The 'temporal effect' in hominids: reinvestigating the nature of support for a chimp-human clade in bone morphology

Alannah Pearson¹, Colin Groves¹, Andrea Cardini^{2, 3*}

¹School of Archaeology & Anthropology, Australian National University, Canberra, ACT 0200, Australia.

²Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, l.go S. Eufemia 19, 41121 Modena, Italy.

³Centre for Forensic Science, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia.

*Corresponding Author

Andrea Cardini

Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, l.go S. Eufemia 19, 41121 Modena, Italy

tel. 0039 059 2055017

email: alcardini@gmail.com, cardini@unimo.it

Key words: allometry; cranial variation; geometric morphometrics; great apes; phylogenetic signal; Procrustes superimposition; shape analysis; size.

ABSTRACT

In 2004, an analysis by Lockwood and colleagues of hard-tissue morphology, using geometric morphometrics on the temporal bone, succeeded in recovering the correct phylogeny of living hominids without resorting to potentially problematic methods for transforming continuous shape variables into meristic characters. That work has increased hope that by using modern analytical methods and phylogenetically informative anatomical data we might one day be able to accurately infer the relationships of hominins, including the closest extinct relatives of modern humans. In the present study, using 3D virtually generated models of the hominid temporal bone and a larger suite of geometric morphometric and comparative techniques, we have re-examined the evidence for a *Pan-Homo* clade. Despite differences in samples, as well as the type of raw data, the effect of measurement error (and especially landmark digitization by a different operator), but also a broader perspective brought in by our diverse set of approaches, our reanalysis largely supports Lockwood and colleagues' original results. However, by focusing not only mainly on shape (as in the original 2004 analysis) but also on size and 'size-corrected' (non-allometric) shape, we demonstrate that the strong phylogenetic signal in the temporal bone is largely related to similarities in size. Thus, with this study, we are not suggesting the use of a single 'character', such as size, for phylogenetic inference, but we do challenge the common view that shape, with its highly complex and multivariate nature, is necessarily more phylogenetically informative than size and that actually size and size-related shape variation (i.e., allometry) confound phylogenetic inference based on morphology. This perspective may in fact be less

generalizable than generally believed. Thus, while we confirm the original findings by Lockwood et al., we provide a deep reinterpretation of their nature and potential implications for hominin phylogenetics and we show how crucial it is not to overlook size in geometric morphometric analyses.

Introduction

Strongly mineralized fossils, like those belonging to the early radiation of the human lineage, are unlikely to provide DNA data and therefore will not allow estimation of the phylogenetic tree of our ancestors using powerful molecular techniques. Phylogenetic analyses on bones have generally failed to recover the correct topology of modern hominids, which has suggested that cranial data and similar types of morphological evidence may be of little help when applied to the fragmentary and generally incomplete fossil record.

Morphological studies often try to maximize the information available for phylogenetic inference. For instance, it has been suggested that accuracy in morphological phylogenetics might be increased by using data from a complex structure and merging information from both cranial and post-cranial material (Oxnard, 2000). It has also been argued (e.g., Gilbert and Rossie, 2007) that by controlling for 'confounding' factors, such as size-related shape changes (i.e. allometry), our ability to detect the phylogenetic signal in morphology can be increased. However, evolutionary developmental biology (or 'evo-devo') has shown that anatomical structures are often made of modules with a variable degree of integration (Klingenberg, 2013). The way the phenotype responds to the environment may vary across modules and this variable

rate of evolutionary change might contribute to a blurring of the phylogenetic signal in morphological characters. A bone which develops early from a cartilage precursor, such as the temporal, could be less susceptible to changes (in particular non-neutral changes, under the control of natural selection) and therefore a better indicator of phylogenetic relationships. This might be one of the reasons why, measuring shape variation in the temporal bone, Lockwood et al. (2004 – henceforth, abbreviated as LEA) succeeded in recovering the topology of the molecular tree of the living hominids, raising hopes that a similar approach might help to reconstruct the phylogeny of our own extinct ancestors accurately.

Lockwood et al.'s (2004) study has also sparked a debate on the generalizability and nature of their findings (von Cramon-Taubadel, 2014). To our knowledge, however, nobody has yet attempted to replicate their results. Using a novel sample of 3D cranial images of hominids and a mix of geometric morphometric (Adams et al., 2013; Cardini, 2013) and comparative methods (Monteiro, 2013), we therefore repeated LEA's analysis employing the same set of anatomical landmarks in order to:

1) Test the significance of the phylogenetic signal in the temporal bone using a recent molecular tree as an independent estimate of the 'correct phylogeny' of the hominids. To this end, three different approaches were used: i) the commonly used but heavily criticized Mantel test, originally employed by LEA; ii) the widely used Klingenberg and Gidazewski's (2010) squared-change parsimony test; iii) Adams' (2014a) multivariate K, which represents the most recent development in this type of analyses and is currently the most accurate index for testing a phylogenetic signal. Only the first of these tests was available at the time of LEA's study and none has been used until now to explore numerically the phylogenetic signal in the morphology of the hominid temporal bone. This first step does not imply phylogenetic inference using

morphology and, even if a signal is found, does not guarantee that the correct topology will be found (Klingenberg and Gidazewski, 2010).

2) Infer phylogenetic trees using size and shape data from the temporal bone. To this end, we followed one of the two distance-based approaches employed by LEA. In fact, current phylogenetic approaches applied to Procrustes data are all potentially problematic, as discussed by Adams et al. (2011, 2013) and Adams (2014a). Results are therefore to be interpreted with great caution and are better seen as exploratory. Bearing this caveat in mind, we investigated how well the data supported the monophyla in the morphological trees. This step was also taken by LEA, who resampled Procrustes shape variables; bootstrapping these types of data, however, is unlikely to be accurate (Caumul and Polly, 2006; Cardini and Elton, 2008a) and we thus reassessed the support for the branches of the morphological trees by using a method developed within the Procrustes framework to estimate accurately node repeatability in relation to sampling error.

3) Numerically measure the congruence among trees based on the morphology of the temporal bone and the molecular tree, and estimate whether differences are smaller than expected by chance. This step is unnecessary if one retrieves the exact same tree topology – as in LEA – but becomes interesting if differences are found.

The ‘replication culture’ (Ioannidis, 2014), which is popular and has proved its usefulness in several fields of biomedicine, is less common in basic science. With our series of analyses we aimed to replicate LEA's study but also wanted to complement and reassess their findings. Indeed, we followed LEA's protocol but expanded it with additional tests and went a step further adding a major element of novelty. This consisted of also considering size and 'size-corrected'

shape variation in all phylogenetic analyses, while LEA focused mainly on shape and tested only the phylogenetic signal in this component of form variation. Using shape alone is not uncommon in geometric morphometrics, and implicitly acknowledges the more reliable information putatively provided by shape as being multivariate in nature, complex and presumably influenced by many genes. In contrast, size and size-related shape variation, or allometry, are considered more evolutionarily labile, strongly affected by selective pressures and therefore less likely to be phylogenetically informative. Indeed, as mentioned, these components of variation in form are often seen as a source of noise in phylogenetic inference (the “potentially confounding effects of allometry on quantitative craniodental characters”, in Gilbert and Rossie's, 2007: 11910, terminology). Strong evidence for this claim may, in fact, be less than generally thought, and we suggest that morphological investigation should not neglect size and its potential covariation with shape, as they provide information complementary to shape data and therefore help to understand better how different aspects of form may affect evolutionary variation. For these reasons, we performed all temporal bone analyses using size, shape and also 'size-corrected' shape, i.e., shape in which allometry is statistically controlled for.

