Multivariate analysis of neognath skeletal measurements: implications for body mass estimation in Mesozoic birds

FRANCISCO JOSÉ SERRANO[1]*, PAUL PALMQVIST[1] and JOSÉ LUIS SANZ[2]

[1]*Área de Paleontología, Departamento de Ecología y Geología, Facultad de Ciencias, Universidad de Málaga, Campus Universitario de Teatinos s/n, 29071, Málaga, Spain*
[2]*Unidad de Paleontología, Departamento de Biología, C/Darwin 2, Universidad Autónoma de Madrid, Cantoblanco, 28049, Madrid, Spain*

The abundant fossils of avian stem taxa unearthed during the last years make it necessary to review and improve the models for estimating body mass used in palaeoecological studies. In this article, single and multiple regression functions based on osteological measurements were obtained from a large data set of extant flying birds for estimating the body mass of 42 Mesozoic specimens from stem taxa Archaeopterygidae, Jeholornithidae, Sapeornithidae, Confuciusornithidae, and Enantiornithes, and basal members of Ornithuromorpha. Traditionally, body mass has been estimated in fossil vertebrates using univariate scaling functions. In contrast, multiple regression functions have been used less frequently. Both predictive methods can be affected by different sources of error from statistics, phylogenetic relationships, ecological adaptations, and bone preservation; however, although some studies have addressed these biases, few have tested them within the context of a single data set. In our data set, we find that the models with greater predictive strength and applicability for new specimens, especially for stem taxa, are those derived from multiple regression analyses. For this reason, we suggest that multiple regression analyses may provide improved predictive strength for stem group specimens. Moreover, the methodology used for selecting variables allowed us to obtain specific sets of predictors for each fossil stem group that presumably minimized the variation resulting from historical contingency (i.e. differences in skeletal morphology arising from phylogeny), locomotor adaptations, and diagenetic compaction. The loss of generalizability in the multiple regression models resulting from collinearity effects was negligible on the body mass estimates derived from our data set. Therefore, the body mass values obtained for Mesozoic specimens are accurate and can be used in future studies in a number of palaeobiological and evolutionary aspects of extinct birds, particularly the first stages of avian flight.

INTRODUCTION

Body mass (BM) correlates with a number of ecophysiological variables of animals (e.g. metabolic rates, heat flux and body temperature, reproduction, structural support and locomotion, development, diet, population density, home range size, and behavioural adaptations; Kleiber, 1947; McMahon & Bonner, 1983; Calder, 1984; Gillooly *et al.*, 2001, 2002), as well as with larger-scale patterns in community structure, biogeography, and macroevolutionary trends (e.g., Marzluff & Dial, 1991; Clauset, Schwab & Redner, 2009; McClain & Boyer, 2009). As a result, BM has obvious implications on resource partitioning and ecological displacement among species, which has important evolutionary consequences for deciphering the autecology of extinct species. In the case of flying birds, BM also poses

*Corresponding author. E-mail: fjsa@uma.es

biomechanical constraints on aerial locomotion, as it is related to metabolic costs (Videler, 2005), mechanical power (Tobalske *et al.*, 2003; Askew & Ellerby, 2007; Pennycuick, 2008), wingbeat frequency (Pennycuick, 2001; Bruderer *et al.*, 2010), muscle activity (Tobalske & Biewener, 2008), migration behaviour (Kvist *et al.*, 2001), and maneuverability and acceleration (Warrick, 1998).

Although the majority of the avian fossil record is fragmentary (Mayr, 2009), many well-preserved Mesozoic fossils of avian stem taxa have been discovered and described during the last two decades, especially from the Jehol Formation, Lower Cretaceous of China (e.g. see Chiappe & Witmer, 2002; Brockelhurst *et al.*, 2012). A previous step to reconstructing the palaeobiology of ancient taxa is to obtain accurate BM estimates based on their skeletal dimensions. In turn, this information will allow investigating a number of aspects related to the palaeoecological properties of extinct lineages, including large-scale ecological and evolutionary patterns.

Body mass estimates for some Mesozoic birds and other extinct animals have been derived from reconstructions of corporal surface (e.g. Seebacher, 2001) and three-dimensional volumetric models (e.g. Henderson, 1999; Allen *et al.*, 2013). These methods have been criticized, however, because: (1) the level of error involved in the estimation of the proportions of tissues and organs is unknown (Grand, 1990); and (2) it is difficult to quantify the effect of air sacs and lungs on body density (Campione & Evans, 2012). Furthermore, reconstruction methods can only be applied confidently to complete, well-preserved avian fossils. This precludes analysing specimens preserved with a low level of skeletal completeness. In addition, it is always easier to measure the skeletal dimensions of a fossil than to reconstruct its body. For these reasons, the use of one or a few metric variables that correlate more closely with BM helps to avoid errors inherent to any reconstruction. Most approaches to estimating BM in fossil birds use allometric equations derived from single regression (SR) analysis on skeletal measurements (Maloiy, Alexander & Jayes, 1979; Prange, Anderson & Rahn, 1979; Alexander, 1983; Yalden, 1984; Olmos, 1988; Campbell & Marcus, 1992; Olmos, Casinos & Cubo, 1996; Nudds, 2007; Elzanowski *et al.*, 2012; Liu, Zhou & Zhang, 2012; Field *et al.*, 2013). This approach has been used in studies of: (1) avian macroevolutionary trends (Sanz *et al.*, 2002; Turner *et al.*, 2007; Butler & Goswami, 2008; Hone *et al.*, 2008; Longrich, Tokaryk & Field, 2011; Balanoff *et al.*, 2013; Benson *et al.*, 2014); (2) intraspecific variation in extinct species (Peters & Peters, 2009; Chinsamy *et al.*, 2013); (3) palaeoautecological reconstruction (Vizcaíno & Fariña, 1999; Palmqvist & Vizcaíno, 2003; Nudds & Dyke,

2010); and (4) aerodynamical characterization of fossil birds (Longrich, 2006; Chiappe *et al.*, 2014).

An alternative method that may potentially provide more accurate BM estimates is multiple regression analysis (MR), which incorporates several osteological variables for predicting BM. This method has been used in several groups of tetrapods, and may provide better results than those obtained with SR (Biknevicius, 1999; Palmqvist *et al.*, 2002; Mendoza, Janis & Palmqvist, 2006; Figueirido *et al.*, 2011; De Esteban-Trivigno & Köhler, 2011), although in some cases the results were similar (e.g. Campione & Evans, 2012; see also the discussion below on Field *et al.*, 2013); however, Smith (2002) suggested that MR had lower generalizability than SR (i.e. MR makes poor predictions when applied to individuals not included in the analyses).

The only study that used MR for predicting BM in fossil birds is a symposium abstract (Atanassov & Strauss, 2002) in which several univariate and multivariate methods were compared for estimating BM in *Archaeopteryx* and pterosaurs, concluding that multivariate models have greater predictive strength than univariate ones. These functions were also used to estimate the BM of the giant Miocene bird *Argentavis magnificens* Campbell & Tonni, 1980 and the four-winged dromaesaur *Microraptor gui* Xu *et al.*, 2000, as a first step before examining them aerodynamically (Chatterjee & Templin, 2007; Chatterjee, Templin & Campbell, 2007).

The use of functions that assume a scaling relationship between BM and a number of skeletal dimensions may also introduce different sources of error in both SR and MR approaches, however, which would result in biased BM predictions. For example, the representativeness of the different taxa included in the data set is important for evaluating the statistical validity and reliability of predictions (see review in Smith, 2002). Unequal distributions of specimens per species, and of species within families, may result from limitations posed by sampling availability and differences in taxonomic diversity, respectively. Mendoza *et al.* (2006) and De Esteban-Trivigno, Mendoza & De Renzi (2008) analysed the effect of taxonomic unevenness on BM predictions in ungulates and other mammals, using taxon weighting, by forcing the regressions to consider an equal contribution from each family. In both cases, results showed a higher predictive strength for the weighted regressions than for the unweighted regressions. In contrast, De Esteban-Trivigno & Köhler (2011) did not find significant differences between weighted and unweighted regressions in a study of BM estimation performed in bovids; however, it should be noted that compared with ungulates, which comprise species from two mammalian orders distributed unevenly among 13 families, the family bovidae shows a narrower morphological range and a more even distribution of species

among tribes. Hence, the testing of regression techniques with a weighting procedure can inform on the generalizability and predictive strength of a sample for obtaining BM predictions, as it allows the effects of taxonomic evenness to be minimized. This has never been tested in a taxonomically broad sample of birds, however.

A second source of error for MR is that morphological linear measurements are generally intercorrelated (Quinn & Keough, 2001; Christiansen, 2002). For this reason, it is important to evaluate how the response variable (e.g. body mass) is affected by the correlations between the predictor variables, which is known as (multi)collinearity. The reason is that MR functions derived from correlated predictors can result in unstable partial regression slopes (Bowerman & O'Connell, 1990), and also in an inflation of the confidence intervals for the parameters of the model. The latter can result in an overall significant equation that includes non-significant partial regression slopes (Mitchell-Olds & Shaw, 1987). Although redundant information should be avoided in any explicatory model, Rawlings, Pantula & Dickey (1998) argued that the use of correlated predictors does not prevent obtaining regression equations with good predictive strength that fit the data well; however, Tabachnick & Fidell (1996) did not agree with this, and Christiansen & Fariña (2004) invoked redundancy for using only two variables in their predictive models for non-avian theropods. Some multivariate methods have been proposed as an alternative to MR for avoiding collinearity, such as principal components regression (PCR; Jolliffe, 1982; Stone & Brooks, 1990; Rawlings et al., 1998). Principal components are orthogonal (i.e. uncorrelated) variables, and thus result in a non-collinear multivariate model (for a cautionary note on the use of PCR, however, see Hadi & Ling, 1998). The results of a comparison of MR and PCR (Atanassov & Strauss, 2002) were ambiguous: a better predictive strength was obtained from MR in the bird sample, but the opposite applied to the remaining samples. In spite of all this, redundancy effects in predictive models derived from MR have not been empirically tested.

Another problem is that the assumption of statistical independence does not apply to a sample of species, given their historical relationships. Felsenstein (1985) developed a comparative method that allows obtaining phylogenetically independent contrasts (PICs) by assuming a correct topology and branch length in the cladogram, as well as a Brownian motion model of character evolution. Although assumptions on Brownian motion have been criticized (e.g. Díaz-Uriarte & Garland, 1996; Hansen & Orzack, 2005), many studies have used PIC methods for controlling sample independence (e.g. see the reviews and applications described in Harvey & Pagel, 1991; Ricklefs & Starck, 1996; Price, 1997;

Clarke, Rothery & Isaac, 2010; Figueirido et al., 2010; Campione & Evans, 2012; Martín-Serra, Figueirido & Palmqvist, 2014; Vanhooydonck et al., 2014). In addition, Revell (2010) indicated that forcing phylogenetic regressions on variables with weak phylogenetic signal could be statistically inappropriate; however, it is worth noting here that we only use PICs to test for the non-independence of data resulting from phylogenetic relationships, as explained below, and our final regressions were all derived from raw data.

In addition, any palaeobiological study assumes a certain degree of extrapolation when regression equations derived from crown-group taxa are applied to stem taxa, as it is impossible to evaluate their accuracy for the stem taxa. The reason is that progressively stemward taxa (in our case, Mesozoic birds) exhibit body plans progressively less similar to those of the crown clade (in this study, Neognathae). This means that the applicability of the regressions is likely to decrease proportionally (e.g. it is an oversimplification to expect that our MR equations derived from a subset of the crown group are wholly representative of a stemward taxon like *Archaeopteryx*, in which the long bony tail, teeth, and absence of sternal keel cannot be accounted for in these equations). We assume, however, that the loads exerted on the appendicular skeleton during flight will be similar in crown and stem taxa.

Given that it is not possible to directly examine the relationship between bone dimensions and BM in fossils, predictive equations derived from extant birds must assume similar scaling patterns (and thus movements and weight-bearing distributions) for extinct species (e.g. Atanassov & Strauss, 2002; Sanz et al., 2002; Palmqvist & Vizcaíno, 2003; Longrich, 2006; Turner et al., 2007; Peters & Peters, 2009; Nudds & Dyke, 2010; Liu et al., 2012; Wang et al., 2012; Balanoff et al., 2013; Chinsamy et al., 2013; Field et al., 2013; Campione et al., 2014). Furthermore, a similar allometric scaling is also assumed for distantly related stem Mesozoic taxa (e.g. for Archaeopterygidae, Jeholornithidae, basal Pigostylia, Enantiornithes, and basal Ornithurimorpha), in spite of their paraphyletic relationships (see Fig. 1). This does not necessarily apply to a number of stem taxa, however, because of the roles played by historical contingency (i.e. phylogenetic legacy) and adaptive changes in the avian lineage. Specifically, models derived from crown taxa that incorporate variables measured in apomorphic traits may result in unreliable predictions for Mesozoic stem taxa. For example, Field et al. (2013) derived an SR equation with a high predictive strength for estimating BM in flying birds, based on the maximum diameter of the humeral articulation facet (HAF) of the coracoid; however, the shoulder joint varies along the avian stem (e.g. *Archaeopteryx*, *Jeholornis*, *Confuciusornis*, and *Sapeornis*), and may be more similar to their maniraptoran
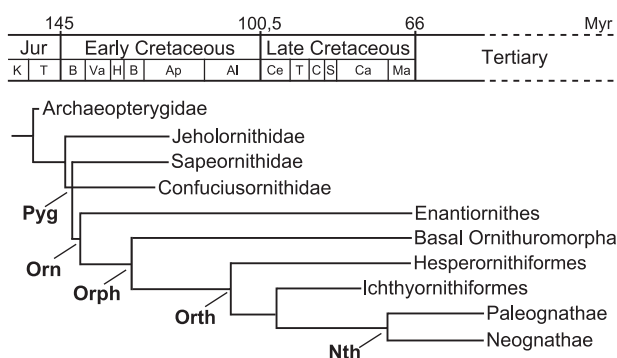
Figure 1. Calibrated phylogeny for the main avian taxa. Tree topology was obtained from O'Connor, Chiappe & Bell (2011) and divergence times are based on a 'literal' interpretation of the fossil record from Brockelhurst *et al.* (2012). Taxa abbreviations: Nth, Neornithes; Orph, Ornithuromorpha; Orn, Ornithothoraces; Orth, Ornithurae; Pyg, Pygostylia.

ancestors (Jenkins, 1993; Chiappe *et al.*, 1999; Chiappe & Witmer, 2002; Chiappe, 2007; for other aspects of skeletal anatomy in stem birds that may be primitive, see Campione *et al.*, 2014). In addition, a number of studies on the functional adaptations of modern birds have shown that the morphology of wing and leg bones is highly influenced by locomotion dynamics (Cubo & Casinos, 1998; De Margerie *et al.*, 2005; Habib & Ruff, 2008; Hinic-Frlog & Motani, 2010; Simmons, 2010; Bell & Chiappe, 2011; Simmons, Hieronymus & O'Connor, 2011). Although this means that BM estimates based on skeletal dimensions can be biased by ecomorphological adaptations, these effects can be adequately tested in a large database (e.g. see Egi, 2001; Campione & Evans, 2012). In addition, differences of bone scaling patterns have not been tested in extant birds of different body size before selecting the best combination of skeletal predictors of BM for specific fossil stem taxa. As explained below, we have evaluated the behaviour of each skeletal measurement with respect to the BM derived from all other measurements. This in turn allowed us to discard variables that showed scaling differences between the fossil taxa and neognaths.

Finally, diagenetic compaction can change the osteological dimensions of fossil birds, particularly their bone width measurements, which may affect BM estimates.

The main objectives of this study are: (1) to obtain allometric scaling models derived from extant crown birds that allow us to estimate BM in fossil specimens of Mesozoic stem groups; and (2) to test for sources of error that affect their accuracy. As explained before, these errors may result from statistical procedures, phylogenetic constraints, ecological adaptions, and bone preservation. In addition, an extensive literature review

of mass estimation in birds, and its application to fossil specimens, was performed in order to compare published results with those obtained here.

METHODS

DATA SET

A total of 42 fossils of presumably flying Mesozoic stem taxa were studied in order to estimate their BM using an extensive data set of modern volant birds (Table S1). The fossil specimens represent a broad sample of the avian clade during the Late Jurassic and Cretaceous, including five members of Archaeopterygidae, two members of Jeholornithidae, four members of Sapeornithidae, seven members of Confuciusornithidae, 14 members of Enantiornithines, and ten basal members of Ornituromorpha (Fig. 1). Most fossils were measured *in situ* with digital calipers (± 0.1 mm) at the Institute of Vertebrate Paleontology and Paleoanthropology (IVPP) of Beijing (China) and at the Universidad Autónoma de Madrid (UAM, Spain). In addition, the fossil data set (Table S2) includes measurements taken from published papers and high-resolution images using tpsDIG 2.17, available at: http://life.bio.sunysb.edu/morph/.

