Evolutionary Biology (2021) 48:190–220
https://doi.org/10.1007/s11692-021-09531-3

RESEARCH ARTICLE

On the Misidentification of Species: Sampling Error in Primates and Other Mammals Using Geometric Morphometrics in More Than 4000 Individuals

Andrea Cardini[1,2] · Sarah Elton[3] · Kris Kovarovic[3] · Una Strand Viðarsdóttir[4] · P. David Polly[5]

Received: 8 October 2020 / Accepted: 8 January 2021 / Published online: 26 February 2021

Abstract

An accurate classification is the basis for research in biology. Morphometrics and morphospecies play an important role in modern taxonomy, with geometric morphometrics increasingly applied as a favourite analytical tool. Yet, really large samples are seldom available for modern species and even less common in palaeontology, where morphospecies are often identified, described and compared using just one or a very few specimens. The impact of sampling error and how large a sample must be to mitigate the inaccuracy are important questions for morphometrics and taxonomy. Using more than 4000 crania of adult mammals and taxa representing each of the four placental superorders, we assess the impacts of sampling error on estimates of species means, variances and covariances in Procrustes shape data using resampling experiments. In each group of closely related species (mostly congeneric), we found that a species can be identified fairly accurately even when means are based on relatively small samples, although errors are frequent with fewer specimens and primates more prone to inaccuracies. A precise reconstruction of similarity relationships, in contrast, sometimes requires very large samples (> 100), but this varies widely depending on the study group. Medium-sized samples are necessary to accurately estimate standard errors of mean shapes or intraspecific variance covariance structure, but in this case minimum sample sizes are broadly similar across all groups (≈ 20–50 individuals). Overall, thus, the minimum sample sized required for a study varies across taxa and depends on what is being assessed, but about 25–40 specimens (for each sex, if a species is sexually dimorphic) may be on average an adequate and attainable minimum sample size for estimating the most commonly used shape parameters. As expected, the best predictor of the effects of sampling error is the ratio of between- to within-species variation: the larger the ratio, the smaller the sample size needed to obtain the same level of accuracy. Even though ours is the largest study to date of the uncertainties in estimates of means, variances and covariances in geometric morphometrics, and despite its generally high congruence with previous analyses, we feel it would be premature to generalize. Clearly, there is no a priori answer for what minimum sample size is required for a particular study and no universal recipe to control for sampling error. Exploratory analyses using resampling experiments are thus desirable, easy to perform and yield powerful preliminary clues about the effect of sampling on parameter estimates in comparative studies of morphospecies, and in a variety of other morphometric applications in biology and medicine. Morphospecies descriptions are indeed a small piece of provisional evidence in a much more complex evolutionary puzzle. However, they are crucial in palaeontology, and provide important complimentary evidence in modern integrative taxonomy. Thus, if taxonomy provides the bricks for accurate research in biology, understanding the robustness of these bricks is the first fundamental step to build scientific knowledge on sound, stable and long-lasting foundations.

Keywords Cranium · Landmark configuration · Multivariate mean · Procrustes shape · Species identification · Taxonomic assessment · Variance–covariance

Introduction

Taxonomy, the naming and classification of organisms, is seen by some as unfashionable and taxonomic expertise is vanishing quickly from natural history museums and other

✉ Andrea Cardini
alcardini@gmail.com; andrea.cardini@unimore.it

Extended author information available on the last page of the article

institutions (Drew 2011). Yet, this ancient branch of biology is today more crucial than ever. We are losing species at a rate comparable to that of the great mass extinctions (Ceballos et al. 2015, 2017) and humans are modifying the planet with the strength of a geological force, with unpredictable but likely negative consequences for most living beings (Lewis and Maslin 2018) including ourselves (Whitmee et al. 2015). Conservationists and ecologists need accurate taxonomic knowledge. They are not alone: a taxonomic foundation underpins all fields of biology and is crucial even for medical doctors who face new diseases emerging from disrupted ecosystems (Olival et al. 2017; Mollentze and Streicker 2020; Rodriguez-Morales et al. 2020). Sir Robert May's famous statement on the centrality of taxonomy in biology remains as current as ever: "without taxonomy to give shape to the bricks [i.e., organisms]…the house of biological science is a meaningless jumble" (May 1990, p. 130). Yet, delimiting taxonomic boundaries remains a complex and sometimes contentious issue, with grey areas which may elude the application of any general species concept (Zachos 2016).

A fundamental operational step in taxonomy is species description and identification. Descriptions were traditionally based on morphology, which is still the main source of information for identification in the field. In palaeontology, taxonomy is overwhelmingly based on morphology, so that the vast majority of fossil species are in fact morphospecies (Simpson 1943, 1951; Harrison 1993). Genetics has become increasingly important for assessing taxonomy, but DNA evidence is only available for modern species and recent subfossils. Ideally, multiple lines of evidence should be taken into account to accurately describe and identify a species. This type of "integrative taxonomy" (Dayrat 2005; Padial et al. 2010) is just 15 years old formally (i.e., since the name has been proposed, although a 'total evidence' approach is much older—e.g., Kluge 1989). However, it has encountered a slowly but constantly growing popularity: searching for references in google scholar (on January 5th 2021) using "integrative taxonomy" AND "species description", the number of entries retrieved for 2005, 2012 and 2019 is 4, 40 and 137 respectively, and the total number, since 2005, when the name was coined, is 975.

Often, integrative taxonomy employs molecular evidence together with morphometric analysis. Geometric morphometrics (Rohlf and Marcus 1993; Zelditch et al. 2012; Cardini and Loy 2013), a combination of image analysis and multivariate statistics, is particularly suitable to this aim, because it is relatively simple but at the same time powerful and effective in data collection and visualization (Adams et al. 2004, 2013). In mammals, one of the taxonomically best studied groups of animals, successful applications of integrative approaches are common. For instance, a combination of molecular and morphometric analyses, together with behavioural and biogeographic data, has brought to the recent discovery of cryptic diversity, and thus a new species, among orangutans, possibly one of the most studied, as well as endangered, genera of primates (Nater et al. 2017). Comparing groups using morphometrics is an important tool in species assessment also in palaeontology. Although fossil species rarely have large comparative samples and are often defined by meristic apomorphies (e.g., the number of molar or premolar cusps, or the presence or absence of foramina), quantitative studies of continuous traits are not uncommon and are in fact routinely used, for instance, in palaeoanthropology. Indeed, the whole field of virtual anthropology originated from the use of a mix of geometric morphometrics and 3D imaging techniques, with extensive applications to reconstruct and compare fragmentary material to produce results unachievable with traditional non-quantitative methods (e.g., Hublin et al. 2009, 2017). More generally, in recent years, taxonomists have turned increasingly frequently to geometric morphometrics to assign living or fossil specimens to species-level taxa using clustering methods, discriminant functions or other morphometric analyses, with variable accuracy depending on the variability and overlap of the morphologies of the species in question as well as the morphometric sample available to the researcher (e.g., Polly and Head 2004; McGuire 2011; Boroni et al. 2017; Fang et al. 2018).

A taxonomy that is as accurate and stable as possible is typically seen as a prerequisite for measuring the loss of biodiversity and for setting conservation priorities. We cannot protect species we do not know and we cannot say if we have lost a species until we describe it: knowing whether, for instance, the Florida panther is a species, subspecies or just a recently isolated population of pumas can make a difference in deciding if and how to preserve it (Culver et al. 2000). Yet, some argue that the relationship between taxonomy and conservation is more complicated than usually depicted (Zachos 2018), and its role in palaeontology may seem even less clear and pressing. So why does it matter that we understand the limits of morphological analysis for taxonomic delimitation in living but also in extinct lineages? The study and naming of fossil species, besides being of intrinsic interest on its own and clearly central, also contributes in a fundamental way to our understanding of the current biodiversity crisis. To assess whether the modern day rate of species extinction is unusually high, we need to compare it to the background extinction rate, which can only be estimated from fossil species, which must therefore be identified and counted in the same way as extant ones (Barnosky et al. 2011). Comparability between taxonomy in the living and fossil records is also needed to reconstruct when the extinction crisis began. For instance, to search for the causes of the end of the Pleistocene megafaunal extinctions, the timing and number of megafaunal species extinctions have

been used to understand whether they coincided with the arrival of humans in a region (Barnosky et al. 2004; Koch and Barnosky 2006). Estimates of species numbers have also allowed to infer the Earth megafaunal carrying capacity and thus demonstrate the biomass trade-off between the rapidly disappearing large species of wild terrestrial vertebrates and the increasing size of the human population and its livestock (Barnosky 2008).

Understanding the impact of humans on the environment and on other species requires reconstructing our own evolutionary history, which in turn depends on finding, studying and classifying our closest extinct relatives (Harrison 1993; Wood 2010; White 2014). This is again partly a taxonomic endeavour, which mostly relies on the assessment of morphospecies and their evolutionary relationships (Wood et al. 2020). Because DNA evidence is lacking for most fossils, both their classification and evolutionary relationships are largely inferred using quantitative analyses of bone morphology. Species diagnosis, in particular, is, in palaeontology, mostly "a phenetically derived morphotype that serves to distinguish the species from all other closely related" ones (Harrison 1993, p. 363). Thus, as mentioned, fossil species may be defined and compared using meristic phenotypic traits but also by employing morphometrics to quantify similarity relationships (i.e., evolutionary grades) and assess whether fossils represent the same or different species or maybe a new, previously unknown, one. Clearly, given the paucity and often fragmentary nature of fossil material (Simpson 1951; Albrecht and Miller 1993; Godfray et al. 2004), scattered across many continents and over an evolutionary timescale of millions of years, this is no easy task. Both taxonomic deflation (Benton 2008) or inflation (Alroy 2002) can happen, with disagreement about the occurrence of one or the other phenomenon even within a single most studied fossil lineage such as the hominins (Tattersall 1986, 1993; Albrecht and Miller 1993; Martin and Andrews 1993; White 2014). In fact, all species are hypotheses (Dayrat 2005), and therefore morphospecies are just a piece of evidence in a much more complex puzzle (Simpson 1943). In this context, if one also bears in mind that species boundaries may be fuzzy and uncertain even in living taxa, it seems likely that taxonomic assessment using a single source of evidence, such as morphology, may be prone to errors both in modern and fossil lineages.

Among the multifarious sources of errors in the assessment of morphospecies, one that afflicts all taxonomic studies is sampling error, that arises because of limited numbers of specimens. This observation is almost tautological, because a morphologically-defined taxonomic species is in fact "an inference…of the morphological species from which a given series of specimens has been drawn" (Simpson 1943, p. 148). Thus, using small and poorly representative samples makes conclusions from morphological studies particularly uncertain and potentially biased (Simpson 1943; Cope and Lacy 1992). An extreme example is the use of a type specimen to describe a species, which, despite the good practical reasons for this convention (Witteveen 2015), misses out the often huge variability in a population (Simpson 1940, 1951; Dayrat 2005). However, even when we adopt a population perspective to taxonomic assessment (Simpson 1940; Newell 1949), we must inevitably draw conclusions from descriptive statistics, based on sample averages, variances etc. (Simpson 1943). These statistics provide the basis for comparing populations, and are therefore central to the assessment of morphospecies, but they are also behind state of the art morphometric analyses in evolutionary and biomedical research ranging from the application of comparative methods (Monteiro 2013) to studies of modularity and integration (Klingenberg 2013), evolutionary trends (e.g., Cardini 2019a) and human evolution (O'Higgins 2000), ecomorphology (e.g., Meloro et al. 2017), forensics (e.g., Franklin et al. 2007) and medicine (e.g., Sanfilippo et al. 2009).

In this research, using cranial landmarks from a total sample of more than 4000 specimens of living mammals, representing a variety of placental orders, we explore the impact of sample size on key morphometric parameters involved in the assessment of morphospecies. As our main interest is the delimitation between closely related species, whose boundaries mark the grey areas of alpha taxonomy, we divided our data set into small clades of closely related species (mostly genera) and used the member species with the largest sample as the focal species (FS). In this species, we investigated how sampling might affect the mean and variance–covariance structure of Procrustes shape data. However, unlike our previous studies on sampling error in Procrustean geometric morphometrics (Cardini and Elton 2007; Cardini et al. 2015), we did not also analyse size. Size is as important as shape, but it is a simpler variable and generally less impacted than shape by sampling (e.g., Cardini et al. 2015) and measurement error (Cardini 2014; Cardini and Chiapelli 2020). More importantly, size is univariate and the ways it might be affected by sampling error are similar to those of other biological variables such as body mass, height, width or length. Procrustes shape coordinates, in contrast, are more complex, not only because of the multivariate nature of shape but also because the Procrustes superimposition alters the covariance structure of the data (see Lele 1991; Rohlf 1998; O'Higgins 2000; Cardini 2019b). Thus, within each clade, we drew many random subsamples of its FS and, for each subsample of the same smaller size, we estimated the mean, variances and covariances of the Procrustes shape coordinates. Instead of directly using these statistics, to assess how taxonomic decisions would be affected by small sample size, and to also enhance comparability of our results to other systems, we develop several 'indices' that express variability in the FS subsample parameters relative to one another and to the

other species of its clade. As the resampling experiments produced a huge set of results, we synthesized the results from all the clades and investigated whether the impacts of sampling error are generalizable or whether and why they differ idiosyncratically from group to group. Finally, as we look for consistent patterns, we will discuss the answer to a most asked question in morphometrics, taxonomic and palaeontological research: how many specimens do I need?

Materials and Methods

Summary of Specific Goals

As the study design is complex, we first outline here briefly what we did and, later in the specific sections of the methods, provide more details, starting with an example to clarify the design and terminology.

We employed a large sample of mammals (Table 1), that included at least one lineage representing each of the four placental superorders. We measured in 3D adult crania using a configuration of 34 anatomical landmarks (Fig. 1). From this configuration, we also selected two reduced configurations, in which we replicated the entire analysis to preliminarily explore the sensitivity of results to the number and specific choice of landmarks. In each lineage we selected the species with the largest sample as the focal study species. In this species, we performed five series of randomized subsampling experiments. In these random subsamples, we calculated six different indices that estimate the impact of sampling error on means, variances and covariances both in relation to interspecific mean differences in a lineage and in relation to individual variability in the specific FS. However, because two of the three series of randomization experiments using the total configuration produced results highly congruent with the third main experiment on this configuration, we will focus on the latter, as well as on the two series using reduced configurations.

Example Explaining the Study Design and Terminology

The FS is the species for which we have the largest sample in a lineage. For instance, for the *Equus* clade the FS is *E. burchellii* (plains zebras), which has a sample size (N) of 103 specimens. From the total FS sample (the 'parent' or total sample), to estimate the effect of sampling error, we extracted 500 random subsamples of progressively smaller size (e.g., 500 subsamples of 50 specimens, then 500 of 20 specimens, then again another 500 of five etc.). As N is reduced, the effect of sampling error on the mean shape, variance, and covariance structure, becomes more pronounced. This is illustrated in Fig. 2, where a cluster analysis and PCA of the mean shapes of the random subsamples and the total samples of the *Equus* clade are shown. For the sake of clarity but also to balance the number of observed and subsample taxa in the graphic, we only show seven subsamples (the same as the number of taxa in the clade) instead of all 500. However, to visualize the full range of variation in the 500 subsamples, we selected the most distinctive means (i.e., those with the largest Procrustes shape distance to the mean of the total sample of *E. burchellii*). When the subsample size is large, they all cluster close to the total plains zebra sample, but, when N drops to five individuals, the error in estimating the mean shape is so large that five out of seven subsamples end up completely separated from all other taxa (including their parent FS, *E. burchellii*). Indeed, at N = 5 the disparity between subsamples becomes almost as great as found in the whole genus. An intermediate subsample size of N = 20 is, however, sufficient for the subsample means to cluster all together with their own FS, although the branch lengths in the dendrograms and convex hulls in PCA ordinations indicate that there is still a fair amount of variability due to sampling error. This figure (Fig. 2), and complementary ones for the other clades found in the supplementary data (Supplementary Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13), provide intuitive summaries of the effects of sample size, that complement the other analytical results presented in this paper.

Because we do not know the true means, variances and covariances of the species, we recognize that we are not assessing true "accuracy" (i.e., how close the estimates are to the truth). We therefore adopt the term relative accuracy to refer to our best proxy for accuracy, namely how close results from the subsamples are to their parent FS (whose N ranges from 44 to 281, with a median sample size of 88) and in relation to the other taxa in the analysis. We reserve the term "precision" to describe how close estimates of the random subsamples of a given size are to one another. Thus, for instance, a group of subsample means that cluster closely with one another but not with their parent FS (e.g., the five most distant means in plains zebras subsamples of N = 5, Fig. 2a) would be described as providing precise estimates of the mean but ones with a very low relative accuracy. On the other hand, if very close to the parent FS but far from one another, they would have higher relative accuracy but lower precision.

Landmarks, Shape Coordinates and Study Samples

The cranial landmark configuration is shown in Fig. 1. A detailed description of the configuration is available in previous studies (Cardini and Polly 2013; Cardini 2019a). Shape coordinates were computed in MorphoJ (Klingenberg 2011) using a Procrustes superimposition (Rohlf and Slice 1990).

