Examination of Sample Size Determination in Integration Studies Based on the Integration Coefficient of Variation (ICV)

Hyunwoo Jung, Mark A. Conaway & Noreen von Cramon-Taubadel

Evolutionary Biology

ISSN 0071-3260

Evol Biol
DOI 10.1007/s11692-020-09514-w

Evolutionary Biology
https://doi.org/10.1007/s11692-020-09514-w

RESEARCH ARTICLE

Examination of Sample Size Determination in Integration Studies Based on the Integration Coefficient of Variation (ICV)

Hyunwoo Jung[1] · Mark A. Conaway[1] · Noreen von Cramon-Taubadel[1]

Received: 13 May 2020 / Accepted: 12 August 2020

Abstract

Although there are various indices available for calculating morphological integration, the integration coefficient of variation (ICV) is most suited for assessing magnitudes of integration within and between morphological variance/covariance (V/CV) matrices. However, it is currently not known what the effects of varying sample sizes are on the reliable estimation of distributions of ICV scores. In this regard, the effects of varying sample size on ICV was examined by simulating parameter V/CV matrices with varying underlying magnitudes of average trait correlation (r^2). ICV distributions were generated using a trait resampling protocol for various sample sizes (11 through 150) within various parameter r^2 values. Next, empirical r^2 values were calculated based on data from 22 skeletal elements of 40 *Macaca fascicularis* specimens to examine whether the results from the simulation corresponded to real biological data. Mean ICV scores of various sample sizes were compared using Mann–Whitney U tests to examine which minimum sample sizes are required to reliably calculate mean ICV. Mann–Whitney U test results based on the simulated data showed that a sample size of 51 may be sufficient even for relatively low r^2 values of 0.05. The empirical macaque data showed that 30─40 individuals may be sufficient to reliably calculate mean ICV scores across skeletal elements. Our results correspond closely with previous assessments by Cheverud and colleagues that argued that a sample size of 40 is necessary to accurately estimate the structure of V/CV matrices.

Keywords Morphological integration · Integration coefficient of variation · Computer simulation · *Macaca fascicularis*

Introduction

Morphological integration and modularity have been studied to examine how complex traits interact in terms of shared developmental pathways and functional demands (Olson and Miller 1958; Hallgrímsson et al. 2009; Armbruster et al. 2014; Goswami et al. 2014; Klingenberg 2014). According to the theoretical framework of integration, strictly independent evolution of traits in living organisms may not be possible due to the correlated responses to selection among traits. Thus, it serves as a reminder of one important principle in evolutionary processes; that not all traits are caused by adaptation or direct selection (Gould and Lewontin 1979).

Morphological integration and modularity can be regarded as “the patterns and processes” of trait interaction and independence, respectively (Armbruster et al. 2014). For instance, a set of traits defined as a ‘module’ may have fewer connections with other anatomical or morphological traits but within-module integration should be higher than others. Thus, integration and modularity are not antonyms but rather complimentary concepts. At the genetic level, integration and modularity can be associated with pleiotropy, epistasis, and linkage disequilibrium, resulting in shared developmental pathways among traits (Cheverud 1984; Hallgrímsson et al. 2009). For instance, a pleiotropic effect can occur when a mutation at a single gene locus causes changes in many phenotypic traits (Cheverud 1984). Functional constraints can also have an impact on morphological integration. For instance, fore- and hindlimb lengths or proportions can be functionally (and developmentally) integrated for the locomotor behaviors, leading to morphological integration

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11692-020-09514-w) contains supplementary material, which is available to authorized users.

✉ Hyunwoo Jung
hjung26@buffalo.edu

[1] Buffalo Human Evolutionary Morphology Lab, Department of Anthropology, University at Buffalo, SUNY, Buffalo, NY, USA

Published online: 24 August 2020

and evolutionary constraint in these anatomical structures (Young et al. 2010; Rolian 2014). Young et al. (2010) showed that fore- and hindlimb lengths or proportion is less integrated in hominoids than other anthropoids, which may have resulted in less evolutionary constraint for the evolution of novel limb proportion and bipedalism in humans. Thus, functional influence later in life (as in, locomotor behaviors in hominoids from Young et al. 2010) can lead to reduction of developmental integration of limb proportions earlier in ontogeny of subsequent generations (as in Zelditch and Carmichael 1989; Kelly et al. 2019).

In studies of morphological integration, the structures of trait correlation or variance/covariance (V/CV) matrices are used as proxies for pattern and magnitude of integration (Cheverud 1984; Ackermann and Cheverud 2000, 2002; Marroig and Cheverud 2004; Porto et al. 2009; Marroig et al. 2009). In other words, multiple correlations or covariations among traits are analyzed to examine overall patterns and magnitudes of integration. For instance, the simple example of A and B in Fig. 1 shows the correlation or covariation between two traits with similar patterns but different magnitudes of integration. Here, β and ΔZ represent the selection vectors and response vectors to selection, respectively. The graphs show that total variation is more concentrated in a single axis in B than A. Thus, in Fig. 1b, the response along selection vector β_2 is more constrained relative to the selection vector β_1 where the selection vector is more closely aligned with the main axis of trait correlation (Of course, the direction of the response and selection vectors may not be in the perfect match in real life). These differences are expressed by different length of responses (ΔZ) between Fig. 1a and b. In this regard, it has been suggested that the pattern and magnitude of integration can constrain or facilitate the evolution of morphology in morphospace depending on the adaptive landscape (Porto et al. 2009, 2013; Marroig et al. 2009; de Oliveira et al. 2009; Hallgrímsson et al. 2009; Shirai and Marroig 2010; Klingenberg 2014; Armbruster et al. 2014; Penna et al. 2017). For instance, the pattern, but not the magnitude of integration has been shown to be fairly consistent in the cranium of mammals (Porto et al. 2009; Marroig et al. 2009). Thus, it appears that morphological diversification may be associated with magnitudes of integration but not necessarily patterns of integration (Porto et al. 2009; Marroig et al. 2009; de Oliveira et al. 2009).

Several indices have been used to calculate the pattern and magnitude of integration or modularity, such as the RV coefficient, covariance ratio (CR) coefficient, partial least square (PLS), Random Skewers (RS) method, coefficient of determination (r^2), and integration coefficient of variation (ICV) (Klingenberg 2009; Adams 2016; Rohlf and Corti 2000; Cheverud and Marroig 2007; Shirai and Marroig 2010). However, only few studies have been conducted regarding the issue of necessary sample sizes for various integration indices (e.g., Adams 2016; Grabowski and Porto 2017). Thus, the purpose of this study is to examine the effects of varying sample sizes on the reliable estimation of distributions of ICV scores since the ICV is more suited for calculating the magnitude of integration of the V/CV matrix (Shirai and Marroig 2010).

The RV coefficient (Escoufier 1973; Klingenberg 2009) and CR coefficient (Adams 2016) are used to quantify patterns of integration or modularity between two or more morphological modules. PLS (Rohlf and Corti 2000) can be used to quantify degree (or magnitude) of integration or modularity between two or more morphological modules. When there are more than two modules, the mean of the calculated indices from all pairs of two modules can be obtained. RS method (Cheverud and Marroig 2007) can be

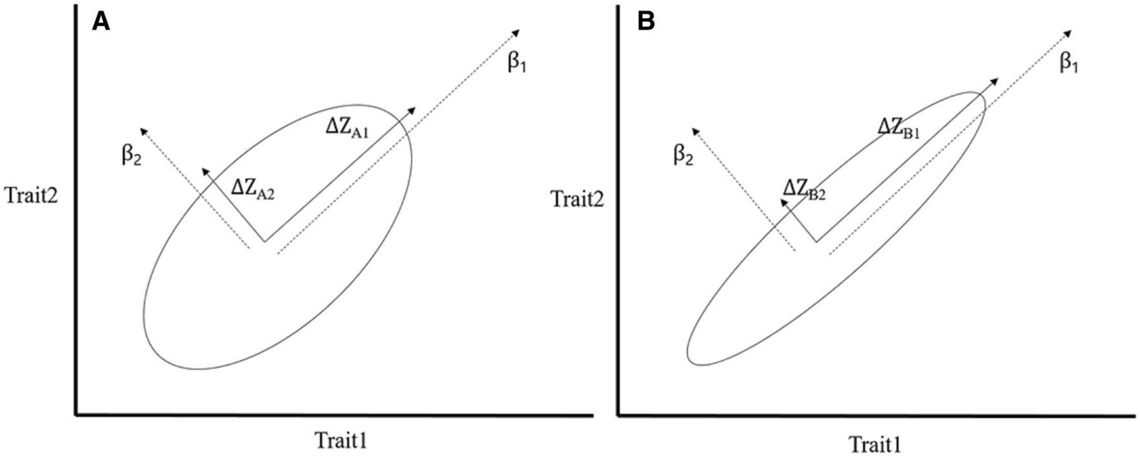

Fig. 1 Two trait correlation or covariation graph with similar patterns of integration but with different magnitudes of integration. A. Traits 1 and 2 are positively but not strongly correlated/covarying. B. Traits 1 and 2 are also positively but much more strongly correlated/covarying. Selection vectors are presented as dashed lines (β) and response vectors to selection vectors are presented as solid lines ΔZ

used for comparing the pattern of integration between two modules. The indices of r^2 and ICV are used to calculate the magnitude of integration in a single correlation or V/CV matrix, respectively (Porto et al. 2009; Shirai and Marroig 2010; Grabowski and Porto 2017).