The data set of modern birds comprises 495 individuals from 164 flying species of Neognathae distributed among 66 families and 22 orders. Only adult individuals, as indicated by epiphyseal closure, were measured. Juveniles were excluded from the data set because their use for generating models can introduce an additional source of error and bias into the equations. The sample covers a broad range of body sizes, of more than three orders of magnitude [from the mute swan (9.4 kg) to the willow warbler (8.8 g)], and includes species from almost all flight types (hummingbirds were not included because of their exceptionally unique hovering abilities) and modes of non-aerial locomotion. Almost all specimens belong to the collection housed in the Burke Museum, University of Washington (UWBM, Seattle, USA), and others come from the Museo Nacional de Ciencias Naturales of Madrid (MNCN, Spain). Each specimen was weighted before death or skeletonization, and 19 linear dimensions of the fore- and hindlimb bones were measured with a digital caliper (Table 1, Fig. 2).

Our database of extant birds comprises only flying members of Neognathae. The sister clade, Paleognathae, includes the non-volant ratites and the flying tinamous (order Tinamiformes). Attempts to collect measurements of tinamous in both the museums visited and in the bibliography were not successful, which would mean that part of the morphological range of the flying Neornithes is absent from our database; however, we think that this absence is not significant for our analy-

Table 1. Definition of osteological limb measurements used in this study

HL	Length of humerus
dpL	Length of deltopectoral crest
bcL	Length of bicipital crest
deHW	Dorsoventral width of distal humerus
dHW	Dorsoventral width at midshaft of humerus
UL	Length of ulna
peUW	Dorsoventral width of proximal ulna
deUW	Dorsoventral width of distal ulna
dUW	Craniocaudal width at midshaft of ulna
RL	Length of radius
dRW	Craniocaudal width at midshaft of radius
CmL	Length of carpometacarpus from the semilunate to the joint with major digit
dCmW	Craniocaudal width at midshaft of major metacarpal
DCmW	Midshaft width from the cranial edge of major metacarpal to the caudal edge of minor metacarpal
FL	Length of femur
dFWcc	Craniocaudal width at femoral midshaft
dFWml	Mediolateral width at femoral midshaft
TL	Length of tibiotarsus
TmL	Length of tarsometatarsus from the crista medial to the trochlea of metatarsus III

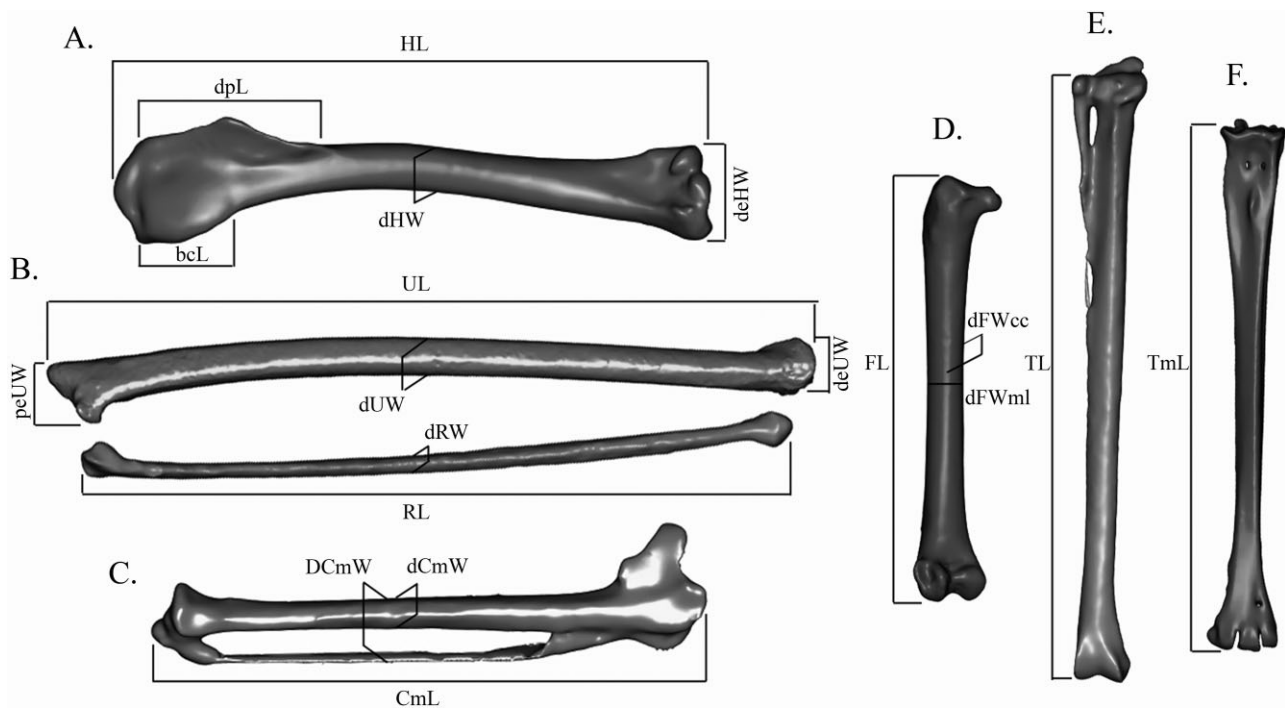Anatomical nomenclature follows Baumel *et al*. (1993).

Figure 2. Illustration of osteological limb measurements used in this study and defined in Table 1: A, cranial aspect of the humerus; B, dorsal aspect of the ulna and the radius; C, dorsal aspect of the carpometacarpus; D, caudal aspect of the femur; E, caudal aspect of the tibiotarsus; F, cranial aspect of the tarsometatarsus.

ses, because a study of 148 morphological traits (Mayr & Clarke, 2003) showed that the skeletal differences between palaeognathans and neognaths lie in the shape of the palatal bones, the number of sternal costal pro-cesses, and the development of the thumb. For this reason, we have assumed that the range of variation in the postcranial anatomy of neognaths covers the mor-phology of tinamous, which represent <0.5% of the

Table 2. Ordinary least-squares regression parameters for three single regressions (SR), for the best multiple regression (MR), and for the principal components regression (PCR)

Regression model	x-variables	N	R^2_{adj}	b (±95% CI)	log a (±95% CI)	VIF	|%MPE|$_{vs}$	|%MPE|
Simple (SR)	HL	486	0.904	1.947 (± 0.057)	−4.005 (± 0.101)	1.0	42.2	40.1
Simple (SR)	dUW	489	0.940	2.715 (± 0.061)	−2.079 (± 0.036)	1.0	32.9	30.6
Simple (SR)	dFWcc	480	0.953	2.330 (± 0.046)	−1.785 (± 0.027)	1.0	30.5	25.8
Simple (SR)	PC 1	382	0.949	0.642 (± 0.015)	−0.572 (± 0.015)	1.0	33.1	28.7
Multiple (MR)	LH	400	0.981	1.012 (± 0.242)	−2.765 (± 0.195)	71.8	16.2	15.9
	Lcb			0.339 (± 0.155)		26.1		
	dH			0.494 (± 0.236)		41.9		
	epU			0.492 (± 0.249)		38.6		
	edU			0.493 (± 0.296)		59.5		
	LR			−1.13 (± 0.198)		43.3		
	DCMC			−0.325 (± 0.183)		19.6		
	dFAP			0.65 (± 0.175)		27.4		
	LT			0.915 (± 0.233)		37.5		
	LTMT			−0.496 (± 0.135)		14.7		
Principal components (PCR)	PC 1	383	0.981	0.642 (± 0.009)	−0.571 (± 0.009)	1.0	17.6	16.4
	PC 3			0.034 (± 0.009)		1.0		
	PC 4			−0.05 (± 0.009)		1.0		
	PC 5			0.071 (± 0.009)		1.0		
	PC 6			−0.045 (± 0.009)		1.0		
	PC 8			−0.03 (± 0.009)		1.0		
	PC 9			0.01 (± 0.009)		1.0		
	PC 10			−0.023 (± 0.009)		1.0		
	PC 12			−0.015 (± 0.009)		1.0		
	PC 13			−0.032 (± 0.009)		1.0		
	PC 14			−0.011 (± 0.009)		1.0		
	PC 16			0.019 (± 0.009)		1.0		

Linearized scaling equations are represented as $\log BM = \log a + b_1 \log x_1 + b_2 \log x_2 \ldots + b_p \log x_p$. 95% confidence intervals for intercepts (a) and slopes (b) appear in brackets. The predictor variables of PCR are those principal components with partial slopes significantly different from zero (F-statistics at the 0.05 level). For abbreviations of variables, see Table 1.
\|%MPE\|, mean percentage prediction error calculated over the whole data set; \|%MPE\|$_{vs}$, mean percentage prediction error calculated over the validating subsample (20% of specimens); VIF, factor of variance inflation, $1/(1 − R^2)$.

diversity of extant Neornithes (see Del Hoyo, Elliott & Sargatal, 1992). It is worth recognizing, however, that Mayr & Clarke (2003) performed a phylogenetic study discussing discrete morphological characters, not relative bone proportions.

It could be argued that in order to avoid data redundancy, the regression analyses should be performed using the mean values of the skeletal predictors in the species analysed instead of all the individuals sampled; however, the use of species averages would mask the morphological differences related to intraspecific variability (e.g. sexual dimorphism and variation resulting from age differences). In addition, our analyses are an improvement on most studies that use means of skeletal measurements and body mass for the species analysed (e.g. Mendoza, Janis, & Palmqvist, 2006; Mendoza *et al*., 2006; Figueirido *et al*.,

2011; Field *et al*., 2013), because the averages usually come from different individuals (e.g. specimens measured in museum collections and mean body masses taken from the bibliography). This introduces an additional source of error and bias in the regression equations (see discussion in Turner & O'Regan, 2002). If the masses used for deriving the regression are estimated from the specimens measured, however, this problem is avoided (Smith, 2002). For example, Field *et al*. (2013: table 2) compared two SR models of BM on HAF derived separately for modern birds, one of them based on mean BM values for species and the other on the actual masses of individuals. They found that the 95% confidence intervals calculated for the slopes and the intercepts of these two regressions overlap. In this article, we have usedspecimens for deriving our regressions models and

species means for comparing with the PIC regression results. Finally, we must acknowledge that body mass shows variations throughout the adult life of the bird, and can also fluctuate throughout the year (because of factors such as migration and breeding); however, given that intraspecific variability exceeds the range of individual variation, as it also comprises differences resulting from age, sexual dimorphism, latitudinal range, and differences in migratory behaviour among populations (Turner & O'Regan, 2002), we chose to use the actual masses of the fresh specimens weighted prior to skeletonization instead of species means.

REGRESSION ANALYSES AND DATA TRANSFORMATION

The data from modern birds were fitted to linear regression functions by means of the ordinary least squares (OLS) method, with BM as the response variable (y) and the skeletal measurements as the predictor variables (x). Both SR and MR functions were derived for estimating BM. A number of investigators (see reviews in Warton *et al.*, 2006; Smith, 2009) have proposed the use of reduced major axis (RMA) instead of OLS to define a line of best fit for a bivariate relationship when the x-variable is measured with error; however, Hansen & Bartoszek (2012) showed that RMA is only reliable when the relationship between x and y is nearly perfect (i.e. when there are neither biological nor observational errors). In addition, RMA is symmetric (i.e. a single line defines the bivariate relationship, regardless of which variable is x and which is y), whereas OLS is asymmetric, so that the slope and resulting interpretation of the data are changed when the variables assigned to x and y are reversed. For this reason, Smith (2009) argued that the symmetry–asymmetry issue should be the criterion by which investigators choose between RMA and OLS. In any case, when the principal aim of the regression analysis is prediction, a relationship of dependence should be assumed for y (i.e. BM in our case) on the x variables (i.e. the osteological measurements), and most studies agree on the suitability of the OLS method (Gingerich, Smith & Rosenberg, 1982; Sokal & Rohlf, 1986; Quinn & Keough, 2001; MacLeod, 2004; Smith, 2009; Campione & Evans, 2012), as used in this article.

Given that raw data for BM and the skeletal measurements show positively skewed distributions, all data were log₁₀ transformed prior to statistical analyses to approximate a normal distribution, a procedure that also minimizes the level of heteroscedasticity (Quinn & Keough, 2001; Cawley & Janacek, 2010; however, see Packard, Boardman & Birchard, 2009). Plots of the residuals against predicted values were also examined to detect outliers; however, the retransformation from predicted log-values to arithmetic units could bias the results. The reason is that although arithmetic and log-transformed values are mathematically equivalent, they are not so for statistical analyses (Smith, 1993). In order to avoid this potential bias, each retransformed value was multiplied by a correction factor known as the ratio estimator (Snowdon, 1991): the ratio between the mean of the observed values and the mean of the predicted anti-log values.

Multiple regressions were obtained using the automatic stepwise procedure for selection of variables. This consists of selecting step-by-step the x-variables that show the highest correlation with the y-variable, in such a way that each step refits the model by introducing or dropping any predictor variable, based on the contribution of its partial regression slope (input F-value fixed at the 0.05 significance level, and output F-value fixed at 0.10) to the multivariate model (for details, see Mendoza *et al.*, 2006). All statistical analyses were carried out using SPSS 20 (http://www.ibm.com/software/es/analytics/spss/).

ACCURACY AND CROSS-VALIDATION OF THE REGRESSION MODELS

In order to evaluate the predictive strength of SR and MR equations, the coefficient of determination (R^2), the percentage standard error of the estimate (%SEE), and the percentage prediction error (%PE) were computed. Both %SEE and %PE are accurate estimators of the predictive strength of SR and MR equations (Smith, 1984). R^2 is also a good indicator of the predictive strength of an equation adjusted for SR, but this does not hold for MR (Smith, 1981; Quinn & Keough, 2001; Mendoza *et al.*, 2006). For this reason, we estimated the value of the adjusted R^2, which takes into account the number of predictor variables incorporated in the model [$R^2_{adj} = 1 - (1 - R^2)(N - 1)/(N - p - 1)$, where p is the number of predictor variables and N is sample size]. The value of %SEE, computed using the standard error of the estimate [$10^{(2 + SEE \cdot 100)}$], allows for estimating the 68% interval of confidence for BM around the mean. The %PE value, calculated as [(actual BM − predicted BM)*100/predicted BM], measures the degree of overestimation (%PE < 0) or underestimation (%PE > 0) for a particular observation. The sample mean of the absolute values of %PE, noted as |%MPE|, is the average percentage deviation between predicted and observed BMs. This allowed us to calculate the average prediction intervals as BM$_{predicted}$ ± |%MPE| values (Smith, 1984; Mendoza *et al.*, 2006; Campione & Evans, 2012; Field *et al.*, 2013).

In MR models, the appropriate number of predictor variables depends on the sample size. We have followed the general rule of a proportion greater than 10 : 1 for sample size on the number of x-variables (Darlington, 1990). It stands to reason that the more

parameters a model has (k), the more precise it will be, in terms of the data on which it was built (in an SR function, $k = 2$; i.e. the y-intercept and the slope); however, this precision comes at the expense of generalizability (i.e. applicability over new samples not included in the analysis; see Smith, 2002). To improve the generalizability of a predictive model is particularly important when the aim is to estimate fossil taxa that are outside the clades upon which the models are based (N. Campione, pers. comm.), as is the case in this article. Standard metrics of model strength and predictability (e.g. R^2, %SEE, and %MPE) are only appropriate when $k_{model1} = k_{model2}$, as they always get better as k increases. There are several methods that allow us to find the optimal subset of predictor variables whilst limiting the required number of parameters, for example the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (see Quinn & Keough, 2001). We opted to use as a criterion for model selection the adjusted R^2 value, as it tends to reproduce empirical data sets better than the models selected with AIC and BIC, which is indicated by its consistently lower sums of squared residuals (Gayawan & Ipinyomi, 2009). It should also be noted, however, that the predictive models chosen with the adjusted R^2 procedure are often more complex than those derived from AIC and BIC, as those derived using the adjusted R^2 procedure usually include a higher number of parameters (Gayawan & Ipinyomi, 2009).

Smith (2002) discussed the importance of using a validating subsample when the prediction equation takes the form of a multiple regression. The reason is that multiple regressions, particularly when predictors are selected by stepwise procedures, are notorious for their lack of generalizability (i.e. they predict poorly on new data). For this reason, the generalizability and predictive strength of our models was further calibrated with a set of preliminary regression functions derived from a subsample of the data set. For doing so, 20% of the specimens of modern birds were randomly selected from the data set and used later as a test subsample for cross-validation, following the procedures described by Hurvich & Tsai (1990), MacNally (2000), and Mendoza *et al*. (2006). The functions derived from the remaining 80% were used for obtaining the BM of the specimens of the validating subsample, and an |%MPE| value was calculated for them (i.e. |%MPE|$_{vs}$). After cross-validation, the specimens of this subsample were incorporated into the preferred regression functions, which included all measured individuals; however, we recognize that validating a model with species that are evolutionarily equivalent to those used to generate it (i.e. extant flying birds) does not inform on how well the model will do with extinct species that are not closely related (i.e. Mesozoic stem taxa). Finally, |%MPE| values were calculated for pub-

lished univariate allometric regressions with our whole sample in order to compare their predictive strength.