Material and Methods

Study samples, data acquisition and molecular tree

The sample comprised 163 adult specimens (77 females and 86 males; see Table 1 for details – species/subspecies, composition according to sex and sample size, as well as information on temporal bone size). Data were obtained using computed tomography (CT) on

samples from collections at The Smithsonian Natural History Museum, the Royal Museum for Central Africa, and the Leipzig Anatomy Institute. Three specimens from the Australian National University were laser scanned by one of us (AP).

Computed tomography (CT) is now a widespread technique in anthropology. CT uses a planar beam of parallel X-rays sent through the object of study and received by a one-dimensional detector array on the opposite side to the beams. As the source detector rotates, it obtains projections from different angles. Following image acquisition, CT slices are inversely reconstructed to form an image stack and generated into a 3D virtual model (Zollikofer and Ponce de Leon, 2005). Conversely, a 3D surface scanner captures light reflected from an object by placing an independent light source of known values onto the object from which depth cues are returned from the reflected light signal. Devices such as the NextEngine scanner, used in this study, involve a set of laser beams scanning along the virtual grid-lines spread over the object.

While Sholts et al. (2011) found problems with the precision of surface models generated from the NextEngine Scanner, one of us (AP) devised a method to offset issues in image reproduction caused by ‘holes’ in the mesh. By first conducting an HD 360 degree rotational scan with eight iterations, followed by a single scan of each surface (such as anterior and posterior regions of the skull), the generated mesh could be reconstructed easily and a customized Poisson reconstruction conducted to fill any minor holes.

CT image segmentation and thresholding was followed by 3D mesh generation in Mimics 17 (Materialise) while surface scans were acquired in ScanStudio HD Pro (NextEngine) before the resulting 3D mesh was imported into Checkpoint (Stratovan) and the 22 LEA landmarks digitized. The landmark configuration is shown in Fig. 1 and described in Table 2, and matches the original configuration published by LEA. Almost all specimens were digitized on the left

temporal bone, preservation permitting; only four were digitized on the right temporal bone. The surface scans and those specimens landmarked on the right temporal bone did not fall outside the species groups, giving confidence that the scanning methods are comparable and that landmark error was modest.

The molecular phylogeny was obtained from the 10KTrees webpage (Arnold et al., 2010; <http://10kTrees.fas.harvard.edu> – June 2014 version) using a consensus time tree of 1000 trees computed using a Bayesian approach and both nuclear and mitochondrial markers (details are provided at <http://10kTrees.fas.harvard.edu> and in Arnold et al., 2010). All 1000 trees provided completely congruent topologies (i.e., identical monophyletic groups). As *Gorilla gorilla diehli* Matschie, 1904 was not included in the 10KTrees phylogeny, we added it as sister to *Gorilla gorilla gorilla* with an estimated divergence of 0.018 millions of years based on results by Thalmann et al. (2011). For brevity, we will henceforth refer to this phylogeny as the 10KT tree.

Most analyses tend to be performed at species level (i.e., pooling subspecies samples). We refer to these as analyses of the 7SP dataset, as seven species are included. However, as LEA also considered putative subspecies, we replicated tests of all main hypotheses also taking subspecies into account. We refer to these analyses as 11SSP, as 11 taxa are analyzed when subspecies are not pooled. In this respect, compared to LEA, there is a small difference, as they included only two subspecies of gorilla, whereas we included all four living subspecies.

Prior to preliminary analysis (see Supplementary Online Material [SOM]), subspecies status was reassessed by one of us (CG) against precise geographic localities from museum records. The two subspecies of Bornean Orangutan (Groves 2001), *Pongo pygmaeus wurmbii* and *Pongo pygmaeus pygmaeus*, were pooled to increase sample size. Finally, using ordinations

and a phenogram (see below), we assessed that there were no evident outliers within each species and subspecies.

Geometric morphometrics

Analyses were performed using Procrustes-based geometric morphometrics (GMM – Adams et al., 2013; Cardini, 2013). Size and shape were obtained from 3D Cartesian coordinates of anatomical landmarks using a Procrustes superimposition (Rohlf and Slice, 1990). For brevity, we refer to centroid size, the conventional measure of the dispersion of landmarks around their baricenter in Procrustes GMM, simply as size, and we similarly use shape to refer either to Procrustes shape coordinates or their reformulation in terms of principal components (PCs – i.e., linear combinations of shape coordinates which are uncorrelated and maximize total variance). We will also use the term 'size-corrected' shape, or non-allometric shape, to refer to shape data from which the covariation with size has been statistically removed. This is done to control for the potential effect of evolutionary allometry (i.e. shape variation correlated to size differences among species and subspecies). Thus, to 'size-correct' shape, we used a common statistical approach, which is based on the multivariate regression of species/subspecies mean shapes on their mean sizes to obtain 'size-corrected' shape residuals (Klingenberg and Marugan-Lobos, 2013; Mitteroecker et al., 2013).

Zelditch et al. (2004, 2012), Cardini (2013) and Adams et al. (2004, 2013) provide, respectively, extensive, medium-length and brief introductions to Procrustes based GMM and the associated toolkit of statistical methods. More information on recent advances in GMM can be found in the 'Yellow Book' (Cardini and Loy, 2013). Software used for these and other statistical analyses is detailed, with references, at the end of the methods section. To avoid repetition,

references are provided only in that section.

Evolutionary allometry

Preliminary to testing the main study hypotheses, we explored whether shape co-varies significantly with size across species of living hominids (evolutionary allometry). This helps to understand whether statistically controlling for allometry ('size-correction', following Klingenberg's [2011] terminology) might have a more or less profound effect on the results of shape analysis. To this end, we followed a regression approach as in Klingenberg and Marugan-Lobos (2013). This is different to allometric studies in traditional morphometrics, using distance measures (Marcus, 1990). As explained by Mitteroecker et al. (2013: 61): “Geometric morphometric studies of allometry require a methodological approach that differs in several aspects from the classic allometry studies. Procrustes shape coordinates are shape variables, not size variables, and they have no natural zero point (they are on interval scales, not ratio scales) and hence cannot be log transformed. Allometry is thus expressed as a – usually linear – function of the Procrustes shape coordinates, estimated by multivariate regression of the shape coordinates on centroid size or the logarithm of centroid size”.

Thus, one should bear in mind two important considerations in the interpretation of our study. The first one is that, in Procrustes-based geometric morphometrics, taxa whose morphologies might represent isometrically scaled versions of the same geometric shape will overlap in the shape space (i.e., their Procrustes shape distance will be zero, or virtually zero). Thus, isometry would be detected by simply analysing shape, which in our dataset is always statistically significantly different among taxa (for species, see SOM; subspecies with samples larger than five were significant at an alpha level < 0.005 in all comparisons except *P*.

trogodytes schweinfurthii versus *P. troglodytes troglodytes*, which was at $p < 0.05$). The second consideration is that as our aim is to control statistically for the effect of size on shape using regression residuals; allometry is therefore, as in all inferential statistics, a property of the specific sample being analyzed. Thus, a common criticism of the use of multivariate regression residuals in morphometrics (i.e., that they fail to identify individuals of the same shape – Jungers et al., 1995) does not strictly apply to Procrustes shape data, as identical shapes would be identified prior to the regression. Similarly, although Procrustes shape is a property intrinsic of the object being analyzed, non-allometric shape cannot be if the aim is to hold statistically constant the effect of a covariate, as in our study. Indeed, as Smith (2005: 249) pointed out in the context of the controversy on the different use of residuals and shape ratios, “the issue is resolved by recognizing the difference between statistical control, in which focus is on a single trait for which a second trait is defined as a confounding effect, and proportionality, which focuses on the relationship between two traits—the relationship that statistical control serves to remove. The calculation of ratios is not effective for statistical control but is essential for the evaluation of proportionality”.