The RV coefficient is the ratio between the covariance of two blocks and their within-block variances (Klingenberg 2009), where blocks refer to matrices describing V/CV patterns either within- or between modules. In order to calculate the RV coefficient, the covariance matrix needs to be structured as follows (Klingenberg 2009), where, for example, S_1 is the within-module V/CV for module 1, while S_{12} is the between-module V/CV for modules 1 and 2:

$$S = \begin{bmatrix} S_1 & S_{12} \\ S_{21} & S_2 \end{bmatrix}$$

when module 1 and module 2 has p and q number of traits, respectively, matrix S has p + q dimensions. Then, calculation of the RV coefficient is as follows (Klingenberg 2009):

$$RV = \frac{trace\left(S_{12}S'_{21}\right)}{\sqrt{trace\left(S_1S'_1\right)trace\left(S_2S'_2\right)}}$$

Trace(S_1S_1') or trace(S_2S_2') is the sum of the squared variance and squared covariance in each block (within-modules). Trace($S_{12}S_{21}'$) is the sum of the squared covariance between two modules. Thus, this formula presents covariation between two blocks which is standardized by the amount of variation within blocks. It is analogous to the (multivariate) correlation coefficient (Klingenberg 2009). Calculated RV coefficients range between zero and one. The RV coefficient is used to quantify how much covariation exists between two modules considering variances within each module (Klingenberg 2009).

However, Adams (2016) has argued that the RV coefficient may be too sensitive to sample size and the number of variables employed. Thus, the covariance ratio (CR coefficient) was suggested instead to calculate patterns of integration or modularity between two or more modules (Adams 2016). The CR coefficient is different from the RV coefficient as calculation of the CR coefficient is conducted with only the off-diagonal matrix as the numerator and denominator in the formula for the RV coefficient above. Thus, S_1, S_2, and S_{12} will have zeroes for their diagonal elements and only the sum of squared covariance will be included, while the sum of squared variance is excluded from the calculation of the CR coefficient (Adams 2016). Hence, the CR coefficient is literally a covariance ratio of the between and within modules covariance, and quantifies whether covariation between modules is larger or smaller than covariation within modules (Adams 2016).

Partial least squares (PLS) is a method for finding a new axis that explains most of the covariation between two or more modules (Rohlf and Corti 2000). Thus, PLS is similar to principal component analysis but the aim of PLS is to maximize covariance patterns between two or more blocks instead of maximizing variance of a single block (Rohlf and Corti 2000).

The Random Skewers (RS) method uses the multivariate Breeder's equation, $\Delta Z = G\beta$, where ΔZ is the evolutionary change in a vector of trait means, G is the additive genetic V/CV matrix, and β is the selection gradient vector (Lande 1979; Cheverud and Marroig 2007). Morphological variation can be analyzed using this Breeder's equation as the additive genetic G-matrix can be substituted with a phenotypic V/CV matrix (P) given that they have been shown to be largely proportional (Cheverud 1996; Roff 1995; de Oliveira et al. 2009). The RS method is applied to two morphological V/CV matrices to compare their evolutionary response to the same set of selection vectors (Cheverud and Marroig 2007). For instance, the same 1000 randomly generated selection vectors are applied to two target matrices and the correlation between their responses to selection vectors is calculated. Thus, using this approach, one can test the similarity of pattern of integration (or structural similarity) between two morphological V/CV matrices (Cheverud and Marroig 2007).

The coefficient of determination (r^2) is simply the mean of squared correlation coefficients between all traits (Porto et al. 2009). Thus, r^2 quantifies the intensity of mean correlations between all traits within a module. Similarly, the integration coefficient of variation (ICV) is an index for calculating magnitude of integration within modules. The ICV is calculated from the standard deviation of eigenvalues divided by the mean eigenvalue of a V/CV matrix (Shirai and Marroig 2010). Thus, high ICV values indicate that most of the shape variation is concentrated within fewer dimensions as $\mathrm{ICV} = \frac{\sigma(\lambda)}{\bar{\lambda}}$, where $\sigma(\lambda)$ is the standard deviation of the eigenvalues and $\bar{\lambda}$ is the mean of those eigenvalues (Shirai and Marroig 2010; Conaway et al. 2018). Moreover, ICV is scale-independent as the standard deviation of eigenvalues is standardized by its mean.

Although there are various indices, the ICV is more suited for analyzing magnitudes of integration within V/CV matrices than r^2, which can be used to quantify integration in correlation matrices (Shirai and Marroig 2010). Moreover, the ICV can practically summarize the capacity of traits to vary in morphospace depending on their overall covariation patterns (Shirai and Marroig 2010; Conaway et al. 2018). Although PLS can be used to quantify degree (or magnitude) of integration (Rohlf and Corti 2000), PLS cannot take account of within-module patterns or magnitudes of integration (Adams 2016). The ICV, on the other hand, can be employed to quantify magnitudes of integration both within- *and* between-modules, by combining traits across modules.

Calculating the magnitude of integration is important as the magnitude of integration has been found to vary in, for example, the mammalian cranium, while the pattern of integration was found to be consistent among the crania of mammals (Marroig et al. 2009; Porto et al. 2009; de Oliveira et al. 2009). On this basis, it was argued that evolution or diversification in cranial morphology of mammals may be constrained and/or facilitated by magnitudes of integration but is less affected by differing patterns. Moreover, some studies have been conducted in the post-cranium and showed that patterns of integration may be consistent in post-cranial skeletal elements of mammalian taxa, such as mustelids, felids, or canids (Arnold et al. 2016; Randau and Goswami 2017; Botton-Divet et al. 2018; Jones et al. 2018). However, it has not been tested whether the magnitude of integration in the post-cranium is also consistent among mammalian taxa or varies like in the cranium (but see, Young et al. 2010). Therefore, given that the ICV index can be used to quantify magnitudes of integration both within- and between-modules across an organism, the ICV is a useful means of testing how (the magnitude of) morphological integration is associated with evolutionary constraint or facilitation in future studies.

Although the ICV is more suitable for calculating magnitudes of integration in V/CV matrices, the effects of varying sample sizes on obtaining reliable estimates of ICV has not yet been explored. Only one study has been conducted regarding the issue of necessary sample sizes, focused on the magnitude of integration (r^2 in Grabowski and Porto 2017). In the study, Grabowski and Porto (2017) argued that a sample size of 108 individuals was required to accurately estimate integration parameters when average trait coefficient of determination (r^2) is 0.05. The minimum required sample size is an important issue in biology as there may not be enough specimens available in museum collections in many occasions. For instance, if the purpose of a study is to analyze patterns and/or magnitudes of integration in the entire skeleton of primates, it may be impossible to achieve the recommended sample size of 108 individuals (Grabowski and Porto 2017), although the required sample size does vary depending on the average trait coefficient of determination (r^2), and the overall number of traits (Grabowski and Porto 2017). These circumstances are exacerbated if a researcher tries to study morphological integration using the post-cranial skeleton, which is often more poorly represented in museum collections than cranial remains.

Given that, the purpose of this study is to explore the effects of sample size on obtaining reliable estimates of the ICV by simulating populations with various magnitudes of integration (as quantified using r^2). This simulation study on the ICV was conducted using a trait resampling method for generating ICV distributions (Conaway et al. 2018). This methodological approach is useful for comparing magnitudes of integration between skeletal elements with different number of traits (Conaway et al. 2018). For instance, two skeletal elements with 50 traits and 100 traits, respectively, would automatically have different ICV scores if all traits were included, even if mean eigenvalues were the same, as the smallest eigenvalue may be more skewed in the eigenvalue distribution of 100 traits, resulting in unintended inflation of ICV scores. Thus, the trait resampling method is an effective way of generating distributions of ICV values for each morphological module, irrespective of the number of traits, that can be compared statistically. To introduce here briefly within and between module ICV calculations, let us assume that there are modules A, B, C, and D with trait numbers of 50, 60, 70, and 80, respectively. In this example, a "random module" created from A−D would have 260 traits in total as all traits from modules A through D are combined together (Conaway et al. 2018). Thereafter, within-module ICV values can be statistically compared to the random module using the resampling method described above, to address the basic question of whether individual modules (A, B, C or D) are statistically more strongly integrated than random sets of traits taken from across all modules. In a similar way, within and between module ICV scores can be compared between two modules A and B with the combination of modules A and B as a "random module", by resampling 10 traits out of 110 traits as there are 50 and 60 traits in module A and B, respectively. Therefore, distributions of ICV values for certain numbers of random sets of vectors would be generated resulting in a mean and standard deviation of ICV values that can be statistically compared across taxa and/or across different morphological modules.