Table 1 Summary of sample composition showing FS sample sizes as well as the average N of other species in the same group; bold is used for emphasizing N in the simulations (with separate sexes for primates and pooled females and males in all other groups)

Superorder	Lineage	Genus	Species	Abbreviation	FS common name	Samples	F	M	U	Total	All species
Euarchontoglires	African colobines	*Piliocolobus*	*ellioti*	Pil_eli	Elliot's red colobus	N	**65**	**44**		109	
			Other 14 sp.			Mean N	**21**	**15**		34	586
	Guenons	*Cercopithecus*	*mitis*	Cer_mit	Blue monkeys	N	**67**	**78**		145	
			Other 21 sp.			Mean N	**19**	**22**		41	1011
	Papionins	*Macaca*	*fascicularis*	Mac_fas	Crab-eating macaque	N	**184**	**281**		465	
			Other 18 sp.			Mean N	**10**	**10**		20	819
		Papio	*anubis*	Pap_anu	Anubis baboon	N	**54**	**123**		177	
			Other 5 sp.			Mean N	**7**	**29**		36	357
	Leporids	*Lepus*	*europaeus*	Lep_eur	European hare	N	45	49	7	**101**	
			Other 19 sp.			Mean N	5	5	8	**14**	367
Laurasiatheria	Erinaceids	*Erinaceus*	*europaeus*	Eri_eur	European hedgehog	N	40	52	34	**126**	
			Other 4 sp.			Mean N	16	10	5	**29**	212
	Equids	*Equus*	*burchellii*	Equ_bur	Plains zebra	N	59	41	3	**103**	
			Other 6 sp.			Mean N	9	6	1	**16**	200
	Canids	*Vulpes*	*vulpes*	Vul_vul	Red fox	N	63	71	19	**153**	
			Other 8 sp.			Mean N	4	5	8	**14**	263
Afrotheria	Hyraxes	*Procavia*	*capensis*	Pro_cap	Rock hyrax	N	6	6	43	**55**	
			Other 3 sp.			Mean N	8	12	18	**38**	168
Xenarthra	Armadillos	*Dasypus*	*novemcinctus*	Das_nov	Nine banded armadillo	N	4	1	54	**59**	
			Other 4 sp.			Mean N		1	2	**2**	66
		Grand total				N	1697	1925	427		4049

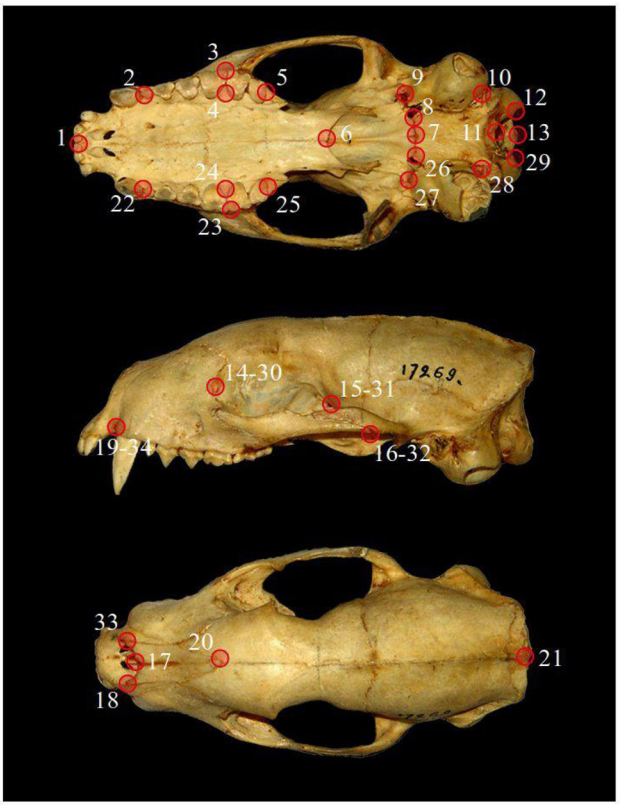

Fig. 1 Landmark configuration. As detailed in Cardini (2019a, b), landmarks were digitized by a single operator (AC) on the left side of the cranium; very small midplane asymmetries were removed; and the right side was reconstructed by mirroring following Cardini (2017)

Study samples and their descriptive statistics are detailed in the supplementary information (Tables S1, S2), but the sample composition is also summarized in Table 1. Each sample consists of adults (ca. 95% of which taken from the wild) of closely related species, mostly genera, of placental mammals. Overall, they belong to ten different clades, with at least one taxon representing each of the four placental superorders. Specifically, we analysed armadillos of the genus *Dasypus* (Xenarthra), hyraxes of all the three living genera (Afrotheria), the Laurasiatheria genera *Equus* (horses and their kin), *Erinaceus* (the European hedgehog and its closest relatives) and *Vulpes* ('true' foxes), the Euarchontoglires genera *Lepus* (hares), *Cercopithecus* (recently split into three closely related genera of guenon monkeys), *Macaca* (macaques), *Papio* (baboons) and *Piliocolobus* (red colobus).

As we mentioned, most of these groups are currently classified as belonging to single genera, but two of them, the former genus *Cercopithecus* and the hyraxes (family Procaviidae), include more than one genus of closely related (family level or below) species. However, these two suprageneric groups originated comparatively recently ca. six and ten millions of years ago (MYA), an evolutionary age that falls within the range of the genus-level groups in our analysis (ca. 2–20 MYA—Upham et al. 2019). Also, regardless of time since common ancestry, and the uncertainties in its estimate, and regardless of taxonomic status, these clades tend to show a fairly conservative cranial morphology.

As in Cardini (2019a), primates, that are strongly sexually dimorphic (Lindenfors et al. 2007; Cardini and Elton 2008a), were analysed separately for each sex. In all other cases, we considered sex differences in cranial size and shape, measured using our specific configuration, as negligible, following the results of Cardini (2019a) on the same taxa.

Simulations

We ran five sets of simulations, each with its own abbreviation, which we describe below together. The first three use the full landmark configuration, and vary either which species are included or how the random subsamples are constructed. The last two simulations use different subsamples of landmarks ('reduced configurations') but otherwise follow the same protocol as the experiment TOTAL.

TOTALobs ('observed'): random subsamples (with all landmarks) were drawn directly from the total bootstrapped FS sample; this constrains the largest subsamples to $N < N_{max}$.

TOTAL: the subsamples were drawn from a simulated set of 1000 individuals from a theoretical population with the same mean shape and VCV as the total FS sample using *mvrnorm()* (Venables and Ripley 2002). All landmarks were used. This strategy uses the multivariate version of a normal distribution to generate a very large sample of uniform size across all the clades, an approach used in previous work on the effect of sampling error on morphospecies assessment (Cope and Lacy 1992). Note, however, that the degrees of freedom in the simulated individuals is fixed by N_{max}, which means that even this large 1000 individual sample underestimates the true variation in the biological population it represents. Because the cranium is the left side mirror reflected and symmetrized (Cardini 2017), the rank of the covariance matrix is 52, which is less than the expected dimensionality of 95 (i.e., three times the number of landmarks minus seven, for the dimensions lost in the superimposition). This means that, in all multivariate normal simulations, with the exception of *P. ellioti* males ($N = 44$) in TOTAL as well as TOTALbig, FS $N_{max} > p$, with p being the number of shape variables. Thus, even if the simulated individuals underestimate the true variation, this is not constrained by N_{max}—$1 < p$, with the only minor exception of *P. ellioti* males.

TOTALbig: as TOTAL, but including only species with $N \geq 10$. This is done in order to assess the sensitivity of TOTAL results to the inclusion of very small samples.

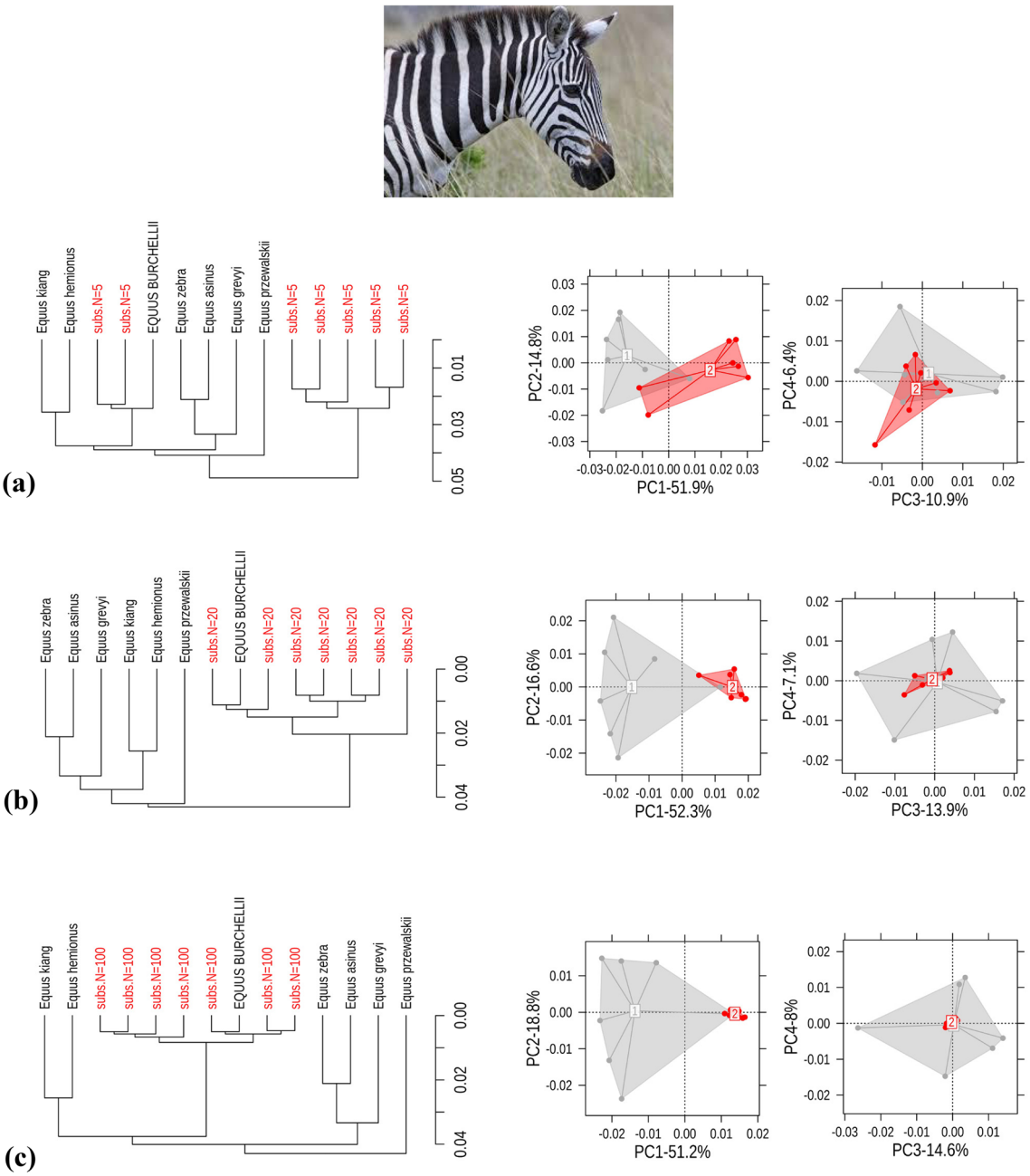

Fig. 2 Example of the effect of sampling error on mean shape estimates in plains zebras. Unweighted Pair Group Mean Average (UPGMA) phenogram and principal component analysis (PCA) are used to summarize shape similarity relationships among observed species means and the means of the FS subsamples. Subsamples N used as examples are **a** 5, **b** 20 and **c** 100. The percentage of variance accounted for by a PC is shown for each axis. In the ordinations, the 'boxed' 1 and 2 refer to the grand mean of respectively the observed species means (grey circles and convex hull) and the means of the FS subsamples (red circles and convex hull). The photo of the focal species is from https://commons.wikimedia.org/wiki/File:Plains_Zebra_(Equus_burchelli).jpg under a Creative Commons Attribution-Share Alike 2.0 Generic licence (Color figure online)

FACE: as TOTAL, but using the subset of facial landmarks (1–6, 14–20, 22–25, 30–33 and 34, in Fig. 1), as an example of a smaller configuration within a specific anatomical region.

HALF: as TOTAL but using half of the total landmarks, with the configuration selected to cover all main cranial regions although with fewer points (landmarks 1, 3, 5, 6, 9, 11, 15, 16, 19, 20–21, 23, 25, 27, 31–32, 34). As with FACE, HALF is another example used to start exploring

the impact of the specific choice and density of the landmark configuration in relation to sampling error.

N of the randomized subsamples was iteratively set at: 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 23, 26, 30, 40, 45, 50, 75, 100, 150, 200, 250, 300, 400, 500. For each N, we randomly drew 500 subsamples from the total FS sample. For TOTALobs, the largest simulated N was limited by the FS N_{max}; so, for instance, in *E. burchellii* with $N_{max} = 103$, the largest subsample had N = 100. For all the other simulations (TOTAL etc.), in contrast, the resampling experiments can be replicated over exactly the same range of sample sizes (from three to 500) in all groups. Thus, we focused principally on TOTAL, FACE and HALF. For TOTAL, however, we first checked that its results mirrored those from TOTALobs and were robust to the inclusion of small ($N < 10$) samples. This was done by computing the correlations between results (i.e., each index in each taxon) obtained in TOTAL and those of TOTALobs over the common range of subsample sizes (i.e., N from 3 to the largest $N < N_{max}$). The same was done, using the full range of N from three to 500, to compare TOTAL and TOTALbig. Similarly, TOTAL was also compared to FACE and HALF, to explore the congruence of the results using all landmarks or the two reduced configurations.

For instance, for *Equus*, we calculated the correlation of WRONG SELF estimated in TOTAL with estimates from TOTALobs (although these included, for TOTAL, only results of subsamples ranging from N = 3 to N = 100). Then, we did the same for all other indices, and finally repeated the whole correlational analysis (over the whole range of Ns from 3 to 500) to compare TOTAL with either TOTALbig, FACE or HALF. These correlations were summarized using the median and the tenth percentile of all correlations across all taxa. The latter was used, instead of the minimum as an estimate of the lowest congruence between two sets of data once extreme cases are excluded. In general and for the same reasons (i.e., reducing the impact of extreme observations), in this study we typically employed trimmed ranges as detailed below.

'Indices' Assessing the Impact of Sampling Error

To assess the effect of sampling error, we categorized results using six 'indices' for the randomization results (i.e., the estimates of means, variances and covariances). The first three indices apply mean shapes and explore the effect of sampling error in the FS on how one would interpret interspecific relationships based on the shape data; the other three are strictly 'micro-evolutionary', in that they compare differences among individuals in the random subsamples to the ones observed in the parent total FS sample. The indices are:

(1) WRONG SELF: the fraction of FS subsamples of a given N, whose mean shape is closer (i.e., it has a shorter Procrustes shape distance) to the mean of another species than to its own parent FS mean. WRONG SELF assesses the risk of misidentification of a FS subsample because its mean shape is so different from the 'true' mean that it groups with another species. WRONG SELF can range from zero (no errors in identifying the correct species) to one (100% of incorrect affiliations). For instance, if we had just seven randomized subsamples as in Fig. 2, WRONG SELF would be 0/7 when N = 20 or 100. However, with N = 5, only two subsample means cluster really close to their own observed species mean, while the other five make up a completely separate cluster. We would have to check one by one those five means to see if, despite being outliers, they are still closer to the observed mean of plains zebras in the parent sample than to any other *Equus* species: if not, WRONG SELF would be 5/7 (> 70% of affiliations to a wrong species)

(2) WRONG SISTER: the proportion of FS subsamples of a given N, whose mean shape does not cluster with the correct phenetic sister species. By "correct phenetic sister species" we mean the species whose mean shape is closest, in terms of Procrustes shape distances, to the observed FS mean shape. Thus, WRONG SISTER provides again information about similarity in mean shapes, but this time is about the risk of an inaccurate inference of interspecific phenetic relationships. Like WRONG SELF, WRONG SISTER can also range from zero (highest relative accuracy) to one (100% of subsamples having the wrong phenetic sister species). For instance, for plains zebras, the phenetic sister species is the kiang: if looking at pairwise distances we found that, in the example of Fig. 2, one subsample mean is in fact closer to, say, the hemion, then WRONG SISTER would be 1/7 (ca 15% of erroneous inferences of the FS nearest neighbour). Unlike these simplified and purely didactic examples, however, in the real analysis the denominator of both WRONG SELF and WRONG SISTER is the total number of random selections of subsamples of a given N, which is typically 500

(3) BG-RV (between group relative variance): the ratio between the multivariate variance of the means of all FS subsamples of a given N and the interspecific multivariate variance of all observed species mean shapes in the same lineage (e.g., seven species, for *Equus*). The numerator is expected to increase in smaller samples, while the denominator is constant and simply used to scale the amount of error in the estimates of the FS mean to the amount of observed interspecific mean shape differences. BG-RV might be interpreted as the proportion of interspecific mean shape space occupied

by the different means of FS subsamples. It can range from zero (no sampling error and perfectly identical means) to one or more, if the variability in FS means is as large as or larger than that of the observed species means in a group. For instance, in the example of Fig. 2b, c, the PCA scatterplots indicate that plains zebra means overlap almost perfectly with the observed mean using subsamples of N = 100 and are still fairly close to it, compared to other species, with N = 20, which would correspond to BG-RV close to zero. In contrast, in Fig. 2a, the area occupied by the means of subsamples of N = 5 is almost as large as the range of interspecific mean differences in *Equus*, which suggests a BG-RV ≈ 1. For BG-RV, as well as for W-RV, we used the sum of the variances of each variable to measure the size of a multivariate shape space. However, this can be done using alternative statistics such as the median of pairwise Procrustes shape distances in a sample or their 90th percentile (the latter being analogous to a trimmed univariate range) (Cardini and Elton 2008b). All three statistics are shown in supplementary Table S2 and are highly correlated (median r = 0.98, minimum r = 0.91), which suggests that using the sum of variances, as we did, or other common alternatives, does not appreciably change results

(4–5) W-RV (within species relative variance): this is analogous to BG-RV but it is based on individual differences within the FS. Variance is computed, as before, as the sum of the variances of the shape coordinates. W-RV of a specific run of a simulation is the ratio between the variance in a subsample of N FS individuals (e.g., N = 10) divided by the observed variance using all individuals (N_{max}) in the parent FS sample. The interpretation of this index is analogous to the one for BG-RV with the difference that, instead of using means, W-RV is within species and thus measure how much of the total parent FS shape space is occupied by one of its random subsamples. Thus, because the numerator varies from run to run, the median (W-RV-median) and its trimmed range (W-RV-range, computed as the absolute difference between the 10th and 90th percentiles of the W-RVs) are used to summarize this index. As we anticipated, trimmed ranges, here and in other instances where ranges are computed, are preferred to the minimum to maximum range, because they are less sensitive to extreme cases. As BG-RV, W-RV can range from zero to one or more

(6) VCVr: the median correlation of variance covariance matrices (VCV) between FS subsamples (of a given N, e.g. N = 20) and the observed total sample (N_{max}) VCV. The correlation does not guarantee identity but can assess proportionality and therefore complements W-RV. VCVr ranges from zero to one.