ANALYSING BIASES IN THE REGRESSION MODELS

Taxon weighting
The sample of modern birds used in this study has a wide taxonomical representation, and each species has a similar number of specimens (between three and five in most cases); however, species are unevenly distributed among families, which means that several families (e.g. Anatidae and Accipitridae, with ten and seven species, respectively) are overrepresented compared with others (e.g. Otididae and Cinclidae, with one species each). This could bias the regressions if the scores of the underrepresented families are not randomly distributed around the regression line. If so, regressions would be biased towards the families that are most represented. Maximizing taxonomic evenness (i.e. a statistical procedure that guarantees that the species of each family contribute equally to the functions; for details, see Mendoza *et al*., 2006) can increase the robustness of the models. For this reason, we adjusted functions weighting the richness of specimens within families, and their %PE and residuals were compared with those from unweighted regressions. For doing that, the weighted value of each specimen from a given family was computed as the mean number of specimens per family divided by the total number of specimens within this family (i.e. $N_{individuals}/N_{families}$; given that the number of specimens per species was similar, this ratio is not biased as a result of intraspecific variation).

The effects of over- and underrepresented families on the regressions were evaluated by comparing the |%MPE| and the residuals. These regressions include: (1) a couple of SR equations derived from the two variables that best correlate with BM (diaphyseal craniocaudal width of ulna and femur); (2) one SR equation derived from humeral length; and (3) the best MR equation obtained with the stepwise procedure for modern birds, which incorporates ten predictor variables.

Multicollinearity
Redundancy effects related to the predictive strength of the MR functions were tested by comparing these equations with non-collinear models, such as SR and PCR. In the case of PCR, a principal components analysis of the predictor variables was performed and the principal components were regressed against BM, retaining those that showed partial slopes differing significantly from zero (see details in Rawlings *et al*., 1998). Multicollinearity was measured by means of the variance inflation factor (VIF), calculated for each variable x_j as $[1/(1 - R^2)]$ from the OLS regression of

x_j against the remaining x-variables. Values of variance inflation factor >10 suggest strong collinearity effects (Kleinbaum, Kupper & Muller, 1997).

Independent contrasts

Phylogenetically independent contrasts (PICs) were performed for the species means of each predictor variable in order to control for effects resulting from phylogenetic non-independence. For doing so, a cladogram for modern birds was constructed in MESQUITE (Maddison & Maddison, 2011), and PICs were obtained using the PDA:PDTREE package (Midford, Garland & Maddison, 2005). Tree topology was based on the relationships of avian families published by Hackett *et al.* (2008), and the relationships within families were obtained from different sources (Fig. S1 and references therein). Branch lengths for neornithine families were obtained averaging the distances (substitutions per site) of those species included in the phylogram of Hackett *et al.* (2008: fig. 3). Within each family, a minimal length of 0.005 substitutions per site was imposed on each internal branch, following a similar procedure to that used by Laurin (2004), Cubo *et al.* (2005), and Butler & Goswami (2008) for calibrating branch lengths in millions of years for successive dichotomies in cladograms of extinct taxa. Phylogenetically independent contrasts assume a Brownian motion for character evolution. Given that the use of variables with a poor phylogenetic signal may be incorrect in statistical terms (Revell, 2010), the effects of phylogeny on the skeletal variables were assessed using MESQUITE. Following Laurin (2004), the squared length of each trait was statistically tested against 10 000 randomly generated trees in which terminal taxa were permuted while holding tree topology and branch lengths constant. Specifically, the effects of phylogenetic signal are significant if the number of steps between the basal node and the terminal branches of the tree analysed is lower than in the random trees (Laurin, 2004).

Selection of variables: scaling patterns in crown and stem groups

An approach was developed to minimize potential biases that could emerge from: (1) differences in morphological scaling patterns between living and extinct taxa; and (2) the effects of diagenetic compaction on fossil bones. The behaviour of each x-variable with respect to the remaining x-variables was analysed, and those that showed values in the fossil taxa outside the trends of Neognathae were discarded. Given that most morphological variance among extant and extinct taxa is explained by size variations (92.5% in the first principal component), we considered that the best way to carry out the selection of the linear measurements was to plot each measurement against the BM predicted from a combination of the remaining x-variables

(Table S3). A regression line was fitted for modern birds in each plot, and variables that scored within the 95% confidence intervals were selected for each fossil group, which allowed us to obtain particular subsets of predictor variables for them. Finally, the stepwise method was used in each subset for obtaining the best multivariate functions. This procedure would also help to detect outliers potentially affected by taphonomy, because it would identify the skeletal measurements that were heavily affected by diagenetic compaction in a given fossil specimen, which would behave as an outlier. In such cases, the regression functions that incorporate these measurements would not be applied to these fossil specimens.

Post-hoc tests for effects of adaptive changes

Once the regression functions were obtained for the fossils, further analyses were performed to evaluate the influence of ecological adaptions on the models. Specifically, the effects of locomotion type on the prediction error of BM (%PE) were graphically represented and statistically tested by means of one-way ANOVAs. Modern birds were classified according to two ecological criteria related to their locomotion modes: mechanical behaviour during flight (i.e. aerial locomotion) and movement in non-atmospheric environment (i.e. non-aerial locomotion: terrestrial or aquatic). Four categories, defined by Pennycuick (2008), for the predominant behaviour during aerial locomotion were used: (1) continuous flapping; (2) flap and gliding; (3) flap and bounding; and (4) soaring. Three basic categories for non-aerial locomotion were also used: (1) aquatic birds; (2) ground-dwelling terrestrial birds; and (3) non-ground-dwelling terrestrial birds. The first category encloses every type of swimming and diving birds described in Hinic-Frlog & Motani (2010). Terrestrial categories follow Pike & Maitland (2004); however, the category of non-ground dwellers comprises birds that primarily climb, perch, or exhibit predatory behaviour (these categories are considered separately by Pike & Maitland, 2004). Furthermore, two intermediate categories were added: (4) aquatic ground-dwellers (e.g. gulls); and (5) perching ground-dwellers (e.g. starlings). Species were classified among categories using behavioural descriptions from Del Hoyo, Elliot & Sargatal (1992) and direct observations. The classifications used in this study largely follow those of Pike & Maitland (2004), Hinic-Frlog & Motani (2010), and Bruderer *et al.* (2010).

RESULTS

STATISTICAL SOURCES OF ERROR ON BM PREDICTION MODELS
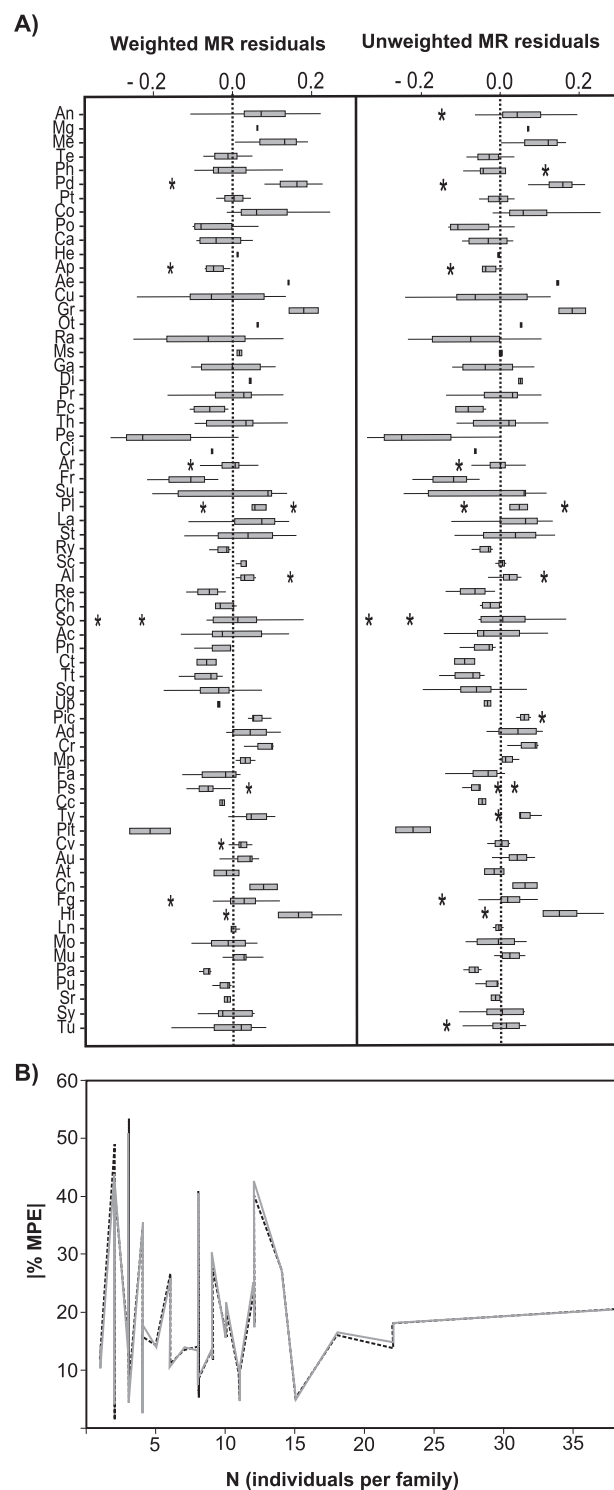
Taxon weighting effects

The distribution of the residuals and the trend for %PE values among families did not change when the

Figure 3. Effects of weighting for the best-fitting multiple regression equation obtained from the modern data set (see Table 2). A, distribution of familiar residuals for unweighted data, and for data weighted by families. Each box plot with whiskers represents one family. Box length shows the interquartile range (25th and 75th percentiles). Horizontal lines indicate the 5–95% confidence limits. Asterisks show outliers. Abbreviations for families: Ac, Accipitridae; Ad, Alcedinidae; Ae, Aegothelidae; Al, Alcidae; An, Anatidae; Ap, Apodidae; Ar, Ardeidae; At, Artamidae; Au, Alaudidae; Ca, Caprimulgidae; Cc, Cacatuidae; Ch, Charadriidae; Ci, Ciconiidae; Co, Columbidae; Cn, Cinclidae; Cr, Coraciidae; Ct, Catharthidae; Cu, Cuculidae; Cv, Corvidae; Di, Diomedeidae; Fa, Falconidae; Fg, Fringillidae; Fr, Fregatidae; Ga, Gaviidae; Gr, Gruidae; He, Hemiprocnidae; Hi, Hirundinidae; La, Laridae; Ln, Lanidae; Me, Meleagridae; Mg, Megapodidae; Mo, Motacillidae; Mp, Meropidae; Ms, Musophagidae; Mu, Muscicapidae; Ot, Otididae; Pa, Paridae; Pc, Pelecanoididae; Pd, Podicipedidae; Pe, Pelecanidae; Ph, Phasianidae; Pic, Picidae; Pit, Pittidae; Pl, Phalacrocoracidae; Pn, Pandionidae; Po, Podargidae; Pr, Procellaridae; Ps, Psittacidae; Pt, Pteroclidae; Pu, Prunellidae; Ra, Rallidae; Re, Recurvirostridae; Ry, Rynchopidae; Sc, Stercoriidae; Sg, Strigidae; So, Scolopacidae; Sr, Sturnidae; St, Sternidae; Su, Sulidae; Sy, Sylviidae, Te, Tetraonidae; Th, Threskiornithidae; Tt, Tytonidae; Tu, Turdidae; Ty, Tyrannidae; Up, Upupidae. B, plot showing the variations of %MPE with the increase of individuals per family. The dashed line represents the unweighted multiple regression analysis (MR) and the grey line represents the weighted MR.

best-fitting MR equation was weighted (Fig. 3). The absence of bias resulting from differences in the number of specimens and species within families indicates that the variables distribute randomly in the families under- and overrepresented in the data set. Furthermore, %MPE values were lower in all SR and MR equations obtained with the unweighted procedure (Table S4). Therefore, maximizing taxonomic evenness in the sample of modern birds neither results in an increase of robustness nor results in an increase in predictive strength. Hence, unweighted regressions were used in the remaining analyses.

Effects of collinearity

As shown in Table 2, the ten predictor variables included in the best MR equation for the modern sample show high VIF values, indicating that the model is affected by redundancy effects; however, this equation shows the highest predictive strength, with an |%MPE| value slightly lower than the PCR, in spite of the fact that the latter analysis included all 19 predictor variables. The SR equations showed higher |%MPE| values than the multivariate functions, and hence less pre-

dictive strength; dFWcc was the best univariate predictor (Table 2, S6). Compared with non-collinear models (i.e. SR and PCR), effects of collinearity on the MR model were reflected in the inflation of the confidence intervals for the partial slopes of the predictor variables and the intercept (Table 2). Although this could

result in an increase of uncertainty for the estimates, it was avoided by using the |%MPE| value for estimating the confidence interval for predictions.

SOURCES OF ERROR FROM PHYLOGENETIC RELATIONSHIPS

Effect of phylogeny in extant species
According to the test described by Laurin (2004), BM and each of the skeletal variables showed a highly significant phylogenetic signal in the neognaths tree ($P < 0.0001$ in all cases; Table 3). This indicates that the use of a phylogenetic regression, like PICs, is quite appropriate (Revell, 2010). Each of the least-squares linear regressions for the PICs on BM was highly significant ($P < 0.01$), as in the case of the regressions with non-standardized variables on mean BM (Table 3). In addition, the confidence intervals for the slopes overlapped in all cases. Accordingly, the morphological patterns depicted by the species sampled can be statistically considered as 'taxon-free' (i.e. the patterns are not strongly driven by phylogeny). Overall, phylogenetically corrected scaling relationships showed lower R^2 values than those adjusted from raw data.

Taxa-specific and unbiased MR functions
After selecting the predictor variables (Figs 4 and S2), seven MR functions were obtained (Table 4), one for each of the six fossil groups and one specific for stem Ornitothoraces with non-preserved hindlimbs, which include *Eocathayornis*, *Otogornis*, *Eoalulavis*, *Apsaravis*, and *Gansus* (specimen CAGS-04-CM-004). The number of predictor variables included in the models ranged from six, in equations adjusted for fossils of Archaeopterygidae, Enantiornithes, and Ornitothoraces, with non-preserved legs, to 12 in equations adjusted for fossils of for Confuciusornithidae. |%MPE| values were ≤20% in all MR equations, which in some cases had greater predictive strength than the best published univariate allometric functions, with the only exception of one SR equation of Field *et al*. (2013) (Table 5). In addition, the comparison of the predictive strength of these MR equations over the validating subsample and over the whole sample shows a minor variation in their |%MPE| values (between 0.1 and 2.4%). This suggests good generalizability for the selected models. The MR equation for Sapeornithidae (Table 4) has the highest predictive strength (|%MPE| = 16.5%). In contrast, the function for

Table 3. Ordinary least-squares single regressions obtained for each predictor variable using the means of species calculated from raw data and the appropriate phylogenetically independent contrasts (PICs)

x-variable	Raw regressions (OLS) $\log BM = \log a + b \log x$			PIC regressions (OLS) $\log BM = \log a + b \log x$ ($a = 0$)	
	log a (95% CI)	b (95% CI)	R^2	b (95% C.I.)	R^2
HL	−3.901 (−4.147 to −3.658)	1.886 (1.756 to 2.018)	0.899	2.102 (1.949 to 2.247)	0.885
dpL	−2.783 (−2.900 to −2.677)	1.844 (1.760 to 1.935)	0.905	1.946 (1.836 to 2.040)	0.900
bcL	−2.479 (−2.583 to −2.378)	2.044 (1.941 to 2.147)	0.921	2.146 (2.006 to 2.275)	0.908
deHW	−3.071 (−3.224 to −2.928)	2.477 (2.338 to 2.618)	0.891	2.394 (2.284 to 2.496)	0.910
dHW	−2.155 (−2.235 to −2.075)	2.457 (2.344 to 2.564)	0.914	2.319 (2.195 to 2.433)	0.910
UL	−4.063 (−4.340 to −3.79)	1.932 (1.788 to 2.075)	0.846	2.026 (1.871 to 2.179)	0.856
peUW	−2.678 (−2.805 to −2.5587)	2.595 (2.454 to 2.736)	0.900	2.475 (2.366 to 2.576)	0.923
deUW	−2.680 (−2.800 to −2.569)	2.570 (2.446 to 2.699)	0.915	2.439 (2.325 to 2.538)	0.918
dUW	−2.055 (−2.125 to −1.986)	2.682 (2.564 to 2.793)	0.933	2.559 (2.424 to 2.672)	0.917
RL	−3.885 (−4.148 to −3.651)	1.865 (1.732 to 2.004)	0.838	1.976 (1.811 to 2.139)	0.845
dRW	−1.109 (−1.162 to −1.047)	1.964 (1.825 to 2.083)	0.905	2.049 (1.923 to 2.179)	0.884
CmL	−3.941 (−4.148 to −3.746)	2.192 (2.073 to 2.321)	0.903	2.313 (2.151 to 2.466)	0.901
dCmW	−1.599 (−1.675 to −1.525)	2.618 (2.450 to 2.784)	0.880	2.453 (2.293 to 2.618)	0.873
DCmW	−2.535 (−2.694 to −2.367)	2.645 (2.444 to 2.836)	0.818	2.472 (2.280 to 2.670)	0.825
FL	−4.788 (−5.045 to −4.528)	2.647 (2.478 to 2.808)	0.831	2.690 (2.535 to 2.864)	0.861
dFWcc	−1.813 (−1.876 to −1.752)	2.360 (2.244 to 2.466)	0.927	2.393 (2.280 to 2.503)	0.932
dFWml	−1.806 (−1.854 to −1.753)	2.343 (2.250 to 2.430)	0.952	2.409 (2.295 to 2.512)	0.939
TL	−4.843 (−5.204 to −4.455)	2.362 (2.154 to 2.555)	0.824	2.391 (2.168 to 2.617)	0.835
TmL	−3.312 (−3.713 to −2.862)	1.764 (1.491 to 1.998)	0.588	1.844 (1.512 to 2.143)	0.618