Thus, we performed a series of multivariate regressions of mean shapes onto mean sizes using separate sexes and either 7SP or 11SSP. Significance was assessed using permutations and analyses performed using both the standard least square regression, which assumes independent observations, and comparative methods, which take non-independence due to phylogeny into account (Monteiro, 2013). Comparative analyses, which for brevity we refer to in this context (i.e., allometry) as phylogenetic regressions, were performed using the 10KT phylogenetic tree with branch lengths proportional to estimated evolutionary times, but also with all branch lengths set to unit lengths. These correspond respectively to weighted and unweighted regressions using

phylogenetic independent contrasts (PIC) of shape and size in MorphoJ (Klingenberg, 2011; Klingenberg and Marugan-Lobos, 2013, and references therein). The test statistic is, as in the standard least square regression in MorphoJ, the percentage of shape variance explained by size (multivariate r square – R^2) compared to the distribution of pseudo- R^2 obtained by randomizing the order of the observations. Evolutionary allometric regressions were also replicated in R using a permutational approach based on Euclidean distances (non-comparative analysis in the package Vegan) and the recently developed (Adams, 2014b; Adams and Collyer, 2015) distance-based phylogenetic generalized least square (comparative method in the package Geomorph).

For the size-correction, as anticipated, we used the residuals from the standard least square regression (Klingenberg, 2011; Klingenberg and Marugan-Lobos, 2013). This was done to avoid testing the phylogenetic signal, and reconstructing and comparing trees, using morphological data whose computation was partly based on information in the 10KT phylogeny.

Main hypotheses

Phylogenetic signal

As for all main hypotheses, analyses of the phylogenetic signal were performed separately in females and males and both in the 7SP and 11SSP datasets. How well or poorly morphological data mirror the pattern of evolutionary transitions suggested by the molecular phylogenetic 10KT tree was assessed using all the three main methods available for both uni- and multivariate data.

We used the Mantel test to measure the correlation between matrices of distances based on morphological and molecular data. For morphology, pairwise Euclidean distances were computed using size, shape and 'size-corrected' shape. For shape, Euclidean distances are

virtually identical to Procrustes shape distances, as the tangent space approximation is almost perfect (see above). Morphological distances were not squared to preserve the similarity relationships observed in the size or shape spaces of the data, although squaring distances, as sometimes considered appropriate for morphological data (e.g., Polly, 2008), did not appreciably change the outcomes of Mantel tests (results not shown). For DNA data, patristic distances were computed from the 10KT phylogeny. The significance of the correlation was assessed using 10000 permutations. This type of use of the Mantel test has been shown to inflate type I error rates and have low statistical power, and it is therefore heavily criticized by several authors (as reviewed by Adams, 2014a). The test was performed only to replicate LEA, and to compare its results with other approaches and other studies using this method.

The second test we used was that developed by Klingenberg and Gidazewski (2010). This is based on squared change parsimony to map uni- and multivariate morphological data onto an independent tree (the 10KT molecular phylogeny, in our case) and estimate how well they fit the tree by computing the total amount of squared change summed over all branches of the tree. This estimate assumes a Brownian motion model of evolutionary change (i.e., increasing variance among taxa in direct proportion to evolutionary time measured by the phylogenetic branch lengths) and is compared to the distribution of sum of square changes obtained by permuting (10000 times) taxa across the tips of the tree. If morphology contains information on phylogeny, the sum of squared change in the data should be significantly smaller than the sum of squared changes from the set of ‘randomized trees’. This approach too has been recently criticized (Adams, 2014a), as simulations suggested that it might lead to unreliable results and problematic interpretations. However, together with the Mantel test, it has been the only test available for multivariate shape data until recently and it is still commonly employed in geometric

morphometrics. For both this test and the PIC regression using permutations in MorphoJ (see above), we cannot forecast whether future theoretical studies may or may not support the recent criticisms, so we decided to keep these tests in the study to present as full a set of results as possible and, where they do not agree, are left open to potential reassessments.

The third test we employed was a multivariate extension (K_{mult}) of Blomberg's K (Blomberg et al., 2003; Revell et al., 2008) recently proposed by Adams (2014a). As in ii), Blomberg's K also takes advantage of a Brownian motion model of evolutionary change and squared differences to estimate the phylogenetic signal in relation to an independent phylogeny (such as our 10KT tree), but uses the ratio between the observed differences relative to the last common ancestor (estimated from the data) and the same quantity computed after correcting from phylogenetic covariance. This ratio (numerator) is in turn divided by another ratio (denominator) obtained with the same computations now applied to data predicted according to a Brownian motion model of evolutionary change. This second division is equivalent to a 'standardization' to express the signal in relation to the expectation under Brownian motion. Thus, K increases with the strength of the phylogenetic signal, as the numerator ratio becomes larger when most of the variance depends on phylogenetic relationships; it is equal to 1 if change accumulates strictly following Brownian motion; it is less or more than 1 if phylogenetic covariation is respectively less or more than under Brownian motion. However, it should be borne in mind that even when $K < 1$ results may be significant, because there can be less signal than expected following Brownian motion but nevertheless still more than expected by chance. Adams (2014a) has shown using simulations that K_{mult} outperforms not only the Mantel test but also Klingenberg & Gidazewski's (2010) squared change parsimony in terms of accuracy and power. This statistic, which is tested using the same randomization approach as in the

Klingenberg and Gidazewski (2010) test, should therefore represent the most reliable estimate of the phylogenetic signal.

Trees based on size and shape

Adams et al. (2011, 2013) briefly reviewed different approaches to phylogenetic inference using Procrustes shape data, discussing their limits and potential pitfalls. Following one of the two methods of LEA, we used neighbor joining (NJ), a distance-based method using a minimum evolution criterion (Felsenstein, 2004). Neighbor joining was applied on pairwise Euclidean distances between taxa in the 7SP and 11SSP datasets, i.e., the same matrices used in the Mantel test. Felsenstein discussed NJ, its advantages (e.g., computational speed) but also the limits of this and other methods using distances, which rely on how good the metric used to compute those distances is as a model for evolutionary change. Because Euclidean distances from Procrustes size and shape are not based on any evolutionary model and simply measure similarity, NJ applied to these data is phenetic in nature but has the advantage of avoiding the constraint of ultrametricity (branches having the same total lengths from the tips to the root of the tree, as in cluster analyses) and also allows one to root the tree. For the sake of clarity, we would like to stress that the NJ analyses mainly serve the purpose of replicating all main steps of LEA and that, for this reason as well as for analytical consistency in methods and results across all types of data, we also estimated trees using size only. However, this does not imply that we advocate phylogenetic inference based on a single (although potentially polygenic – [Hemani et al., 2013]) trait such as size.

To root the tree, we used as an outgroup one of the two species of orangutans, which are known to be basal to other hominids in the radiation of this lineage. We chose *P. pygmaeus*

simply because mean estimates of size and shape are likely to be more accurate in this species than in *P. abelii*, whose sample size was about ¼ of that of *P. pygmaeus*. Also, NJ programs do not permit the use of multiple species as an outgroup; although it is not impossible that one could write a software for forcing two outgroups to cluster in a NJ tree, one would lose a 'degree of freedom' that can say something about the strength of the phylogenetic signal. By not forcing *P. pygmaeus* to be sister to *P. abelii* with the whole genus basal to all others, it is possible to check whether *P. abelii* is in fact next to *P. pygmaeus* in morphological trees. Thus, an apparent paraphyletic position of *Pongo* in our NJ trees is actually indicative of a correct correspondence between morphological and genetic information, as becomes obvious if one simply re-roots a tree between *Pongo* and the other hominids (an operation which does not change the length of the tree). In contrast, if one (or more) of the other hominids is interposed between *P. abelii* and *P. pygmaeus* in a NJ tree, this indicates a failure of morphology to recover the correct phylogenetic relationship of the two orangutans.