The definitions of the indices as well as all main abbreviations specific to this study are briefly repeated in Table 2, which is provided as an aid for the reader and should be used as a quick reference to consult when in doubt.

Graphical Summaries, Tables, and '10% Error Threshold'

Producing an effective summary of our results is not straightforward, as the set of numbers generated by each simulation is vast. Just for TOTAL, for instance, there are 27 subsample Ns by six indices by 14 taxa, which makes a total of more than 2250 values (from an overall set of more than 2250 × 500 ca. = 1.1 million numbers). The main trends were therefore visually assessed using profile plots (index vs FS subsample N); summarized with medians and ranges (trimmed using the 5th and 95th percentile of values across all taxa); and further explored using a '10% error threshold'.

The 10% threshold is arbitrary (we could have chosen 5% or 20% or anything else) but it seems reasonable to us (not too small and not too large), and the approach has already been adopted in other morphometric studies of sampling error to summarize results (Stec et al. 2016). In practice, the threshold means that, for WRONG SELF and WRONG SISTER, we selected the minimum sample size for having no more than 10% of runs misidentifying respectively the FS or its phenetic sister species; for BG-RV, we selected the minimum N for relative variance to be < 0.1 (i.e., 10% of the size of the interspecific shape space of the means); for W-RV-median, the threshold was not computed, as this index turned out to be almost completely unbiased (i.e., ca. = 1 regardless of N); for W-RV-range, we selected the smallest N for estimates of the magnitude of FS variance in subsamples to remain within ca. ± 10% of the value observed in the total sample; finally, for VCVr, we looked for the sample size corresponding to a median correlation ≥ 0.9 (i.e., less than 10% smaller than a perfect correlation of one).

Results

Summary Explanation of Indices

We remind readers to consult Table 2 for brief definitions, but, before presenting the results, we summarize here briefly and informally what the different indices measure. To start, the first three (1–3) are based on sample mean shapes and the second three (4–6) on individuals within the FS samples. In all instances, sampling error is assessed in the FS species either in relation to other species in its clade (1–3) or in relation to the total parent FS sample (4–6).

WRONG SELF is informative about the risk of affiliating a sample mean to the wrong species, whereas WRONG

Table 2 Main abbreviations specific to our study (see main text for details)

Topic	Abbreviation	Definition
General terms	FS	Focal species: species in which we assess the effect of sampling error by using randomized subsamples. The parent or total sample is the one including all (i.e., N_{max}) measured specimens
Indices	WRONG SELF	Fraction of FS subsamples of a given N, whose mean shape is closer to the mean of another species than to its own parent FS mean. With no sampling error, it should be zero, whereas with a strong effect of sampling error it will get closer to one
	WRONG SISTER	Proportion of FS subsamples of a given N, whose mean shape does not cluster with the correct phenetic sister species, which is the species whose mean is most similar (thus, closest) to the total FS mean shape. As the previous index, it ranges from zero (no impact of sampling error) to one (when 100% of the time the phenetic sister species is wrongly inferred)
	BG-RV	Between group relative variance: the ratio between the multivariate variance of the means of all FS subsamples of a given N and the interspecific variance of all observed species mean shapes in the same lineage. It should be close to zero if sampling error has little impact but will become closer to or even larger than one if sampling error introduces so much variation in the FS estimates of mean shapes that they vary more than found among different species in that group
	W-RV	Ratio between the variance in a subsample of N FS individuals (e.g., N = 10) divided by the observed variance using all individuals in the parent FS sample. With no effect of sampling error, estimates should be identical and the ratio equal to one
	W-RV-median	Median of W-RVs in 500 simulated samples of a given N
	W-RV-range	Absolute difference between the 10th and 90th percentiles of the W-RVs in 500 simulated samples of a given N
	VCVr	Median of the correlations between variance covariance matrices (VCV) of FS subsamples of a given N and the observed total sample VCV: it should be close to one if sampling error is small
Randomized experiments	TOTAL	Full configuration with subsamples drawn from a simulated set of 1000 individuals from a theoretical multivariate normal population with the same mean shape and VCV as the total FS sample
	TOTALbig	As above but including only species with larger samples (10 or more individuals) in the interspecific analyses
	TOTALobs	Full configuration with subsamples of FS randomly drawn from the bootstrapped total sample
	FACE	As TOTAL but using only facial landmarks
	HALF	As TOTAL but using half of the landmarks in the total configuration (see main text for the list of landmarks included)

SISTER is about how often, because of sampling error, one might wrongly infer what species is most similar to the FS. Thus, the closer WRONG SELF or WRONG SISTER are to zero, the smaller the impact of sampling error.

BW-RV is the portion of the 'box' (fraction of the shape space), containing interspecific mean differences, that is occupied by uncertainties in estimates of the FS mean shape when samples are smaller than in the total parent sample: like the standard error of univariate means, one wants this uncertainty to be as small as possible and definitely much smaller than interspecific differences (i.e., < 1 and as close to zero as possible). W-RV is analogous to BW-RV but it is within species and thus compares variability among individuals (like a univariate standard deviation) in smaller samples of the FS to that observed including all FS individuals. If W-RV-median is about the average relative accuracy (defined as explained above), W-RV-range is about precision in estimates of the magnitude of within species variability: if accurate and precise, variance in subsamples should be the same as in the total sample, which implies W-RV-median = 1 and W-RV-range = 0. Finally, after two indices concerning the magnitude of the variability in a sample, VCVr is about the direction of shape differences and thus compares the covariance structure in subsamples with that of the total parent sample: this index should be close to 1 when sampling error has a negligible effect.

Congruence of Results

The congruence between TOTAL and other sets of simulations is very high (Table 3). If W-RV-median is excluded (because this index is almost always ≈ 1, and therefore negligible fluctuations around this constant value lower the correlations), the range of median correlations across all indices is 0.97–1.00, with the lower boundary (10th percentile) ranging from 0.73 to 1.00. More precisely, only WRONG SISTER shows two instances with r < 0.9 and both occur when TOTAL is compared to the two reduced configurations (10th percentile = 0.73–0.82 respectively for HALF and FACE). Overall, however, correlations had a median r > 0.95 more than 80% of the times (100% if W-RV-median is excluded)

Table 3 Summary of correlations between the results of TOTAL and those of the other sets of simulations

TOTAL vs.	Index	Median	10th percentile
TOTALbig	WRONG SELF	1.00	0.98
TOTALobs	WRONG SELF	0.99	0.96
FACE	WRONG SELF	0.99	0.95
HALF	WRONG SELF	0.99	0.91
TOTALbig	WRONG SISTER	0.99	0.97
TOTALobs	WRONG SISTER	0.98	0.95
FACE	WRONG SISTER	0.96	0.82
HALF	WRONG SISTER	0.97	0.73
TOTALbig	BG-RV	1.00	1.00
TOTALobs	BG-RV	1.00	1.00
FACE	BG-RV	1.00	1.00
HALF	BG-RV	1.00	1.00
TOTALbig	W-RV-median	0.73	0.34
TOTALobs	W-RV-median	0.59	0.40
FACE	W-RV-median	0.70	0.51
HALF	W-RV-median	0.72	0.53
TOTALbig	W-RV-range	1.00	0.99
TOTALobs	W-RV-range	1.00	0.99
FACE	W-RV-range	1.00	0.99
HALF	W-RV-range	1.00	0.99
TOTALbig	VCVr	1.00	1.00
TOTALobs	VCVr	1.00	1.00
FACE	VCVr	1.00	1.00
HALF	VCVr	1.00	1.00

with a lower 10th percentile $r > 0.95$ in more than 70% of the comparisons (85% excluding W-RV-median).

Besides correlations, plots and summary statistics (not shown) all indicate that TOTAL and other sets of simulations produce very similar results, which are in fact almost identical when TOTAL is compared to TOTALobs and TOTALbig. This demonstrates that simulated data in TOTAL are an excellent approximation of TOTALobs, within the range of sample sizes in common between the two sets of analyses, and that TOTAL is robust to the inclusion of small samples. Therefore, in the rest of the paper, we focus on results from TOTAL, together with those of FACE and HALF, and omit those of TOTALobs and TOTALbig, as they are redundant. The two reduced configurations are also largely congruent with TOTAL, when assessed using correlations (Table 3), but also suggest a few small but potentially interesting differences.

Graphical Summaries

Profile plots in Figs. 3 and 4 summarize the results from TOTAL, FACE and HALF. Figure 3A, B show the profile plots for the first three indices based on interspecific differences in mean shapes. These plots either (A) include all FS subsamples within the range affected by sampling error or (B) focus on a specific segment of this range to provide more detail using representative cases. Figure 4A, B provides the same information for the remaining three 'intraspecific' indices (within FS variances and covariances). Figure 5 shows profile plots of the minimum N for the 10% threshold for each index and set of simulation. The corresponding values are shown in Table 4, which also provides summary statistics of minimum Ns for each index across the different groups. In this figure and table, smaller Ns imply the requirement of fewer specimens for the same approximate level of relative accuracy (better performance and less serious issues with sampling error) and larger N indicate the need of larger samples for achieving that relative accuracy (worse performance and stronger impact of sampling error). In Fig. 5 we included also a profile plot for the number of species in each study group and the observed ratios (computed from Table S2) of between species mean variance and within FS total sample variance.

Detailed Results for Each Index

WRONG SELF (Frequency of Incorrect Affiliation of FS Using Mean Shapes)

For WRONG SELF (first column of Fig. 3A, B), 40 individuals guarantee the highest relative accuracy in all taxa and datasets, with all FS subsample means correctly affiliated (i.e., closer) to FS. With $N < 40$, relative accuracy rapidly deteriorates in most species, so that in the smaller samples ($N \leq 10$) chances of affiliating subsample mean shapes to another species increase to 10–20% and up to 40–50% when $N = 3$. Using the 10% threshold to suggest what the minimum N might be for a reasonable relative accuracy (Table 4; Fig. 5a), we find that for WRONG SELF < 0.1 there must be typically between three and 10 specimens, with an average of 6–7; the main exception is the males of *P. ellioti* with > 10% of incorrect affiliations even when $N = 10$. However, if we aimed at an even lower occurrence of wrong affiliations and set the relative accuracy threshold to < 5%, we would need ca. $N = 20$ in virtually all species and datasets.

The '10% threshold' is useful also to confirm which taxa might be particularly sensitive to small N (Table 4; Fig. 5a). Thus, for WRONG SELF, we find that with $N = 3$ nine to 10 of the 14 groups (i.e., ca. 2/3) have more than 10% of wrong affiliations, but with $N = 10$ this happens only in 1–2 taxa. Primates are particularly impacted in all sets of simulations (TOTAL, FACE and HALF). All the FS with the highest error rate (up to 40–50% of subsample means affiliated to the wrong species) are primates, which overall constitute 95% of the taxa with errors > 10% using $N = 3$–5. The main non-primate exception among the poor performers is the

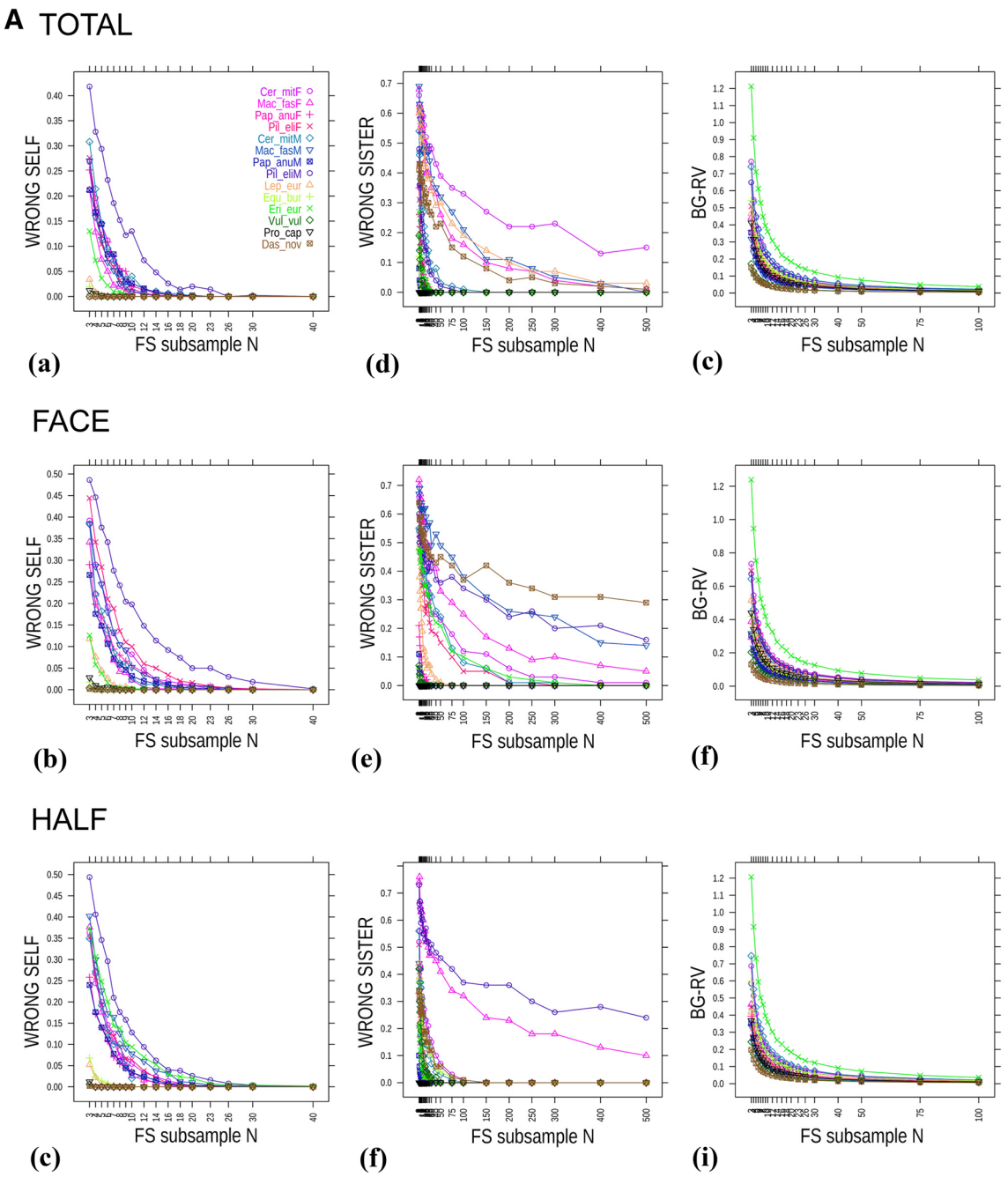

Fig. 3 **A**, **B** Profile plots for WRONG SELF (first column, **a**–**c**), WRONG SPECIES (second column, **d**–**f**) and BG-RV (third column, **g**–**i**), subdivided according to the dataset (TOTAL, FACE, HALF). In this and the next figures, species abbreviations are those shown in Table 1 (using the first three letters of the genus and species scientific names, and followed by F, for females, or M, for males, when analyses are done with separate sexes). **A** Shows the range of subsample Ns with non-negligible sampling error; **B** (next page) focuses on a few Ns, taken as examples within the range of Ns showing the largest effect of sampling error

European hedgehog. Its error rate is generally below that of primates, but still always larger than 10% in the smallest samples (N = 3). In fact, using the HALF configuration, WRONG SELF in the European hedgehog is among the highest of all taxa. The other non-primate species with WRONG SELF above the 10% threshold is the European hare, but it is only slightly above it (11.8%) and this happens only with N = 3 in FACE.

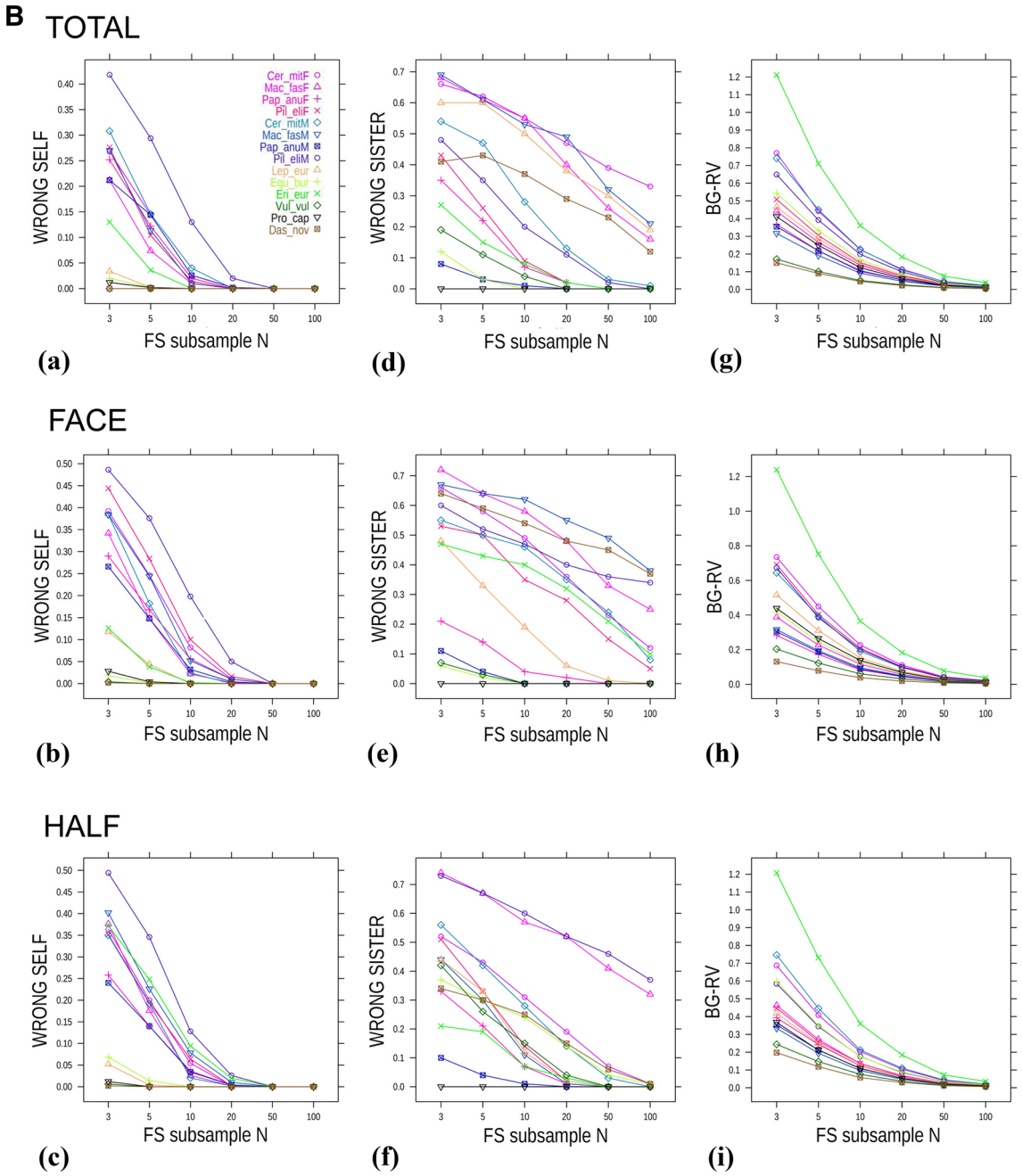

Fig. 3 (continued)

WRONG SISTER (Frequency of Wrong Identification of the FS Phenetic Sister Species Using Mean Shapes)

Misidentifications of the correct phenetic sister species in smaller samples (WRONG SISTER) are much more serious and can happen frequently even with large N. Inaccuracies vary broadly depending on the group and landmark configuration (Fig. 3A, B, second column).

