E-170	1.00E-170
4	3	2219.34		637.26	4.17E-139	4.17E-139
5	8	2380.83		324.35	3.69E-71	3.69E-71
6	17	2395.76		312.75	1.22E-68	1.22E-68
7	22	2557.25	-5069.93	0.00	1.00	1.000
8	3	2153.94	-4301.87	768.06	1.65E-167	1.65E-167
9	8	2226.56	-4437.03	632.90		3.69E-138
10	17	2257.63	-4480.93	589.01		1.26E-128
11	22	2330.25	-4615.93	454.00		2.60E-99
12	3	2172.35	-4338.69	731.24		1.63E-159
13	10	2246.04	-4471.95	597.98		1.41E-130

14	30	2417.44	-4773.85	296.09	오늘 있는 것을 같은 것을 많은 것을 많 같이 같은 것을 많은 것을 많	5.07E-65
15	37	2491.12	-4906.68	163.26		3.54E-36
16	3	2079.47	-4152.93	917.00		7.50E-200
17	5	2214.56	-4419.08	650.85	ng dia tanàna amin'ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny fari Ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny faritr'o amin'ny faritr'	4.67E-142
18	5	2109.73	-4209.43	860.51	t is a set of the set	1.39E-187
19	7	2244.82	-4475.57	594.36	8.62E-130	8.62E-130
20	3	2262.47	-4518.93	551.01	2.24E-120	2.24E-120
21	4	2265.54	-4523.05	546.88	1.76E-119	1.76E-119
22	5	2324.39	-4638.75	431.18	2.34E-94	2.34E-94
23	6	2327.46	-4642.87	427.06	1.84E-93	1.84E-93
24	6	2286.11	-4560.17	509.76	2.03E-111	2.03E-111
25	8	2348.03	-4679.99	389.95	2.11E-85	2.11E-85
26	3	2181.12	-4356.23	713.70	1.05E-155	1.05E-155
27	4	2181.12	-4354.23	715.71	3.85E-156	3.85E-156
28	5	2204.15	-4398.27	671.66	1.42E-146	1.42E-146
29	6	2204.15	-4396.26	673.67	5.17E-147	5.17E-147
30	6	2195.90	-4379.76	690.18	1.35E-150	1.35E-150
31	8	2218.93	-4421.78	648.15	1.80E-141	1.80E-141

	Adult	Adult	Sub-Adult	Juvenile	Infant	
	Females	Males	(M2 erupted)	(M1 erupted)	(Deciduous only)	
Module 1	0.43	0.46	0.43	0.44	0.55	
Module 2	0.77	0.77	0.81	0.76	0.67	
Module 3	0.24	0.35	0.40	0.19	0.22	
Module 4	0.15	0.18	0.14	0.16	0.15	
Module 5	0.12	0.23	0.14	0.17	0.23	
Module 6	0.28	0.29	0.30	0.30	0.28	
M1 to M2	0.10	0.13	0.13	0.13	0.13	
M1 to M3	0.22	0.29	0.35	0.21	0.31	
M1 to M4	0.18	0.22	0.14	0.14	0.20	
M1 to M5	0.21	0.21	0.22	0.22	0.29	
M1 to M6	0.19	0.17	0.22	0.20	0.28	
M2 to M3	0.13	0.22	0.08	0.08	0.12	
M2 to M4	0.14	0.08	0.12	0.08	0.14	
M2 to M5	0.07	0.09	0.10	0.13	0.10	

Table 3: Optimal values of ρ within the six modules and for the 15 inter-module correlations estimated in Model 7 for each of the macaque data sets partitioned by ontogenetic stage.

M2 to M6	0.12	0.27	0.08	0.17	0.08
M3 to M4	0.11	0.15	0.11	0.11	0.13
M3 to M5	0.16	0.12	0.16	0.09	0.16
M3 to M6	0.11	0.12	0.15	0.10	0.14
M4 to M5	0.14	0.15		0.12	0.13
M4 to M6	0.13	0.12		0.11	0.11
M5 to M6	0.17	0.17		0.16	0.15



## A. No modularity



## B. Two modules



## C. Six modules