Methods and Materials

Simulations were conducted in three steps (Fig. 2). First, a variance/covariance (V/CV) matrix based on a multivariate normal distribution with known parameter values of coefficient of determination (r^2) of V/CV matrix was generated using the 'genPositiveDefMat' function in the clusterGeneration *R* package (Joe 2006; Qiu et al. 2006). The generation method used was "c-vine" and the range of variance was between 0.5 and 0.6 with reference to Grabowski and Porto (2017). The generated V/CV matrix had 300 dimensions, representing 300 traits. Grabowski and Porto (2017) showed that generating V/CV matrices using the "c-vine" method may sometimes underestimate the effect of sample size for integration indices due to extremely small values of the smallest eigenvalue. Thus, it was suggested that V/CV matrices with too much skewness in terms of log-eigenvalue

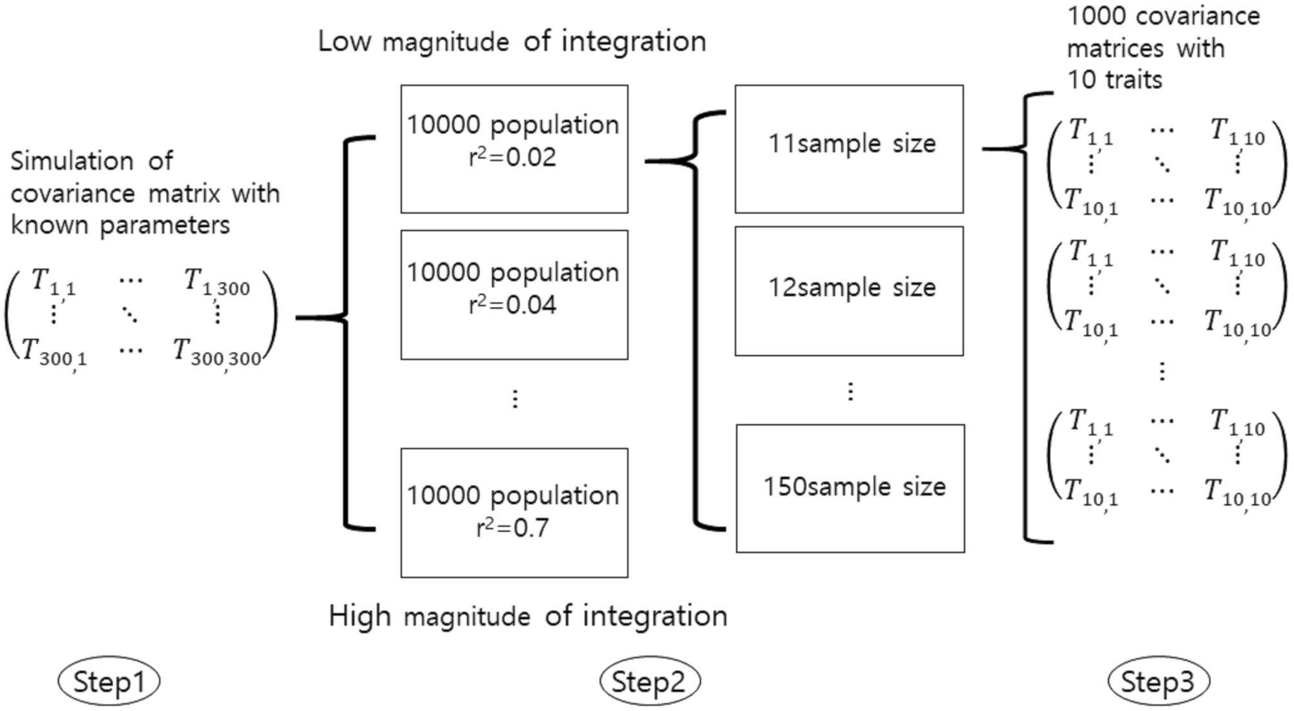
Low magnitude of integration
Simulation of covariance matrix with known parameters
10000 population r²=0.02
10000 population r²=0.04
10000 population r²=0.7
11sample size
12sample size
150sample size
1000 covariance matrices with 10 traits
High magnitude of integration
Step1
Step2
Step3

Fig. 2 Summary of the three steps used in this study to conduct simulations

distribution be filtered out and to choose a V/CV matrix with 'proper' skewness when generating V/CV matrix with parameter values. However, there was no detectable relationship found between skewness of log-eigenvalue distribution and ICV values (Supplementary Fig. 1). Moreover, there was no detectable relationship in the scatter plot of the smallest eigenvalue and ICV values (Supplementary Fig. 2). Thus, the subsequent simulation was conducted without filtering V/CV matrices. In this regard, a V/CV matrix was generated once and used to apply the same pattern of integration to all subsequent simulation procedures as patterns of integration were found to have no effect on sampling effort in a previous study (Grabowski and Porto 2017). The first eigenvalue of the generated V/CV matrix was scaled to adjust parameter r^2 values ranging from 0.02 to 0.7. Thus, the parameter r^2 of the V/CV matrix was allowed to vary, while the pattern of the V/CV matrix remained the same in this study. Accordingly, calculations of ICV distributions were based on the same pattern of integration but differential parameter r^2 values, and therefore, differing magnitudes of integration.

Next, populations of size 10,000 were generated based on the simulated V/CV matrix with a mean of zero and various parameter r^2 values ranging from 0.02 to 0.7 with about 0.02 intervals (Fig. 2). Distributions of ICV values were generated using a resampling method whereby 10 traits were sampled at random out of 300 traits 1000 times (Kazi-Aoual et al. 1995; Conaway et al. 2018; Fig. 2). The resampling method was conducted based on the generated population size of 10,000. The resampling method randomly generates vectors with ten elements from the original vector of length 300. Next, ICV scores were calculated from these randomly generated vectors with length of 10. This procedure is reiterated 1000 times each for differential sample sizes of between 11 and 150 specimens. Samples of between 11 to 150 specimens were randomly selected from the population of 10,000 based on the simulated V/CV matrix with various parameter r^2 values. Sample size started at 11 in order to generate sample V/CV matrices with full rank (i.e., more samples than the number of traits) as 10 traits were resampled in this simulation (Grabowski and Porto 2017). Thus, distributions of ICV values were generated for each sample size to examine the effect of varying sample sizes between 11 and 150. As a result, there were 140 ICV distributions (with 1000 ICV scores in each distribution) generated for each simulated V/CV matrix with specific parameter r^2 values. Boxplots illustrating ICV distributions for varying sample size were examined to determine which minimum sample sizes were required to generate 'stable' ICV calculations, under different assumptions of average trait r^2 values based on previous empirical estimates, such as $r^2 = 0.05$ for the human cranium, $r^2 = 0.08$ for the hominoid cranium, and $r^2 = 0.12$ for the cranium of New and Old World monkey (Marroig et al. 2009; Porto et al. 2009; de Oliveria et al. 2009). Moreover, for future reference, simulations were also conducted with r^2 values of 0.2 and 0.35. Mann–Whitney U tests were employed to statistically compare mean ICV scores between sample size distributions intervals of 10 (e.g., 11 vs. 21 and 21 vs. 31) within each r^2 value tested. Bonferroni adjustment was applied due to multiple comparisons and the resultant alpha level was 0.0038 (i.e., 13 comparisons within each r^2 value).

It was also possible that ICV values may be affected by the number of traits (in this case 300) used in the starting V/CV matrix, as skeletal elements are likely to have different number of traits based on the number of landmarks or measurements available. The effects of changing the total trait numbers on ICV values was examined by altering starting trait numbers to 75, 150, and 300. Moreover, the number of resampled traits was also altered from 10 to 29 traits, sample sizes were set as either 30 or 100, and parameter r^2 values were set as 0.1 or 0.5 to simultaneously examine the effect of the number of total traits to choose from, the number of resampled traits, sample size, and the r^2 (magnitude of integration) of the V/CV matrix. Based on the results of this simulation, it is possible to test whether different skeletal elements (e.g., cranium or mandible) would show systematically differential ICV values due to differing total number of traits. Alternatively, it may be the case that only r^2 and/or the number of resampled traits matters for calculating ICV using the resampling method as predicted above and that it is not necessarily dependent on the total number of traits in terms of magnitude of integration.

For real biological data, sample r^2 values were empirically quantified for 22 different skeletal elements using samples of n = 40 *Macaca fascicularis.* Skeletal elements were scanned using a HDI-120 and a Macro R5X structured-light scanner (LMI technologies INC., Vancouver, Canada). 18 wild specimens of *M. fascicularis* were from the Museum of Comparative Zoology at Harvard University and 22 captive specimens were from the Department of Anthropology of University at Buffalo, SUNY. For the limb bones (scapula, humerus, radius, os coxa, femur, tibia), only 35 specimens (18 wild and 17 captive) out of total 40 specimens were available. The following 22 skeletal elements were quantified: cranium, mandible, 13 elements of the vertebral column (C1, C2, C3, C5, C7, T1, T4, T7, T10, T12, L1, L4, L7), sacrum, scapula, humerus, radius, os coxa, femur, and tibia (there were 3 specimens with T13 as the last thoracic vertebra, 4 specimens with L6 as the last lumbar vertebra, and 1 specimen with 4 sacral vertebrae). In the case of bilateral elements, the left side was landmarked. When the left side was damaged, the right side was landmarked instead. All landmarks were digitized using the software *Landmark* (Wiley et al. 2005) on the 3D scanned skeletal elements. Descriptions of the landmarking protocol for each skeletal element can be found in supplementary data (Supplementary Table 1–10). Traits were generated for each skeletal element by calculating all possible Euclidean distances between pairs of landmarks on each bone.