More specifically, FACE performs particularly poorly, with *D. novemcintus*, male *P. ellioti* and both female and male *M. fascicularis* misidentifying the phenetic sister species ca. 10–35% of times even when N = 250. In general, with this configuration, when N = 20–50, WRONG SISTER ranges for most species between > 0.1 and > 0.5 (i.e., more than 10–50% of errors). In the smallest subsamples (N = 3–5), however, it can be up to 0.5–0.7 and this happens

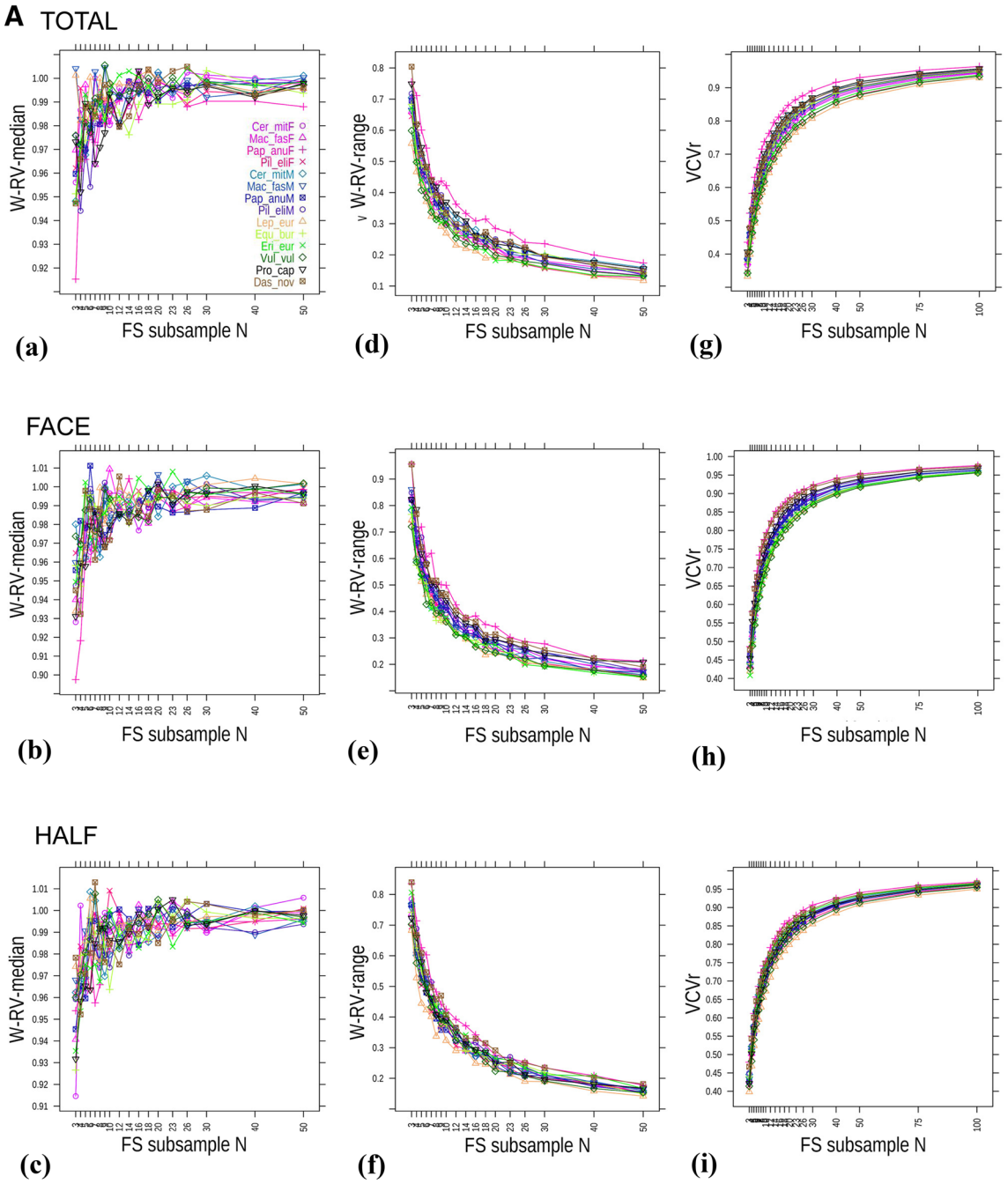

Fig. 4 **A**, **B** Same as Fig. 3A, B but now for W-RV-median (first column, **a**–**c**), W-RV-range (second column, **d**–**f**) and VCVr (third column, **g**–**i**)

in most primates, as well as of *D. novemcintus* and the European hare and hedgehog.

HALF in contrast, despite also suggesting large inaccuracies even in relatively large samples, has only two species performing extremely poorly in terms of identification of the correct phenetic sister species. These are male *P. ellioti* and female *M. fascicularis* with more than 10% of misidentifications even in samples with several hundreds of individuals. All other species, in comparison, show a relatively modest error rate, with WRONG SPECIES always < 0.1 when $N \geq 50$.

The full configuration (TOTAL) performs somewhat in between the worst (FACE) and least (HALF) affected by sampling error, with few species of primates together with *D. novemcintus* and the European hedgehog showing WRONG SISTER error rates of 10–40% even when $N = 100$.

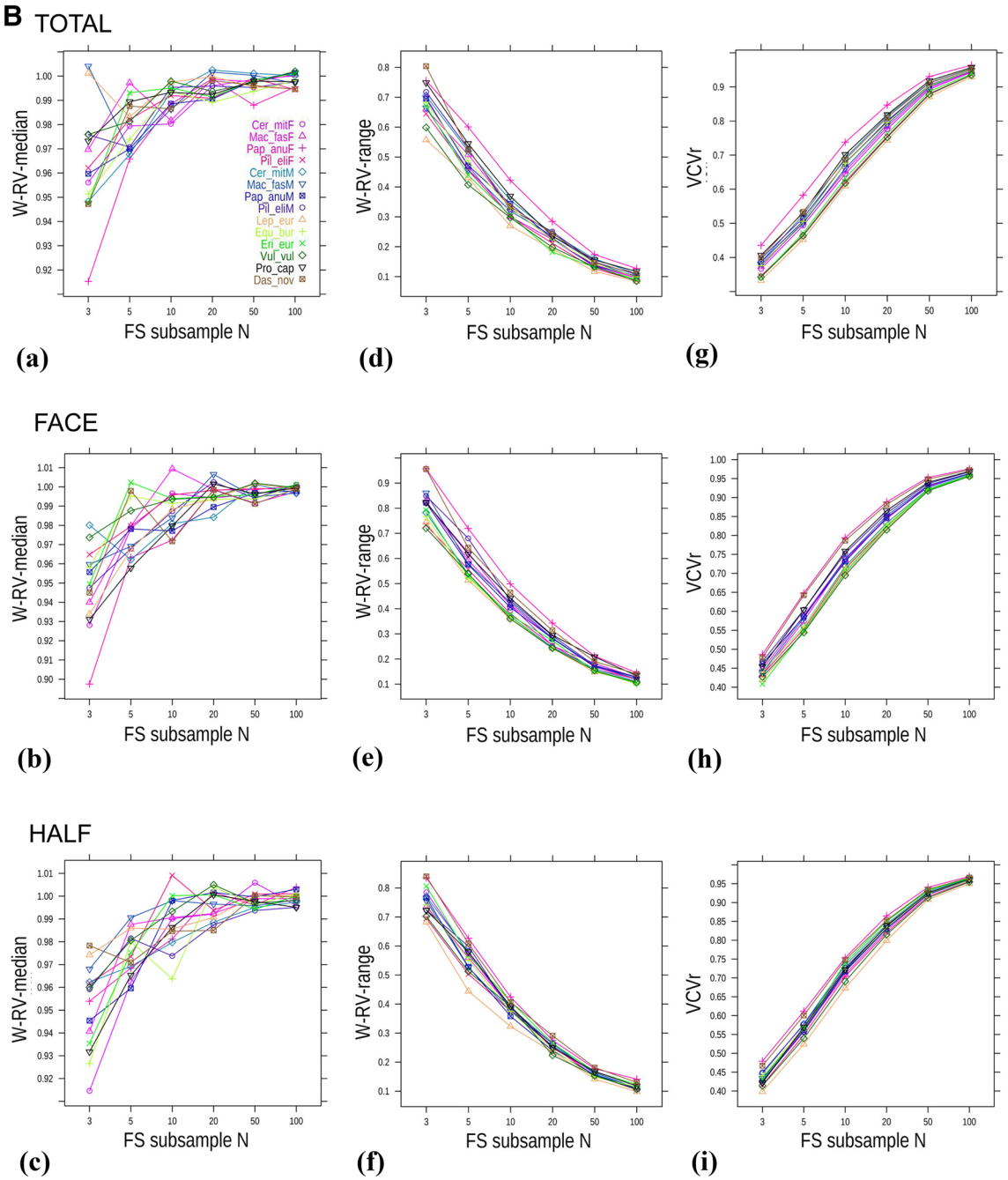

Fig. 4 (continued)

In the smallest samples (N = 3–5), the range of TOTAL WRONG SISTER error rates is broad, going from no errors in *Procavia capensis* to almost 70% in macaques.

Using the 10% error threshold (WRONG SISTER < 0.1) to summarize results, we found that we need samples of ca. 15 specimens on average in TOTAL and HALF and much larger ones (N ≈ 90 on average) in FACE for this level of relative accuracy (Table 4; Fig. 5b). Yet, in a few cases, not even using the largest samples the 10% accuracy threshold can be reached. This happens in one case in both TOTAL and HALF, as well as in three taxa in FACE, and more specifically in *D. novemcintus* (FACE) and a few of the primate groups (all other cases). Unlike all other indices, the trimmed range of N variation for WRONG SISTER < 0.1 is very large in all sets of simulations, going from three to several hundreds. Even using a slightly more liberal threshold, such as a 20% relative accuracy in the assessment of which species is phenetically sister to the focal sample, TOTAL

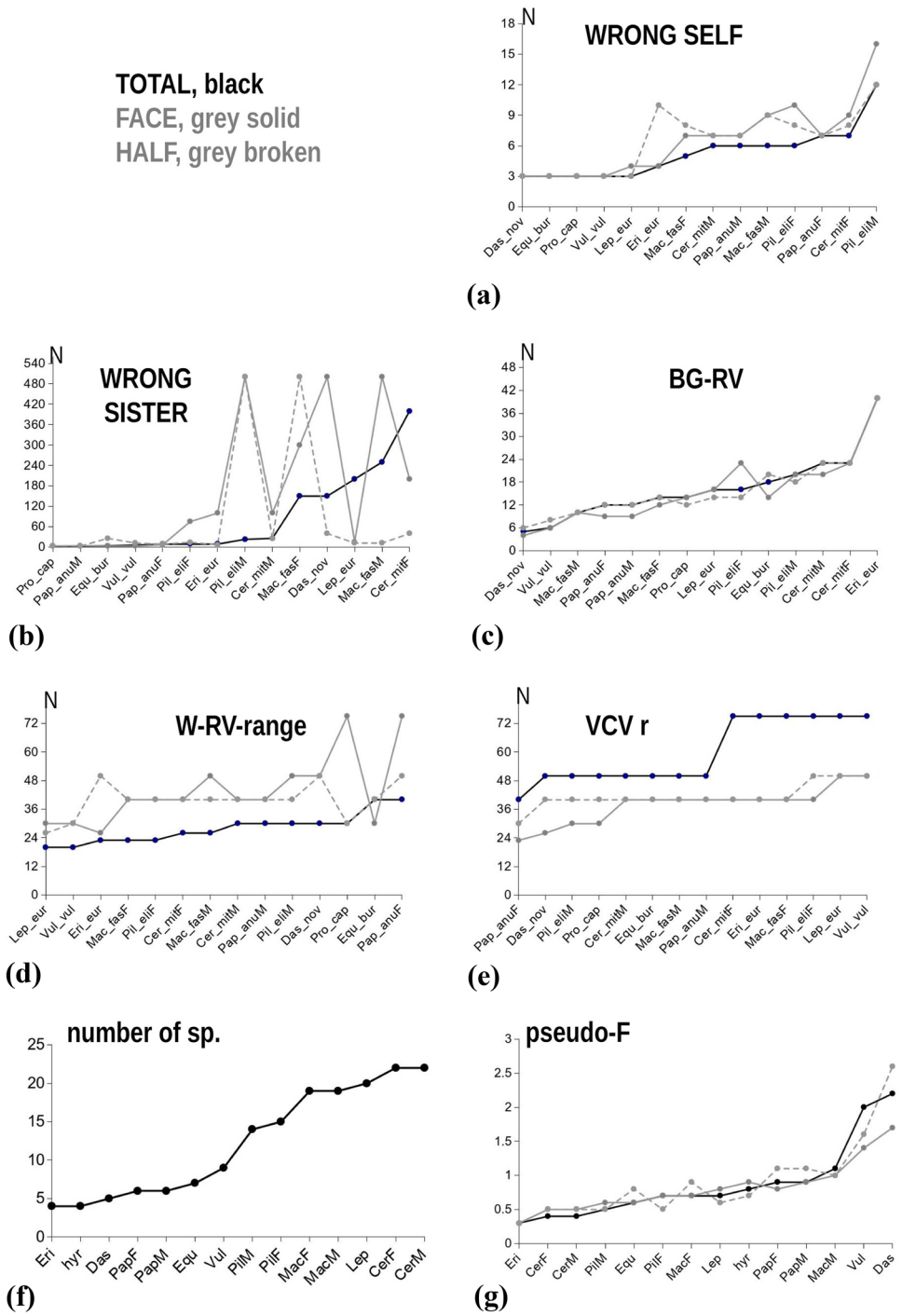

Fig. 5 **a**–**e** Profile plots of the minimum Ns for the '10% relative accuracy threshold' in TOTAL, FACE and HALF: **a** WRONG SELF, **b** WRONG SISTER, **c** BG-RV, **d** W-RV-range, **e** VCVr. **f**, **g** Profile plots for the number of species in each study taxon (**f**) and the pseudo-F ratio (**g**) in TOTAL, FACE and HALF. Species in the plots are in increasing order of the plotted value (e.g., increasing minimum Ns); when multiple datasets are present, the values for TOTAL are used to order the species. It is easy to observe that the order for BG-RV is almost exactly the reverse as for pseudo-F, as expected given their high correlation

and HALF suggest the requirement of about 8–9 individuals on average (and more than 50 using FACE), although this can be much less (N = 3) or much more (between ca. 150 and almost 400) depending on the taxon.

Therefore, to summarize the results of the first two indices, although one can predict the right affiliation (low WRONG SELF) even with small samples of the FS, for accurately discovering its shape similarity relationships with other species (low WRONG SISTER), one requires much larger N. Besides, whereas WRONG SELF produces very similar results regardless of the configuration (Fig. 5a), WRONG SISTER is highly variable in relation to the choice of landmarks, which is particularly evident in the species with the largest errors (Fig. 5b). What species, if any, are most strongly affected by sampling error (which we refer to as 'taxonomic bias') is also less clear for WRONG SISTER compared to WRONG SELF. With the latter, primates tend to be impacted more severely by errors in small samples.