PICs assume a regression through the origin ($a = 0$). The statistical significance of each regression was tested by means of ANOVA. For abbreviations of variables, see legend of Table 1. Body mass (BM) and all skeletal variables showed significant phylogenetic signal at $P < 0.001$. All regressions are significant at $P < 0.01$.
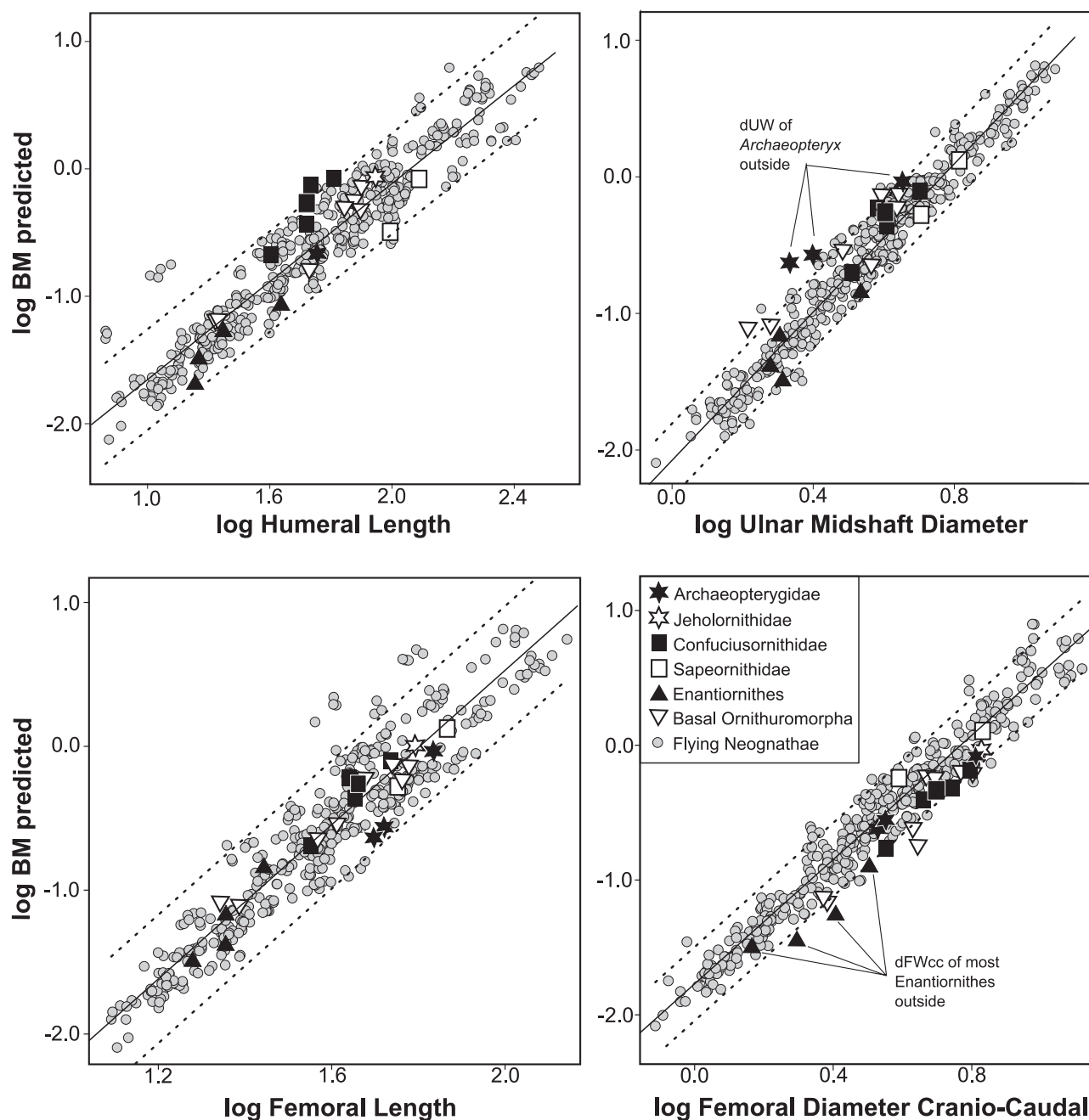
Figure 4. Biplots used in the selection of four variables for each fossil group of Mesozoic birds: humeral length (HL), femoral length (FL), diaphyseal craniocaudal width of ulna (dUW), and diaphyseal craniocaudal width of femur (dFWcc). Each predictor variable was plotted against a combination of the 14 remaining variables used for estimating body mass (BM; Table S3). The biplots show the regression line fitted for extant birds (grey circles) with the 95% confidence intervals for BM predictions (dotted lines). Following the procedure of selection of variables (see text), HL and FL were used for generating functions that can be applied to all fossil avian taxa. In contrast, dUW and dFWcc could not be incorporated into the functions adjusted for estimating BM in Archaeopterygidae and Enantiornithes, respectively.

Ornithothoraces with non-preserved hindlimbs is the poorest predictor (|%MPE| = 20.2%), as expected from the fact that it exclusively incorporates wing measurements.

SOURCES OF ERROR FROM ADAPTIVE CHANGES

Figure 5A and Table 6 show that the SR for the aerial locomotion groups have appreciable differences in %PE

Table 4. Best ordinary least-squares multiple regressions obtained for fossil taxa, adjusted with the stepwise method from the subset of selected variables

Multiple regression functions	$\log BM = \log a + b_1 \log x_1 + b_2 \log x_2 \ldots + b_p \log x_p$	N	R^2_{adj}	%SEE	$|\%MPE|_{vs}$	$|\%MPE|$
ARCH	−2.947 (± 0.388) + 1.508 (± 0.455) HL + 1.068 (± 0.313) dHW − 1.165 (± 0.407) RL + 0.582 (± 0.357) dFWcc + 0.837 (± 0.472) TL − 0.458 (± 0.268) TmL	422	0.978	25.5	17.3	18.8
JEHO	−1.933 (± 0.505) + 1.486 (± 0.421) HL + 0.416 (± 0.318) bcL − 0.965 (± 0.451) dHW − 0.36 (± 0.599) deHW − 1.536 (± 0.418) UL − 0.635 (± 0.415) peUW − 0.397 (± 0.394) FL + 0.834 (± 0.431) dFWml + 0.302 (± 0.247) TL	417	0.980	24.0	17.4	17.3
SAPE	−2.876 (± 0.243) + 0.952 (± 0.232) HL + 0.352 (± 0.149) bcL + 0.424 (± 0.228) dHW + 0.967 (± 0.955) UL − 0.615 (± 0.283) deUW − 0.340 (± 0.301) dUW − 1.891 (± 0.908) RL − 0.343 (± 0.181) DCmW + 0.589 (± 0.177) dFWcc + 0.878 (± 0.231) TL − 0.446 (± 0.134) TmL	387	0.983	22.3	14.1	16.5
CONF	−2.557 (± 0.502) + 0.271 (± 0.376) HL + 0.275 (± 0.326) bcL − 0.362 (± 0.316) dpL + 0.296 (± 0.532) dHW − 0.303 (± 0.628) deHW + 0.362 (± 0.542) peUW + 1.051 (± 0.657) dUW − 0.378 (± 0.441) DCmW − 0.187 (± 0.374) FL + 1.01 (± 0.4) dFWcc + 0.963 (± 0.5) TmL − 0.488 (± 0.289) TmL	411	0.978	24.7	18.1	16.9
ENAN	−2.626 (± 0.284) + 1.528 (± 0.436) HL + 0.34 (± 0.312) bcL + 0.828 (± 0.427) dHW − 1.451 (± 0.386) UL + 0.811 (± 0.448) dFWml + 0.378 (± 0.201) TL	426	0.978	25.0	19.0	19.4
ORPH	−2.26 (± 0.275) + 1.407 (± 0.418) HL − 0.414 (± 0.57) deHW + 0.854 (± 0.486) dHW − 1.36 (± 0.452) UL + 0.77 (± 0.409) peUW + 0.306 (± 0.404) CmL + 0.83 (± 0.31) dFWcc	430	0.979	24.9	18.6	19.4
ORNnl	−2.392 (± 0.259) + 1.799 (± 0.422) HL − 0.355 (± 0.32) bcL + 1.014 (± 0.427) dHW − 1.003 (± 0.546) dUW − 1.475 (± 0.384) RL + 0.363 (± 0.395) DCmW	438	0.975	27.0	19.5	20.2

The 95% confidence intervals for intercepts and slopes appear in brackets. Abbreviations for fossil taxa: ARCH, Archaeopterygidae; CONF, Confuciusornithidae; ENAN, Enantiornithes; JEH, Jeholornithidae; ORPH, Ornithuromorpha; ORNnl-Ornitothoraces with non-preserved legs; SAPE, Sapeornithidae. For abbreviations of variables, see legend to Table 1; |%MPE|, mean percentage prediction error calculated over the whole data set; |%MPE|$_{vs}$, mean percentage prediction error calculated over the validating subsample (20% of specimens); N, number of individuals; %SEE, percentage standard error of the estimate.

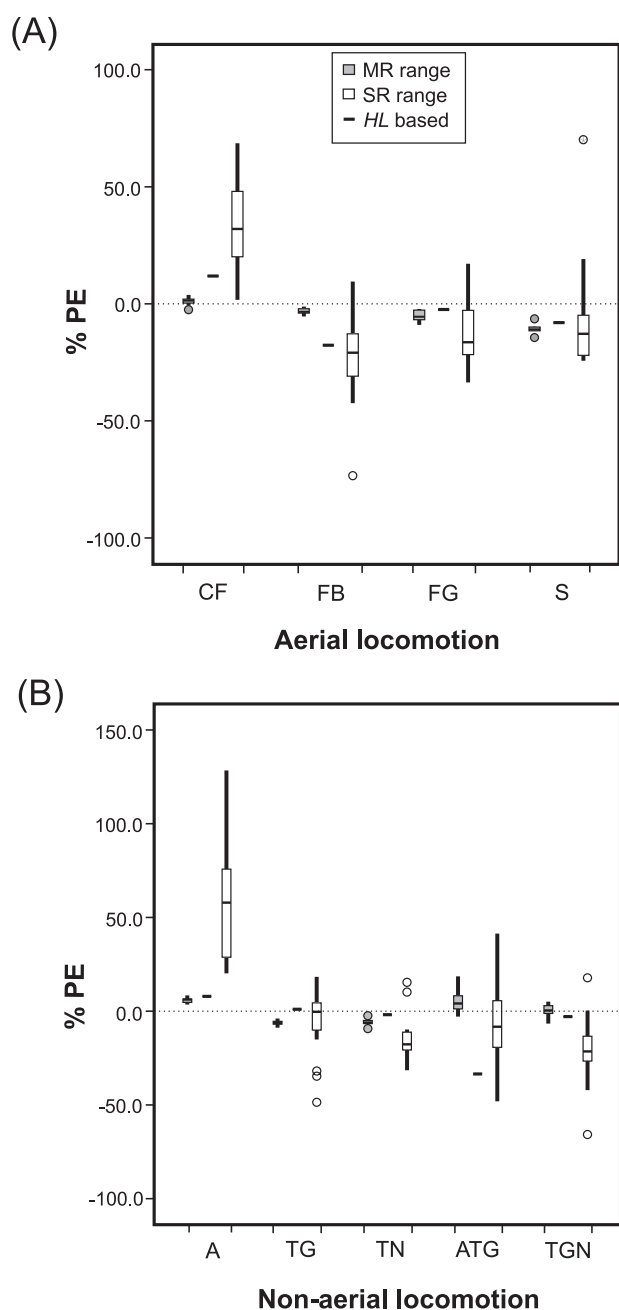**Table 5.** Published univariate scaling equations used for estimating body mass (BM) in fossil birds

Source	$y = ax^b$	N	R^2	$|\%MPE|$
Maloiy *et al.*, 1979	dFW = 6.4 BM$^{0.44}$	8	–	35.0
	FL = 76 BM$^{0.28}$		–	288.9
	TL = 115 BM$^{0.43}$		–	53.0
	TmL = 81 BM$^{0.49}$		–	135.3
Prange *et al.*, 1979	HL = 118.03 BM$^{0.482}$	22	0.980	53.7
	FL = 61.64 BM$^{0.359}$		0.980	51.3
Alexander, 1983	dFWcc = 6.3 BM$^{0.43}$	≈ 70	–	35.0
	FL = 36 (BM / 0.19)$^{0.37}$		–	59.1
	TL = 64 (BM / 0.22)$^{0.37}$		–	65.1
	TmL = 41 (BM / 0.17)$^{0.34}$		–	333.9
Yalden, 1984	BM = 13.25 dFW$^{2.353}$	≈ 100	–	36.1
	BM = 8.398 dHW$^{2.5}$		–	39.0
	BM = 0.143 HL$^{2.189}$		–	71.3
	BM = 0.047 UL$^{2.118}$		–	53.7
Olmos 1988	FL = 0.5659 BM$^{0.3424}$	65*	0.865	51.6
Olmos *et al.* 1996	HL = 0.46 BM$^{0.43}$	101	0.903	43.7
	dHW = 0.06 BM$^{0.368}$	108	0.922	51.3
	UL = 0.6 BM$^{0.393}$	113	0.810	52.0
	dUW = 0.05 BM$^{0.351}$	115	0.922	30.1
	FL = 0.56 BM$^{0.342}$	105	0.865	50.3
	dFW = 0.04 BM$^{0.383}$	106	0.941	34.1
	TL = 1.05 BM$^{0.376}$	104	0.865	153.1
Nudds 2007	total arm = 0.3 BM$^{0.42}$	306	0.884	58.0
	HL = 0.1 BM$^{0.45}$	748	0.884	44.3
	UL = 0.1 BM$^{0.43}$	743	0.792	51.9
	L$_{manus}$ = 0.09 BM$^{0.39}$	306	0.846	58.8
Liu *et al.* 2012	BM = 0.117 HL$^{1.733}$	422	0.851	66.4
	BM = 17.26 dTW$^{2.159}$		0.865	–
Field *et al.* 2013	BM = 2 HAF$^{2.44}$	863	0.988	13.0†

N and R^2 refer to the sample sizes used to generate the model in the original studies and values published for the coefficient of determination, respectively. $|\%MPE|$ values were calculated by applying these functions to our data set of modern birds. For abbreviations of the variables, see legend to Table 1. N, number of specimens; *number of species; $|\%MPE|$, mean percentage prediction error calculated over the whole data set at this work; †MPE published by the authors.

values. This results in an important underestimation of BM for continuous flapping flyers (from 1.7 to 68.6%), as well as an overestimation in the flap and bounding group (from −73.4 to −0.2%). Body mass is also overestimated by most SR functions in flap-and-gliding [from −33.6 to −2.4%, except 17.1% for length of tibiatarsus (TL)] and soaring birds [from −24.3 to −3.3%, except 19.1% for TL and 70.0% for length of tarsometatarsus (TmL)]. In comparison, MR equations show %PE values that were much closer to zero in all cases. Continuous flapping flyers exhibit a range of %PE between −2.5 and 3.8%, and all remaining groups are only slightly overestimated (flap and gliding from −9.0 to −2.2%; flap and bounding from −5.4 to −1.2%; soaring from −14.4 to −6.4%). The results of the ANOVA (Table 6) show that the differences in %PE values among aerial groups obtained with the regres-

sions for Archaeopterygidae and Sapeornithidae are nonsignificant (and hence their BM predictions are independent from flight type). These differences are significant in the rest of the MR and SR, which means that their BM predictions are influenced by locomotion type; however, the lower values of mean %PE, standard deviations of %PE, and F-statistics obtained for each aerial group (Table 6) indicate that the remaining MR functions (i.e. those derived for Jeholornithidae, Confuciusornithidae, Enantiornithes, basal Ornithuromorpha, and Ornithothoraces, with nonpreserved hindlimbs) minimize the influence of locomotion type better than any SR.