Support for nodes in a morphological tree is generally assessed by bootstrapping characters (Felsenstein, 2004, and references therein). This was also the approach taken by LEA on shape variables and more recently by Gilbert (2011). Procrustes shape data (coordinates or their PCs) are, however, not independent, as they are a function of both the choice of the superimposition, which is based on statistical convenience (Cardini, 2013, and references therein), and the set of taxa included in the analysis (Caumul and Polly, 2006; Adams et al., 2011). Thus, LEA's approach was problematic.

Sampling error (i.e., what individuals were available for the study in each taxon) is another source of uncertainty in trees reconstructed using continuous morphological characters. In fact, especially in small samples, the variation among individuals used to estimate species or

subspecies averages, when used to infer trees, can influence topology even more profoundly than character sampling (Cardini and Elton, 2008a). Thus, for these reasons, and because we aimed at replicating LEA's analysis using exactly the same landmark configuration and, at the same time, providing estimates of node repeatability that do not violate the assumptions of the Procrustes paradigm (Adams et al., 2013), we estimated NJ tree branch support based on individual sampling. This was achieved following Cole et al. (2002), as in Caumul and Polly (2006) and Cardini and Elton (2008a): first we bootstrapped individuals in each species or subspecies sample (7SP/11SSP datasets), then we recomputed mean sizes or shapes in the resulting pseudo-samples, and used these pseudo-means to reconstruct NJ trees using Euclidean distances; finally, we repeated these operations 1000 times and summarized the resulting 1000 trees using a 50% majority rule consensus tree.

Tree congruence

The last aim of the study was to measure numerically the differences between molecular and morphological trees, and estimate their significance. Testing differences among trees without reference to the underlying data has a strong limitation, in that it ignores the structure of the data which generated those trees (Felsenstein, 2004); once this is acknowledged, however, it does provide some useful information. Indeed, unless tree topology is identical, deciding by eye to what extent differences are negligible or substantial becomes hard and potentially arbitrary.

As in Adams et al. (2011), to estimate differences we used a series of measures of the congruence between the topology of the molecular tree and those of each morphological tree (size and shape of the temporal bone in females and males of the 7SP and 11SSP datasets). As the aim here was to assess whether trees showed similar or different groupings of

species/subspecies, we focused only on tree topology (unrooted trees with no information on branch lengths – note that rooting all trees, including the molecular one to match the 'constraint' of NJ, using *P. pygmaeus* as the outgroup does not appreciably change the outcomes – results not shown). The following three types of dissimilarity measures were used following the definitions by Page (1993): a) the partition measure, which counts each cluster that is found in only one or other of the trees.; b) the number of leaves to prune to get greatest agreement subtree; c) the nearest neighbor interchange metric, which measures the minimum number of nearest neighbor interchanges required to change one tree into another.

After measuring tree differences, we assessed whether they were smaller than expected by chance. This was done by comparing the observed score with the distribution of the same type of dissimilarity measure computed when the 10KT tree is compared either to all possible alternative topologies for N taxa (7SP) or, when this number was too high (11SSP), to a subsample of 10000 random trees from all possible trees. If, relative to these trees, the observed measure was smaller than 95% of them, the morphological and molecular trees were considered more similar than expected by chance (i.e., significantly similar). For instance, for female shape, the partition score was zero for 7SP and six for 11SSP. With seven taxa, there are 945 possible unrooted trees and only one has a zero partition metric compared to the 10KT tree, which corresponds to $p = 1/945 = 0.00106$). With 11 taxa, the number of all possible trees is too large to consider all of them, but in a subsample of 10000 randomly selected trees, only 68 had a partition metric equal to or smaller than six, which means $p = 68/10000 = 0.0068$.

Software

Geometric morphometrics and statistical analyses were performed using the following free

programs, the majority of which can be found following the links at <http://life.bio.sunysb.edu/morph/toc-software.html>: TPSSmall 1.28 (Rohlf, 2015 – superimposition and tangent space approximation); MorphoJ 1.06b (Klingenberg, 2011 – preliminary analyses including PCAs, permutation tests of group differences, phylogenetic regression and 'size-correction', squared-change parsimony test of the phylogenetic signal); PAST 2.17c (Hammer et al., 2001 – preliminary analyses, summary statistics for size, neighbor joining analyses); NTSYSpc 2.2r (Rohlf, 2013 – Mantel tests; bootstrapped cluster and neighbor joining analyses); R-project (R Core Team, 2014) and in particular the package Geomorph (Adams et al., 2013 – phylogenetic regressions and tests for the phylogenetic signal using multivariate K) and Vegan (Oksanen et al., 2013 – regressions and tests of species differences using permutations); Mesquite 2.75 (Maddison & Maddison, 2011; tree manipulation and drawing, consensus trees) and Component (Page, 1993; measures of tree similarity and tests of their differences).

Results

Evolutionary allometry

Results of the regressions are in Table 3. Half of the analyses are at least marginally significant ($p < 0.1$), little less than half are significant ($p < 0.05$), and only ca. 15% are highly significant ($p < 0.0021$, with a Bonferroni correction). R^2 , the amount of variance in shape explained by size, ranged from ca. 10% to almost 56%. Results varied remarkably depending on whether analyses were using standard or comparative methods and also depended on whether branch lengths were (weighted) or were not (unweighted) considered. MorphoJ PICs and the Geomorph R-package

phylogenetic regression based on distances produced identical results in terms of R^2 , as expected with methods which are computationally different but equivalent, although when p values are examined, only standard regressions (i.e., those not using comparative methods) are in perfect agreement. Comparative analyses, in contrast, tend to differ considerably, with MorphoJ generally showing higher significance for larger R^2 and Geomorph R-package showing the opposite. Thus, overall, the analysis of evolutionary allometry produces somewhat contrasting evidence about how important the covariation of shape with size is in living hominids. However, as allometry explains on average about $\frac{1}{4}$ of shape variance, there seem to be good reasons to test all main hypotheses using size-corrected shape also.

Phylogenetic signal

Results of the tests for the phylogenetic signal of the different components of morphological variation in the temporal bone morphology are summarized in Tables 4-5. We mainly focus on K_{mult} , as it is likely to be the most reliable statistic. K_{mult} is at least marginally significant in most tests of size, shape and size-corrected shape. However, when multiple tests are taken into account, high significance is achieved in 62% of tests of shape, only 12% of those of size, and none of those of size-corrected shape. The results of tests using squared-change parsimony are generally in good agreement with K_{mult} but suggest slightly less evidence of a signal in size-corrected shape (marginally significant or significant in about only half of the tests and never highly significant) and stronger evidence of a signal in size (significant in all tests and highly significant in 50% of them). Finally, matrix correlation is, as predicted because of the expected low power in Mantel tests, marginally significant or significant in only about half of the tests of size, highly significant in 12% of them, but never significant for both shape and size-corrected

shape. When significance is reached in Mantel tests, it is relatively higher using unit branch lengths in the 10KT. Overall, although generally slightly larger in shape, size and shape seem to retain a comparably strong phylogenetic signal, whereas size-corrected shape has a modest or negligible (if Bonferroni corrected) signal.

Trees based on size and shape

Neighbor joining trees for size, shape and size-corrected shapes are shown in Figures 2 and 3 together with the corresponding 7SP and 11SSP 10KT trees. We summarize below the main findings from the NJ trees without reference to sex unless results for females and males differ substantially.