Table 4 Minimum N (minN) required for the '10% threshold' of RA (W-RV not shown because unbiased; prop. = proportion): for each index, the three least and most impacted cases are emphasized using respectively an italics and bold

Set of landmarks	Focal species	WRONG SELF < 0.1		WRONG SISTER < 0.1*		BG-RV < 0.1		W-RV-range < 0.2		VCVr > 0.9	
	(F = female; M = male)	minN	Prop.	minN	Prop.	minN	Fraction	minN	Fraction	N	r
TOTAL	*Dasypus novemcinctus*	*3*	*0.000*	150	0.080	*5*	*0.090*	30	0.193	50	0.911
	Procavia capensis	*3*	*0.012*	*3*	*0.000*	14	0.091	30	0.195	*50*	*0.918*
	Equus burchellii	3	0.020	*4*	*0.070*	18	0.092	**40**	**0.167**	50	0.904
	Erinaceus europaeus	4	0.072	9	0.080	**40**	**0.093**	*23*	*0.183*	**75**	**0.920**
	Vulpes vulpes	*3*	*0.000*	6	0.090	*6*	*0.084*	*20*	*0.198*	75	0.915
	Lepus europaeus	3	0.034	**200**	**0.100**	16	0.089	*20*	*0.192*	75	0.908
	F Cercopithecus mitis	**7**	**0.066**	**400**	**0.130**	**23**	**0.096**	26	0.197	**75**	**0.925**
	M Cercopithecus mitis	6	0.096	26	0.100	**23**	**0.097**	30	0.197	*50*	*0.912*
	F Macaca fascicularis	5	0.074	150	0.100	14	0.097	23	0.196	**75**	**0.929**
	M Macaca fascicularis	6	0.068	**250**	**0.080**	*10*	*0.092*	26	0.195	50	0.901
	F Papio anubis	**7**	**0.080**	9	0.100	12	0.092	**40**	**0.199**	*40*	*0.916*
	M Papio anubis	6	0.084	*3*	*0.080*	12	0.090	30	0.194	50	0.905
	F Piliocolobus ellioti	6	0.074	9	0.100	16	0.095	23	0.189	75	0.916
	M Piliocolobus ellioti	**12**	**0.072**	23	0.100	20	0.099	30	0.196	50	0.905
Set of landmarks	Focal species	WRONG SELF < 0.1		WRONG SISTER < 0.1*		BG-RV < 0.1		W-RV-range < 0.2		VCVr > 0.9	
	(F = female; M = male)	minN	Prop.	minN	Prop.	minN	Fraction	minN	Fraction	N	r
FACE	*Dasypus novemcinctus*	*3*	*0.002*	**500**	**0.290**	*4*	*0.096*	**50**	**0.190**	*26*	*0.906*
	Procavia capensis	3	0.028	*3*	*0.000*	14	0.095	**75**	**0.158**	*30*	*0.904*
	Equus burchellii	*3*	*0.018*	*3*	*0.060*	14	0.089	*30*	*0.197*	40	0.905
	Erinaceus europaeus	4	0.058	100	0.100	**40**	**0.094**	*26*	*0.199*	40	0.901
	Vulpes vulpes	*3*	*0.004*	*3*	*0.070*	*6*	*0.100*	*30*	*0.195*	**50**	**0.918**
	Lepus europaeus	4	0.076	16	0.100	16	0.096	30	0.198	**50**	**0.917**
	F Cercopithecus mitis	**9**	**0.092**	200	0.060	**23**	**0.093**	40	0.192	40	0.915
	M Cercopithecus mitis	7	0.086	100	0.080	20	0.096	40	0.182	40	0.909
	F Macaca fascicularis	7	0.068	300	0.100	12	0.089	40	0.195	40	0.916
	M Macaca fascicularis	**9**	**0.092**	**500**	**0.140**	10	0.096	50	0.172	**40**	**0.916**
	F Papio anubis	7	0.094	7	0.100	9	0.095	**75**	**0.170**	*23*	*0.902*
	M Papio anubis	7	0.072	4	0.060	*9*	*0.098*	40	0.177	40	0.915
	F Piliocolobus ellioti	**10**	**0.100**	75	0.100	**23**	**0.086**	40	0.178	40	0.903
	M Piliocolobus ellioti	**16**	**0.092**	**500**	**0.160**	20	0.099	50	0.175	30	0.901
Set of landmarks	Focal species	WRONG SELF < 0.1		WRONG SISTER < 0.1*		BG-RV < 0.1		W-RV-range < 0.2		VCVr > 0.9	
	(F = female; M = male)	minN	Prop.	minN	Prop.	minN	Fraction	minN	Fraction	N	r
HALF	*Dasypus novemcinctus*	*3*	*0.002*	40	0.060	*6*	*0.096*	**50**	**0.181**	*40*	*0.921*
	Procavia capensis	*3*	*0.012*	*3*	*0.000*	12	0.087	*30*	*0.197*	40	0.909
	Equus burchellii	3	0.068	26	0.100	20	0.088	40	0.182	40	0.909
	Erinaceus europaeus	**10**	**0.094**	*7*	*0.090*	**40**	**0.091**	**50**	**0.167**	*40*	*0.916*
	Vulpes vulpes	*3*	*0.006*	12	0.100	*8*	*0.089*	*30*	*0.191*	**50**	**0.913**
	Lepus europaeus	3	0.052	12	0.090	14	0.090	*26*	*0.190*	**50**	**0.908**
	F Cercopithecus mitis	8	0.072	**40**	**0.100**	**23**	**0.088**	40	0.179	40	0.907
	M Cercopithecus mitis	7	0.100	26	0.100	**23**	**0.096**	40	0.189	40	0.912
	F Macaca fascicularis	8	0.072	**500**	**0.100**	14	0.098	40	0.175	40	0.904
	M Macaca fascicularis	**9**	**0.092**	12	0.090	*10*	*0.096*	40	0.187	40	0.905
	F Papio anubis	7	0.070	9	0.100	12	0.098	**50**	**0.178**	*30*	*0.907*

Table 4 (continued)

Set of landmarks	Focal species	WRONG SELF < 0.1		WRONG SISTER < 0.1*		BG-RV < 0.1		W-RV-range < 0.2		VCVr > 0.9	
	(F = female; M = male)	minN	Prop.	minN	Prop.	minN	Fraction	minN	Fraction	N	r
	M Papio anubis	7	0.078	*3*	*0.100*	12	0.087	40	0.173	40	0.907
	F Piliocolobus ellioti	8	0.094	14	0.050	14	0.096	40	0.184	**50**	**0.917**
	M Piliocolobus ellioti	**12**	**0.094**	**500**	**0.240**	18	0.099	40	0.183	40	0.911
	Summary statistics	minN		minN		minN		minN		minN	
TOTAL	Median	6		16		15		28		50	
	10th perc	3		3		7		21		50	
	90th perc	7		235		23		37		75	
FACE	median	7		88		14		40		40	
	10th perc	3		3		7		30		27	
	90th perc	10		500		23		68		47	
HALF	median	7		13		14		40		40	
	10th perc	3		4		9		30		40	
	90th perc	10		362		23		50		50	

*Unless specified differently and emphasized by underscoring

With WRONG SISTER, in contrast, unless N is really big, errors are large in almost all species. For instance, with 10 individuals, depending on the configuration, only 30–50% of taxa have inaccuracies < 10%. Among these, some in fact require Ns as large as 150–400. Specifically, using WRONG SISTER, the species worst affected by sampling error in at least one of the three sets of simulations are *D. novemcintus*, the European hare and European hedgehog, as well as both sexes of *M. fascicularis* and *C. mitis* and females of *P. ellioti*. Not only these species require very large samples to keep WRONG SISTER below the 10% threshold. With the exception of the European hedgehog, they also have in at least one of the three configurations huge error rates of more than 50% when N is very small (N = 3–10). In FACE, in particular, these very large inaccuracies using N ≤ 10 are almost the rule, as they occur in half of the FS.

BG-RV (Proportion of the Interspecific Mean Shape Space 'Occupied' By Variation in FS Means Due to Sampling Error)

The effect of N on BG-RV seems fairly similar in all groups and datasets except the European hedgehog, which is more negatively impacted by sampling error (Fig. 3A, B, third column). The region of the shape space accounted for by errors in estimates of FS mean shape is ca. less than 20% of that capturing interspecific mean differences when N = 20 and becomes almost completely negligible in most species when N ≥ 50. In contrast, when N = 10, between ca. 60% and almost 80% of the taxa have BG-RV > 0.1 and up to 0.36, which indicates that the magnitude of variance due to sampling error in estimating the FS mean shape is as large as approximately 1/10 to 1/3 of the observed variance of interspecific means. With even smaller samples, when N ≤ 5, the variability in FS means can be as large as 10–80% of the interspecific mean shape variation and even larger than that in the European hedgehog. In this species, sampling error with N = 3 produces means that vary 1.2 times more than observed interspecific mean shapes. This does not happen in any of the other species, although with N = 3 either one or both sexes of *C. mitis* and *P. ellioti*, as well as the plains zebra and the European hare, can have at least one set of simulations with a variance among subsample means as large as ca. 50–80% of that of interspecific means (i.e., BG-RV ≈ 0.5–0.8).

Using the 10% threshold (BG-RV < 0.1), one needs, on average, N ≈ 15, which is about the same average sample size as for WRONG SISTER in TOTAL and HALF (Table 4; Fig. 5c). However, with BG-RV results are largely congruent across all three sets of simulations and different taxa (trimmed range = 7–23), suggesting that, regardless of the configuration, ca. 10–20 individuals may be enough for the error in estimates of a species mean shape to be fairly small relative to the differences between species in its lineage.

In terms of potential taxonomic biases, despite some primates being often among the species with larger BG-RV for a given N, this index does not suggest a particularly strong effect of sampling on a specific lineage, with the main exception of hedgehogs.

W-RV-Median and Range (Proportion of the Observed Total FS Intraspecific Shape Space 'Occupied' By Individual Variation Due to Sampling Error)

W-RV-median shows very little variation with N being ≈ 1 (Fig. 4A, B, first column), thus suggesting on average almost identical estimates of the magnitude of intraspecific shape variance in FS subsamples as in the total sample of FS. Thus, the magnitude of within species variance is virtually unbiased, as expected with a variance, except in the smallest samples ($N \leq 5$), where it is slightly underestimated (up to ca. 10% compared to the observed magnitude of variance in FS).

In contrast, W-RV-range is much broader in smaller samples (Fig. 4A, B, second column), increasing in all datasets form ca. 0.1 to up to 0.7–0.8 in the smallest samples. As W-RV has a roughly symmetric distribution around the median, this translates into over- or under-estimates of within FS variance of about ± 5% and up to ± 35–40%. Thus, for the W-RV-range to stay between 1.1 and 0.9 (i.e., ± 10% of observed total sample variance—Table 4; Fig. 5d), one needs on average about 30 individuals for TOTAL and about 40 for FACE and HALF, although in the worst cases, such as using facial landmarks in *P. capensis* or female *P. anubis* baboons, more than 70 specimens are necessary for precise (within ± 10% of observed) estimates of within FS variance. However, in general, different FS are approximately similarly impacted by low precision in smaller (< 30–40) samples and therefore there seems to be no evident taxonomic bias for this index.

VCVr (Proportionality of FS Shape Variance Covariance Matrices)

Finally, VCVr, the median correlation of VCVs in subsamples with the parent FS VCV, is high (ca. 0.9 or larger) when N is ca. 20–50 or larger (Fig. 4A, B, third column; Fig. 5e; Table 4). More precisely, the reduced configurations (FACE and HALF) require on average slightly fewer specimens ($N = 40$) than TOTAL ($N = 50$) for VCVr ≥ 0.9. However, variability in VCVr is generally modest: as an example, with $N = 20$, VCVr ranges, depending on the FS and set of simulation, between 0.74 and 0.89; and with $N = 50$, it ranges between 0.87 and 0.95. With N ca. < 20–30 the correlation becomes rapidly smaller (and especially so in the complete landmark configuration) and, with N ca. < 10, VCVr ranges between 0.6 and 0.7 and just ca. 0.4. Yet, overall, except for the slightly worse performance of TOTAL, the trend in change of VCVr with N is broadly similar in all datasets and taxa.

Discussion

Main Findings

The impact of sampling error on cranial shape data, as assessed by the six indices we used, varies across groups but a strong and consistent taxonomic bias is found only for WRONG SELF, with primates requiring comparatively larger samples for the same relative accuracy of other taxa. In general, across all groups, 40 individuals guarantee that a mean shape is closer to its own species than to any other in that genus or supra-generic lineage, thus making no errors in taxonomic identification. However, with fewer than 10 specimens, erroneous affiliations may occur 10–50% of the time, despite an average requirement of just 6–7 individuals for staying within a 10% error threshold.

To achieve this same low error rate in the identification of the correct phenetic sister species (WRONG SISTER < 10%), in contrast, N must be on average between ca. twice to ten times larger than in WRONG SELF, which corresponds to about 15–90 individuals. The FACE dataset is more strongly impacted by sampling error, so that, even with $N = 20–50$, there can be 10–50% of erroneous identifications of the phenetic sister species. The taxonomic bias is, however, much less obvious with WRONG SISTER compared to WRONG SELF, as some of the largest errors (for a given N) are found not only among some of the primates but also in *D. novemcintus* and European hares. Thus, this index suggests that primates may be more problematic, but the issue is general and less predictable. If these results are generalizable, they suggest that to correctly identify a placental morphospecies and its phenetic sister species one needs, respectively, ca. 10–40 and 20 to several hundred individuals. 'Guessing' the right species seems therefore feasible with means based on relatively small samples (yet, still larger than what is often available with fossil mammals), but reconstructing mean similarity relationships may be much harder unless really large samples are available. This is congruent with studies that use random subsampling of tip taxa to produce bootstrap values on trees built from morphometric data. For example, using continuous traits maximum-likelihood and tip taxa with $4 \leq N \leq 13$, the proportion of tree nodes recovered from mandible, cranial, and tooth shape data was sometimes as low as 16.5% and never higher than 96.5% in a study on marmots with approximately the same taxonomic scope as our clades (Caumul and Polly 2005). Nodes linking sister taxa were wrong 4–38% of the time at sample sizes of fewer than ten, which is comparable to the range of WRONG SISTER error in our simulations. Similarly, by bootstrapping samples with N ranging between four and 20, Pearson et al. (2015) found that most nodes in trees built using neighbour-joining applied to Procrustes shape distances between means

of subspecies of great apes received a support of less than 70%, with several cases below a 50% threshold, whereas species trees (with on average larger N and larger interspecific differences) fared much better.

The average N required for the variability in FS mean shapes to be no more than 10% of the magnitude of between species mean variance (BG-RV) is ca. 15, which is about the same average sample size as for WRONG SISTER in TOTAL and HALF. This might indicate that misidentifications of the phenetic sister species become more likely when sampling error increases the inaccuracy of the estimates of the FS mean shape to the extent that their uncertainty accounts for more than 1/10 of the interspecific mean shape space in that lineage. However, unlike WRONG SISTER, BG-RV does not vary much across datasets and taxa, and mostly suggests that 10–20 individuals are enough to have high relative accuracy. Indeed, with the main exception of hedgehogs, with their consistently larger BG-RV for reasons we will discuss later, all taxa perform about equally using this index.

Thus, the minimum N required for relative accuracy in mean shape estimates is, in this study, slightly smaller but fairly close to findings by Cardini et al. (2015), whose resampling experiments on Procrustes shape data from horse premolars suggested that accurate means might require at least 20 individuals. Our 'desirable' Ns for relative accuracy in species comparisons are in good agreement also with another simulation study based on parameters from real samples (Cope and Lacy 1992), despite their different approach (using the coefficient of variation to decide if a sample represented one or more species) and nature of the data (univariate on teeth). In this analysis, they showed that species of *Cercopithecus* can be well discriminated on average even in samples of just 5–10 individuals but more than 20 are required to achieve adequate power and reduce the rate of false positives (i.e., results suggesting multiple species when only one is present) to less than 10–40% (Cope and Lacy 1992; Cope 1993). Also using traditional morphometrics but another different method to investigate sampling error, Wood and Constantino (2007) showed that the number of specimens necessary for the average of cranio-mandibular measurements of *Paranthropus boisei* to stabilize on a given value is ca. 7–15 specimens. This fits well with Cardini et al.'s (2015) finding that premolar size in horses might be fairly accurately estimated with just about ten specimens, which in turn was about the same as in Cardini and Elton's (2007) analysis of monkey skulls. Overall, these findings from previous research suggest that univariate estimates of means, such as the centroid size of a set of landmarks or linear measurements, may indeed require slightly smaller samples for relative accuracy, but their Ns of at least ca. 5–20 individuals are not too far from our ca. 7–15 to 40–50 for achieving respectively the < 10% threshold or a next to zero inaccuracy in WRONG SELF and BG-RV.

Relative accuracy in within species estimates of variance and covariance, in contrast, might need even larger samples in both size and shape data. Cardini and Elton (2007) and Cardini et al. (2015) showed that, unlike univariate means, size variance cannot be similarly accurately estimated unless some 20–40 individuals are available within a species, which is about the same range of sample sizes we found for relative accuracy in the estimates of shape variances and covariances. Similarly, using the same type of resampling experiments as Cardini et al. (2015) in cranial data of voles, Schlis-Elias (2020) found that mean size was accurately predicted with just five individuals, but its variance required some 40 specimens, thus reproducing in rodent skulls almost exactly the findings from horse teeth. Indeed, also for univariate measurements, the minimum N for relative accuracy likely varies depending not only on the group but also on the test statistics. For instance, in an in depth assessment of sampling error in bivariate allometric regressions using cranial measurements of *Alligator mississippiensis*, Brown and Vavrek (2015) found that accurate estimates of static allometries in adults require N > 20. This, compared to the studies of size variables we have just mentioned, is larger than the ca. ten individuals needed for the mean and about the same or slightly fewer than the 20–40 required for the variance.

Precision in estimates of within FS shape variance magnitude (i.e., a W-RV-range of 0.9–1.1, which corresponds to the magnitude of FS variance being ± 10% of that observed in the total sample) requires on average slightly larger samples (30–40 individuals). There is also more variability across taxa, but again no clear taxonomic bias. These results, for W-RV-range, are congruent with those for the correlation of VCV (subsamples vs. total FS sample), which suggest the requirement of N ≈ 20–50 (sometimes more) for r > 0.9, with some variability depending on the group, but no evident taxonomic bias. Thus, within species, 25–50 individuals could be appropriate for a good relative accuracy in cranial shape estimates of variance and covariance, although ca. 100 individuals may be more appropriate to keep errors really low.

If for mean shape relative accuracy Cardini et al. (2015) were slightly more pessimistic in terms of minimum N, for within species variance covariance they were almost the same or at most slightly more optimistic than we are in this study. They suggested that at least ca. 15–20 individuals are needed, which is an estimate very close to Polly's (2005) analysis of molars in the common shrew. In that study, he found that VCV estimates are inaccurate with N < 15, which is almost identical to Kryštufek et al.'s (2016) findings using skull shape in *Bandicota indica* rats, even if these authors only focused on the magnitude of shape variance. Thus,

overall, it seems that, despite small differences, results of all these studies on adult mammal cranio-dental variation are fairly similar. Because, like ours, these previous analyses were largely based on resampling experiments using sample estimates as a baseline to assess relative accuracy, they are also likely to produce underestimates of minimum Ns, as we discuss in more depth in the next section. At the same time, however, the generally high congruence we found (with N for relative accuracy in variance and covariance of ca. 15–20 in those analyses vs ca. 20–50 in ours) is surprising and even more so in consideration of the different study structures and taxa, as well as the differences in methods.

Conservative and Optimistic Results?

Do the minimum sample sizes we found look like a difficult target especially for palaeontological studies of closely related species or analyses of groups with very few museum specimens, as, for instance, rare samples of small island populations? Clearly, the appropriate N in a study depends on the specific question, test statistic and size of the effect being measured. We investigated sampling error in the context of the taxonomic assessment of mammalian morphospecies but, even in this narrow field, a robust generalizable answer on desirable Ns for the accuracy of means, variances and covariances clearly requires more research. In terms of internal validity, however, it is likely that our findings are in fact conservative and tend to err on the optimistic side.