Similar results were obtained for groups of nonaerial locomotion (Fig. 5B; Table 7). All univariate regressions underestimate BM in aquatic birds (%PE from 7.9 to 128.4%). In contrast, non-terrestrial birds tend

(A)

Aerial locomotion

(B)

Non-aerial locomotion

Figure 5. Box plots with whiskers showing the variation among locomotion modes of percentage prediction errors (%PE) for the multiple regression analysis (MR; grey) and the single regression analysis (SR; white) equations adjusted in modern birds. Box length shows the interquartile range (25th and 75th percentiles). The horizontal line within boxes indicates the median. Vertical lines show the 5–95% confidence limits. Points indicate values outside these limits (i.e. outliers). Black horizontal lines between the grey and white boxes show the %PE estimates obtained with the equation based on humeral length (HL), the osteological variable less affected by ecological groupings (as shown by its lower F-statistic values; see Tables 6 and 7). %PE values greater than zero indicate an underestimation and %PE values lower than zero indicate an overestimation. A, %PE variation with respect to groups of aerial locomotion. Abbreviations: CF, continuous flapping; FG, flapping and gliding; FB, flapping and bounding; S, soaring. B, %PE variation with respect to groups of non-aerial locomotion. Abbreviations: A, aquatic; TG, terrestrial ground-dwelling; TN, terrestrial non-ground-dwelling; ATG, aquatic and ground-dwelling; TGN, perching and ground-dwelling.

◄ _____

and perching) show also narrow ranges of %PE values, but these results must be interpreted with caution because they have low sample sizes. ANOVA tests showed the absence of significant differences between %PE values for non-aerial groups in the SR of humeral length and in the MR function for fossils with non-preserved hindlimbs (Table 7), which means that the predictions of these equations are not biased by non-aerial locomotion type. As in the analysis of aerial locomotion, the remaining MR equations showed low values for the within-group mean %PE values, and their standard deviations, as well as low F-statistics (Table 7). Compared with SR, this again indicates a lesser influence of non-aerial locomotion type on the final predictions.

These results show that the use of a combination of several x-variables in an MR predictive model minimizes the effects of morphological differences among modern birds resulting from locomotion mode. Based on the current data set, results support MR models as more robust for estimating BM than SR functions in neognaths.

BODY MASS OF MESOZOIC STEM TAXA

The MR functions obtained (Table 4) were used to estimate BM in the fossil specimens (Table 8). The range of BM for *Archaeopteryx* shows that a specimen from Solnhofen is particularly large (1088 g, range 883–1292 g) compared with the others analysed (all ≤450 g). The two specimens of *Jeholornis* also show noticeable differences in BM, 983 g (range 813–1153 g) and

to be overestimated, although they show the lowest error range (%PE from −31.5 to 15.4%). Finally, ground-dwelling birds have intermediate %PE values (between −48.6 and 18.3%). As in the analyses of birds distributed among groups of aerial locomotion, %PE values from MR errors are closer to zero in each of the non-aerial groups. Aquatic birds are slightly underestimated (from 3.5 to 8.4%) by MR equations, whereas both terrestrial dwelling and non-dwelling birds are slightly overestimated (from −8.8 to −3.9% and from −9.3 to −2.4%, respectively). Intermediate groups (i.e. aquatic and ground-dwelling, and terrestrial-dwelling

Table 6. One-way ANOVAs for the mean prediction error (%PE) values in the groups of aerial locomotion for each single regression (SR) and multiple regression (MR) analysis

Function	CF			FG			FB			S			ANOVA	
	N	%PE	SD	N	%PE	SD	N	%PE	SD	N	%PE	SD	F	P
HL	205	11.9	55.5	118	−2.4	75.9	98	−17.7	22.6	65	−8.0	45.6	6.956	**
dpL	207	33.3	68.2	119	−15.9	40.0	103	−14.5	28.7	66	−23.7	26.9	39.817	**
bcL	207	21.7	50.2	119	−30.3	28.7	103	−0.2	34.7	66	−4.9	33.5	41.930	**
deHW	207	66.0	100.1	119	−16.4	20.9	103	−28.2	15.5	66	−22.2	30.9	69.845	**
dHW	206	41.9	80.4	118	−18.6	28.6	100	−19.0	18.6	66	−22.0	21.5	49.144	**
UL	192	48.3	86.6	112	−21.4	50.5	99	−31.9	22.8	62	−24.3	32.4	52.525	**
peUW	202	52.7	74.8	119	−14.4	23.3	103	−32.2	16.8	65	−24.0	24.2	89.854	**
deUW	202	39.0	76.1	119	−17.2	20.4	101	−12.8	17.7	66	−4.7	34.1	39.510	**
dUW	206	27.7	52.9	118	−15.3	24.4	100	−11.9	19.1	66	−12.8	36.4	42.964	**
RL	200	48.0	91.8	112	−21.7	50.4	95	−29.3	25.6	61	−21.5	36.0	44.473	**
dRW	200	1.7	51.9	113	−29.4	36.0	94	9.5	59.7	65	−7.4	37.5	13.676	**
CmL	203	26.7	73.8	113	−25.0	28.7	99	−6.8	26.1	66	−17.1	32.1	28.464	**
DCmW	200	68.6	135.1	108	−33.6	23.9	99	−20.9	24.4	66	−8.5	46.0	39.974	**
FL	202	32.0	107.8	113	−2.8	39.9	101	−30.9	16.6	65	−3.3	58.0	16.634	**
dFWcc	201	20.1	54.4	113	2.4	35.0	101	−26.2	16.3	65	−16.7	34.6	32.044	**
dFWml	201	3.5	39.3	113	−2.0	32.2	102	−14.2	22.1	65	−10.1	37.0	6.970	**
TL	193	3.8	54.2	110	17.1	55.9	97	−42.4	30.5	65	19.1	96.9	22.491	**
TmL	207	16.1	85.2	112	0.8	58.0	103	−73.4	17.0	63	70.0	252.2	25.162	**
ARCH	181	−2.5	25.3	99	−2.2	28.0	88	−1.2	17.3	54	−11.6	19.7	2.542	−
JEHO	173	3.8	24.0	100	−2.8	24.8	87	−4.5	15.6	57	−11.4	17.5	7.926	**
SAPE	168	−0.6	22.0	88	−2.6	21.9	85	−1.9	15.1	54	−6.4	20.1	1.145	−
CONF	173	2.1	23.5	94	−5.5	21.2	87	−3.3	14.2	57	−10.2	19.3	6.152	**
ENAN	176	1.8	27.9	103	−9.0	25.1	89	−3.4	17.1	58	−11.0	18.4	6.480	**
ORPH	181	1.2	26.3	101	−5.6	25.7	89	−5.4	15.2	59	−14.4	21.1	6.997	**
ORNnl	190	1.9	28.0	101	−8.0	25.1	88	−2.3	20.8	59	−9.7	21.9	5.101	**

Flying mode abbreviations: CF, continuous flapping; FG, flapping and gliding; FB, flapping and bounding; S, soaring. For abbreviations of variables, see legend to Table 1. For abbreviations of MR functions specific for fossil taxa, see legend to Table 4. $**P < 0.01$; $^{-}P > 0.05$. SD indicates the standard deviations for the mean values.

1516 g (range 1254–1779 g). Of these specimens, the smallest (V 13353) is considered as a juvenile (Zhou & Zhang, 2003). Among the members of *Sapeornis*, DNHM 3078 (considered as a juvenile by Gao *et al*., 2012) has the lowest estimate (491 g, range 410–572 g), whereas the BM range for the three adults is 752–1049 g. The estimates obtained for the juveniles of *Jeholornis* and *Sapeornis* should be considered with caution. In the case of Confuciusornithidae, the most basal *Eoconfuciusornis* is the smallest specimen, with an estimate of 217 g (range 181–254 g). In contrast, BM ranges from 308 to 987 g in the later genus *Confuciusornis*. Among Enantiornithes, BM is estimated at 70 g (range 57–84 g) for the oldest known record of *Protopteryx*. Similarly, the remaining members of this clade do not exceed 90 g, except in three cases: *Otogornis* (171 g, range 137–206 g); *Longipteryx* (193 g, range 155–230 g); and the heaviest specimen of *Pengornis* (437 g, range 352–521 g). Ordered by age, the earliest members of Ornithuromorpha,

Archaeorhynchus, *Longicrusavis*, and *Hongshanornis*, have BM estimates of 272 g (range 219–325 g), 89 g (range 72–106 g), and 88 g (range 71–105 g), respectively. In contrast, subsequent members have larger estimates, 321 g (range 259–383 g) for *Yixianornis*, 498–921 g for *Yanornis*, and 820 g (range 661–979 g) in the case of *Jianchangornis*. In the youngest genera, BM is estimated to range between 126 and 195 g for the two specimens of *Gansus*, and is 184 g (range 147–221 g) for *Apsaravis*. The differences with the values published for all these stem taxa (Table 8) are discussed below.

DISCUSSION

Body size represents a major physical constraint on the morphology of organisms (McMahon & Bonner, 1983). Among vertebrates, limb bone morphology is considered to reflect a trade-off between energetic requirements and functional constraints resulting from

Table 7. One-way ANOVAs for the mean prediction error (%PE) values in the groups of non-aerial locomotion in each single regression (SR) and multiple regression (MR) analysis

Function	A			TG			TN			ATG			TGN			ANOVA	
	N	%PE	SD	N	%PE	SD	N	%PE	SD	N	%PE	SD	N	%PE	SD	F	P
HL	103	7.9	40.1	166	1.0	61.7	180	−1.9	62.9	21	−33.5	19.0	16	−2.9	30.2	1.84	−
dpL	104	26.7	52.9	171	18.3	69.4	182	−20.7	37.3	22	−9.6	23.7	16	0.3	29.7	16.81	**
bcL	104	28.8	54.5	171	2.9	44.2	182	−9.8	33.5	22	−48.1	24.6	16	−5.9	26.4	22.19	**
deHW	104	117.5	115.7	171	−0.3	35.3	182	−24.1	23.5	22	−0.7	10.9	16	−26.6	13.1	98.89	**
dHW	103	81.5	94.7	169	−3.7	29.6	181	−24.6	21.8	21	5.6	17.3	16	−21.5	5.7	77.9	**
UL	99	60.0	75.0	157	4.4	82.5	176	−19.0	46.5	19	−37.9	21.6	14	−25.5	28.2	23.36	**
peUW	103	75.8	94.8	167	−4.4	39.4	181	−12.6	30.2	22	−8.3	11.4	16	−34.0	9.4	50.35	**
deUW	104	55.2	94.1	164	7.1	40.1	182	−11.2	28.0	22	−15.6	13.6	16	−18.6	12.5	25.16	**
dUW	104	43.4	65.0	169	1.1	30.8	180	−16.1	25.3	22	16.2	21.4	15	−19.9	14.0	40.19	**
RL	103	57.9	75.3	158	8.1	90.8	174	−17.7	46.8	19	−40.2	20.5	14	−22.7	32.2	19.88	**
dRW	103	20.2	48.8	159	−3.9	55.6	176	−22.1	40.7	20	−19.0	29.5	14	17.8	50.5	13.41	**
CmL	103	23.2	41.3	167	7.3	78.4	177	−10.6	37.3	20	−38.4	18.2	14	−13.4	15.5	8.26	**
DCmW	101	128.4	167.2	165	1.9	40.9	172	−31.5	20.9	20	−0.9	19.6	15	−24.6	10.8	60.13	**
FL	99	114.9	108.6	168	−31.9	22.8	180	−17.9	37.6	18	25.6	17.4	16	−28.5	8.6	127.5	**
dFWcc	99	60.3	53.3	168	−15.1	21.6	179	−18.2	27.9	18	39.7	28.2	16	−19.3	9.1	116.0	**
dFWml	99	25.5	46.9	168	−10.1	23.6	180	−17.3	22.1	18	41.4	39.4	16	−5.4	15.9	44.09	**
TL	101	40.9	58.1	158	−34.7	37.7	174	10.2	69.8	18	−4.6	26.9	14	−42.1	25.3	33.41	**
TmL	104	69.2	73.2	171	−48.6	51.3	177	15.4	163.4	18	−19.3	39.2	15	−65.8	35.5	21.66	**
ARCH	94	6.6	26.6	143	−6.2	19.9	159	−9.3	23.7	15	18.6	23.5	11	5.1	6.7	11.45	**
JEHO	89	6.2	27.4	138	−6.4	17.6	163	−2.4	22.8	15	2.3	23.5	12	1.8	6.0	4.59	**
SAPE	90	5.1	23.7	133	−4.9	18.3	147	−4.9	19.9	14	6.0	18.2	11	0.4	7.1	4.91	**
CONF	91	8.4	24.7	139	−5.9	17.6	153	−6.2	20.7	16	4.1	15.3	12	−6.6	11.1	9.36	**
ENAN	93	6.2	31.1	141	−8.8	18.8	165	−5.0	24.6	15	−2.9	20.3	12	−2.4	4.4	5.59	**
ORPH	91	4.3	29.9	146	−7.5	19.6	167	−7.0	23.1	14	10.5	27.4	12	4.1	7.5	6.01	**
ORNnl	97	3.5	32.1	150	−3.9	21.7	161	−6.1	25.3	17	0.1	21.4	13	0.0	9.6	2.36	−

Abbreviations: A, aquatic birds; TG, terrestrial ground-dwelling birds; TN, terrestrial non-ground-dwelling birds; ATG, aquatic and ground-dwelling birds; TGN, ground-dwelling and perching birds. For abbreviations of variables, see legend to Table 1. For abbreviations of MR functions specific for fossil taxa, see legend to Table 4. **$P < 0.01$; $^{-}P > 0.05$; SD indicates the standard deviations for the mean values.

locomotion and body size (Rubin & Lanyon, 1982; Blob & Biewener, 2001; Garcia & Da Silva, 2006; Clemente *et al.*, 2011; Campione & Evans, 2012). Therefore, the factors involved in this trade-off must be considered when generating models for size estimation. Our results indicate that the combined use of the limb measurements that correlated best with BM provides robust predictive models. Moreover, although approaches based on body reconstructions provide a valuable tool for estimating BM in extinct taxa (Henderson, 1999; Seebacher, 2001; Allen *et al.*, 2013), the equations generated from skeletal measurements do not share the same sources of error inherent to body reconstruction, such as the need to accurately estimate the bulk densities and proportions of the different body tissues (Grand, 1990; Campione & Evans, 2012). In addition, there are two advantages of the regression approach: (1) it is easier to perform a regression analysis using a set of linear measurements than to recon-

struct an animal body; and (2) a regression equation can be used with any fossil provided that the skeletal variable(s) included in the equation can be measured. Reconstruction methods can only be applied accurately to well-preserved fossils.

The use of scaled bone functions introduces other potential sources of error and bias, however, as a result of statistical limitations, phylogenetic constraints, ecological adaptations, and diagenetic compaction of fossil skeletons. Although several studies have taken some of these factors into account when generating BM estimation models (e.g., Egi, 2001; Mendoza *et al.*, 2006; De Esteban-Trivigno *et al.*, 2008; Figueirido *et al.*, 2011; Campione & Evans, 2012), this is the first time in which all have been tested and controlled for simultaneously. In addition, this analysis is particularly novel with regards to investigating diagenetic effects and gives credit to studies that have looked at the other factors in tandem.