Within-sex mean standardization was conducted to remove the potentially confounding effect of sexual dimorphism and to control for size differences within and between traits of different skeletal elements (Conaway et al. 2018). For instance, larger bones may have higher ICV scores if larger interlandmark distances explain most of the variation and show larger variance. For each trait, the mean was centered to zero but variance was not scaled within each sex. Thus, the variance and covariance structure was similarly maintained while the effect of sexual dimorphism in size difference was controlled. Next, a MANOVA was conducted on the mean standardized traits for each skeletal element to remove the possibility of artificial inflation of variance due to inclusion of both wild and captive specimens, by extracting trait residuals which were used in further analyses. For each of the 22 elements, the average r^2 values were calculated based on the MANOVA residuals for all possible traits available for that element.

The means and standard deviations (SD) of ICV scores for *M. fascicularis* skeletal elements were calculated with varying sample sizes to examine how r^2 values and sample size affect ICV distributions based on real empirical data. Distributions of ICV scores were calculated using the resampling method described above. For the vertebral column, only the first and last vertebra for the cervical, thoracic, and lumbar region was examined (i.e., six vertebrae). Sample sizes were set to 10, 20, 30, and 40 (or 35) for comparison. For instance, for a sample size of 10 for the cranium, 10 individuals from 40 individuals, and 10 traits out of 595 interlandmark distances were randomly drawn with resampling 1000 times. Hence, there were 1000 ICV scores in each ICV distribution. To test the correlation between the empirical r^2 value and the mean and SD of ICV scores, a Spearman's rank correlation test was conducted for each sample size category. For statistical comparison, Mann–Whitney U tests were conducted between different sample sizes (e.g., 10 vs. 20) with Bonferroni adjustment. Test results were considered to be statistically significant when p-values were less than 0.0125. All simulations and statistical tests were conducted with r 3.4.4. and ICV scores were calculated using the 'CalcICV' function in the *evolqg* package (Melo et al. 2015) in r 3.4.4.

Results

Simulation Study

The results showed that means of ICV distributions could be reliably estimated for r^2 values of 0.05 or greater with sample sizes of about 100, for r^2 values of 0.08 or greater with sample sizes of about 55−60, and for r^2 values of 0.12 with sample sizes of about 40−45 (Figs. 3, 4, and 5). In

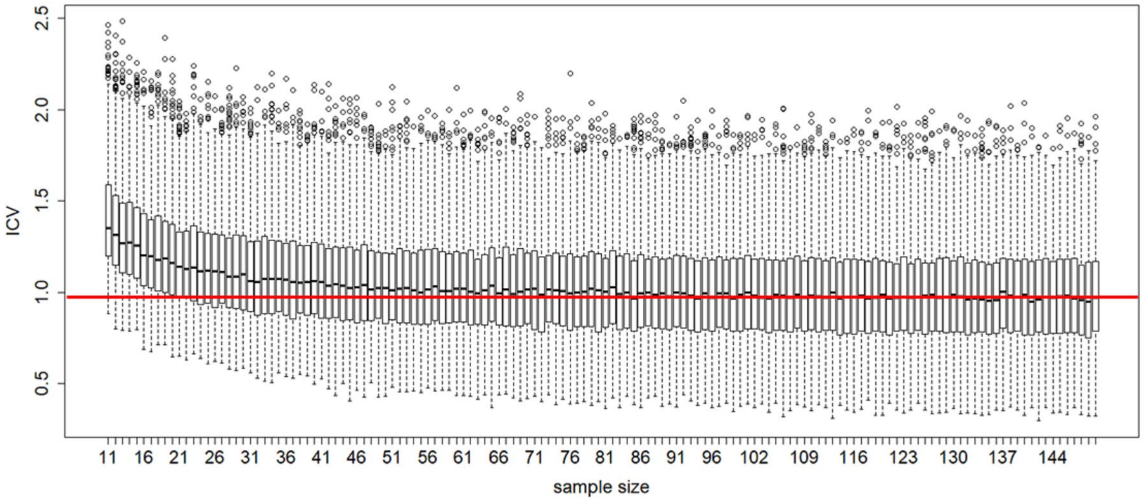

Fig. 3 Distributions of ICV values based on 10 traits resampled at random from 300 traits with varying sample sizes (n = 11–150) from a multivariate normal population of 10,000 individuals when average among trait correlation is $r^2 = 0.05$ (similar to the empirical r^2 for human cranial traits in Porto et al. 2009). Red straight line shows where sample size starts to approach an asymptote ICV value

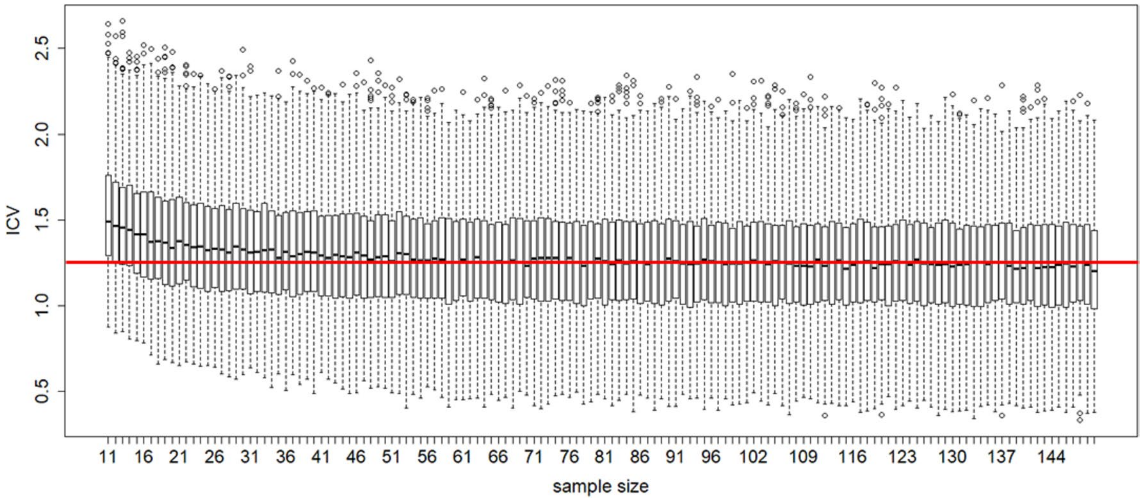

Fig. 4 Distributions of ICV values based on 10 traits resampled at random from 300 traits with varying sample sizes (n = 11–150) from a multivariate normal population of 10,000 individuals when average among trait correlation is $r^2 = 0.08$ (similar to the empirical r^2 for hominoid cranial traits in Porto et al. 2009). Red straight line shows where sample size starts to approach an asymptote ICV value

other words, when the mean trait r^2 value is relatively low, larger sample sizes are required to accurately estimate ICV values, and vice versa. On the other hand, a sample size of 40 individuals is large enough to accurately calculate the magnitude of integration when the average trait r^2 value is over 0.08. Moreover, with fairly high r^2 values ($r^2 > 0.2$ or > 0.35), about 20 or 30 individuals are sufficient for calculating reliable ICV means (Figs. 6 and 7). For the parameter r^2 value of 0.05, Mann–Whitney U tests were not statistically significant once sample sizes were over 51 (Table 1). For r^2 values of 0.08 and 0.12, Mann–Whitney U tests were not statistically significant once sample sizes were over 31. For r^2 values of 0.2 and 0.35, Mann–Whitney U tests were not statistically significant once sample sizes were over 11 (Table 1). Mann–Whitney U test results showed that even sample sizes with means above the ‘asymptote’ red line in Figs. 3, 4, 5, 6, and 7 are sufficient to reliably estimate mean ICV scores for some r^2 values.

Boxplots of ICV distributions based on various parameter r^2 values show that the mean ICV based on various sample sizes increased logarithmically with increasing parameter r^2 values (Fig. 8). Conversely, the standard deviation (SD) of mean ICV exponentially decreased with increasing parameter r^2 values (Fig. 9). Thus, there is a clear relationship

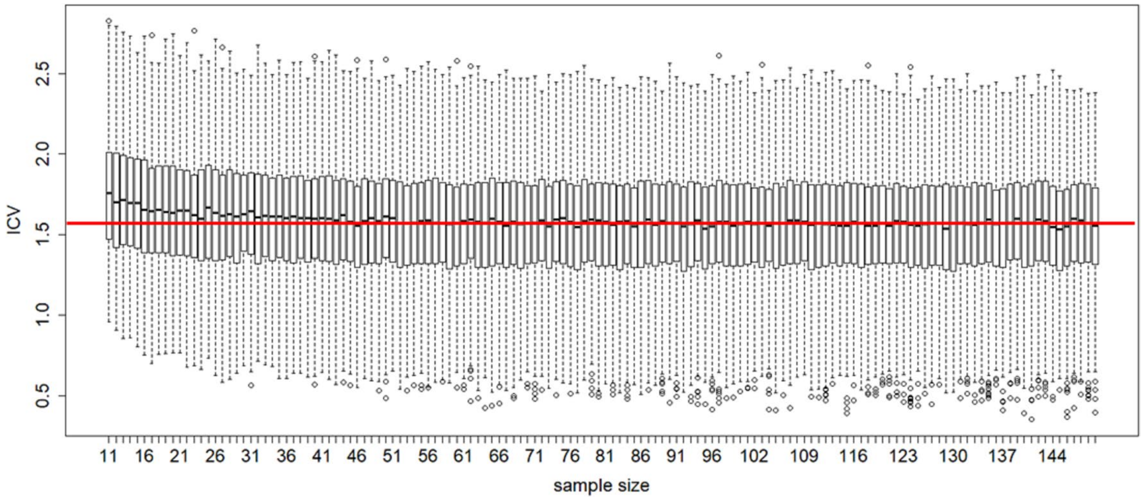

Fig. 5 Distributions of ICV values based on 10 traits resampled at random from 300 traits with varying sample sizes (n = 11–150) from a multivariate normal population of 10,000 individuals when average among trait correlation is $r^2 = 0.12$ (similar to the empirical r^2 for Old and New World monkey cranial traits in Porto et al. 2009). Red straight line shows where sample size starts to approach an asymptote ICV value