To appreciate why, we can take BG-RV as an example. To compute this index, after drawing 500 random FS subsamples of a given size (e.g., N = 10), we divide the multivariate variance of their 500 means by the multivariate variance of the observed species means in that group (including the mean of the total FS sample). The numerator quantifies the effect of sampling error on the estimate of the FS mean shape 'scaled' by the magnitude of observed variation in that taxon (the denominator). However, this ratio is almost certainly an underestimate, because species samples are small and this tends to inflate distances between means, thus overestimating the interspecific differences used as denominator. This has been observed empirically in geometric morphometric studies (Cardini et al. 2015) and is expected from de Moivre's equation, which states that the standard error of the mean increases as N decreases (Wainer 2007). Besides, the numerator is probably underestimated not only because the total FS sample from which subsamples are drawn is a fraction of its overall population size in the wild (and likely affected by problems such as uneven sampling across the distribution range—Cardini 2020a). The numerator is an underestimate also, and simply, because those FS means originate from random subsamples of a bigger sample, which introduces a degree of autocorrelation and makes them more similar than expected in truly independent samples of a population. This second issue likely applies to all indices. Furthermore, because we simulate sampling error only in the FS, holding the others constant, we do not assess the simultaneous effect of sampling error in each of the species, which would almost certainly increase uncertainties. Thus, rather than being pessimistic, our results are probably overoptimistic and, assuming their generalizability is demonstrated, a really cautious morphometrician should probably aim at samples larger than the minimum Ns we tentatively suggest. Nevertheless, there might be a few instances where within FS sample variance was in fact inflated because the focal taxon might in fact include more than one species, as we discuss later using *P. ellioti* as an example.

Even if our results are generalizable in the context of the taxonomic assessment of morphospecies, sampling error, like measurement error (Cardini 2014; Fruciano 2016), is always relative to the amount of 'true' differences in a study. The same absolute error, which would invalidate a study on a small amount of group variation, could be tolerable in a different comparison, focusing on much larger differences. Nevertheless, even studies at higher taxonomic levels often involve a hierarchy of variance partitioning. Therefore, one must be sure that errors are negligible at all levels of the analysis. This suggests that larger samples not only help to increase accuracy, but may also allow more flexibility in terms of applications.

Differences Between Taxa and Indices

Specific groups were sometimes more strongly impacted by sampling error. If there is a taxonomic bias, however, its severity varies with the type of index and therefore depends on what is being quantified. In contrast, the specific landmark configuration seems less important, as results are, for the majority, congruent between the full and reduced configurations.

Primates performed particularly poorly in terms of WRONG SELF, requiring larger samples than other taxa. Primates also comprise many of the cases with highest values for WRONG SISTER, although European hares and *D. novemcintus* were also relatively poor at discovering the correct phenetic sister species when N is small. For BG-RV, European hedgehogs clearly stood out as the group with the largest variability due to error in estimates of FS mean shapes. Unlike these indices, which are mainly assessing the relative accuracy of mean shapes in relation to interspecific differences, the other three indices, focusing on within FS variation, do not generally show evident taxonomic biases. W-RV-median is almost unbiased, whereas the range of estimates of within species variance (W-RV-range), as well as the degree of correlation between matrices of variance covariance (VCVr), are similarly strongly affected by sampling, regardless of the species and the specific landmark configuration. For W-RV, our results

are congruent with the effect on the coefficient of variation of sampling error in empirical datasets of craniodental measurements (Cope 1993): the coefficient is moderately underestimated in the smallest samples (close or equal to N = 5) but this is accompanied by a remarkable increase in the range of estimates and thus a strong reduction in precision.

Thus, it seems that predicting a minimum N for relative accuracy of mean shapes in a group of closely related species (as assessed by WRONG SELF/SISTER and BG-RV) is more difficult and tends to be specific of the study group, which implies that results are less easily generalizable. In contrast, within species (W-RV and VCVr) results could be more robust and general, as they are less dependent on the choice of the study taxon. This apparent difference in the larger or smaller variability of the effect of sampling error in relation to the between *vs* within species level of the analysis seems like a potentially intriguing conclusion, but also one that clearly requires to be confirmed in future studies. However, bearing in mind this caveat, we explore in the next sections why the different indices might vary. Because, in this respect, the 'supra-specific' indices, based on mean shapes, show more differences, we will mostly focus on them. Thus, we start with BG-RV, more easily comparable to results from previous studies and more homogeneous across taxa; go on with WRONG SELF, with its clear taxonomic bias; and finally conclude discussing WRONG SISTER, highly variable across taxa and, to some extent, datasets.

BG-RV: How Much Do Mean Shapes Vary Because of Sampling Error and Why Does That Change Among Taxa?

BG-RV is a numerical version of the graphical approach shown in Fig. 2 and used in previous research (Cardini et al. 2015): it quantifies the amount of interspecific mean shape space 'occupied' by variability in estimates of a species mean due to sampling error. This metric is also related to the ratio of Procrustes shape distances between means employed in our first study on sampling error in geometric morphometrics (Cardini and Elton 2007). In that paper, we computed the Procrustes shape distance between the observed FS sample mean shape and either the means of its random subsamples (averaged) or the mean of a closely related single 'outgroup' species. These two quantities are analogous to respectively the numerator and denominator of BG-RV. Cardini and Elton's (2007) FS was vervet monkeys (*Chlorocebus*), with all currently recognized species treated as a single superspecies, and the 'outgroup' the blue monkey, *C. mitis*. By progressively reducing the size of the random subsamples of vervets, they showed (Cardini and Elton 2007) that with 10–30 individuals (p. 129) "the error in the mean shape estimate can be on average as large as 37–20% of the interspecific distance between mean shapes of *C. aethiops* and *C. mitis*, two species that diverged about 8 million years ago (Tosi et al. 2005) and which have profound differences in their ecology and behaviour".

Cardini and Elton's (2007) finding, that means of subsamples of ten individuals, within a species, have an average distance of almost 40% of its distance to a different species, seems even worse than our results using BG-RV, which suggests that differences among FS means from subsamples of just ca. 15 individuals on average account for 10% or less of the observed interspecific variance. Thus, with similar Ns, the magnitude of the error in Cardini and Elton (2007) appears almost four times larger than in our study. However, results are not strictly comparable, because BG-RV is related to but somewhat different than the ratio of Cardini and Elton (2007). For this reason, in TOTAL, FACE and HALF, we also computed the same type of ratio as in Cardini and Elton (2007) and called it BG-RV2. The index is redundant, because it is related to BG-RV1 and measures, in a slightly different way, the same aspect of the impact of sampling error. However, it allows to compare more directly our study with previous findings and helps to provide a better contextualization. Thus, we briefly summarize in the Discussion the results of BG-RV2, which was obtained using the median distance between random subsample means and their parent FS mean in the numerator and the median of the distances of the other species to the same FS mean (denominator). With BG-RV2, we found that, somewhat surprisingly, despite the differences in their study (larger configuration of 86 unilateral landmarks, different FS and their use of a single interspecific distance), our current analysis reproduces almost perfectly their main finding: for a BG-RV2 of ca. 0.4 (i.e., 40% of interspecific distances 'accounted for' by within FS sampling error), one needs no fewer than ca. 15 specimens, with a trimmed range of five to 25 depending on the species and dataset. This result also mirrors almost exactly the range of N for BG-RV ≤ 0.1, but clearly provides a different and less optimistic perspective on the impact of sampling error: the minimum N for the uncertainty in FS means to be constrained within a small portion of the volume of the interspecific mean shape space (BG-RV < 0.1) corresponds to errors in estimates that can be > 1/3 of the average interspecific mean difference in that group (BG-RV2 ≈ 0.4). This is why Cardini and Elton (2007) argued that 30 or more specimens may be needed to reduce inaccuracies in species mean shapes. In our current study, to keep the percentage of interspecific mean distance accounted for by errors in a species mean estimate below 30–20-10% (i.e., BG-RV2 < 0.3, 0.2 and 0.1), one would need on average respectively ca. 25–60-240 specimens in the FS, which is again in very good agreement with Cardini and Elton (2007). At the opposite extreme of variation in sample size, with just three-four specimens, as not unusual with fossils, the median distance of subsample means to the total FS mean becomes

about as large as the median interspecific mean distance in a clade (BG-RV2 ≈ 1) in 40% of the analyses.

These last observations, from the comparison of BG-RV and BG-RV2, have a potential implication for the interpretation of the results of WRONG SISTER. If, even with an average N = 15, uncertainties in FS means due to sampling error can be comparatively large, it seems likely that the position of the mean shape of FS relative to those of other species can vary widely. Then, means of randomized FS subsamples might frequently end up close to the mean of a species which is not its observed phenetic sister, thus leading to frequent inaccuracies (10–35% of incorrect sister species identifications). This type of inaccuracy is also likely to become more common with more species in a lineage and especially if interspecific differences are small compared to within species variation.

Discovering why an index varies in relation to the study group is a challenge. However, one can explore the correlations between the minimum Ns for reaching the 10% threshold (e.g., BG-RV < 0.1) and the main descriptive statistics of the study samples (Table S2). For BG-RV, in all datasets, the sample size required for BG-RV < 0.1 has a high negative correlation (− 0.76, − 0.70 and − 0.77, respectively in TOTAL, HALF and FACE) with the observed magnitude of interspecific variance in mean shapes. The negative correlation is even slightly stronger (− 0.78, − 0.82 and − 0.77, respectively in TOTAL, HALF and FACE) if the mean interspecific variance is divided by the observed within FS total sample variance. Like an F ratio (although clearly not the same as!), this ratio says something about how big observed mean interspecific differences are compared to those among individuals within the FS. For brevity, we call it pseudo-F. Then, it makes sense that a taxon with a large amount of between species differences relative to the variance within the FS (i.e., a large pseudo-F) will not be strongly impacted by sampling error, because the variation within the FS occupies a relatively small portion of the interspecific mean shape space. On the other hand, if FS individuals vary a lot and interspecific mean differences are small (small pseudo-F), the effect of sampling error will be stronger and large Ns will be required to keep BG-RV small. Indeed, using TOTAL as an example, we find that the species requiring larger samples (N > 16 and up to 40) to keep BG-RV < 0.1 are precisely those (the European hedgehog, *C. mitis*, *P. ellioti*, plains zebra and the European hare) with the lowest pseudo-Fs (≈ 0.3–0.5). At the opposite extreme, *D. novemcintus* and the red fox, with pseudo-F ≈ 2, require only 5–6 specimens for the same relative accuracy.

Because phenotypic divergence generally increases with time, one might also expect larger disparity, and thus smaller impacts of sampling error, in older lineages. Indeed, despite the large uncertainties around our crude approximations of evolutionary age (Upham et al. 2019), the pseudo-F tends to increase with the lineage evolutionary age (e.g., *Dasypus* with the largest pseudo-F is the oldest taxon and *Erinaceus* with the smallest is one of the youngest), but the correlation is moderate (r = 0.55–0.59), which is probably why evolutionary age is not a good predictor (average r ≈ − 0.33) of the minimum N for BG-RV < 0.1. In contrast, evolutionary age correlates negatively with the minimum N for WRONG SELF < 0.1 (with r ranging from − 0.42 to − 0.61), which may partly explain why primates, on average younger than other groups, are so strongly impacted by sampling error in species identification. Yet, this should imply that interspecific differences are smaller in younger groups, but the correlation between age since the last common ancestor and interspecific mean variance is in fact weak (r ≤ 0.2). This is not surprising, because we know that interspecific shape differences do not always increase in a simple linear way with evolutionary time, as the rate of morphological evolution varies widely within and between taxa (Tattersall 1986). For instance, it is very slow in 'living fossils' and typically very fast in insular species (Millien 2006). Therefore, although one might predict a smaller effect of sampling error when the study group contains older species, what really matters is how much bigger the shape divergence among its species is relative to within species variation (i.e., the magnitude of the pseudo-F).

WRONG SELF: When Should We Expect Inaccurate Species Affiliation?

WRONG SELF is similar across datasets but varies among taxa. Also for this index, exploring the correlation between the minimum N for WRONG SELF < 0.1 and the total sample descriptive statistics (Table S2) helps to provide clues on what causes the differences among the groups. Minimum N correlates positively, and consistently in all configurations, with within FS variance (r = 0.51–0.66). Although weaker, it has also, like BG-RV, a negative correlation (average r ≈ − 0.44) with pseudo-F. The reasons for these correlations are partly the same as for BG-RV. A larger variance within the FS implies a higher probability that small subsamples will produce 'unusual' mean shapes, which may be farther from their own species than from other species means. Indeed, the primate FS species, which perform particularly poorly in terms of correct species identification, have on average at least 50% more variance compared to the average in other clades.

Why then does the observed within FS variance vary three folds in magnitude across the different species? It could be simply because some FS samples are larger than others and therefore capture more differences in a population, but this is only a partial and unsatisfactory explanation,

as shown by the small correlation between variance and FS total N (r = 0.25–0.36). It is also possible that within species variance is slightly inflated by a relative larger measurement error in smaller animals (Polly 1998), which, together with their recent evolutionary divergence, is a likely reason why the small European hedgehog has a large pseudo-F and is strongly impacted by sampling error. Yet, large variance might instead reflect genuine variability in relation to the specific evolutionary history, pattern of distribution and breadth of ecological adaptations or degree of plasticity of a species. However, variance could also be biased by how well the museum samples cover the full geographic range of a species (Albrecht and Miller 1993; Cope 1993; Harrison 1993; Cardini 2020a). For *D. novemcintus*, a small variance is almost certainly an artefact of sampling. Although detailed information was missing, this is an unusual sample, as the majority of specimens originate from the same collection and likely are closely related zoo animals. In contrast, why primates tend to have more within FS variance than found in species of other placental orders is less easy to say. It might be simply because primate data were collected in more museums than those used for the other groups, which included only the main European museums as funds were limited and specific for those institutions. As specimens from different institutions generally originate from different time and localities, they will be less autocorrelated and more representative of the overall species geographical range. Regardless of the reason for having more variability within FS species, it is probably largely because of this, that primates require at least 2–3 times more specimens for the same relative accuracy in the identification of the correct species.

If it is reasonable to have a larger minimum N with a larger FS variance, one might also expect that species identification improves using more landmarks to capture more information (something which is often claimed but is potentially misleading—Cardini 2020b). Yet, in our datasets, doubling the number of landmarks (in TOTAL compared to FACE or HALF) produces a rather negligible improvement in species predictive accuracy: on average WRONG SELF < 0.1 is achieved with N = 6 in TOTAL vs. N = 7 in FACE and HALF. With fewer landmarks, however, some species do need larger samples, as suggested by the upper extreme of the trimmed range of N going from 7 (TOTAL) to 10 (FACE and HALF). For now, any conclusion on whether more landmarks could help to mitigate issues with sampling error is premature and likely the answer to this question will change from case to case. Also, one should be careful because, even if in a specific study there were good reasons for increasing the number (p) of morphometric descriptors, there is a trade-off between N and p. Several recent studies (Kocovsky et al. 2009; Bookstein 2017; Björklund 2019; Cardini et al. 2019) have drawn attention to a well known statistical problem, mentioned in every introductory text on multivariate statistics (e.g., Hair et al. 1998): when N is not adequately large compared to p (which generally means N >> p), methods such as PCA, DA and between group PCA, and some other multivariate techniques, may have serious problems and potentially produce spurious findings. In general, the choice of measurements, and therefore landmarks, must be functional to the specific study hypothesis (Oxnard and O'Higgins 2009), and quantity is clearly not a simple substitute for quality (Cardini 2020a, b).

Overall, a median minimum sample size of 6–7 for a species mean shape to be correctly assigned to its own species at least 90% of the time (i.e., WRONG SELF < 0.1) seems quite modest, which is good news but (for the reasons we already discussed) likely to be overoptimistic. In fact, with rare fossils, one may often have samples even smaller than that and it is not uncommon to have just one individual, whose taxonomic affiliation needs to be assessed (Simpson 1943). Besides, if all from the same site and stratum, and thus likely to be close relatives, individuals from a palaeontological excavation may be highly autocorrelated. They could also be more similar than real simply because, if fragmentary, they have been partly virtually reconstructed using computerized methods (Gunz et al. 2009). In all these instances, taxonomic assessment in relation to putatively closely related species should therefore be particularly cautious (Shea et al. 1993; White 2014).

With palaeontological data, there are also other potential problems. Sex is often an important confounding factor (Martin and Andrews 1993; Cameron 1997), unlike in our study, where we knew in advance which species show strong sexual dimorphism and the majority of individuals were of known sex. In theory this issue can be mitigated by selecting traits showing little or no dimorphism (Cope 1993), but these may be hard to find (or even absent) and greatly reduce the range of morphological evidence available for taxonomic assessment. Besides, variability over time can also bias estimates, with effects that may be difficult to predict but, at least using assumptions of gradual change, likely to increase differences over tens of thousands of years (Hunt 2004), which is well within the duration of the average lifespan of ca. 0.5–2 million years of a mammal species (Regan et al. 2001; Ceballos et al. 2015). How this inflated variance over time interacts with evidence from samples that inevitably represent point localities of a larger distribution range is hard to predict, but clearly another reason for increased caution in palaeontological taxonomic assessment (Cope and Lacy 1992; Albrecht and Miller 1993; Harrison 1993; Godfray et al. 2004).

WRONG SISTER: What About Inaccurate Similarity Relationships Among Species?

The considerations made in the conclusion of the previous paragraph are similar and even more important when the aim is to infer the similarity relationship of a sample, and more precisely of its mean, to other closely related species. We exemplified this aim by focusing on the errors in the prediction of the correct phenetic sister species (WRONG SISTER), which is concerned only with similarity using a specific set of morphometric descriptors and may or may not be informative on phylogenetic relatedness. Results for this index are more complex to interpret. The minimum N required for errors to occur less than 10% of the time has a low correlation ($r < 0.5$) with virtually all of the main descriptive statistics (Table S2). The exception is the positive correlation with the total number of species in a taxon, but even this is inconsistent, as it is high (0.66) in TOTAL and rather small in the other two sets of simulations ($r < 0.3$). However, it is reasonable that predicting the right phenetic sister species becomes more difficult with higher diversity in a lineage. Nevertheless, this effect is almost completely negligible in the reduced configurations. One would also expect that, with larger differences between interspecific means and/or less variation within FS, sampling error should be reduced (as in WRONG SELF) but this does not happen, as correlations with observed variances (as well as their pseudo-F ratios) are always small ($-0.2 \leq r \leq 0.4$) and sometimes even inconsistent in sign.