Table 8. Body mass estimates for specimens of Mesozoic stem taxa (in kg) obtained from multiple regressions in this study (Table 4)

Species	Museum code	BM	%MPE L.I.	%MPE U.I.	Published BM (kg)
Archaeopteryx lithographica[a]	BMNH 37001	0.450	0.366	0.535	0.497[1] 0.214–0.242[4] 0.770[6] 0.468[9]
A. lithographica[a]	HMN 1880/1881	0.343	0.279	0.408	0.363[1] 0.250[2] 0.138–0.172[4] 0.530[6] 0.066–0.132[8] 0.276[9] 0.188[10]
A. lithographica[a]	JM 2257	0.176	0.143	0.209	163[1] 0.057–0.066[4] 0.174[9]
A. lithographica[a]	BMMS 500	1.088	0.883	1.292	0.265–0.330[4] 0.500[5] 0.992[10]
A. lithographica[a]	WDC-CSG-100	0.264	0.214	0.313	0.47[6]
Jeholornis prima	IVPP 13274	1.516	1.254	1.779	0.601–1.117[7] 2.598[10]
J. prima	IVPP 13353	0.983	0.813	1.153	0.285–0.856[7] 0.411[†]
Sapeornis chaoyangensis	IVPP 12698	0.900	0.752	1.049	0.726[5] 0.759–0.856[7] 1.900[10]
S. chaoyangensis	IVPP 13275	1.041	0.869	1.212	0.681–1.225[7]
S. chaoyangensis	IVPP 13276	1.038	0.866	1.209	0.713–0.712[7]
S. chaoyangensis juv.[b]	DNHM 3078	0.491	0.410	0.572	–
Eoconfuciusornis zhenghi[c]	IVPP 11977	0.217	0.181	0.254	0.21[6] 0.167[10]
Confuciusornis sp	IVPP 11370	0.844	0.702	0.987	0.260–0.582[7]
Confuciusornis sanctus	IVPP 11372	0.371	0.308	0.434	0.164–0.291[7]
C. sanctus	IVPP 11374	0.520	0.432	0.608	–
C. sanctus	IVPP 11375	0.429	0.356	0.501	–
C. sanctus	IVPP 11640	0.762	0.633	0.891	–
C. sanctus	MCFO-0374	0.550	0.457	0.643	–
Pengornis houi	V 15336	0.437	0.352	0.521	0.235–0.557[7] 0.159–0.311[8]
Cathayornis yandica	V 9169	0.063	0.051	0.076	0.061[3] 0.062–0.095 [7] 0.023[10]
Cuspirostrisornis houi	V 10897	0.070	0.056	0.083	–
Eocathayornis walkeri	V 10916	0.056	0.045	0.067	0.042[7]
Eoenantiornis buhleri	V 11537	0.080	0.065	0.096	0.061[7] 0.109[10]
Longchengornis sanyanensis	V 10530	0.086	0.070	0.103	0.084[10]
Otogornis genghisi	V 9607	0.171	0.137	0.206	0.085[7]
Longirostravis hani	V 11309	0.039	0.032	0.047	0.043–0.054[7] 0.031[10]
Longipteryx chaoyangensis	V 12325	0.193	0.155	0.230	0.127–0.510[7] 0.130[10]
Rapaxavis pani[d]	DNHM D2522	0.047	0.038	0.056	0.063[7] 0.016[10]
Protopteryx fengningensis	V 11665	0.070	0.057	0.084	0.036[10]
Vescornis hebeiensis[e]	NIGPAS 130722	0.050	0.040	0.059	0.043–0.114[7] 0.037[10]
Concornis lacustris	LH-2814	0.070	0.056	0.084	0.079[3] 0.046[10]
Eoalulavis hoyasi	LH-13500	0.045	0.036	0.054	0.055[10] 0.031[†]
Archaeorhynchus spathula	V 14287	0.272	0.219	0.325	0.171–0.258[7] 0.167[10] 0.171[†]
Yixianornis grabaui	V 12631	0.321	0.259	0.383	0.100[5] 0.142–0.213[7] 0.136–0.266[8] 0.315[10] 0.208[†]
Zhongjianornis yangi	V 15900	0.570	0.459	0.680	0.279–0.363[7] 0.392[†]
Yanornis martini	V 12558	0.772	0.622	0.921	0.336–0.557[7] 1.492[10]
Y. martini	V 10996	0.618	0.498	0.738	–
Jianchangornis microdonta	V 16708	0.820	0.661	0.979	0.304–0.739[7]
Hongshanornis longicresta	V 14533	0.088	0.071	0.105	0.049–0.077[7] 0.031[10]
Longicrusavis houi[f]	PKUP IVPP1069	0.089	0.072	0.106	–
Gansus yumenensis	CAGS 04-CM-003	0.156	0.126	0.187	0.109[10] 0.397[†]
G. yumenensis	CAGS 04-CM-004	0.162	0.130	0.195	–
Apsaravis ukhaana[g]	IGM 100/1017	0.184	0.147	0.221	0.100[5] 0.115[10] 0.361[†]

Prediction intervals were calculated as $BM_{predicted} \pm |\%MPE|$. In the fossils not studied directly, superscript letters indicate the source of the images on which the measurements were taken: [a]Wellnhofer (2008); [b]Gao *et al*. (2012); [c]pictures provided by L. Chiappe; [d]pictures provided by J. O'Connor; [e]Zhang, Ericson & Zhou (2004); [f]O'Connor, Gao & Chiappe (2010); [g]Clarke and Norell (2002). Sources of BM published for fossils are indicated by superscript numbers: [1]Yalden (1984); [2]Seebacher (2001), [3]Sanz *et al*. (2002), [4]Atanassov and Strauss (2002), [5]Turner *et al*. (2007), [6]Wang *et al*. (2012), [7]Liu, Zhou & Zhang (2012), [8]Allen *et al*. (2013), [9]Elzanowski (2002), [10] Benson *et al*. (2014). [†]BM calculated from the allometric function based on HAF, derived by Field *et al*. (2013).

STATISTICAL ERRORS IN THE SAMPLE OF MODERN BIRDS

The representativeness of the different taxa included in the data set is important for evaluating the statistical validity and reliability of the predictions (Smith, 2002). On the one hand, a couple of studies focused on mammals have shown that an unequal distribution of species within the families of a sample results in worse predictor functions than those obtained from the same sample weighted for taxonomic evenness (Mendoza *et al*., 2006; De Esteban-Trivigno *et al*., 2008). On the other hand, De Esteban-Trivigno & Köhler (2011) did not obtain differences of predictive strength between weighted and unweighted equations, which they attributed to using a data set with a smaller taxonomic range than others had used in previous works. The broad taxonomic range of our data set, however, which includes volant species from 22 orders of Neognathae, recovered a reduction in model support when they were weighted for taxonomic evenness (Table S4). In any case, the absence of trends between the residuals for weighted and unweighted equations (Fig. 3) indicates that the robustness of the predictions does not increase with the taxonomic evenness of the sample. This may suggest that, compared with mammals, the avian postcranial skeleton is morphologically conservative in terms of the skeletal measurements studied in this paper, which may reflect functional constraints posed by aerial locomotion; however, it is worth noting that a comparison of the SR equations of Campione & Evans (2012) for mammals and reptiles with those of Field *et al*. (2013) for birds shows that a number of measurements, such as stylopodial circumferences, are more conserved in quadruped tetrapods, compared with birds. In any case, neither Campione & Evans (2012) nor Field *et al*. (2013) tested explicitly for the effects of taxonomic evenness; however, it is worth recognizing that an indirect test was performed in both studies, as they statistically compared clades with varying sample sizes.

Our MR equations incorporate combinations of measurements of length and diameter taken in the distal and proximal limb bones, which results in greater accuracy for the models (Gingerich, 1990). The consistency of the BM predictions based on these equations provides an additional indication of the predictive strength of our MR models compared with the use of SR equations. This effect is even greater in our study, as we use a combination of wing and hindlimb measurements, which are under different loading regimes (see De Margerie *et al*., 2005).

Univariate allometric equations have been largely used for estimating BM in extinct birds (see references in Table 5) and other vertebrates (e.g. Damuth & McFadden, 1990; Palmqvist *et al*., 2011). Compared with SR analyses, MR functions have been less used for estimating BM, especially in the case of birds, in spite of the fact that MR tends to increase predictive strength (Biknevicius, 1999; Atanassov & Strauss, 2002; Christiansen & Fariña, 2004; Mendoza *et al*., 2006; Figueirido *et al*., 2011; De Esteban-Trivigno *et al*., 2008; De Esteban-Trivigno & Köhler, 2011). An exception is the study on quadrupedal tetrapods by Campione & Evans (2012), who did not report on differences between both methods. Smith (2002) argued that MR equations have a notorious lack of generalizability for new data, especially when stepwise methods for selection of variables are employed, as in this study, thus requiring a validating sample. In our case, the mean percentage prediction errors obtained with the MR functions for the validating subsample (Tables 2 and 4) are similar to those derived from the whole data set, which supports their generalizability. Moreover, all specimens from two families (Otididae and Ciconiidae) as well as from two species [*Tadorna tadorna* (Linnaeus, 1758) and *Tringa glareola* (Linnaeus, 1758)] were excluded from the validating sample. In all of these cases, the prediction errors were similar to (or even lower than) the ones obtained for the remaining families and species. Our results also agree with the studies cited above, which indicate that MR functions increase their predictive strength compared with SR approaches: our MR equations have lower |%MPE| values that our SR functions (Tables 2 and 4). Similarly, they have more predictive strength than previously published SR functions (Tables 4 and 5). The only exception is the study of Field *et al*. (2013), who provided an SR equation with higher predictive strength than the MR functions of this study, as evidenced by a slightly lower |%MPE| value.

Finally, morphological variables almost always show common, interdependent relationships (Quinn & Keough, 2001; Christiansen, 2002), which can introduce redundancy effects that result in: (1) unstable models (i.e. small changes in the sample of species or in the skeletal measurements can change the regression parameters considerably; Bowerman & O'Connell, 1990); and (2) an inflation of the confidence intervals for the model parameters (Quinn & Keough, 2001). Collinearity effects are a problem if the main goal of the functions is to provide a descriptive model (Mitchell-Olds & Shaw, 1987; Tabachnick & Fidell, 1996; Rawlings *et al*., 1998); however, evaluating their effects on predictive models is more controversial [e.g. compare Tabachnick & Fidell, (1996) with Rawlings *et al*. (1998)]. An alternative to MR, PCR has been proposed for avoiding redundancy, as it is a multivariate non-collinear model (Jolliffe, 1982; Stone & Brooks, 1990; Rawlings *et al*., 1998). Atanassov & Strauss (2002) showed that predictive strength was slightly greater for PCR than for MR in their sample, although the result for their bird

subsample was the opposite. Our results show that MR has a slightly higher predictive strength than PCR, in spite of the correlations between the variables included in the equations. In addition, although the confidence intervals for the MR parameters are greater than those of SR (Table 2), which could result in an increase of uncertainty for the estimates (Quinn & Keough, 2001), this was avoided using the |%MPE| value for estimating the confidence interval for predictions (Table 8). Therefore, our results support those of Rawlings *et al.* (1998) in that redundancy present in our MR models does not affect its accuracy in volant birds. We chose MR instead of PCR for estimating BM in fossils for two reasons: (1) compared with the regression coefficients for the skeletal variables, the regression coefficients for the principal components are more difficult to interpret (Mitchell-Olds & Shaw, 1987), because each component is a linear combination of all *x*-variables used as predictors; (2) PCs may fail to account for the regression fit (i.e. it is theoretically possible that the first PCs, which can account for almost 100% of the variance, contribute nothing to the fit, whereas the response variable may fit perfectly the last PCs, which are always ignored by the PCR methodology; see Hadi & Ling, 1998); (3) the fossils must be incorporated in the PCA before obtaining their PC scores, which implies a lack of consistency for the model (as it could only be applied to those specimens that were initially included in the analysis, and any addition of new data will result in a different PCA); and (4) all measurements included in the PCA need to be available in the fossil taxa, which is unfortunately not always the case.

AVOIDING BIASES FROM THE PHYLOGENETIC RELATIONSHIPS BETWEEN MODERN BIRDS

The use of phylogenetic comparative methods is now essential in any statistical approach for analysing comparative data (Harvey & Pagel, 1991). The method most used, Felsenstein's PIC, has been criticized for assuming Brownian motion for character evolution (e.g. Díaz-Uriarte & Garland, 1996; Hansen & Orzack, 2005); however, many studies have used this method for controlling sample independence (e.g. see reviews in Ricklefs & Starck, 1996; Price, 1997; Nudds, 2007; Clarke *et al.*, 2010; Figueirido *et al.*, 2010; Wang *et al.*, 2012; Allen *et al.*, 2013; Martín-Serra *et al.*, 2014; Vanhooydonck *et al.*, 2014). Given that the regressions based on PICs are highly significant and the confidence intervals of their slopes overlap with those derived from raw data on species means (Table 3), we suggest that phylogeny does not play an essential role in the scaling patterns of the major limb bones among flying birds, which can be considered as independent for statistical purposes (see Campione & Evans, 2012). This agrees with

the results obtained by Field *et al.* (2013), who tested their SR equations for differences in prediction errors among avian orders; however, it is worth mentioning that Field *et al.* (2013) did not directly test for phylogenetic effects.

AVOIDING BIASES WHEN ESTIMATING BM IN FOSSILS

Scaling methods are almost universally accepted as a means for accurately estimating BM in extinct taxa from crown groups, but have certainly been criticized when applied to more distantly related stem taxa (Henderson, 1999; Carrano, 2001; Smith, 2002; Packard *et al.*, 2009; Campione & Evans, 2012; Field *et al.*, 2013). Although our modern bird sample has a broad BM range and the bone dimensions of each fossil fall within the range of extant birds, it should be noted that Mesozoic birds are stem groups. This means that the predictions incur 'statistical phylogenetic extrapolation' from Neognathae. Therefore, scaling patterns between osteological dimensions and BM should be taken into account. Given that a direct examination of the relationship between bone dimensions and BM in fossils is not possible, similar scaling patterns to those of extant birds have usually been assumed (e.g. Atanassov & Strauss, 2002; Sanz *et al.*, 2002; Palmqvist & Vizcaíno, 2003; Longrich, 2006; Turner *et al.*, 2007; Peters & Peters, 2009; Nudds & Dyke, 2010; Liu *et al.*, 2012; Wang *et al.*, 2012; Balanoff *et al.*, 2013; Chinsamy *et al.*, 2013; Field *et al.*, 2013). We did not uphold this assumption, however, because historical contingency and adaptive changes in the avian lineage have affected skeletal morphology (e.g. see in Jenkins, 1993; Chiappe *et al.*, 1999; Baier, Gatesy & Jenkins, 2007), which would invalidate the applicability of the equations obtained from extant birds. For example, the recent study of Field *et al.* (2013) provided a robust SR model for estimating BM in the avian crown group, based on the maximum diameter of the HAF of the coracoid; however, as they acknowledge, their applicability to stem taxa is more problematic. The morphological changes in the shoulder joint throughout avian evolution, which affect the orientation and dimensions of the major axis of the glenoid facet, are well known. *Archaeopteryx* and *Confuciusornis* have shoulder anatomy that differs from the shoulder anatomy of extant birds, and is similar to that of their running maniraptoran ancestors, with the scapula and coracoid fused at an angle of 90° and a laterally oriented glenoid facet. This orientation is intermediate between the posteroventrally oriented glenoid of coelurosaurs and the dorsolaterally oriented glenoid of Neornithes. In contrast, the scapula–coracoid joint is not fixed in modern birds and displays an acute angle, with a concavo–convex glenoid facet and a bulbous humeral head well adapted for sustained flapping flight (Jenkins, 1993; Chiappe *et al.*,

1999; Chiappe, 2007). Other stem taxa such as *Jeholornis*, *Sapeornis*, and *Otogornis* show intermediate morphologies (Hou, 2001; Baier *et al*., 2007). This is accompanied by a noticeable morphological variation of the scapula and coracoid among major avian lineages (see Chiappe & Witmer, 2002). As a result, mass estimates obtained for fossils using an apomorphic trait would be unrealistic. Therefore, we consider that the variable HAF of Field *et al*. (2013) is a good predictor for BM in the avian crown group, but is not appropriate for estimating BM in stem taxa. We adopt a methodology whereby variables were selected for each fossil group that showed similar scaling patterns to those found in modern birds (Fig. 4). This procedure generated seven clade-specific functions, which we used to predict BM in a sample of extinct birds.

A number of studies on the functional adaptations of modern birds have shown that the skeletal morphology of the wings and hindlimbs is highly constrained by locomotion dynamics (Cubo & Casinos, 1998; De Margerie *et al*., 2005; Habib & Ruff, 2008; Hinic-Frlog & Motani, 2010; Simmons, 2010; Bell & Chiappe, 2011; Simmons *et al*., 2011; Dececchi & Larsson, 2013; Field *et al*., 2013). For this reason, there are dramatic differences between BM estimates for a given Mesozoic bird using different univariate models derived from the same sample (Table S5; also see Liu *et al*., 2012: table 3). In contrast, MR results showed that the combined use of variables highly correlated with BM minimized the effects of mechanical adaptions on bone scaling patterns, both for flight and for non-aerial locomotion (Fig. 5; Tables S3 and S4). It is worth noting here that the variable HAF of Field *et al*. (2013) seems to be very conservative among avian orders, which could imply that it is minimally affected by locomotion types, although this has not been directly tested.

Diagenetic compaction of fossil skeletons can alter their osteological dimensions, in particular epiphyseal and diaphyseal width variables. Their use could result in unrealistic BM predictions for Mesozoic birds. Many studies have circumvented this problem using scaling equations that incorporate only one longitudinal measurement (e.g. Sanz *et al*., 2002; Longrich, 2006; Turner *et al*., 2007; Peters & Peters, 2009; Wang *et al*., 2012; Balanoff *et al*., 2013; Chinsamy *et al*., 2013), in spite of the fact that these functions have less predictive strength than those derived from diaphyseal variables (Campione & Evans, 2012; Campione *et al*., 2014; Tables 2 and 5). Following the selection of variables (Figs 3 and S2), measurements of fossils that were severely altered by diagenetic compaction were discarded; however, a medium-to-low degree of diagenetic compaction is not easy to detect, but it could have noticeable effects on BM predictions based on a single diaphyseal variable. In the MR functions, however, the individual effect of each variable is less marked. Hence, the use of MR will minimize the effects of diagenetic compaction on bone diameter.