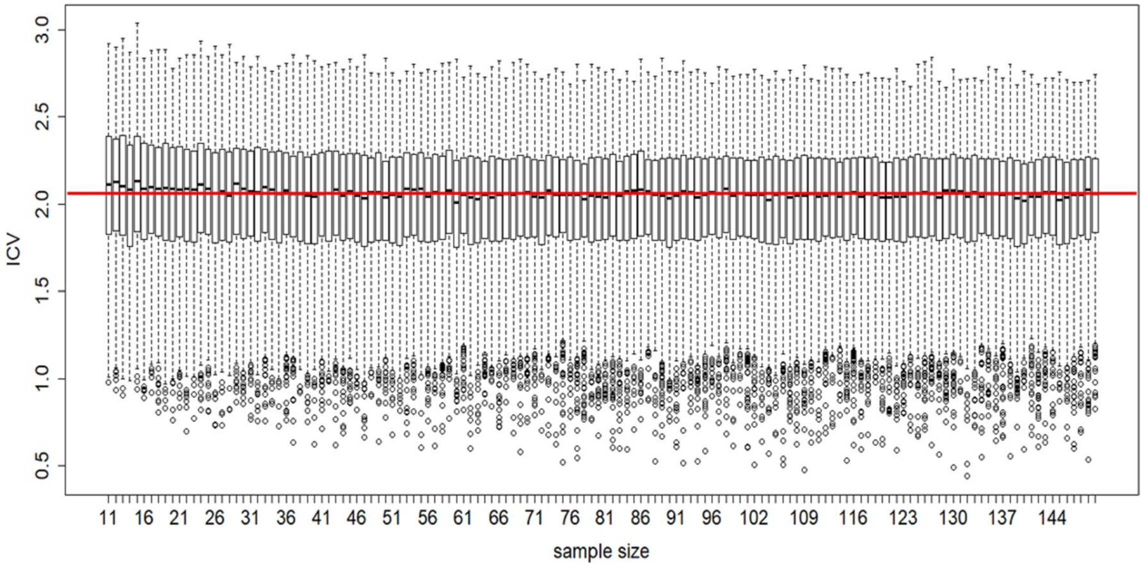

Fig. 6 Distributions of ICV values based on 10 traits resampled at random from 300 traits with varying sample sizes (n = 11–150) from a multivariate normal population of 10,000 individuals when average among trait correlation is $r^2 = 0.2$. Red straight line shows where sample size starts to approach an asymptote ICV value

between parameter r^2 values and the means and standard deviations of resultant ICV estimates, which explains why only relatively small sample sizes are required to reliably calculate ICV distribution using the resampling method if overall trait r^2 values are reasonably high.

The results of the simulations of the effects of total trait number, sample size, and number of resampled traits on mean ICV showed that mean ICV increased with more resampled traits irrespective of the starting total number of traits, sample size, or parameter r^2 value (Fig. 10). However, the rate at which mean ICV increased as the number of resampled traits increased was faster when the r^2 value was higher. Moreover, the mean ICV was mostly influenced by r^2 values and number of resampled traits, rather than sample size and the total number of traits (Fig. 10). Thus, it appears that the total number of traits per skeletal element does not affect the calculation of ICV distributions as long as the number of resampled traits is held constant between elements being compared.

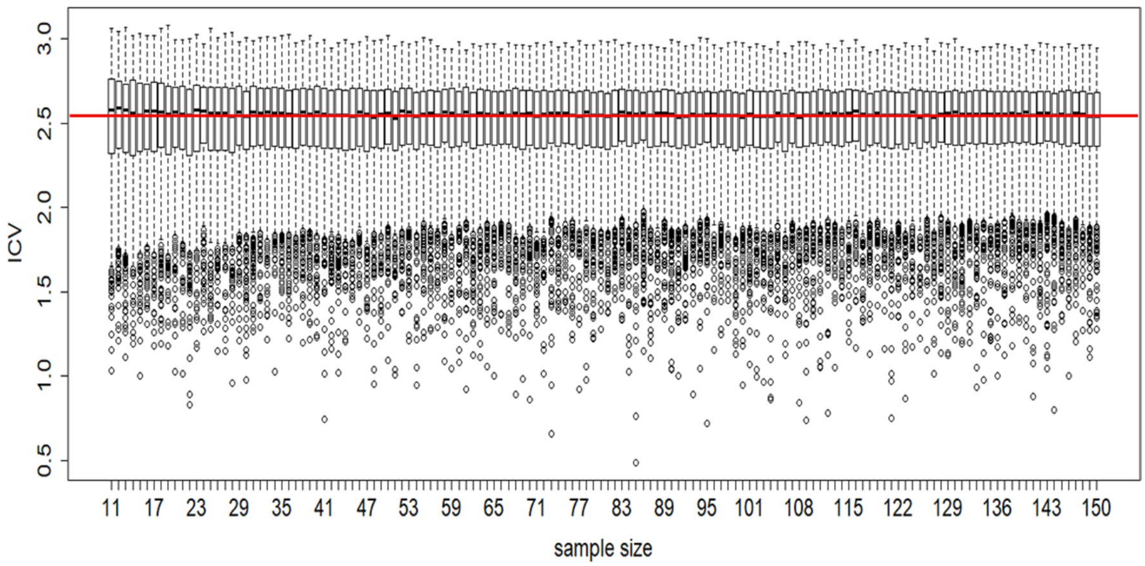

Fig. 7 Distributions of ICV values based on 10 traits resampled at random from 300 traits with varying sample sizes (n = 11–150) from a multivariate normal population of 10,000 individuals when average among trait correlation is $r^2 = 0.35$. Red straight line shows where sample size starts to approach an asymptote ICV value

Table 1 Mann–Whitney U test results between different sample sizes under different assumptions of r^2 values

	r^2 values				
Comparison of sample sizes	0.05	0.08	0.12	0.2	0.35
11 vs. 21	**U = 718,330; p < 0.0001**[*]	**U = 630,000; p < 0.0001**	**U = 608,050; p < 0.0001**	U = 530,910; p = 0.0166	U = 516,950; p = 0.1893
21 vs. 31	**U = 593,780; p < 0.0001**	**U = 547,050; p < 0.0001**	**U = 559,720; p < 0.0001**	U = 530,070; p = 0.0198	U = 528,270; p = 0.0286
31 vs. 41	U = 530,500; p = 0.0182	U = 519,410; p = 0.1328	U = 513,290; p = 0.3035	U = 484,220; p = 0.2216	U = 477,640; p = 0.0833
41 vs. 51	**U = 548,320; p = 0.002**	U = 514,950; p = 2469	U = 536,140; p = 0.005	U = 499,690; p = 0.9809	U = 504,820; p = 0.7087
51 vs. 61	U = 512,090; p = 0.3493	U = 505,760; p = 0.6554	U = 485,950; p = 0.2767	U = 513,800; p = 0.2853	U = 512,470; p = 0.3341
61 vs. 71	U = 516,650; p = 0.1974	U = 502,390; p = 0.8531	U = 532,310; p = 0.012	U = 490,560; p = 0.4647	U = 489,330; p = 0.4086
71 vs. 81	U = 509,790; p = 0.4483	U = 503,560; p = 0.7826	U = 487,720; p = 0.3417	U = 493,720; p = 0.6265	U = 515,450; p = 0.2315
81 vs. 91	U = 521,260; p = 0.01	U = 513,300; p = 0.3031	U = 503,060; p = 0.8126	U = 505,140; p = 0.6909	U = 492,580; p = 0.5655
91 vs. 101	U = 488,920; p = 0.3908	U = 491,430; p = 0.5071	U = 505,790; p = 0.6537	U = 515,580; p = 0.2277	U = 519,220; p = 0.1366
101 vs. 111	U = 488,380; p = 0.3681	U = 516,920; p = 0.1901	U = 504,480; p = 0.7284	U = 481,070; p = 0.1427	U = 480,950; p = 0.1402
111 vs. 121	U = 518,070; p = 0.1616	U = 497,420; p = 0.8416	U = 496,250; p = 0.7714	U = 505,730; p = 0.6573	U = 504,240; p = 0.7427
121 vs. 131	U = 504,580; p = 0.7229	U = 509,830; p = 0.4464	U = 506,270; p = 0.6274	U = 492,600; p = 0.5665	U = 486,360; p = 0.2909
131 vs. 141	U = 509,020; p = 0.4849	U = 520,400; p = 0.1141	U = 500,330; p = 0.9799	U = 520,390; p = 0.1144	U = 502,250; p = 0.8618