Thus, WRONG SISTER and WRONG SELF have some commonalities, but also clear differences. The estimated minimum Ns for accurate (< 10% errors) predictions of a phenetic sister species are on average at least twice larger than those of WRONG SELF, but the most striking difference is that, for WRONG SISTER, they vary widely depending on the lineage (from a few specimens to several hundreds). It is likely that WRONG SISTER is influenced by many factors, including the possibility that one or more species are about as close as the observed phenetic sister to the observed FS mean. Thus, when the mean of the FS changes, even if that may be a slight change in a big sample, that is enough to move it closer to the wrong phenetic sister species. If so, the precise geometry of the space of the species mean shapes can be even more important than the number of species in the lineage or the magnitude of between and within species variance.

Overall, Then, When Should We Expect a Stronger Impact of Sampling Error?

Discussing the three most variable indices has provided the first clues on what may contribute to change the severity of the impact of sampling error. This was specific to the estimates of a species mean shape, one of the most important descriptive statistics in the taxonomic assessment of a morphospecies. Of the other three indices, focusing on within species variances and covariances, one (W-RV-median) is almost unbiased and the other two (W-RV-range and VCVr) suggest for all taxa very similar requirements of minimum sample size for high relative accuracy (ca. 30–50 individuals on average). Correlational exploratory analyses, using minimum Ns ('< 10% error threshold'), as with the indices based on mean shapes, did not discover any consistent and large correlation between W-RV-range or VCVr and the main descriptive statistics (Table S2).

The heterogeneity of sample size requirements for mean shapes in relation to interspecific differences, and the fair homogeneity of patterns in the case of within species variance and covariance, seem to indicate differences in how sampling error affects parameters depending on the level of the analysis. This is a first preliminary but potentially important conclusion of our work. However, even if there might be differences between indices using species means and those using individuals within a species, we decided to explore further the overall results from all indices in order to look for a possible general explanation of our findings. Thus, excluding the almost unbiased W-RV-median, we transformed the minimum Ns of Table 4 (obtained using the '< 10% error threshold') into standardized z-scores and averaged them across the three main sets of results (TOTAL, FACE and HALF). Compared to the raw results, z-scores preserve the relative differences in Ns while rescaling them more uniformly. This avoids that a few very large N in WRONG SISTER might dominate the averaged results. Indeed, the averaged z-scores are as a sort of ranking, where lower (negative) scores imply a smaller impact of sampling error and larger (positive) ones indicate a stronger impact. From this, two well separated clusters emerge: 1) the least impacted ($-0.49 \leq z \leq -0.18$), which are *V. vulpes*, *P. capensis*, *D. novemcinctus*, *E. burchellii*, *L. europaeus*, and male and female *P. anubis*; 2) the most impacted ($z \geq 0.05$), which are all other primates plus the European hedgehog, with this species as well as female *C. mitis* and male *P. ellioti* being the most badly affected by sampling error ($z \geq 0.48$ *vs* $0.05 \leq z \leq 0.21$ in the remaining species).

As with the raw indices, one can eventually investigate the correlations between the averaged z-scores and the main descriptive statistics (Table S2, plus the pseudo-F ratios—Fig. 5g). This shows that there are only a few strong correlations (ranging from − 0.58 to − 0.66), which are consistently those with the pseudo-F ratios of the different configurations. Although smaller, and positive ($r = 0.45$), there is also a potentially interesting correlation with the number of species in a group. Thus, a fairly simple, partial and very preliminary explanation seems to emerge for the variable severity of the problems with sampling error, as assessed

in our study. With larger between species differences and smaller within species variance, relatively small N can still produce fairly accurate estimates of means, variances and covariances. However, when within species variance is big compared to interspecific differences, one then needs really big samples for accuracy. For instance, using an hypothetical example, one could do well with a dozen, or a few dozens, of specimens, in a study of a population with small variability and a highly distinctive shape, such as an insular species that has gone through genetic bottlenecks and may have an accelerated rate of morphological evolution (Millien 2006). However, during an adaptive radiation on a continent, with at least some species having large populations, and thus big variability among individuals, as well as potentially variable degrees of interspecific divergence, accurate estimates of means and variance covariance structure will require much larger samples. Besides, with more species, the problem might become even more serious, but this appears, at least in our analysis, as a much less relevant factor compared to the pseudo-F ratio of variances. Indeed, even with few species, if interspecific differences are small and within species variation large, the impact of sampling error may be serious, as convincingly shown by hedgehogs.

The Influence of Taxonomic Uncertainties and Fuzzy Interspecific Boundaries

Primates and especially *C. mitis* and *P. ellioti* were often strongly affected by sampling error even in relatively large samples. On average, the primate clades we used in the study are younger than those from other orders but this has not produced smaller interspecific divergence in cranial morphology, as in fact the variance of mean shapes is slightly bigger in this group (Table S2). However, primate FS also have larger within species variance, so much larger that their pseudo-F ratios are almost 30% smaller than in FS of other placental orders. Because relatively larger within species variance compared to interspecific differences seems the best predictor for a stronger impact of sampling error, so that often primates required larger Ns for the same relative accuracy, it is important to understand the reasons for the larger FS variability of the primates.

Besides the possibility briefly mentioned above that this is largely an effect of better sampling across more museums, taxonomic uncertainties may also have played a role. Indeed, at least some of the primate FS might belong to the 'grey area' of taxonomy where populations may be considered either species or subspecies, depending on the criterion used to establish these taxonomic categories (Zachos 2016). Thus, there could be a degree of taxonomic inaccuracy and cryptic diversity that inflates within species variance in cranial shape. As an example, we will focus on the case of *P. ellioti*, which, like most other red colobus, is characterized by an unstable taxonomy and a complex pattern of evolutionary divergence and potentially incomplete reproductive isolation (Oates and Ting 2015).

As Zachos (2016, p. 143) discusses, we undertake zoological research in a "continuous world with fuzzy boundaries". In Table S1 we outline that the taxonomic schemes we adopt in our analyses are not the only ones available to us, and there is considerable discussion about the composition of the species we include. For instance, in the past two decades, *P. ellioti* has been included in the genus *Procolobus* and subgenus *Piliocolobus* (although it is now suggested that *Piliocolobus* should be raised to a full genus) and assigned, as a subspecies, to several different species (*oustaleti*, *badius, pennantii* and *rufomitratus*—see review in Maisels and Ting 2020). Indeed, Grubb et al. (2003) considered *ellioti* a subspecies but did not assign it to a species. The scheme we use in this paper follows Grubb et al. (2003), which is the classification used by most museums at the time of data collection, but raises *ellioti* to a full species. However, others, such as the IUCN Red List (Hart et al. 2020; Maisels and Ting 2020) do not consider *ellioti* a valid taxon. Instead, the red colobus we describe as *ellioti* is split into two species, *Piliocolobus langi* and *Piliocolobus semlikiensis* (the geographic distribution of our *ellioti* sample covers the ranges of both taxa). Consideration of the underlying biology helps to make sense of this taxonomic instability. It appears that the red colobus in the '*ellioti*' range (in northeast Democratic Republic of Congo) comprise a 'hybrid swarm' of three potential taxa, *langi, semlikiensis* and *oustaleti* (Groves 2007). All have been reported to be externally phenotypically similar (e.g. in pelage and skin colour) and such variation is largely continuous (Groves 2007), although Struhsaker (2010) remarked that variation may be more extensive than previously supposed. In cranial morphology, *P. ellioti* and *P. oustaleti* cluster together, with no significant differences in size and shape between the taxa, and a relatively small shape distance (Cardini and Elton 2009). The results we report here bring the 'fuzzy boundaries' of these taxa into sharp relief. As anticipated, we use *P. ellioti* as an example, not least because primates are among the best studied of all the mammals we analysed, and in consequence have experienced more taxonomic revision. Nonetheless, the issues highlighted are likely to be applicable across several other mammals, given the evidence for hybridisation (Taylor and Larson 2019) and the longstanding debates over how to recognise and demarcate species (Zachos 2016). For now, our pragmatic solution was to follow an older taxonomic scheme and accept that there will always be uncertainty in where we draw boundaries between species. However, this decision means that taxonomic uncertainty might have affected some of our analytical results by inflating intraspecific variance in a 'compound' FS such as *P. ellioti*. Yet, in most other cases, as we discussed before, the variability in the wild populations

of the FS has been almost certainly underestimated, which makes our results in terms of the severity of sampling error likely to be conservative, with minimum Ns suggested for relative accuracy smaller than truly desirable.

Indeed, in the red colobus as well as in all other clades, the choice of the FS was dictated by a simple practical consideration: the importance of using the largest available sample to better approximate accuracy. One might wonder, however, to what extent this selection might have influenced results. The example case of *P. ellioti* highlights the issue of uncertain or contested taxonomy, which is actively being debated for this taxon but is certainly not unique. *Cercopithecus mitis* is another example, which like *P. ellioti* is widely distributed and contains many subspecies (Grubb et al. 2003), some of which are sometimes elevated to species status (Upham et al. 2019). The trade-off between taxonomic uncertainty and sample size is an important one to consider. Our results demonstrate that larger sample sizes are better for estimating mean shapes and covariance patterns, but the drive to accumulate a large sample of most mammals based on museum collections almost always requires combining individuals sampled from across the species' geographic range. This strategy, however, risks combining subgroups that are genetically distinct and thus different in their mean shape and covariance structure. Our primate results may well reflect this issue: the relatively larger minimum Ns required for primates could have been the result of a bias caused by the choice of a taxonomically mixed and therefore potentially inaccurate FS. As already discussed, we cannot completely rule out this possibility but we can explore what would have happened if we had, for instance, chosen the second largest sample available for either the red colobus or *Cercopithecus-Chlorocebus* monkeys. Thus, we replicated the analyses using as FS *P. badius* and *P. oustaleti*, for respectively female and male red colobus, and *C. pygerythrus* for both sexes in the *Cercopithecus-Chlorocebus* clade. Results were almost identical, with the median N for the '10% threshold' ranging between 20 and 40 in *badius-oustaleti* and between 23 and 40 in *pygerythrus* (compared to 14–40 in *P. ellioti* and 26–40 in *C. mitis*). The findings of the main analysis, including the requirement on average larger samples for relative accuracy in primates, are therefore confirmed and were not biased by selecting *P. ellioti* and *C. mitis*.

In fact, with the same approach, using whenever possible the second largest samples ($N \geq 40$, with a median FS $N = 56$) of each clade as a FS to replicate the simulations, we find an excellent agreement with the results of the original analysis (Table 2). Although this limits the comparison of desirable Ns to *E. roumanicus*, *Vulpes lagopus*, and *Lepus timidus*, among the non-primates, relative to *C. pygerythrus*, males of *P. cynocephalus* and females and males of respectively *P. badius* and *P. oustaleti*, among primates, the upward bias of the primates is well supported with an overal median N of 18 for non-primates and 30 for primates (vs. respectively 20 and 26, for the same groups, in the original results of Table 4). The re-analysis also shows an increase in both the median and 90th percentile for the minimum Ns of all FS, across all indices, from respectively 23–40 (original analysis of Table 2) to 26–45 (re-analysis using different FS). Thus, it does seem that at least using the clades we selected for this study, results might be robust and the main patterns supported more generally by the data we analysed.

Conclusions

This study is probably one of the broadest on sampling error in geometric morphometrics. The total sample size is large and there is at least one taxon for each of the four placental superorders. Nevertheless, it is still a small portion of the overall diversity of placental mammals and we were limited to analyses of adults, leaving unexplored the effect of ontogenetic growth on sample size issues. Also, we only considered cranial shape, which has its own peculiarities of genetic underpinnings, developmental processes, dimorphism, and variation, that may produce intra- and inter-specific variance patterns that are different from other morphological systems (e.g., Caumul and Polly 2005; Polly et al. 2013). Thus, in our study, the accuracy with which morphometrics can be used to identify specimens to species-level taxa depended mainly on the interaction between within-group variability and between-group differences, which in turn determined the size of the sample needed for a correct classification. The relative amount of between to within variation depends on at least three factors: precision of measurement (imprecision adds to apparent within-group variation and subtracts from apparent between-group difference), genetic variation within species versus genetic divergence between them, and non-genetic environmental variation (e.g. bone remodeling in response to stress). These factors differ by species, by morphological system, and even by individual. Highly variable but shallowly diverged species will be more prone to classification error than will deeply diverged species all other things being equal. Genetically and developmentally complex morphologies with less non-genetic environmental variation (e.g., molar shape) should in contrast provide a more accurate classification than morphologies with simpler genetics and development but more ecophenotypic variance (e.g., mammalian long bones). For this reason, and in spite of the good correspondence of many of our conclusions about the minimum N for relative accuracy with those of previous studies in mammals, it is too early to claim generalizability even about evaluating morphospecies using adult cranial data of closely related taxa at the boundary between micro- and macro-evolution (i.e., within species and in

relation to closely related ones). Besides, as we stressed multiple times, even in terms of internal validity of the analysis, our estimates are almost certainly overoptimistic.

Bearing in mind these caveats, our results tentatively suggest that:

(1) A minimum sample of 10–15 adult specimens (per sex if the species is dimorphic) is required to estimate mean shape and to have a low standard error of the mean relative to the variance among members of a mammalian genus. Such a sample size gives a good chance of numerically identifying the species correctly. However, note that with this sample size the estimated mean could differ from the true mean by about 40% of the distances separating species within a genus. Thus, N of at least 40–50 will give much better approximations
(2) For reconstructing accurate similarity relationships and finding the correct phenetic sister species, samples must be typically larger (ca. 15–90 on average but up to more than 100–200 in some cases) and requirements of minimum N will vary considerably depending on the taxon
(3) For reasonable estimates of the magnitude of total variance and VCV, 30–50 specimens may be enough on average, but to increase confidence in these estimates one should aim at even larger N.

Overall, we conclude that ca. 25–40 specimens (depending on using either the median or 90th percentile of Ns in the results of all our simulations shown in Table 4) is the best sample size across the board based on the 10% threshold of all our indices and datasets. But even this should not be taken as a general and definitive conclusion, because there is large variability depending on the taxon, configuration and the parameter being assessed. Therefore, for some situations, even larger samples may be required to produce robust results and, clearly, there is no universal recipe for controlling for sampling error. Moreover, a morphospecies is just a morphospecies and, regardless of the size of the sample and quality of the data, one should always remember that it is an important but small piece of taxonomic evidence (Simpson 1943; Jolly 1993), even smaller when based just on a given ontogenetic stage and a specific set of anatomical traits (such as crania or other skeletal parts—Godfray et al. 2004).

Taxonomic accuracy is central to all biological research both in living and extinct organisms. Resampling experiments, such as those we used in this and previous studies, have limitations but allow to start exploring the sensitivity of results to sampling error. This type of analysis should be encouraged, if we want to improve the assessment of morphospecies by providing information on the confidence one can have in her/his results. This might help not only to avoid overstatements and reduce the risk of taxonomic inflation, but also to make the classification more stable and useful. Even when working with fossils, that rarely offer large samples, one can easily explore the problem using rarefaction analyses in modern living relatives (Cope and Lacy 1992, 1995; Roth 1992; Jolly 1993; Martin and Andrews 1993; Plavcan 1993; Plavcan and Cope 2001), if available and under a uniformitarian assumption of roughly similar evolutionary patterns, or with numerical simulations (Kelley and Plavcan 1998; Plavcan and Cope 2001). Without taxonomy, biology is indeed a "meaningless jumble" (May 1990, p. 130), but it can be a chaotic jungle if taxonomy is inaccurate and its uncertainties are not acknowledged.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11692-021-09531-3.

Acknowledgements We are very much in debt with Nathan Upham for his essential and prompt help with part of the phylogenetic and taxonomic background of this study. A special thank also to Frank Zachos and Bernard Wood for their insightful comments on species concepts and population boundaries, and to the anonymous reviewers, whose suggestions greatly improved the original manuscript. Crucial to this research were also all museum curators and collection managers of the many institutions AC visited to collect the data, as well as the funders who provided fundamental economic support for some of the various parts of this project, and mainly the Leverhulme Trust, Durham COFUND and SYNTHESYS. The research aim of this paper has had, for AC, a long personal history, which begins with a naive question to Marco Corti (1950–2007) about the 'right' number' of specimens to study the similarity relationships of marmots using geometric morphometrics; that happened 22 years ago, while AC was a PhD candidate under the supervision of Paolo Tongiorgi (1936–2018): for their great mentoring, friendship and example, we wish to dedicate this work to Marco and Paolo, but also to all students who asked the same question and all supervisors who ever had to grapple with the answer.

Funding The funding was supported by Leverhulme Trust (GB) and COFUND.

Data Availability The full list of museum specimens is available upon request from the corresponding author.

Compliance with Ethical Standards

Conflict of Interest The author declares that they have no conflict of interest to disclose.

References

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution.' *The Italian Journal of Zoology, 71,* 5–16.

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix, 24,* 7.

Albrecht, G. H., & Miller, J. M. A. (1993). Geographic variation in primates. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 123–161). Boston, MA: Springer.