In the 7SP study, size produces trees which misplace *P. abelii* but correctly recover all other clades including a well-supported sister relationship between *Homo* and *Pan*. Shape trees correctly reproduce the topology of the molecular tree with strongly supported nodes. The congruence of DNA and shape trees is immediately evident if one mentally re-roots the tree between the two orangutans and the other hominids. We remind readers that this was not done a priori not only because NJ requires a single species as an outgroup, but also to allow the algorithm to place *P. abelii* according to the morphological data instead of forcing it together with *P. pygmaeus*. Using this expedient, we could show, for instance, that size does not correctly infer the monophyly of *Pongo*, as *P. abelii* does not branch out next to *P. pygmaeus* and rerooting between a monophyletic *Pongo* and the other genera is therefore impossible. Finally, size-corrected shape recovers the basal position of *Pongo* but produces clades with an incorrect mix of species of different genera: a striking example is humans, which are sister of one of the two gorillas. In conclusion, in the 7SP analyses, both size and shape allow us to correctly infer

most of the molecular clades; in contrast, after the effect of size on shape is statistically removed, the phylogenetic signal is largely lost and the corresponding morphological trees are a poor proxy for the molecular phylogeny.

When subspecific differences are taken into account (11SSP), size again misplaces *P. abelii*, but correctly groups gorillas and also retrieves a well-supported human-*Pan* clade, although in this clade *P. troglodytes* is not monophyletic. Shape correctly places *P. abelii* basally and correctly infers the monophyly of *Gorilla*, but, in females, humans are well supported as sister of *Pan* plus *Gorilla*, thus misplacing gorillas, and *P. troglodytes* is not monophyletic. In males, shape correctly places *P. abelii* basally, followed by *Gorilla* which is sister of humans plus *Pan*, with bonobos sister to chimpanzees. These nodes are well supported and the topology of the tree is very similar to that of the 10KT tree, although *Pan troglodytes verus* is incorrectly placed next to gorillas, and neither of the two species of gorillas is recognized as monophyletic. Finally, as in the 7SP analysis, the phylogenetic signal is largely lost after controlling for allometry, with trees where polytypic genera are not monophyletic and *Homo* is sister to a species of gorilla. Thus, although the phylogenetic signal is slightly degraded when subspecies are taken into consideration, results from the 11SSP analyses largely mirror those of the 7SP dataset, with size and, especially, male shape recovering at least partly the main molecular clades. As in the 7SP analyses, controlling for size-related shape variation seems to mostly remove the phylogenetic information in the shape data.

Tree congruence

When the congruence between the molecular tree and the morphological ones is measured using the three different metrics for tree dissimilarity which we introduced in the methods, numerical

results support our visual comparison of trees. We remind readers that significance in the tests for tree congruence means that the topological congruence of molecular and morphological trees is better than expected by chance. Thus, size is either marginally significant ($p < 0.1$) or significant ($p < 0.05$) in at least half of the tests, as summarized in Table 6, but never highly significant ($p < 0.004$ after a Bonferroni correction). Shape in contrast is almost always significant and in more than half of the tests is also highly significant. Size-corrected shape, in contrast, never produces trees which are significantly more congruent with the DNA phylogeny than expected by chance.

Discussion

With the development of full-genome sequencing, the taxonomy of the great apes has received further support and clarifications (Wall, 2013). Yet, morphological-based analyses have often failed to produce phylogenies congruent with molecular trees. This is especially true for analyses using cranial measurements, although some (e.g., Begun, 1992, Strait and Grine, 2004, who both included fossils), but not all (e.g., Collard and Wood, 2000, who did not analyze fossils), cladistic analyses using meristic coding did manage to find support for a hominin-chimpanzee clade. Overall, however, the support LEA found for the *Pan-Homo* clade was interesting and unexpected, especially because they did not include fossils, which might help to bridge the huge 'gap' between chimpanzees and modern humans (Strait and Grine, 2004, and references therein). Also, they adopted a largely phenetic approach based on overall similarity to make direct use of the most accurate type of morphometric data currently available for the morphological study of continuous characters (i.e., Procrustes-based geometric morphometrics – Adams et al., 2013). Following up from that work, we thus addressed both species-level (7SP) and subspecies support

(11SSP) for a phylogenetic signal in the hominid temporal bone using a variety of approaches and both size and shape components of morphological variation. As LEA, we found that this bone contains interesting information for inferring phylogeny but we came to a radically new interpretation of what makes it phylogenetically informative.

LEA did not test the phylogenetic signal in size, as we did, but they assessed whether the topology of their trees based on shape may have been influenced by size because of allometry. To this end, instead of using the more common regression model that we used, they performed a series of Mantel tests between size and shape distances. They found significance only after excluding humans and concluded (Lockwood et al. 2004: 4359) that “size contributes significantly to shape variation among great apes. However, it [i.e., their test of allometry] demonstrates that the *Pan–Homo* affinity is not due to size”.

We used a more conventional regression approach to test allometry and performed the analyses using both standard and comparative methods (which LEA did not employ). In this context, allometry, estimated using a regression, provides a statistical description of the covariation between size and shape but does not imply that size variation is the mechanism that causes shape differences. Thus, when we use expressions such as “variance explained by size” or “allometry ... effect on ... shape”, we simply refer to the description of a pattern without any necessary direct link to the underlying processes. Within this statistical framework aimed at testing the association between size and shape, allometry was significant in two thirds of the regressions using subspecies but in only one sixth of those using species. The analysis of species likely suffered from low power, as there are only seven observations. This and the large percentage of variance explained by size (i.e., allometric shape) do suggest that allometry has an important effect on temporal bone shape. Consistent with this, we found: (1) a significant

phylogenetic signal not only in shape but also in size; (2) trees based on size more similar to the molecular 10KT tree than expected by chance; and (3), after controlling for allometry, an almost complete loss of phylogenetic signal. Therefore, our study supports LEA's findings that the temporal bone contains important phylogenetic information. Indeed, despite differences in samples, taxonomic compositions, type of raw data and inter-operator variability in landmarking, the correlations between our data (either size or shape) and those of LEA are about 0.9 or larger; nonetheless, despite this high congruence among data, contrary to their conclusion, results from the whole battery of our analyses consistently and strongly indicate that the phylogenetic signal is largely related to size and allometry.

The presence of evolutionary allometry in the skeletal anatomy of the great apes is not unprecedented (Shea, 1983, 1988) nor unexpected, but it should always be thoroughly examined. This is particularly the case here, as there are many fewer studies of the temporal than of the whole cranium. For instance, studies in primates generally show that non-allometric shape has about the same phylogenetic signal as total shape, if not even more (Cardini and Elton, 2008a, b). This, however, is not what our temporal bone study showed, which leads to the next question: why is this so?

It is difficult to find an answer, as analyses such as ours and LEA's are largely correlational and are therefore useful in describing and assessing patterns but do not allow one to directly infer processes. The discussion which follows is thus largely speculative, but we hope it will help to stimulate further studies on the origin and nature of the phylogenetic signal in hominid morphology.