[*] Significant results are in bold when p-value of Mann–Whitney U test is < 0.0038

Skeletal elements of Macaca fascicularis

The empirical r^2 values for the 22 skeletal elements of *M. fascicularis* tested are presented in Table 2. Among skeletal elements, mean trait r^2 values ranged between 0.22 and 0.51 (Table 2). In general, postcranial elements showed higher r^2 values than the cranium and mandible. The lowest r^2 value was 0.22 in the first cervical vertebra (C1), while the highest r^2 value was 0.51 in the tibia. The average r^2 value across skeletal elements was 0.35. The correlation between r^2 values and mean ICV scores was significant for all sample sizes (sample size of 10: r = 0.758, p = 0.002; sample

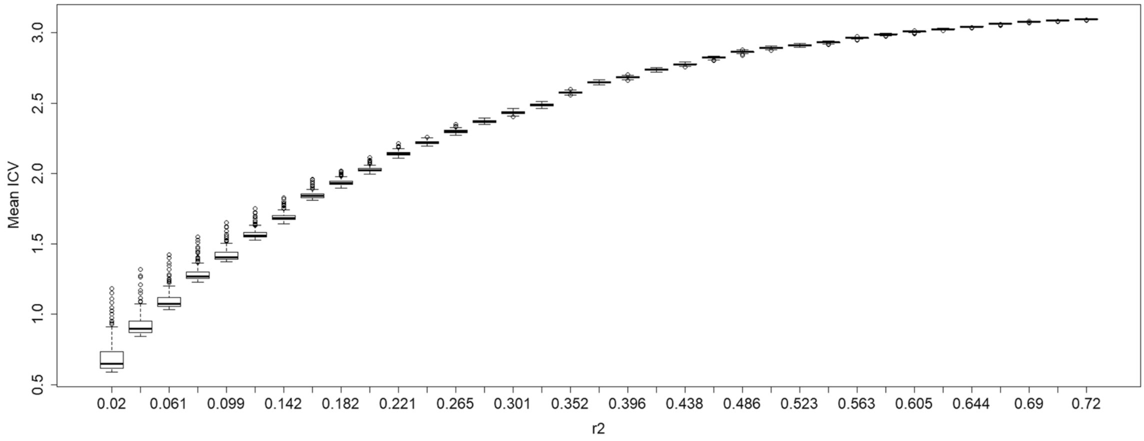

Fig. 8 Distributions of mean ICV of varying sample sizes (n = 11–150) based on various r^2 values (0.02–0.7)

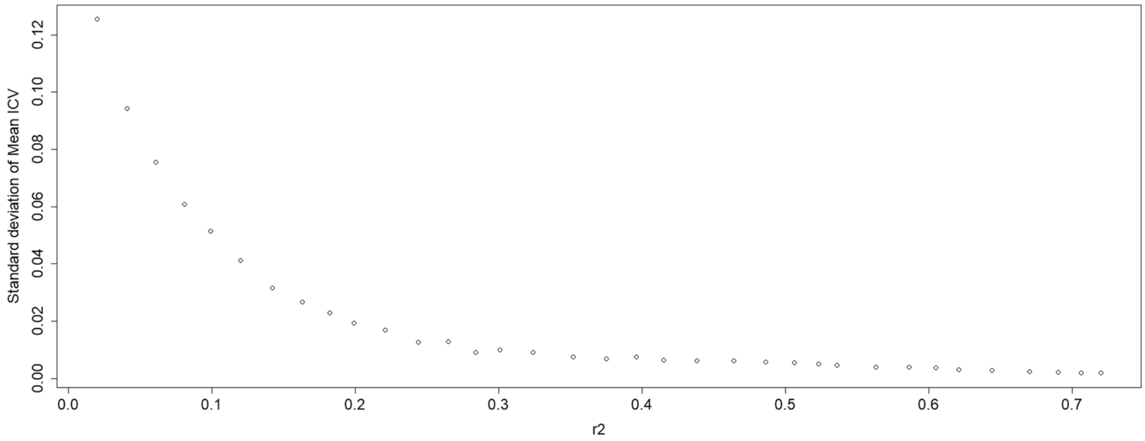

Fig. 9 Standard deviation of mean ICV of varying sample sizes (n = 11–150) based on various r^2 values (0.02–0.7)

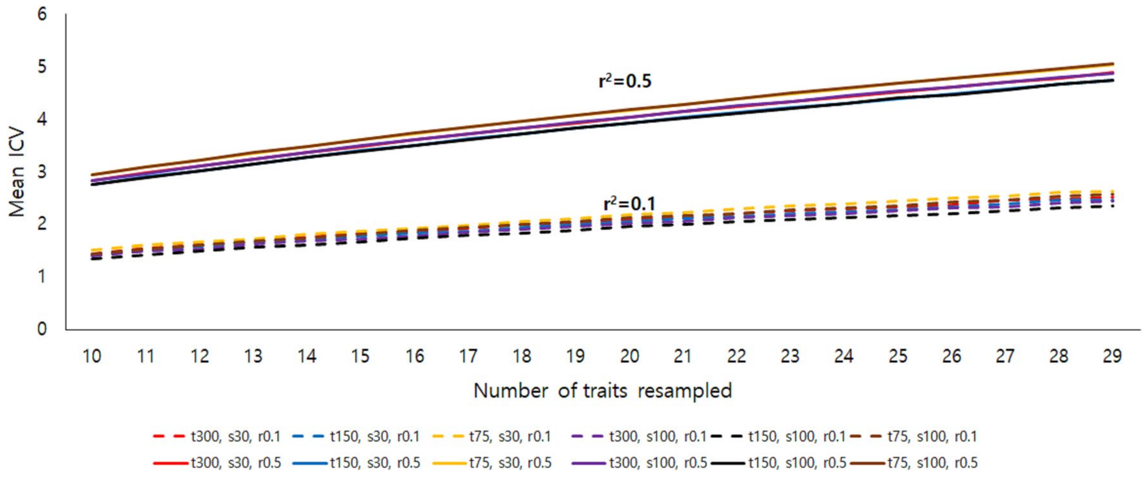

Fig. 10 Effect of total trait number, sample size, and number of traits resampled on mean ICV. *t* total trait number, *s* sample size, *r* r^2 value

Table 2 Coefficient of determination (r^2) values for skeletal elements in *Macaca fascicularis*

Skeletal element	r^2 value
Cranium	0.24
Mandible	0.29
C1	0.22
C2	0.31
C3	0.33
C5	0.31
C7	0.29
T1	0.33
T4	0.38
T7	0.35
T10	0.38
T12	0.44
L1	0.40
L4	0.41
L7	0.43
Sacrum	0.40
Scapula	0.30
Humerus	0.35
Radius	0.40
Os coxa	0.24
Femur	0.39
Tibia	0.51

size of 20: r = 0.725, p < 0.001; sample size of 30: r = 0.766, p < 0.001; sample size of 40 (or 35): r = 0.771, p < 0.001). The correlation between r^2 values and standard of deviations (SD) of ICV scores was also significant for all sample sizes (sample size of 10: r = − 0.547, p = 0.035; sample size of 20: r = − 0.58, p = 0.023; sample size of 30: r = − 0.649, p = 0.009; sample size of 40 (or 35): r = − 0.63, p = 0.012). Thus, mean ICV scores increased but SD of ICV scores decreased with increasing r^2 values.

The mean ICV scores for different skeletal elements of *M. fascicularis* generated using various sample sizes showed that for all skeletal elements there was no significant difference in ICV scores between sample sizes of 30 and 40 (or 35) (Table 3). In all skeletal elements, except for the radius and tibia, there was a significant difference in ICV between sample sizes of 10 and 20. Therefore, even very small samples were sufficient to accurately estimate the ICV for the radius and tibia, which reflect the fact that these skeletal elements had the highest r^2 values (0.4 and 0.51, respectively) among skeletal elements. For all other skeletal elements, barring the femur, scapula, and humerus, there was also a significant difference in ICV scores between sample sizes of 20 and 30, but for all bones tested, increasing the sample size above 30 did not have any significant impact on the estimation of distributions of ICV. Thus, in general, the results from the empirical analysis of *M. fascicularis* skeletal elements corroborate the computer simulation results in

Table 3 Means and standard deviations of ICV distributions generated using resampling method (1000 resamples of 10 traits) with various sample sizes for skeletal elements of *Macaca fascicularis*