Alroy, J. (2002). How many named species are valid? *Proceedings of the National Academy of Sciences, 99,* 3706–3711.
Barnosky, A. D. (2008). Megafauna biomass trade-off as a driver of quaternary and future extinctions. *Proceedings of the National Academy of Sciences, 105,* 11543–11548.
Barnosky, A. D., Koch, P. L., Feranec, R. S., Wing, S. L., & Shabel, A. B. (2004). Assessing the causes of late pleistocene extinctions on the continents. *Science, 306,* 70–75.
Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., et al. (2011). Has the Earth's sixth mass extinction already arrived? *Nature, 471,* 51–57.
Benton, M. J. (2008). Fossil quality and naming dinosaurs. *Biology Letters, 4,* 729–732.
Björklund, M. (2019). Be careful with your principal components. *Evolution, 73,* 2151–2158.
Bookstein, F. L. (2017). A newly noticed formula enforces fundamental limits on geometric morphometric analyses. *Evolutionary Biology, 44,* 522–541.
Boroni, N. L., Lobo, L. S., Romano, P. S. R., & Lessa, G. (2017). Taxonomic identification using geometric morphometric approach and limited data: An example using the upper molars of two sympatric species of Calomys (Cricetidae: Rodentia). *Zoologia, 34,* 1–11.
Brown, C. M., & Vavrek, M. J. (2015). Small sample sizes in the study of ontogenetic allometry; implications for palaeobiology. *PeerJ, 3,* e818.
Cameron, D. W. (1997). Sexual dimorphic features within extant great ape faciodental skeletal anatomy and testing the single species hypothesis. *Zeitschrift für Morphologie und Anthropologie, 81,* 253–288.
Cardini, A. (2014). Missing the third dimension in geometric morphometrics: How to assess if 2D images really are a good proxy for 3D structures? *Hystrix, The Italian Journal of Mammalogy, 25,* 73–81.
Cardini, A. (2017). Left, right or both? Estimating and improving accuracy of one-side-only geometric morphometric analyses of cranial variation. *Journal of Zoological Systemetics and Evoloutionary Research, 55,* 1–10.
Cardini, A. (2019a). Craniofacial allometry is a rule in evolutionary radiations of placentals. *Evolutionary Biology*. https://doi.org/10.1007/s11692-019-09477-7.
Cardini, A. (2019b). Integration and modularity in procrustes shape data: Is there a risk of spurious results? *Evolutionary Biology, 303,* 2747–2765.
Cardini, A. (2020a). Modern morphometrics and the study of population differences: Good data behind clever analyses and cool pictures? *Anatomical Record, 303,* 2747–2765.
Cardini, A. (2020b). Less tautology, more biology? A comment on "high-density" morphometrics. *Zoomorphology, 139,* 513–529.
Cardini, A., & Chiapelli, M. (2020). How flat can a horse be? Exploring 2D approximations of 3D crania in equids. *Zoology, 139,* 125746.
Cardini, A., & Elton, S. (2007). Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology, 126,* 121–134.
Cardini, A., & Elton, S. (2008a). Variation in guenon skulls (II): Sexual dimorphism. *Journal of Human Evolution, 54,* 638–647.
Cardini, A., & Elton, S. (2008b). Variation in guenon skulls (I): Species divergence, ecological and genetic differences. *Journal of Human Evolution, 54,* 615–637.
Cardini, A., & Elton, S. (2009). The radiation of red colobus monkeys (Primates, Colobinae): Morphological evolution in a clade of endangered African primates. *Zoological Journal of the Linnean Society, 157,* 197–224.
Cardini, A., & Loy, A. (2013). On growth and form in the "computer era": From geometric to biological morphometrics. *Hystrix, The Italian Journal of Mammalogy, 24,* 1–5.
Cardini, A., O'Higgins, P., & Rohlf, F. J. (2019). Seeing distinct groups where there are none: Spurious patterns from between-group PCA. *Evolutionary Biology, 46,* 303–316.
Cardini, A., & Polly, P. D. (2013). Larger mammals have longer faces because of size-related constraints on skull form. *Nature Communications, 4*(2458), 1–7.
Cardini, A., Seetah, K., & Barker, G. (2015). How many specimens do I need? Sampling error in geometric morphometrics: Testing the sensitivity of means and variances in simple randomized selection experiments. *Zoomorphology, 134,* 149–163.
Caumul, R., & Polly, P. D. (2005). Phylogenetic and environmental components of morphological variation: Skull, mandible and molar shape in marmots (*Marmota*, Rodentia). *Evolution, 59,* 2460–2472.
Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances, 1,* e1400253.
Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences, 114,* E6089–E6096.
Cope, D. A. (1993). Measures of dental variation as indicators of multiple taxa in samples of sympatric cercopithecus species. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 211–237). Boston, MA: Springer.
Cope, D. A., & Lacy, M. G. (1992). Falsification of a single species hypothesis using the coefficient of variation: A simulation approach. *American Journal of Physical Anthropology, 89,* 359–378.
Cope, D. A., & Lacy, M. G. (1995). Comparative application of the coefficient of variation and range-based statistics for assessing the taxonomic composition of fossil samples. *Journal of Human Evolution, 29,* 549–575.
Culver, M., Johnson, W. E., Pecon-Slattery, J., & O'Brien, S. J. (2000). Genomic ancestry of the American puma (Puma concolor). *Journal of Heredity, 91,* 186–197.
Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society, 85,* 407–417.
Drew, L. W. (2011). Are we losing the science of taxonomy? As need grows, numbers and training are failing to keep up. *BioScience, 61,* 942–946.
Fang, Z., Fan, J., Chen, X., & Chen, Y. (2018). Beak identification of four dominant octopus species in the East China Sea based on traditional measurements and geometric morphometrics. *Fisheries Science, 84,* 975–985.
Franklin, D., Oxnard, C. E., O'Higgins, P., & Dadour, I. (2007). Sexual dimorphism in the subadult mandible: Quantification using geometric morphometrics*. *Journal of Forensic Sciences, 52,* 6–10.
Fruciano, C. (2016). Measurement error in geometric morphometrics. *Development Genes and Evolution, 226,* 139–158.
Godfray, H. C. J., Knapp, S., Forey, P. L., Fortey, R. A., Kenrick, P., & Smith, A. B. (2004). Taxonomy and fossils: A critical appraisal. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 359,* 639–653.
Groves, C. P. (2007). The taxonomic diversity of the Colobinae of Africa. *Journal of Anthropological Sciences, 85,* 7–34.
Grubb, P., Butynski, T. M., Oates, J. F., Bearder, S. K., Disotell, T. R., Groves, C. P., & Struhsaker, T. T. (2003). Assessment of the diversity of African primates. *International Journal of Primatology, 24*(6), 1301–1357.

Gunz, P., Mitteroecker, P., Neubauer, S., Weber, G. W., & Bookstein, F. L. (2009). Principles for the virtual reconstruction of hominin crania. *Journal of Human Evolution, 57,* 48–62.

Hair, J. F., Anderson, R. E., Tatham, R. L., & William, C. (1998). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall.

Harrison, T. (1993). Cladistic concepts and the species problem in hominoid evolution. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 345–371). Boston, MA: Springer.

Hart, J., Laudisoit, A., Struhsaker, T. T., & Oates, J. F. (2020). *Piliocolobus langi* (amended version of 2019 assessment). *The IUCN Red List of Threatened Species*. https://doi.org/10.2305/IUCN.UK.2020-1.RLTS.T18261A166605018.en.

Hublin, J.-J., Weston, D., Gunz, P., Richards, M., Roebroeks, W., Glimmerveen, J., & Anthonis, L. (2009). Out of the North Sea: The Zeeland Ridges Neandertal. *Journal of Human Evolution, 57,* 777–785.

Hublin, J.-J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., et al. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature, 546,* 289.

Hunt, G. (2004). Phenotypic variation in fossil samples: Modeling the consequences of time-averaging. *Paleobiology, 30,* 426–443.

Jolly, C. J. (1993). Species, subspecies, and baboon systematics. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 67–107). Boston, MA: Springer.

Kelley, J., & Plavcan, J. M. (1998). A simulation test of hominoid species number at Lufeng, China: Implications for the use of the coefficient of variation in paleotaxonomy. *Journal of Human Evolution, 35,* 577–596.

Klingenberg, C. P. (2011). MorphoJ: An integrated software package for geometric morphometrics. *Molecular Ecology Resources, 11,* 353–357.

Klingenberg, C. P. (2013). Cranial integration and modularity: Insights into evolution and development from morphometric data. *Hystrix, The Italian Journal of Mammalogy, 24,* 43–58.

Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Biology, 38*, 7–25. https://doi.org/10.1093/sysbio/38.1.7.

Koch, P. L., & Barnosky, A. D. (2006). Late quaternary extinctions: State of the debate. *Annual Review of Ecology Evolution and Systematics, 37,* 215–250.

Kocovsky, P. M., Adams, J. V., & Bronte, C. R. (2009). The effect of sample size on the stability of principal components analysis of truss-based fish morphometrics. *Transactions of the American Fisheries Society, 138*(3), 487–496.

Kryštufek, B., Janžekovič, F., Hutterer, R., & Klenovšek, T. (2016). Morphological evolution of the skull in closely related bandicoot rats: A comparative study using geometric morphometrics. *Hystrix, 27,* 1–7.

Lele, S. (1991). Some comments on coordinate-free and scale-invariant methods in morphometrics. *American Journal of Physical Anthropology, 85,* 407–417.

Lewis, S. L., & Maslin, M. A. (2018). *The human planet: How we created the anthropocene* (p. 413). London: Penguin Books.

Lindenfors, P., Gittleman, J. L., & Jones, K. E. (2007). Sexual size dimorphism in mammals. In D. J. Fairbairn, W. U. Blanckenhorn, & T. Szekely (Eds.), *Sex, size and gender roles: Evolutionary studies of sexual size dimorphism*. Oxford: Oxford University Press.

Maisels, F., & Ting, N. (2020). *Piliocolobus semlikiensis*. *The IUCN Red List of Threatened Species*. https://doi.org/10.2305/IUCN.UK.2020-1.RLTS.T92657343A92657454.en.

Martin, L. B., & Andrews, P. (1993). Species recognition in middle Miocene hominoids. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 393–427). Boston, MA: Springer.

May, R. M. (1990). Taxonomy as destiny. *Nature, 347,* 129–130.

McGuire, J. L. (2011). Identifying California Microtus species using geometric morphometrics documents quaternary geographic range contractions. *Journal of Mammalogy, 92,* 1383–1394.

Meloro, C., Hunter, J., Tomsett, L., Miguez, R. P., Prevosti, F. J., & Brown, R. P. (2017). Evolutionary ecomorphology of the Falkland Islands wolf *Dusicyon australis*. *Mammal Review, 47,* 159–163.

Millien, V. (2006). Morphological evolution is accelerated among island mammals. *PLoS Biology, 4,* e321.

Mollentze, N., & Streicker, D. G. (2020). Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proceedings of the National Academy of Sciences, 117,* 9423–9430.

Monteiro, L. R. (2013). Morphometrics and the comparative method: Studying the evolution of biological shape. *Hystrix, The Italian Journal of Mammalogy, 24,* 8.

Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., et al. (2017). Morphometric, behavioral, and genomic evidence for a new orangutan species. *Current Biology, 27,* 3487-3498.e10.

Newell, N. D. (1949). Types and hypodigms. *American Journal of Science, 247,* 134–142.

Oates, J., & Ting, N. (2015). Conservation consequences of unstable taxonomies: The case of the red colobus monkeys. In A. M. Behie & M. F. Oxenham (Eds.), *Taxonomic tapestries: The threads of evolutionary, behavioural and conservation research* (pp. 321–343). Canberra: ANU Press.

O'Higgins, P. P. (2000). The study of morphological variation in the hominid fossil record: Biology, landmarks and geometry. *Journal of Anatomy, 197,* 103–120.

Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L., & Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature, 546,* 646–650.

Oxnard, C., & O'Higgins, P. (2009). Biology clearly needs morphometrics. Does morphometrics need biology? *Biological Theory, 4,* 84–97.

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology, 7,* 16.

Pearson, A., Groves, C., & Cardini, A. (2015). The 'temporal effect' in hominids: Reinvestigating the nature of support for a chimp-human clade in bone morphology. *Journal of Human Evolution, 88,* 146–159.

Plavcan, J. M. (1993). Catarrhine dental variability and species recognition in the fossil record. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 239–263). Boston, MA: Springe.

Plavcan, J. M., & Cope, D. A. (2001). Metric variation and species recognition in the fossil record. *Evolutionary Anthropology: Issues, News, and Reviews, 10,* 204–222.

Polly, P. D. (1998). Variability in mammalian dentitions: Size-related bias in the coefficient of variation. *Biological Journal of the Linnean Society, 64,* 83–99.

Polly, P. D. (2005). Development and phenotypic correlations: The evolution of tooth shape in *Sorex araneus*. *Evolution and Development, 7,* 29–41.

Polly, P. D., & Head, J. J. (2004). Maximum-likelihood identification of fossils: Taxonomic identification of quaternary marmots (Rodentia, Mammalia) and identification of vertebral position in the pipesnake Cylindrophis (Serpentes, Reptilia). In A. M. T. Elewa (Ed.), *Morphometrics-applications in biology and paleontology* (pp. 197–222). Heidelberg: Springer.

Polly, P. D., Polyakov, A. V., Ilyashenko, V. B., Onischenko, S. S., White, T. A., Shchipanov, N. A., et al. (2013). Phenotypic

variation across chromosomal hybrid zones of the common shrew (*Sorex araneus*) indicates reduced gene flow. *PLoS ONE, 8*(7), e67455.
Regan, H. M., Lupia, R., Drinnan, A. N., & Burgman, M. A. (2001). The currency and tempo of extinction. *The American Naturalist, 157,* 1–10.
Rodriguez-Morales, A. J., Bonilla-Aldana, D. K., Balbin-Ramon, G. J., Rabaan, A. A., Sah, R., Paniz-Mondolfi, A., et al. (2020). History is repeating itself: Probable zoonotic spillover as the cause of the 2019 novel coronavirus epidemic. *Infez Med, 28,* 3–5.
Rohlf, F. J. (1998). On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology, 47,* 147–158.
Rohlf, F. J., & Slice, D. (1990). Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Zoology, 39,* 40–59.
Rohlf, F. J., & Marcus, L. F. (1993). A revolution morphometrics. *Trends in Ecology and Evolution, 8,* 129–132.
Roth, V. L. (1992). Quantitative variation in elephant dentitions: Implications for the delimitation of fossil species. *Paleobiology, 18,* 184–202.
Sanfilippo, P. G., Cardini, A., Hewitt, A. W., Crowston, J. G., & Mackey, D. A. (2009). Optic disc morphology: Rethinking shape. *Progress in Retinal and Eye Research, 28,* 227–248.
Schlis-Elias, M.C. (2020). Ecological release and allometry explain insular gigantism and shape variation in a widespread North American rodent. Master Thesis. https://aspire.apsu.edu/handle/20.500.11989/6700
Shea, B. T., Leigh, S. R., & Groves, C. P. (1993). Multivariate craniometric variation in Chimpanzees. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 265–296). Boston, MA: Springer.
Simpson, G. G. (1940). Types in modern taxonomy. *American Journal of Science, 238,* 413–431.
Simpson, G. G. (1943). Criteria for genera, species, and subspecies in zoology and paleozoology. *Annals of the New York Academy of Sciences, 44,* 145–178.
Simpson, G. G. (1951). The species concept. *Evolution, 5,* 285–298.
Stec, D., Gąsiorek, P., Morek, W., Kosztyła, P., Zawierucha, K., Michno, K., et al. (2016). Estimating optimal sample size for tardigrade morphometry. *Zoological Journal of the Linnean Society, 178,* 776–784.
Struhsaker, T. T. (2010). *The Red Colobus monkeys: Variation in demography, behavior, and ecology of endangered species*. Oxford: Oxford University Press.
Tattersall, I. (1986). Species recognition in human paleontology. *Journal of Human Evolution, 15,* 165–175.
Tattersall, I. (1993). Speciation and morphological differentiation in the genus *Lemur*. In W. H. Kimbel & L. B. Martin (Eds.), *Species, species concepts and primate evolution* (pp. 163–176). Boston, MA: Springer.
Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology and Evolution, 3,* 170–177.
Tosi, A. J., Detwiler, K. M., & Disotell, T. R. (2005). X-chromosomal window into the evolutionary history of the guenons (Primates: Cercopithecini). *Molecular Phylogenetics and Evolution, 36,* 58–66.
Upham, N. S., Esselstyn, J. A., & Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biology, 17,* e3000494.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
Wainer, H. (2007). The most dangerous equation. *American Scientist, 95,* 249.
White, T. D. (2014). Delimitating species in paleoanthropology. *Evolutionary Anthropology: Issues, News, and Reviews, 23,* 30–32.
Whitmee, S., Haines, A., Beyrer, C., Boltz, F., Capon, A. G., de Souza Dias, B. F., et al. (2015). Safeguarding human health in the anthropocene epoch: Report of The rockefeller foundation-lancet commission on planetary health. *The Lancet, 386,* 1973–2028.
Witteveen, J. (2015). Naming and contingency: The type method of biological taxonomy. *Biology and Philosophy, 30,* 569–586.
Wood, B. (2010). Colloquium paper: Reconstructing human evolution: Achievements, challenges, and opportunities. *Proceedings of the National Academy of Sciences of the USA, 107,* 8902–8909.
Wood, B., & Constantino, P. (2007). Paranthropus boisei: Fifty years of evidence and analysis. *American Journal of Physical Anthropology, 134,* 106–132.
Wood B, Doherty D, & Boyle E. (2020). Hominin taxic diversity. Oxford research encyclopedia of anthropology. https://oxfordre.com/anthropology/view/. Accessed August 2020.
Zachos, F. E. (2016). *Species concepts in biology: Historical development, theoretical foundations and practical relevance*. New York: Springer.
Zachos, F. E. (2018). Mammals and meaningful taxonomic units: The debate about species concepts and conservation. *Mammal Review, 48,* 153–159.
Zelditch, M. L., Swiderski, D. L., & Sheets, H. D. (2012). *Geometric morphometrics for biologists: A primer* (p. 490). Cambridge: Academic Press.

Authors and Affiliations

Andrea Cardini[1,2] · Sarah Elton[3] · Kris Kovarovic[3] · Una Strand Viðarsdóttir[4] · P. David Polly[5]

[1] Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi, 103-41125 Modena, Italy

[2] School of Anatomy, Physiology and Human Biology, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

[3] Department of Anthropology, Durham University, Dawson Building, South Road, Durham DH1 3LE, UK

[4] The Biomedical Center of the University of Iceland, Læknagarður, Vatnsmýrarvegi 16, 101 Reykjavík, Iceland

[5] Earth and Atmospheric Sciences, Biology, and Anthropology, Indiana University, 1001 E. 10th Street, Bloomington, IN 47405, USA