The phylogenetic signal in the temporal bone is, we suggest, mostly a by-product of a

highly derived condition in *Gorilla* compared to a relatively plesiomorphic one in the *Pan-Homo* branch. This implies that evolution modelled cranial regions at a different pace, and size variation was the main driver of change specifically in this anatomical region. This modifies the perspective by making the temporal bone a fairly good phylogenetic indicator neither because of its complexity (LEA) nor because of its low plasticity (von Cramon-Taubadel, 2014, and references therein). High complexity and low plasticity could make a structure less sensitive to the environment and to strong directional change, thus reducing the effect of factors that may mask the phylogenetic signal. Contrary to this general view, we suspect that the main reason why the temporal bone may be a good proxy for hominid phylogeny is in fact because its shape was largely modelled by strong selection promoting larger size in gorillas, while humans and *Pan* retained a less derived and smaller size. Thus, autapomorphic size increases and the consequent allometry pushes *Gorilla* far from *Pan* and *Homo*, which become relatively close, using overall similarity measures such as size and shape distances, because they share similarly smaller size and therefore are less susceptible to allometric differences. As the phylogenetic signal, however, is slightly greater in shape than in size, there is likely also a potential component of synapomorphic variation in shape, but this is small and not enough to preserve the signal once allometric shape is regressed out. In this respect, orangutans, which are moderately larger than *Pan* and humans in terms of temporal bones, are an exception. *Pongo abelii* clusters close to the outgroup, *P. pygmaeus*, as expected according to phylogeny and probably partly because of size similarities, but in this case also, and possibly mainly because, of shared synapomorphies in shape. This last observation is inferred by the insensitivity of this finding to the 'size-correction' with the relative position of the two species of orangutans unchanged regardless of whether allometry is controlled for.

If the effect of evolutionary allometry is pervasive, however, one might wonder why similarities in size between humans and *Pan* did not make their overall cranial morphology alike. Clearly, other cranial regions did not respond to size in the same way as the temporal bone. Thus, modifications in shape were to some extent uncoupled from changes in size, and humans could evolve a highly autapomorphic cranial shape with a very short face and a massive braincase despite the predictions of evolutionary allometry in closely related species (Cardini and Polly, 2013; Cardini et al., 2015). As these rapidly evolving derived features broadly dominate cranial morphology, it is not surprising that studies of crania as a whole, unlike those focusing just on the temporal region, tend to single out humans and put together prognathic and comparatively small brained genera such as *Pan* and *Gorilla*. By preserving intermediate morphologies, adding fossils to the analyses of modern species may help to make the transition from prognathic-small brained taxa to orthognathic-large brained humans less abrupt (Begun, 1992; Strait and Grine, 2004). This mitigates the issue of low diversity in modern hominids and provides additional evolutionary information of phylogenetic interest. However, why different cranial regions may respond differently to natural selection and other evolutionary forces is still an open question, and one which is a central issue to evolutionary developmental biology (Klingenberg, 2013).

The trees that we presented correspond to some extent to a decline in masticatory robusticity as one moves toward the tips of the phylogeny: for example, *Pongo* and *Gorilla* show greater degrees of masticatory hypertrophy than *Pan* and especially *Homo*, where masticatory robusticity declines sharply. These size-related shape changes might especially affect the temporomandibular joint, where pull from mastication would be greater, effectively changing the morphology and in particular the depth of this joint. It is possible that this functional aspect

contributed to some of the patterns in temporal bone morphology we observed. Interestingly, in fossil hominins, the genus *Paranthropus* has large masticatory regions associated with sagittal and nuchal crest superstructures (Wood and Lieberman, 2001). The relationships of *Paranthropus* with other hominins tend to be unstable in phylogenetic reconstructions (Strait et al., 2007): depending on the authors, it appears either before *Australopithecus*, nested within it, or after it and immediately before early *Homo* (Wood, 1992; Lieberman et al., 1996; Strait et al., 1997). Species of *Paranthropus*, like *Gorilla*, possess large cranial superstructures which significantly impact on the degree of cranial robusticity present (Wood and Constantino, 2007). Because of the dominance of size and robusticity, it would be tempting to explore the use of non-allometric shape to infer the evolutionary relationships of *Paranthropus*. However, we showed that guessing whether allometric variation confounds or actually increases the phylogenetic signal is not straightforward. Results will be specific to the lineage and, even in a given lineage, they might change depending on what species are or are not included. As in all inferential methods, statistically controlling for allometry using a regression approach produces results that are sample-dependent (Jungers et al., 1995; Smith, 2005). This is another example of how variables derived from the statistical manipulations of Procrustes shape coordinates may lack the property of character invariance, which is fundamental in phylogenetic inference (Adams et al., 2011). In fact, other methods, such as Gilbert et al. (2009, their Fig. 9), share this same issue, as they also depend on the slopes of regression lines, with the added complication of splitting continuous characters into meristic ones. As we anticipated, phylogenetic inference was not the aim of this study and trees were estimated mainly as a step in replicating LEA's analysis. The resulting topologies, however, were in agreement with the tests of the phylogenetic signal, showing higher congruence between morphological and molecular trees for data with a stronger

signal.

To understand better how the covariation between size and shapes might affect the information on phylogenetic relationships that one could obtain from measurements of cranial variation would require studying a number of lineages within modern primates, and, more generally, other mammals. This would also allow one to explore the sensitivity of results to more or less inclusive analyses in terms of taxonomic sampling. In fact, findings might change if fossils or hylobatids, for instance, are included in the analysis of the temporal bone. For this specific study, however, that is less relevant, as our aim was to replicate LEA's work using their same landmark configuration in living great apes and investigate size as well as shape. In fact, we may have also considered form and replicated all analyses in the Procrustes form space as well (Mitteroecker et al., 2013). It is not unlikely that form could, in this specific dataset, have an even higher phylogenetic signal than size and shape on their own. Form spaces tend to be dominated by size variation, when size differences are large. This is likely to be the case in our study, where the average correlation of Procrustes form distances with size distances is 0.841. More importantly, a higher signal in form would not have changed the conclusion that size is the main source of phylogenetic information in the extant great ape temporal bone. Thus, regardless of whether this finding is or is not common in other groups and with all the limitations we discussed (such as sample dependence, use of a measure of overall similarity, and lack of character invariance), size should be always analyzed in parallel to and in association with (allometry) shape.

Our work presents a number of differences compared to the original study by LEA. To start, they employed a subset of the methods we used in this study. This was partly because of their different approach and focus but also because some of the techniques we employed were

not available at that time (e.g., reliable tests for phylogenetic signal). The most obvious difference, however, is, as we mentioned in the previous paragraph, that LEA did not directly use size information in their analysis of phylogenetic signal and tested the association between size and shape only in Mantel tests, which are known to have low power (Adams, 2014a). Including the size component is generally customary in geometric morphometrics. Nevertheless, morphometricians sometimes see size as uninteresting or obvious, and partly or completely overlook this aspect. By showing that the strongest source of phylogenetic information in the hominid temporal bone is actually size and size-related shape variation, we changed the original interpretation and also exemplified how the study of size is no less crucial than that of shape for a deep understanding of morphological change in evolution.

On the methodological side, another important aspect is that testing hypotheses using multiple approaches may be revealing and help to understand better their advantages and disadvantages. This is shown by the differences, sometimes remarkable, we found between comparative allometric regressions using the PIC approach in MorphoJ and the distance-based method recently proposed by Adams (2014b). These are equivalent methods, and indeed their multivariate R^2 were identical, but puzzlingly p values differed. This is likely related to how the randomization is performed (Adams and Collyer, 2015): in MorphoJ, PICs are calculated once, and then shuffled for significance testing; in Geomorph, the data are shuffled first, and then the whole comparative regression analysis is performed on each iteration. Although MorphoJ results might look more intuitive, as larger R^2 are associated with smaller p -values, the method implemented in Geomorph may be preferable, as it randomizes the dependent-independent variables relationship but does not alter the phylogenetic information (i.e., retain the error variance-covariance structure of the phylogeny). However, it might also be that differences in the

way the null hypothesis being tested is formulated account for the apparent incongruence of the two approaches (Klingenberg, pers. comm.) and further studies will be required to understand the issue better and decide which method is more accurate.