Skeletal element	Sample size	Mean and standard deviation of ICV scores
Cranium*,#	10	1.95 (0.312)
	20	1.85 (0.267)
	30	1.80 (0.246)
	40	1.79 (0.227)
Mandible*,#	10	2.16 (0.302)
	20	2.09 (0.259)
	30	2.06 (0.246)
	40	2.04 (0.228)
C1*,#	10	1.79 (0.272)
	20	1.67 (0.219)
	30	1.62 (0.187)
	40	1.60 (0.165)
C7*,#	10	2.11 (0.283)
	20	2.01 (0.229)
	30	1.99 (0.203)
	40	1.97 (0.177)
T1*,#	10	2.34 (0.301)
	20	2.29 (0.258)
	30	2.28 (0.223)
	40	2.25 (0.207)
T12*,#	10	2.32 (0.261)
	20	2.27 (0.193)
	30	2.26 (0.160)
	40	2.25 (0.140)
L1*,#	10	2.23 (0.298)
	20	2.16 (0.242)
	30	2.16 (0.209)
	40	2.14 (0.195)
L7*,#	10	2.37 (0.243)
	20	2.28 (0.189)
	30	2.28 (0.157)
	40	2.28 (0.131)
Sacrum*,#	10	2.41 (0.313)
	20	2.36 (0.260)
	30	2.36 (0.227)
	40	2.36 (0.203)
Scapula*	10	2.35 (0.273)
	20	2.28 (0.239)
	30	2.26 (0.236)
	35	2.26 (0.222)
Humerus*	10	2.64 (0.274)
	20	2.61 (0.242)
	30	2.61 (0.225)
	35	2.60 (0.231)

Table 3 (continued)

Skeletal element	Sample size	Mean and standard deviation of ICV scores
Radius	10	2.97 (0.129)
	20	2.99 (0.096)
	30	2.99 (0.084)
	35	2.99 (0.078)
Os coxa*,#	10	2.08 (0.313)
	20	2.01 (0.276)
	30	1.96 (0.251)
	35	1.96 (0.248)
Femur*	10	2.93 (0.144)
	20	2.93 (0.146)
	30	2.94 (0.089)
	35	2.94 (0.083)
Tibia	10	3.07 (0.077)
	20	3.07 (0.052)
	30	3.07 (0.049)
	35	3.07 (0.049)

*Significant difference between sample size 10 and 20 when p-value of Mann–Whitney U test is <0.0125

#Significant difference between sample size 20 and 30 when p-value of Mann–Whitney U test is <0.0125

suggesting that sample sizes of 30−40 are sufficient to reliably calculate mean ICV scores in skeletal elements where the average r^2 values among traits are reasonably high.

Discussion

Interpretation of the simulation results

In general, the results corresponded to the previous study of Grabowski and Porto (2017) as both sample size and r^2 were important for calculating and estimating reliable indices of morphological integration. In this regard, it is advisable that a 'universal' sample size for different anatomical regions or skeletal elements be based on the lowest r^2 obtained for any anatomical region or skeletal element. For instance, if the cranium shows lower r^2 than post-cranial skeletal elements, the sampling effort should be based on the r^2 of the cranium, although post-cranial skeletal elements may show relatively higher r^2 values than the cranium. The simulation results showed that calculation of reliable mean ICV mostly depends on the value of the parameter r^2. When r^2 is 'moderate' (i.e., $r^2>0.08$), a sample size of 40 individuals is large enough to calculate the mean ICV using a trait resampling method. Moreover, standard deviations of mean ICV exponentially decreased as r^2 values increased (Fig. 9). Thus, the accuracy of the calculation of mean ICV is fairly reliable with high r^2 values. In contrast, when r^2 is low as is the case with the human cranium ($r^2=0.05$) found in previous studies (Porto et al. 2009; de Oliveira et al. 2009), a larger sample size is required to calculate stable mean ICV.

Nevertheless, the Mann–Whitney U test results showed that a sample size of 51 may be sufficient even with a low r^2 value of 0.05 for calculating reliable mean ICV scores using the trait resampling method (Table 1). Thus, for the human cranium ($r^2=0.05$), 51 individuals may be the bare minimum required to calculate reliable mean ICV based on the results of the present study, while at least 108 individuals are required to calculate reliable integration indices (e.g., r^2) based on the results of Grabowski and Porto (2017). The discrepancy between the results of Grabowski and Porto (2017) and the present study may depend on the methods used to quantify errors. In the present study, we used Mann–Whitney U tests to statistically compare significant differences in mean ICV between sample sizes to examine how many individuals are required to calculate a reliable mean ICV for each r^2 value. In contrast, Grabowski and Porto (2017) used three measures of error; bias, imprecision, and inaccuracy. For instance, inaccuracy was calculated as "Inaccuracy = Imprecision + Bias^2," where imprecision is "the distance of repeated measurements to each other", and bias is "the difference between the expected value of a parameter and the true parameter value." Or, inaccuracy can be calculated as "the distance of a measured value to its parameter value" (Grabowski and Porto 2017). They considered an inaccuracy value of 0.05 as a "cut-off" for calculating reliable integration indices in their simulation (Grabowski and Porto 2017). Hence, Grabowski and Porto's (2017) findings that much larger sample sizes are required may be related to their fairly strict criteria for calculating reliable integration indices. If we determine required minimum sample sizes based on the 'asymptote' lines in Figs. 3, 4, 5, 6, and 7, the present study demonstrates similar conclusions to the results of Grabowski and Porto (2017) (i.e., more than 100 individuals are required when $r^2=0.05$). However, the results of the Mann–Whitney U tests show a less strict criterion for sample size determination that may be applicable to empirical studies of morphological integration. Thus, in terms of required sample size, our results correspond better to those of Cheverud and colleagues (e.g., Ackermann 2009), which showed that sample sizes of 40 are necessary for accurately estimating the structure of V/CV matrices or calculating integration indices. At the very least, our simulation results suggest that a sample size of 30−40 when trait r^2 value is over 0.08 (e.g., cranium of hominoids) or a sample size over 51 when r^2 value is over 0.05 (e.g., cranium of humans) are the bare minimums required to accurately calculate magnitudes of integration using mean ICV scores (Table 1).

Moreover, as shown in previous studies (Grabowski and Porto 2017) and the present study, the number of (resampled) traits can significantly affect the index of morphological integration as different numbers of traits will change the average trait r^2 of a certain morphology. For instance, mean ICV increased when the number of resampled traits increased in this study (Fig. 10). Nevertheless, it was also shown that the total number of traits did not affect mean ICV values when the number of resampled traits remained the same (Fig. 10). This result shows the merit of the resampling method as different skeletal elements have different numbers of traits, which may artificially alter ICV values. For instance, two skeletal elements with smaller and larger number of traits, respectively, would automatically have different ICV scores as the smallest eigenvalue may be more skewed in an eigenvalue distribution from a larger number of traits, resulting in unintended inflation of ICV scores even if mean eigenvalues are the same between skeletal elements with different number of traits. Thus, calculating distributions of ICV scores using a resampling method is an effective way to compare magnitudes of integration between skeletal elements with different number of traits, such as different developmental and/or functional modules.

Empirical Data of *Macaca fascicularis*

The analysis of skeletal elements in *M. fascicularis* showed that average trait r^2 values ranged between 0.22 and 0.51 (Table 2). The correlation coefficient between r^2 values and mean ICV scores of skeletal elements ranged between 0.725 in the sample size of 20 and 0.771 in sample size of 40 (or 35). Although there is a clear relationship between r^2 and ICV measures of morphological integration, the ICV is more appropriate for calculating magnitudes of integration in V/CV matrices, while r^2 is better for the same purpose when using correlation matrices as suggested by Shirai and Marroig (2010). The average trait r^2 value for the cranium ($r^2 = 0.24$) was slightly higher than that found in a previous study ($r^2 \approx 0.17$; de Oliveira et al. 2009). This difference most likely reflects the different landmarks used to capture cranial traits and differences in sample composition used in the present and previous studies. In particular, the total number of traits for the cranium in this study was 595 interlandmark distances, which was much larger than trait sets used in previous studies. The r^2 value of the post-cranium was generally higher than the cranium except for C1, while the os coxa showed the same r^2 value as the cranium (Table 2). In this regard, the r^2 value for the cranium of mammals in previous studies (e.g., Marroig et al. 2009; Porto et al. 2009; de Oliveira et al. 2009) can be used as a guideline for deciding the minimum sample size required to reliably calculate the magnitude of integration using mean ICV scores. Based on the simulation results above, only about 10–20 specimens are required to analyze magnitudes of integration for the skeletal elements of *M. fascicularis* presented here (Table 1 and 2). However, it was found that sample sizes of 10–30 are required to reliably calculate mean ICV scores based on the results of the Mann–Whitney U tests comparing the effect of sample size on mean ICV for the skeletal elements of *M. fascicularis* (Table 3). The difference between the results of the simulation and real biological data is likely due to the structure of the data as a parameter V/CV matrix was generated in the simulation with multivariate normality but this is not always likely to be the case for real biological data. Thus, if one wants to examine mean ICV scores in all skeletal elements of *M. fascicularis*, a minimum sample size of 30 is recommended (Table 3).