BODY MASS ESTIMATIONS FOR ANCIENT BIRDS

The BM estimates obtained for fossil birds using MR functions (Table 8) were higher than those predicted previously from body reconstructions (Seebacher, 2001; Allen *et al*., 2013) and from a femoral length scaling equation used by Turner *et al*. (2007), which was derived from the volumetric displacement of reconstructed models (Christiansen & Fariña, 2004). Given that the fossil specimens analysed in the studies cited before were also measured here, the differences in size estimates cannot result from individual variation in these species. Therefore, this disagreement probably results from errors in the estimates of body density and tissue proportions used in the reconstructions, as discussed before. This is further suggested by the fact that BM estimates obtained for the same fossils using various volumetric methods (e.g. the specimen of *Archaeopteryx* from Berlin and *Yixianornis* in Table 8) did not agree.

In the case of those osteological variables for which the fossils showed values that agree with their overall skeletal dimensions (i.e. *x*-variables for which fossils are placed near the regression line for modern birds). BM estimates derived from MR equations were closer to those obtained using SR functions (Tables 6 and S3). For example, our results showed BMs that are close to those obtained from femoral diaphyseal width (dFW) by Yalden (1984) and Elzanowski (2002) for *Archaeopteryx*, and from femoral length (FL) by Wang *et al*. (2012) for *Eoconfuciusornis* and by Sanz *et al*. (2002) for two enantiornithines, *Concornis* and *Cathayornis* (Fig. 4). The SR equations that use variables in which fossils score distantly from the regression line (even within the confidence intervals for Neognathae), however, tend to underestimate or overestimate BM, as in the case of the three specimens of *Archaeopteryx* studied by Wang *et al*. (2012), for which the BM estimates derived from FL were higher than our MR predictions (Table 8). The reason is that, compared with modern flying birds, the members of Archaeopterygidae have higher values of FL than expected for their body size (Fig. 4). For this reason, we think that the widely used SR equation based on FL for estimating BM in *Archaeopteryx* (e.g. Sanz *et al*., 2002; Longrich, 2006; Turner *et al*., 2007; Nudds & Dyke, 2010; Wang *et al*., 2012; Balanoff *et al*., 2013) is not an accurate model. This supports the results of Campione & Evans (2012) and Field *et al*. (2013), in which FL, in general, was found to be a non-accurate BM predictor.

Compared with our BM predictions (Table 8), the univariate scaling equations from humeral length (HL) and tibial width at midshaft (dTW), adjusted by Liu

et al. (2012), underestimate BM for almost all Mesozoic birds relative to our models. Specifically, HL provides very low estimates, except for enantiornithines. Surprisingly, these underestimations were independent of the relationship observed between HL and BM (Fig. 4), and were also below the BM obtained by our SR based on HL (Table S5). The predictions of Liu *et al.* (2012) based on dTW were also generally lower than our own estimates, but closer to those obtained from HL. Unfortunately, the variable dTW was not incorporated into our data set, and so we cannot know how it scales with body size. In spite of the huge data set used by Liu *et al.* (2012), they do not provide error estimates and, for this reason, the accuracy of their predictions can only be evaluated from their R^2 values.

The BM estimated by Atanassov & Strauss (2002) for four specimens of *Archaeopteryx* using multivariate functions were notably lower than our estimates. Unfortunately, we cannot analyse the reason for this disagreement, because they did not provide the functions and did not specify the variables used. Therefore, their results should be taking with caution.

As discussed above, the recent study of Field *et al.* (2013) provided a highly predictive SR function for the avian crown group, based on the variable HAF. For this reason, we used this equation with seven fossils of the stem taxa analysed here in which the variable HAF was measured unambiguously (Table 6), in order to compare the results obtained with our own BM estimates. Of the seven specimens, the two more derived specimens (*Apsaravis* and *Gansus*) provided BMs that doubled our estimates. In contrast, this comparison showed that BM was clearly underestimated in the other five fossils, which occur lower on the avian stem (*Jeholornis*, *Eoalulavis*, *Archaeorrhynchus*, *Yixianornis*, and *Zhongjianornis*). These differences may result from structural variations in the anatomy of the shoulder joint for avian stem taxa, given the differences in the shape of the glenoid facet discussed above. In such a case, the glenoid facet of the basalmost stem taxa would be shorter than in the avian crown group, which means that consistent scaling cannot be assumed for this variable.

Finally, the recent study of Benson *et al.* (2014) provides BM estimates for a number of Mesozoic birds based on their femoral anteroposterior and mediolateral shaft diameters. Compared with our BM values, most of these estimates are lower, except for *Jeholornis*, *Sapeornis*, and *Yanornis*, for which BM is largely overestimated. This discrepancy probably arises from the fact that Benson *et al.* (2014) derived their equations from a database of quadruped tetrapods that did not include any birds (Campione & Evans, 2012). The reason is that the femur bears different loads in bipeds and quadrupeds: in birds, in which the femur is more horizontally oriented than in most mammals, the an-

imal's weight is supported exclusively by the hindlimb; in contrast, weight loads during terrestrial locomotion are shared by the fore- and hindlimbs in quadrupeds (see Carrano, 1998; De Margerie *et al.*, 2005; Farke & Alicea, 2009). These differences tentatively suggest that the BM estimates provided for Mesozoic birds in the present study are more accurate; however, it should be noted that our MPEs greatly overlap with those of Benson *et al.* (2014), which are based on Campione & Evans (2012), who provided a mean MPE of 25%. Given that MPEs represent an average of the residuals around a common trend, this suggests that both approaches rely on common ground, despite vastly differing methods and data sets. In addition, it is worth noting that our mass estimate for the large specimen of *Archaeopteryx* is very close to the one obtained by Campione *et al.* (2014), based on the minimum circumference of stylopodial elements in a large data set of quadruped tetrapods.

In any case, the harsh reality of estimating BM for extinct species is that a number of assumptions must be made, in particular for stem taxa in which the skeletal morphology differs greatly from that of the crown taxa. This means that it is not possible to conclude confidently that a given BM estimate derived from a given model is more accurate than others. In spite of this, it is always possible to make comparisons between the estimates obtained using different models. If they are consistently different (i.e. one model systematically under- or overestimates relative to the other, as in the comparison between this study and those of Atanassov & Strauss, 2002; Liu *et al.*, 2012; and Field *et al.*, 2013), a general explanation can be suggested. If the differences are inconsistent (i.e. one model overestimates in some cases and underestimates in others, as in the comparison with Benson *et al.*, 2014), however, then they can be case-specific.

The BM models from Campione & Evans (2012), used by Benson *et al.* (2014), were meant to be general, offering a consistent method with which to estimate mass across multiple lineages and body plans. In contrast, our study developed a specific BM model for each avian stem taxa. This means that the accuracy of both approaches will be contingent on the base assumption about whether stem birds followed primitive or derived scaling patterns between skeletal measurements and body mass (N. Campione, pers. comm.); however, given the absence of birds in the data set of Campione & Evans (2012) and the fact that our models are stem-taxa specific, it stands to reason that our BM estimates for members of Aves are probably more precise.

CONCLUSION

We consider that the MR models obtained in this study are the most accurate for predicting BM in extinct birds,

especially for stem taxa, because: (1) the models derived from osteological measurements avoid the assumptions of the reconstruction methods; (2) the predictive strength and generalizability of the MR equations derived here is higher than those of the SR and PCR functions, and is also higher than that of other SR equations published based on appendicular variables; (3) the statistical problems that are usually attributed to regression models (i.e. taxonomic unevenness and non-independence of data as a result of phylogenetic relationships), and particularly to MR models (i.e. multicollinearity), have negligible effects on our MR equations; (4) the methodology used for selecting variables allowed us to obtain accurate subsets of predictor variables for estimating BM in each fossil stem group by minimizing the effects of differential scaling patterns with respect to the avian crown group that result from historical contingency or adaptive changes; (5) the combined use of several morphological variables in the MR equations allowed us to detect the presence of specimens heavily affected by diagenetic compaction; and (6) the MR models obtained were not affected by differences in locomotion mechanics.

Furthermore, although the several SR equations published are also robust to taxonomic differences, as shown by Campione & Evans (2012) for tetrapods and Field *et al*. (2013) for birds, this is the first time that statistical, phylogenetic, ecological, and taphonomic effects have been considered jointly for estimating BM in avian crown and stem taxa. Therefore, we suggest that the BM estimated here for avian stem fossils are probably the most useful published to date (although this cannot be evaluated directly for stem group taxa), and that the MR functions can be used for estimating BM in other fossils of presumably flying birds. Body mass estimates obtained in this study can be used in future studies for a number of palaeobiological and evolutionary aspects of Mesozoic stem taxa, including inferences on their flight abilities and trends throughout the avian lineage (Sanz *et al*., 2002; Palmqvist & Vizcaíno, 2003; Longrich, 2006; Turner *et al*., 2007; Butler & Goswami, 2008; Peters & Peters, 2009; Hone *et al*., 2008; Nudds & Dyke, 2010; Balanoff *et al*., 2013; Chinsamy *et al*., 2013; Benson *et al*., 2014; Chiappe *et al*., 2014). Furthermore, the methodology developed here for obtaining MR approaches can be used for controlling potential errors in BM predictions of other vertebrate groups.

ACKNOWLEDGEMENTS

REFERENCES

Alexander RM. 1983. Allometry of the leg bones of moas (Dinornithes) and other birds. *Journal of Zoology* **200:** 15–23.

Allen V, Bates K, Li Z, Hutchinson J. 2013. Linking the evolution of body shape and locomotor biomechanics in bird-line Archosaurs. *Nature* **497:** 104–107.

Askew GN, Ellerby DJ. 2007. The mechanical power requirements of avian flight. *Biological Letters* **3:** 445–448.

Atanassov MN, Strauss R. 2002. How much did *Archaeopteryx* and *Quetzalcoatlus* Weigh? Estimation of the mass of fossil organisms by multivariate analysis of bone dimensions. *Journal of Vertebrate Paleontology* **22:** 33A. (abstract).

Baier DB, Gatesy SM, Jenkins FA. 2007. A critical ligamentous mechanism in the evolution of avian flight. *Nature* **445:** 307–310.

Balanoff AM, Bever GS, Rowe TB, Norell MA. 2013. Evolutionary origins of the avian brain. *Nature* **50:** 93–96.

Baumel JJ, King AS, Breazile J, Evans H, Vanden Berge JC. 1993. *Handbook of avian anatomy: Nomina Anatomica Avium*. Cambridge: Massachusetts.

Bell A, Chiappe L. 2011. Statistical approach for inferring ecology of Mesozoic birds. *Journal of Systematic Palaeontology* **9:** 119–133.

Benson RBJ, Campione NE, Carrano MT, Mannion PD, Sullivan C, Upchurch P, Evans DC. 2014. Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. *PLoS Biology* **12:** e1001853.

Biknevicius AR. 1999. Body mass estimation in armoured mammals: cautions and encouragements for the use of parameters from the appendicular skeleton. *Journal of Zoology* **248:** 179–187.

Blob RW, Biewener AA. 2001. Mechanics of limb bone loading during terrestrial locomotion in the green iguana (*Iguana iguana*) and American alligator (*Alligator mississippiensis*). *Journal of Experimental Biology* **204:** 1099–1122.

Bowerman BL, O'Connell RT. 1990. *Linear statistical models: an applied approach*. California: Duxbury Press.

Brockelhurst N, Upchurch P, Mannion P, Brockelhurst OCJ. 2012. The completeness of the fossil record of mesozoic birds: implications for early avian evolution. *PLoS ONE* **7:** e39056.

Bruderer B, Peter D, Boldt A, Liechti F. 2010. Wing-beat characteristics of birds recorded with tracking radar and cine camera. *Ibis* **152:** 272–291.

Butler RJ, Goswami A. 2008. Body size evolution in Mesozoic birds: little evidence for Cope's rule. *Journal of Evolutionary Biology* **21:** 1673–1682.

Calder WA. 1984. *Size, function, and life history*. Cambridge: Harvard University Press.

Campbell KE, Marcus L. 1992. The relationships of hindlimb bone dimensions to body weight in birds. *Natural History Museum of Los Angeles County Science Series* **36:** 395–412.

Campione N, Evans D. 2012. A universal scaling relationship between body mass and proximal limb bone dimensions in quadrupedal terrestrial tetrapods. *BMC Biology* **10:** 60.

Campione NE, Evans DC, Brown CM, Carrano MT. 2014. Body mass estimation in non-avian bipeds using a theoretical conversion to quadruped stylopodial proportions. *Methods in Ecology and Evolution* **9:** 913–923.

Carrano MT. 1998. Locomotion in non-avian dinosaurs: integrating data from hindlimb kinematics, in vivo strains, and bone morphology. *Paleobiology* **24:** 450–469.

Carrano MT. 2001. Implications of limb bone scaling, curvature and eccentricity in mammals and non-avian dinosaurs. *Journal of Zoology* **254:** 41–55.

Cawley GC, Janacek GJ. 2010. On allometric equation s for predicting body mass of dinosaurs. *Journal of Zoology* **280:** 355–361.

Chatterjee S, Templin RJ. 2007. Biplane wing planform and flight performance of the feathered dinosaur *Microraptor gui*. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 1576–1580.

Chatterjee S, Templin RJ, Campbell KE. 2007. The aerodynamics of *Argentavis*, the world's largest flying bird from the Miocene of Argentina. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 12398–12403.

Chiappe LM. 2007. *Glorified dinosaurs: the origin and early evolution of birds*. John Wiley & Sons. Inc., Hoboken, NJ.

Chiappe LM, Shu´An J, Qiang J, Norell MA. 1999. Anatomy and systematics of the Confuciusornithidae (Theropoda, Aves) from the late Mesozoic of northeastern China. *Bulletin of the American Museum of Natural History* **242:** 1–89.

Chiappe LM, Witmer LM. 2002. Mesozoic birds: above the heads of dinosaurs. Berkeley.

Chiappe LM, Zhao B, O'Connor JK, Chunling G, Wang X, Habib M, Marugán-Lobón J, Qingjin M, Cheng X. 2014. A new specimen of the Early Cretaceous bird *Hongshanornis longicresta*: insights into the aerodynamics and diet of a basal ornithuromorph. *PeerJ* **2:** e234.

Chinsamy A, Chiappe LM, Marugán-Lobón J, Chunling G, Fengjiao Z. 2013. Gender identification of the Mesozoic bird Confuciusornis sanctus. *Nature Communications* **4:** 1381. doi:10.1038/ncomms2377

Christiansen P. 2002. Locomotion in terrestrial mammals: the influence of body mass, limb length and bone proportions on speed. *Zoological Journal of the Linnean Society* **136:** 685–714.

Christiansen P, Fariña R. 2004. Mass prediction in theropod dinosaurs. *Historical Biology* **16:** 85–92.

Clarke A, Rothery P, Isaac NJ. 2010. Scaling of basal metabolic rate with body mass and temperature in mammals. *Journal of Animal Ecology* **79:** 610–619.

Clarke J, Norell M. 2002. The morphology and phylogenetic position of *Apsaravis ukhaana* from the late cretaceous of mongolia. *American Museum Novitates* **3387:** 1–46.

Clauset A, Schwab DJ, Redner S. 2009. How many species have mass *M*? *American Naturalist* **173:** 256–263.

Clemente CJ, Withers PC, Thompson G, Lloyd D. 2011. Evolution of limb bone loading and body size in varanid lizards. *Journal of Experimental Biology* **214:** 3013–3020.

Cubo J, Casinos A. 1998. Biomechanical significance of cross-sectional geometry of avian long bones. *European Journal of Morphology* **36:** 19–28.

Cubo J, Ponton F, Laurin M, De Margerie E, Castanet J. 2005. Phylogenetic signal in bone microstructure of sauropsids. *Systematic Biology* **54:** 562–574.

Damuth J, MacFadden BJ. 1990. *Body size in Mammalian Paleobiology: estimation and biological implications*. Cambridge: Cambridge University Press.

Darlington RB. 1990. *Regression and linear models*. New York: McGraw-Hill.

De Esteban-Trivigno S, Köhler M. 2011. New equations for body mass estimation in bovids: testing some procedures when constructing regression functions. *Mammalian Biology* **76:** 755–761.

De Esteban-Trivigno S, Mendoza M, De Renzi M. 2008. Body mass estimation in Xenarthra: a predictive equation suitable for all quadrupedal terrestrial placentals? *Journal of Morphology* **269:** 1276–1293.

De Margerie E, Sanchez S, Cubo J, Castanet J. 2005. Torsional resistance as a principal component of the structural design of long bones: comparative multivariate evidence in birds. *Anatomical Record Part A* **282A:** 49–66.

Dececchi TA, Larsson HCE. 2013. Body and limb size dissociation at the origin of birds: uncoupling allometric constraints across a macroevolutionary transition. *Evolution* **67:** 2741–2752.