Another difference between our study and LEA's analysis is that LEA mainly focused on subspecies, for which they had fairly large samples, whereas we did all analyses not only at subspecies but also at species level. Subspecies, by definition, are not completely separated from each other, so may be expected to be unstable in a phylogenetic tree, which may not in fact be the best type of representation for intraspecific relationships (Posada and Crandall, 2001).

Although using Procrustes shape data for phylogenetic inference is problematic, as we have already mentioned, and not our main aim, pooling subspecies and analysing species allowed us to mitigate some of the problems with phylogenetic analyses and, more importantly, produced larger samples and therefore more robust statistical estimates (Cardini and Elton, 2007; Cardini et al., 2015). This is why the outcome of species analysis is generally more reliable, whereas our subspecific analyses should be interpreted with great caution. Thus, we dedicated only a few paragraphs (below) to the discussion of subspecies and focused most of this section to the broader picture emerging from our study and the consequent reinterpretation of LEA's findings.

Examination of *P. troglodytes verus* illustrates how a superficial reading of our findings is potentially misleading, if one does not bear in mind the multiplicity of issues we anticipated. Morin et al. (1994, 2001) examined molecular evidence for *P. t. verus* showing apparent full speciation both in terms of behavior and molecular structure. Subsequently, it has been amply confirmed that the western chimpanzee was the first to diverge in the *Pan troglodytes* clade (Becquet et al., 2007). It would therefore be interesting to discover morphological support for

these differences, and it does seem that at least in one of our shape trees (male dataset – Fig. 2c right) *P. t. verus* is far from other members of *Pan*. LEA, in contrast, using neighbor joining and a larger sample, consistently found the same topology as in the 10KT molecular tree, with *P. t. verus* basal to *Pan troglodytes*; they suggested that this chimpanzee subspecies might in fact have some similarities in temporal anatomy to gorillas and orangutans, yet in only one of their least-squares trees, *P. t. verus* did not cluster together with other subspecies of common chimpanzee. In our own analysis, the sample of *P. t. verus* is in fact so small that it is not unlikely that mean estimates from it might look like outliers, which tend to be farther from other taxa and cannot therefore be relied upon (Cardini, 2013). This means that, for the time being, a hypothesis with monophyletic *P. troglodytes*, and *P. t. verus* basal to other subspecies, is the more credible one, and it is largely supported by LEA as well as by DNA data (10KT tree and Becquet et al. 2007).

Suggesting the monophyly of the species of *Gorilla* based on our analysis of female shape would be similarly unwise, and this is, in fact, contradicted by the male tree. Unfortunately, in this case, we cannot compare our findings to LEA, as they did not distinguish subspecies for this genus. Thus, for gorillas, the temporal bone does not provide conclusive evidence on whether there is congruence with the molecular data (10KT tree and Thalmann et al., 2011). In contrast, however, the monophyly of *Pongo* (on this point, we remind readers to mentally re-root the tree, which we did not do to faithfully reproduce the tree from the neighbor joining using *P. pygmaeus* as the only outgroup) and that of *Gorilla* are both firmly supported by size and full shape data.

In conclusion, our revaluation of the landmark study by LEA has produced both

confirmatory (strong phylogenetic signal in the temporal bone of hominids) and novel (dominating role of size and allometry in producing the signal) results. This leads to a better understanding of the nature of the phylogenetic signal in hominid bones and emphasizes the importance of carefully assessing size variation and its effects on shapes.

It also suggests a future direction for broad comparative analyses modeling phylogenetic patterns in modern species across primates and, more generally, other mammals. To this aim, we have provided examples of a variety of approaches, which can be employed to fruitfully develop this type of framework, and have discussed their limitations. In the narrower context of human evolution, the most important group to include in the study of the phylogenetic signal is the family Hylobatidae. Lockwood et al. (2004: 4357) did not use gibbons and their kin, as they saw them as “a problematic outgroup for resolving the phylogeny of great apes and humans because of the great phenetic and size difference between it and the ingroup taxa”. Indeed, hylobatids could be too autapomorphic to be used as an outgroup for the hominids; instead, they may be analyzed on their own or together with hominids using macaques or another Old World monkey as the outgroup.

As a matter of fact, one can approach the problem of searching for a reliable morphological proxy that may find applications to infer the phylogeny of extinct hominins from at least two different but non-mutually exclusive views. The first is the type of sample LEA analysed and we replicated, where the focus was on relatively young lineages whose variation may better model the degree of divergence within the hominins. These analyses could be redone in other primates (e.g., hylobatids, papionins or just macaques, guenons, or a clade of colobines to see whether any skeletal structure is found that might tend to consistently and accurately preserve strong phylogenetic information across groups of similar evolutionary age. The second

approach is to do more inclusive analyses which progressively incorporate a larger number of species and lineages (e.g., hominoids, Catarrhini, Haplorrhini). This second strategy, however, seems less promising, as large ecological shifts and adaptive radiations may almost completely erase the phylogenetic signal in morphology (Polly, 2008). Regardless of the strategy, it seems that it will still be a long time before morphological data from geometric morphometrics are fruitfully and convincingly incorporated into phylogenetic inference, available methods need to be better assessed and new and more accurate methods might be required (Adams et al., 2014), and even when the phylogenetic signal is significant, there is no guarantee that the correct tree is inferred (Klingenberg and Gidaszewski, 2010).

Acknowledgements

We are most grateful to many colleagues who helped in a variety of ways. Dean Adams, Emma Sherratt and Mike Collyer contributed with their precious advice on Geomorph and especially with the interpretation of the differences in the phylogenetic regressions using MorphoJ and Geomorph; Chris Klingenberg also provided important feedback on this same issue, and why it might still be open to debate which approach is best. Paul O'Higgins, David Polly, Jim Rohlf and David Sheets greatly helped with their comments on what regression residuals mean in Procrustes-based geometric morphometrics. We would like to thank also Stephen Wroe for help with CT and software, and Matthew Tocheri at the Smithsonian Natural History Museum and Emmanuel Gilissen at Museums for Central Africa for their willingness to share CT data with us. Finally, we are in debt to William Jungers, and two other anonymous reviewers, as well as an

anonymous Associate Editor and the Editor, Sarah Elton, for a most stimulating, insightful and careful review of the first version of the manuscript: thanks to their advice we made substantial improvements to the methodological discussion, the interpretations and the bibliography (with many apologies for several important references we initially missed!).