A Cautionary Tale: The Relationship Between r^2 Values and Landmarking Protocol

As shown in this study, a posterior calculation of average trait r^2 values will be important and necessary for validating necessary sample sizes for analyzing magnitudes of integration using the ICV. Having said this, the results of the present study also suggest that how traits are determined, such as the landmarking protocol used, may also be an important issue for morphometric integration studies, in terms of estimation of the average r^2 value (Conaway et al. 2019). In previous studies and in the present study, one of the main ways in which researchers can decide on an appropriate sampling effort for reliably calculating integration indices is to evaluate the average trait r^2 value. Thus, it is also important to consider whether different (geometric) morphometric methods applied to skeletal elements have substantial effects on the calculation of r^2 values. In this respect, Conaway et al. (2019) reported that the r^2 can vary based on the landmarking protocol used to define morphometric traits, such as the use of different numbers of landmarks and/or inclusion of semi-landmarks in the analysis. In other words, the magnitude of integration could be different even within the same skeletal region due to differing landmarking protocols and, resultant varying r^2 values. Therefore, sampling effort should be considered not only prior to the study based on r^2 values from previous studies but also after the (preliminary) sampling is completed when r^2 can be recalculated posteriorly based on the specific traits (i.e., specific landmarking protocol) being employed in each analysis. In this regard, it is recommended to prepare to sample more than the minimum required sample sizes for specific skeletal elements of certain species (e.g., more than 51 individuals for the human cranium) as it is impossible to know a priori what the precise r^2 will be based on the specific morphometric protocol being employed and the sample composition. For instance, it may be that the actual r^2 value is lower than 0.05, which would require a much larger sample size than 51 individuals as the standard deviation of mean ICV exponentially increase with lower r^2 values (Fig. 9). In a nutshell, while the results of this simulation provide guidelines for minimum sample sizes based on average trait r^2 values, sampling effort should be tailored to the specific purpose (i.e., hypothesized developmental and/or functional module), trait definition method (i.e., landmarking protocol), and material (i.e., skeletal element) of each study.

Acknowledgement The authors are grateful to Mark Omura at the Museum of Comparative Zoology at Harvard University for granting assess to the primate collection for this study. We thank the SUNY Research Foundation for funding to support this research. This material is based upon work supported by the National Science Foundation under grant number BCS-1830745.

Author contributions HJ, MC, and NVC conceived the ideas and designed methodology. HJ and MC collected data. HJ ran computer simulation, analyzed data, and wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Availability Data available from the Dryad Digital Repository https://doi.org/10.5061/dryad.p8cz8w9m6 (Jung, Conaway, & von Cramon-Taubadel, 2020).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ackermann, R. R. (2009). Morphological integration and the interpretation of fossil hominin diversity. *Evolutionary Biology, 36*(1), 149–156.

Ackermann, R. R., & Cheverud, J. M. (2000). Phenotypic covariance structure in tamarins (genus *Saguinus*): A comparison of variation patterns using matrix correlation and common principal component analysis. *American Journal of Physical Anthropology, 111*(4), 489–501.

Ackermann, R. R., & Cheverud, J. M. (2002). Discerning evolutionary processes in patterns of tamarin (genus *Saguinus*) craniofacial variation. *American Journal of Physical Anthropology, 117*(3), 260–271.

Adams, D. C. (2016). Evaluating modularity in morphometric data: Challenges with the RV coefficient and a new test measure. *Methods in Ecology and Evolution, 7*(5), 565–572.

Armbruster, W. S., Pélabon, C., Bolstad, G. H., & Hansen, T. F. (2014). Integrated phenotypes: Understanding trait covariation in plants and animals. *Philosophical Transactions of the Royal Society B, 369*(1649), 20130245.

Arnold, P., Forterre, F., Lang, J., & Fischer, M. S. (2016). Morphological disparity, conservatism, and integration in the canine lower cervical spine: Insights into mammalian neck function and regionalization. *Mammalian Biology-Zeitschrift für Säugetierkunde, 81*(2), 153–162.

Botton-Divet, L., Houssaye, A., Herrel, A., Fabre, A. C., & Cornette, R. (2018). Swimmers, diggers, climbers and more, a study of integration across the mustelids' locomotor apparatus (Carnivora: Mustelidae). *Evolutionary Biology, 45*(2), 182–195.

Cheverud, J. M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology, 110*, 155–171.

Cheverud, J. M. (1996). Developmental integration and the evolution of pleiotropy. *American Zoologist, 36*(1), 44–50.

Cheverud, J. M., & Marroig, G. (2007). Comparing covariance matrices: Random skewers method compared to the common principal components model. *Genetics and Molecular Biology, 30*(2), 461–469.

Conaway, M. A., Jung, H., & von Cramon-Taubadel, N. (2019). The effects of morphometric protocol on morphological integration statistics: A case study in scapulae. *American Journal of Physical Anthropology, 168*, 47–47.

Conaway, M. A., Schroeder, L., & von Cramon-Taubadel, N. (2018). Morphological integration of anatomical, developmental, and functional postcranial modules in the crab-eating macaque (*Macaca fascicularis*). *American Journal of Physical Anthropology, 166*(3), 661–670.

de Oliveira, F. B., Porto, A., & Marroig, G. (2009). Covariance structure in the skull of Catarrhini: A case of pattern stasis and magnitude evolution. *Journal of Human Evolution, 56*(4), 417–430.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics, 29*, 751–760.

Goswami, A., Smaers, J. B., Soligo, C., & Polly, P. D. (2014). The macroevolutionary consequences of phenotypic integration: From development to deep time. *Philosophical Transactions of the Royal Society B, 369*(1649), 20130254.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society, London B, 205*(1161), 581–598.

Grabowski, M., & Porto, A. (2017). How many more? Sample size determination in studies of morphological integration and evolvability. *Methods in Ecology and Evolution, 8*(5), 592–603.

Hallgrímsson, B., Jamniczky, H., Young, N. M., Rolian, C., Parsons, T. E., Boughner, J. C., et al. (2009). Deciphering the palimpsest: Studying the relationship between morphological integration and phenotypic covariation. *Evolutionary Biology, 36*(4), 355–376.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis, 97*(10), 2177–2189.

Jones, K. E., Benitez, L., Angielczyk, K. D., & Pierce, S. E. (2018). Adaptation and constraint in the evolution of the mammalian backbone. *BMC Evolutionary Biology, 18*(1), 172.

Kazi-Aoual, F., Hitier, S., Sabatier, R., & Lebreton, J. D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics & Data Analysis, 20*(6), 643–656.

Kelly, E. M., Marcot, J. D., Selwood, L., & Sears, K. E. (2019). The development of integration in marsupial and placental limbs. *Integrative Organismal Biology, 1*(1), oby13.

Klingenberg, C. P. (2009). Morphometric integration and modularity in configurations of landmarks: Tools for evaluating a priori hypotheses. *Evolution & Development, 11*(4), 405–421.

Klingenberg, C. P. (2014). Studying morphological integration and modularity at multiple levels: Concepts and analysis. *Philosophical Transactions of the Royal Society B, 369*(1649), 20130249.

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

Marroig, G., & Cheverud, J. M. (2004). Cranial evolution in sakis (*Pithecia*, Platyrrhini) I: Interspecific differentiation and allometric patterns. *American Journal of Physical Anthropology, 125*(3), 266–278.

Marroig, G., Shirai, L. T., Porto, A., de Oliveira, F. B., & De Conto, V. (2009). The evolution of modularity in the mammalian skull

II: Evolutionary consequences. *Evolutionary Biology, 36*(1), 136–148.

Melo, D., Garcia, G., Hubbe, A., Assis, A. P., & Marroig, G. (2015). EvolQG-An R package for evolutionary quantitative genetics. *F1000Research, 4*, 925.

Olson, E. C., & Miller, R. L. (1958). *Morphological integration*. Chicago: University of Chicago Press.

Penna, A., Melo, D., Bernardi, S., Oyarzabal, M. I., & Marroig, G. (2017). The evolution of phenotypic integration: How directional selection reshapes covariation in mice. *Evolution, 71*(10), 2370–2380.

Porto, A., de Oliveira, F. B., Shirai, L. T., De Conto, V., & Marroig, G. (2009). The evolution of modularity in the mammalian skull I: Morphological integration patterns and magnitudes. *Evolutionary Biology, 36*(1), 118–135.

Porto, A., Shirai, L. T., de Oliveira, F. B., & Marroig, G. (2013). Size variation, growth strategies, and the evolution of modularity in the mammalian skull. *Evolution, 67*(11), 3305–3322.

Qiu, W., Joe, H., & Qiu, M. W. (2006). The clusterGeneration package.

Randau, M., & Goswami, A. (2017). Morphological modularity in the vertebral column of Felidae (Mammalia, Carnivora). *BMC Evolutionary Biology, 17*(1), 133.

Roff, D. A. (1995). The estimation of genetic correlations from phenotypic correlations: A test of Cheverud's conjecture. *Heredity, 74*(5), 481.

Rohlf, F. J., & Corti, M. (2000). Use of two-block partial least-squares to study covariation in shape. *Systematic Biology, 49*(4), 740–753.

Rolian, C. (2014). Genes, development, and evolvability in primate evolution. *Evolutionary Anthropology, 23*(3), 93–104.

Shirai, L. T., & Marroig, G. (2010). Skull modularity in neotropical marsupials and monkeys: Size variation and evolutionary constraint and flexibility. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 314*(8), 663–683.

Wiley, D. F., Amenta, N., Alcantara, D. A., Ghosh, D., Kil, Y. J., Delson, E., et al. (2005). Evolutionary morphing.

Young, N. M., Wagner, G. P., & Hallgrímsson, B. (2010). Development and the evolvability of human limbs. *Proceedings of the National Academy of Sciences, 107*(8), 3400–3405.

Zelditch, M. L., & Carmichael, C. (1989). Ontogenetic variation in patterns of developmental and functional integration in skulls of *Sigmodon fulviventer*. *Evolution, 43*(4), 814–824.