Del Hoyo J, Elliott A, Sargatal J. 1992. *Handbook of the birds of the world. Vol. 1–16*. Barcelona: Lynx Ediciones.

Díaz-Uriarte R, Garland T. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. *Systematic Biology* **45:** 27–47.

Egi N. 2001. Body mass estimates in extinct mammals from limb bone dimensions: the case of North American hyaenodontids. *Palaeontology* **44:** 497–528.

Elzanowski A. 2002. Archaeopterygidae (Upper Jurassic of Germany). In: Chiappe LM, Witmer LM, eds. *Mesozoic birds: above the heads of dinosaurs*. Berkeley: The University of California Press, 129–159.

Elzanowski A, Bieńkowska-Wasiluk M, Chodyń R, Bogdanowicz W. 2012. Anatomy of the coracoid and diversity of the Procellariiformes (Aves) in the Oligocene of Europe. *Palaeontology* **55:** 1199–1221.

Farke AA, Alicea J. 2009. Femoral strength and posture in terrestrial birds and non-avian theropods. *The Anatomical Record* **292:** 1406–1411.

Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist* **125:** 1–15.

Field DJ, Lynner C, Brown C, Darroch SAF. 2013. Skeletal correlates for body mass estimation in modern and fossil flying birds. *PLoS ONE* **8:** e82000.

Figueirido B, PéREZ–Claros JA, Hunt RM, Palmqvist P. 2011. Body mass estimation in amphicyonid carnivoran mammals: a multiple regression approach from the skull and skeleton. *Acta Palaeontologica Polonica* **56:** 225–246.

Figueirido B, Serrano-Alarcón FJ, Slater GJ, Palmqvist P. 2010. Shape at the cross-roads: homoplasy and history in the evolution of the carnivoran skull towards herbivory. *Journal of Evolutionary Biology* **23:** 2579–2594.

Gao C, Chiappe L, Zhang F, Pomeroy D, Shen C, Chinsamy A, Walsh M. 2012. A subadult specimen of the Early Cretaceous bird *Sapeornis chaoyangensis* and a taxonomic reassessment of sapeornithids. *Journal of Vertebrate Paleontology* **32:** 1103–1112.

Garcia GJ, Da Silva JK. 2006. Review: interspecific allometry of bone dimensions: a review of the theoretical models. *Physics of Life Reviews* **3:** 188–209.

Gayawan E, Ipinyomi RA. 2009. A comparison of Akaike, Schwarz and R square criteria for model selection using some fertility models. *Australian Journal of Basic and Applied Sciences* **3:** 3524–3530.

Gillooly JF, Brown J, West G, Savage V, Charnov E. 2001. Effects of size and temperature on metabolic rate. *Science* **293:** 2248–2251.

Gillooly JF, Charnov E, West G, Savage V, Brown J. 2002. Effects of size and temperature on developmental time. *Nature* **417:** 70–73.

Gingerich P, Smith B, Rosenberg K. 1982. Allometric scaling in the dentition of primates and prediction of body weight from tooth size in fossils. *American Journal of Physical Anthropology* **58:** 81–100.

Gingerich PD. 1990. Prediction of body mass in mammalian species from long bone lengths and diameters. *Contributions from the Museum of Paleontology University of Michigan* **28:** 79–92.

Grand TI. 1990. The functional anatomy of body mass. In: Damuth J, MacFadden BJ, eds. *Body size in Mammalian Paleobiology: estimation and biological implications*. Cambridge: Cambridge University Press, 39–47.

Habib MB, Ruff C. 2008. The effects of locomotion on the structural characteristics of avian limb bones. *Zoological Journal of the Linnean Society* **153:** 601–624.

Hackett S, Kimball R, Reddy S, Bowie R, Braun E, Braun M, Chojnowski J, Cox W, Han K, Harshman J, Huddleston C, Marks B, Miglia K, Moore W, Sheldon F, Steadman D, Witt C, Yuri T. 2008. A Phylogenomic study of birds reveals their evolutionary history. *Science* **320:** 1763–1767.

Hadi AS, Ling RF. 1998. Some cautionary notes on the use of principal components regression. *The American Statistician* **52:** 15–19.

Hansen TF, Bartoszek K. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology* **61:** 413–425.

Hansen TF, Orzack SH. 2005. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. *Evolution* **59:** 2063–2072.

Harvey PH, Pagel MD. 1991. *The comparative method in evolutionary biology*. Oxford: Oxford University Press.

Henderson DM. 1999. Estimating the masses and centers of mass of extinct animals by 3-D mathematical slicing. *Paleobiology* **25:** 88–106.

Hinic-Frlog S, Motani R. 2010. Relationship between osteology and aquatic locomotion in birds: determining modes of locomotion in extinct Ornithurae. *Journal of Evolutionary Biology* **23:** 372–385.

Hone DW, Dyke GJ, Haden M, Benton MJ. 2008. Body size evolution in Mesozoic birds. *Journal of Evolutionary Biology* **21:** 618–624.

Hou L. 2001. *Mesozoic birds of China*. Taiwan: Phoenix Valley Bird Park, Lugu Hsiang.

Hurvich CM, Tsai CL. 1990. The impact of model selection on inference in linear regression. *American Statistician* **44:** 214–217.

Jenkins FA. 1993. The evolution of the avian shoulder joint. *American Journal of Science* **293:** 253–253.

Jolliffe IT. 1982. Note on the use of principal components in regression. *Applied Statistics* **31:** 300–303.

Kleiber M. 1947. Body size and metabolic rate. *Physiological Reviews* **27:** 511–541.

Kleinbaum DG, Kupper LL, Muller KE. 1997. *Applied regression analysis and other multivariable methods*. Halethorpe: Duxbury Press.

Kvist A, Lindstrom A, Green M, Piersma T, Henk Visser G. 2001. Carrying large fuel load during sustained bird flight is cheaper than expected. *Nature* **413:** 730–731.

Laurin M. 2004. The evolution of body size, Cope's rule and the origin of amniotes. *Systematic Biology* **53:** 594–622.

Liu D, Zhou Z, Zhang Y. 2012. Mass estimate and evolutionary trend in Chinese Mesozoic fossil birds. *Vertebrata Palasiatica* **50:** 39–52.

Longrich N. 2006. Structure and function of hindlimb feathers in *Archaeopteryx lithographica*. *Paleobiology* **32:** 417–431.

Longrich NR, Tokaryk T, Field DJ. 2011. Mass extinction of birds at the Cretaceous–Paleogene (K–Pg) boundary. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 15253–15257.

MacLeod N. 2004. Palaeo-math 101: regression 2. *The Palaeontological Association, Newsletter* **56:** 60–71.

MacNally R. 2000. Regression and model-building in conservation biology, biogeography and ecology: the distinction between –and reconciliation of– 'predictive'and 'explanatory' models. *Biodiversity & Conservation* **9:** 655–671.

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. Available at: http://mesquiteproject.org

Maloiy G, Alexander RM, Jayes A. 1979. Allometry of the legs of running birds. *Journal of Zoology* **187:** 161–167.

Martín-Serra A, Figueirido B, Palmqvist P. 2014. A three-dimensional analysis of morphological evolution and locomotor performance of the carnivoran forelimb. *PLoS ONE* **9:** e85574.

Marzluff JM, Dial KP. 1991. Life history correlates of taxonomic diversity. *Ecology* **72:** 428–439.

Mayr G. 2009. *Paleogene fossil birds*. Berlin: Springer-Verlag.

Mayr G, Clarke J. 2003. The deep divergences of neornithine birds: a phylogenetic analysis of morphological characters. *Cladistics* **19:** 527–553.

McClain CR, Boyer AG. 2009. Biodiversity and body size are linked across metazoans. *Proceedings of the Royal Society of London, B Biological Sciences* **276:** 2209–2215.

McMahon TA, Bonner JT. 1983. *On size and life*. New York: Scientific American Library.

Mendoza M, Janis CM, Palmqvist P. 2006. Estimating the body mass of extinct ungulates: a study on the use of multiple regression. *Journal of Zoology* **270:** 90–101.

Midford PE, Garland T, Maddison WP. 2005. PDAP Package of Mesquite. Version 1.16.

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

Nudds RL. 2007. Wing-bone length allometry in birds. *Journal of Avian Biology* **38:** 515–519.

Nudds RL, Dyke GJ. 2010. Narrow primary feather rachises in confuciusornis and archaeopteryx suggest poor flight ability. *Science* **328:** 887–889.

O'Connor J, Chiappe LM, Bell A. 2011. Pre-modern birds: avian divergences in the mesozoic. In: Dyke G, Kaiser G, eds. *Living Dinosaurs*. Oxford: Wiley-Blackwell, 39–114.

O'Connor J, Gao K, Chiappe LM. 2010. A new Ornithuromorph (Aves: Ornithothoraces) bird from the Jehol Group indicative of higher-level diversity. *Journal of Vertebrate Paleontology* **30:** 311–321.

Olmos M. 1988. Investigaciones sobre la alometría funcional de las extremidades de las aves. PhD dissertation, Universidad de Barcelona.

Olmos M, Casinos A, Cubo J. 1996. Limb allometry in birds. *Annales des Sciences Naturelles-Zoologie et Biologie Animale* **17:** 39–49.

Packard GC, Boardman TJ, Birchard GF. 2009. Allometric equations for predicting body mass of dinosaurs. *Journal of Zoology* **279:** 102–110.

Palmqvist P, Martínez-Navarro B, Pérez-Claros JA, Torregrosa V, Figueirido B, Jiménez-Arenas JM, Espigares P, Ros-Montoya S, De Renzi M. 2011. The giant hyena Pachycrocuta brevirostris: modelling the bone-cracking behavior of an extinct carnivore. *Quaternary International* **243:** 61–79.

Palmqvist P, Mendoza M, Arribas A, Gröcke DR. 2002. Estimating the body mass of Pleistocene canids: discussion of some methodological problems and a new 'taxon free' approach. *Lethaia* **35:** 358–360.

Palmqvist P, Vizcaíno S. 2003. Ecological and reproductive constraints of body size in the gigantic Argentavis magnificens (Aves, Theratornithidae) from the Miocene of Argentina. *Ameghiniana* **40:** 379–385.

Pennycuick CJ. 2001. Speeds and wingbeat frequencies of migrating birds compared with calculated benchmarks. *Journal of Experimental Biology* **204:** 3283–3294.

Pennycuick CJ. 2008. *Modelling the flying bird. AP theoretical ecology series elsevier*. Oxford: Academic Press.

Peters WS, Peters DS. 2009. Life history, sexual dimorphism and 'ornamental' feathers in the Mesozoic bird *Confuciusornis sanctus*. *Biological Letters* **5:** 817–882.

Pike A, Maitland D. 2004. Scaling of bird claws. *Journal of Zoology* **262:** 73–81.

Prange HD, Anderson JF, Rahn H. 1979. Scaling of skeletal mass to body mass in birds and mammals. *American Naturalist* **113:** 103–122.

Price T. 1997. Correlated evolution and independent contrasts. *Philosophical Transactions of the Royal Society of London Ser. B Biological Sciences* **352:** 519–529.

Quinn GP, Keough MJ. 2001. *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press.

Rawlings JO, Pantula SG, Dickey DA. 1998. *Applied regression analysis; a research tool*. New York: Springer -Verlag.

Revell LJ. 2010. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* **1:** 319–329.

Ricklefs RE, Starck JM. 1996. Applications of phylogenetically independent contrasts: a mixed progress report. *Oikos* **77:** 167–172.

Rubin CT, Lanyon LE. 1982. Limb mechanics as a function of speed and gait: a study of functional strains in the radius and tibia of horse and dog. *Journal of Evolutionary Biology* **101:** 187–211.

Sanz JL, Álvarez JC, Soriano C, Hernández-Carrasquilla F, Pérez-Moreno B, Meseguer J. 2002. Wing loading in Primitive Birds. In: Zhou Z, Zhang F, eds. *Proceedings of the 5th Symposium of the Society of Avian Paleontology and Evolution*. Beijing: Science Press, 253–258.

Seebacher F. 2001. A new method to calculate allometric length–mass relationships of dinosaurs. *Journal of Vertebrate Paleontology* **21:** 51–60.

Simmons EL. 2010. Forelimb skeletal morphology and flight mode evolution in pelecaniform birds. *Zoology* **113:** 39–46.

Simmons EL, Hieronymus TL, O'Connor PM. 2011. Cross sectional geometry of the forelimb skeleton and flight mode in pelecaniform birds. *Journal of Morphology* **3:** 958–971.

Smith RJ. 1981. Interpretation of correlations in intraspecific and interspecific allometry. *Growth* **45:** 291–297.

Smith RJ. 1984. Allometric scaling in comparative biology: problems of concept and method. *American Journal of Physiology Regulatory Integrative and Comparative Physiology* **246:** 152–160.

Smith RJ. 1993. Logarithmic transformation bias in allometry. *American Journal of Physiological Anthropology* **90:** 215–228.

Smith RJ. 2002. Lead review: estimation of body mass in paleontology. *Journal of Human Evolution* **43:** 271–287.

Smith RJ. 2009. Use and misuse of the reduced major axis for line-fitting. *American Journal of Physiological Anthropology* **140:** 476–486.

Snowdon P. 1991. A ratio estimator for bias correction in logarithmic regressions. *Canadian Journal of Forest Research* **21:** 720–724.

Sokal RR, Rohlf JF. 1986. *Introduction to biostatistics*. San Francisco: W.H. Freeman and Company.

Stone M, Brooks RJ. 1990. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society Ser. B* **52:** 237–269.

Tabachnick B, Fidell L. 1996. *Using multivariate statistics*. New York: Harper and Rowm.

Tobalske BW, Biewener A. 2008. Contractile properties of the pigeon supracoracoideus during different modes of flight. *Journal of Experimental Biology* **211:** 170–179.

Tobalske BW, Hedrick TL, Dial KP, Biewener A. 2003. Comparative power curves in bird flight. *Nature* **421:** 363–366.

Turner A, O'Regan HJ. 2002. The assessment of size in fossil Felidae. *Estudios Geológicos* **58:** 45–54.

Turner AH, Pol D, Clarke JA, Erickson G, Norell M. 2007. A basal dromaeosaurid and size evolution preceding avian flight. *Science* **317:** 1378–1381.

Vanhooydonck B, James RS, Tallis J, Aerts P, Tadic Z, Tolley KA, Herrel A. 2014. Is the whole more than the sum of its parts? Evolutionary trade-offs between burst and sustained locomotion in lacertid lizards. *Proceedings of the Royal Society B: Biological Sciences* **281:** 20132677.

Videler JJ. 2005. Chapter 9: Comparing the metabolic costs of flight. In: Birkhead TR, ed. *Avian flight*. Oxford Ornithology Series. Oxford: Oxford University Press, 212–232.

Vizcaino S, Fariña R. 1999. On the flight capabilities and distribution of the giant Miocene bird *Argentavis magnificens* (Teratornithidae). *Lethaia* **32:** 271–278.

Wang X, Nudds RL, Palmer C, Dyke GJ. 2012. Size scaling and stiffness of avian primary feathers: implications for the flight of Mesozoic birds. *Journal of Evolutionary Biology* **25:** 547–555.

Warrick DR. 1998. The turning and linear maneuvering performance of birds: the cost of efficiency for coursing insectivores. *Canadian Journal of Zoology* **76:** 1063–1079.

Warton DI, Wright IJ, Falster DS, Westoby M. 2006. Bivariate line-fitting methods for allometry. *Biological Reviews* **81:** 259–291.

Wellnhofer P. 2008. Archaeopteryx*, der urvogel von Solnhofen*. München: Verlag Dr. Friedrich Pfeil.

Yalden D. 1984. What size was Archaeopteryx? *Zoological Journal of the Linnean Society* **82:** 177–188.

Zhang F, Ericson P, Zhou Z. 2004. Description of a new enantiornithine bird from the Early Cretaceous of Hebei, northern China. *Canadian Journal of Earth Sciences* **41:** 1097–1107.

Zhou Z, Zhang F. 2003. *Jeholornis* compared to *Archaeopteryx*, with a new understanding of the earliest avian evolution. *Die Naturwissenschaften* **90:** 220–225.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Figure S1. Cladogram for Neognathae used in this study.

Figure S2. Biplots used in the procedure for selection of different variables from those represented in Figure 4.

Table S1. Data set for living birds used in the analyses.

Table S2. Data set for Mesozoic stem birds used in the analyses.

Table S3. Ordinary least-squares multiple regressions used in the procedure for selection of variables, in which each predictor variable was plotted against a combination of the remaining skeletal predictors (see Material and methods).

Table S4. Differences resulting from taxon weighting for the best OLS multiple regression (ten predictor variables) and for three single regressions, one from a longitudinal measurement (HL) and two from diaphyseal measurements (dUW, dFWcc).

Table S5. Body mass (BM) obtained for Mesozoic birds using three single regressions, two adjusted with the variables that correlate best with BM (dUW and dFWcc) and the third with HL.

Table S6. Ordinary least-squares single regressions obtained from the variables used in this study.