References

- Adams, D.C., 2014a. A Generalized K Statistic for Estimating Phylogenetic Signal from Shape and Other High-Dimensional Multivariate Data. *Syst. Biol.* 0, 1–13.
- Adams, D.C., 2014b. A Method for Assessing Phylogenetic Least Squares Models for Shape and Other High-Dimensional Multivariate Data. *Evolution*. 86:9, 2675-2688.
- Adams, D.C., Otárola-Castillo, E., 2013. Geomorph: an r package for the collection and analysis of geometric morphometric shape data. *Methods. Ecol. Evol.* 4, 393–399.
- Adams, D.C., Collyer, M.L., 2015. Permutation tests for phylogenetic comparative analyses of high-dimensional shape data: what you shuffle matters. *Evolution*. 69:3, 823-829.
- Adams, D.C., Rohlf, F.J., Slice, D.E., 2004. Geometric morphometrics: ten years of progress following the ‘revolution’. *Ital. J. Zoo.* 71, 5–16.
- Adams, D.C., Cardini, A., Monteiro, L.R., O’Higgins, P., Rohlf, F.J., 2011. Morphometrics and phylogenetics: Principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. *J. Hum. Evol.* 60, 240–243.
- Adams, D., Rohlf, F., Slice, D., 2013. A field comes of age: geometric morphometrics in the 21st century. *Hystrix*, 24, 7-14.
- Arnold, C., Matthews, L.J., Nunn, C.L., 2010. The 10kTrees website: A new online resource for

- primate phylogeny. *Evol. Anthropol. Issues, News, and Reviews* 19, 114–118.
- Becquet, C., Patterson, N., Stone, A.C., Przeworski, M., Reich, D. 2007. Genetic structure of chimpanzee populations. *PloS Genet.* 3:4, 617–626.
- Blomberg S.P., Garland T., Ives A.R., 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57, 717–745.
- Cardini, A., 2013. Geometric morphometrics. *Encyclopedia Of Life Support Systems (EOLSS)*, Developed under the Auspices of the UNESCO, Eolss Publishers, Paris, France.
- Cardini, A., Elton, S. 2007. Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology* 126, 121–134.
- Cardini, A., Elton, S. 2008a. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. *Biol. J. Linn. Soc.* 93, 813–834.
- Cardini, A., Elton, S. 2008b. Variation in guenon skulls (I): species divergence, ecological and genetic differences. *J. Hum. Evol.* 54, 615–637.
- Cardini, A., Loy, A., 2013. On growth and form in the "computer era": from geometric to biological morphometrics. *Hystrix*, 24, 1–5.
- Cardini, A., Seetah, K., Barker, G., 2015. How many specimens do I need? Sampling error in geometric morphometrics: testing the sensitivity of means and variances in simple randomized selection experiments. *Zoomorphology*. 1–15.
- Caumul, R., Polly, P.D., 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). *Evolution* 59, 2460–2472.
- Cole, III T.M., Lele, S., Richtsmeier, J.T., 2002. A parametric bootstrap approach to the detection of phylogenetic signals in landmark data. In: MacLeod, N., Forey, P.L. (Eds.),

- Morphology, Shape and Phylogeny. Taylor and Francis, London, pp. 194–219.
- von Cramon-Taubadel, N., 2014. Evolutionary insights into global patterns of human cranial diversity: population history, climatic and dietary effects. *J. Anthropol. Sci* 92: 43–77.
- Dunn, J., Cardini, A., Elton, S., 2013. Biogeographic variation in the baboon: dissecting the cline. *J. Anat.* 223, 337–352.
- Felsenstein, J., 2004. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland.
- Gilbert, C.C., 2011. Phylogenetic analysis of the African papionin basicranium using 3-D geometric morphometrics: The need for improved methods to account for allometric effects. *Am. J. Phys. Anthropol.* 144, 60–71.
- Gilbert, C.C., Rossie, J.B., 2007. Congruence of molecules and morphology using a narrow allometric approach. *Proc. Natl. Acad. Sci.* 104, 11910–11914.
- Gilbert, C.C., Frost, S.R., Strait, D.S., 2009. Allometry, sexual dimorphism, and phylogeny: A cladistic analysis of extant African papionins using craniodental data. *J. Hum. Evol.* 57, 298–320.
- Hammer, O., Harper, D., Ryan, P., 2001. PAST: Paleontological statistics software package for education and data analysis. *Paleontol. Electron.* 4 (1), 1–9.
- Hemani, G., Yang, J., Vinkhuyzen, A., Powell, J.E., Willemsen, G., Hottenga, J.-J., Abdellaoui, A., Mangino, M., Valdes, A.M., Medland, S.E., Madden, P.A., Heath, A.C., Henders, A.K., Nyholt, D.R., de Geus, E.J.C., Magnusson, P.K.E., Ingelsson, E., Montgomery, G.W., Spector, T.D., Boomsma, D.I., Pedersen, N.L., Martin, N.G., Visscher, P.M., 2013. Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *Am. J. Hum. Genet.* 93, 865–875.
- Ioannidis, J.P., 2014. How to make more published research true. *PLoS medicine*. 11, e1001747.

- Klingenberg, C.P., 1996. Multivariate allometry. In: Marcus, L.F., Corti, M., Loy, A., Naylor, G. J. P., Slice, D. E. (Eds.) *Advances in Morphometrics*. Plenum Press, New York,.
- Klingenberg, C.P., 2011. MorphoJ: an integrated software package for geometric morphometrics. *Mol. Ecol. Resour.* 11, 353–357.
- Klingenberg, C.P., 2013. Cranial integration and modularity: insights into evolution and development from morphometric data. *Hystrix*, 24, 43-58.
- Klingenberg, C.P., Gidaszewski, N.A., 2010. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. *Syst. Biol.* 59, 245 –261.
- Klingenberg, C.P., Marugán-Lobón, J., 2013. Evolutionary covariation in geometric morphometric data: Analyzing integration, modularity, and allometry in a phylogenetic context. *Syst. Biol.* 62, 591–610.
- Lieberman, D.E, Wood, B.A., Pilbeam, D.R., 1996. Homoplasy and early *Homo*: an analysis of the evolutionary relationships of *H. habilis sensu stricto* and *H. rudolfensis*. *J. Hum. Evol.* 30, 97-120.
- Lockwood, C.A, Kimbel, W.H, Lynch, J.M., 2004. Morphometrics and hominoid phylogeny: Support for a chimpanzee–human clade and differentiation among great ape subspecies. *Proc. Natl. Acad. Sci.* 101, 4356–4360.
- Maddison, W.P., Maddison, D., 2001. Mesquite: a modular system for evolutionary analysis. Version 3.02 <http://mesquiteproject.org>
- Marcus, L.F., 1990. Traditional morphometrics. In: Rohlf F.J., Bookstein F.L. (Eds.). *Proc. Michigan Morphometrics Workshop*. Univ. Michigan Museums, Ann Arbor, Michigan, 77–122.

- Mitteroecker, P., Gunz, P., Windhager, S., Schaefer, K., 2013. A brief review of shape, form, and allometry in geometric morphometrics, with applications to human facial morphology. *Hystrix*, 24, 59–66.
- Monteiro, L.R., 2013. Morphometrics and the comparative method: studying the evolution of biological shape. *Hystrix*, 24, 8.
- Morin, P.A., Moore, J.J., Chakraborty, R., Jin, L., Goodall, J., Woodruff, D.S., 1994. Kin selection, social structure, gene flow and the evolution of chimpanzees. *Science* 265, 1193-1201.
- Morin, P.A., Chambers, K.E., Boesch, C., Vigilant, L., 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). *Mol. Ecol.* 10, 1835-1844.
- Oksanen J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., Suggests, M., 2007. The vegan package. Community ecology package. Available from <http://cran.r-project.org>, <https://github.com/vegandevs/vegan>.
- Oxnard, C.E., 2000. Morphometrics of the primate skeleton and the functional and developmental underpinnings of species diversity. Linnean Society Symposium Series 20: 207–234.
- Page, R.D., 1993. COMPONENT: Tree comparison software for Microsoft Windows, version 2.0. Natural History Museum, London.
- Polly, P.D. 2008., Adaptive zones and the pinniped ankle: A three-dimensional quantitative analysis of carnivoran Tarsal evolution. In: Sargis, E.J, Dagosto, M. (Eds.) *Vertebrate Paleobiology and Paleoanthropology Series. Mammalian Evolutionary Morphology*.

- Springer, Netherlands, pp. 167–196.
- Posada, D., Crandall, K.A., 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45.
- R Core Team., 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.
- Revell, L.J., Harmon, L.J., Collar, D.C., 2008. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* 57, 591–601.
- Rohlf, F.J., 2013. NTSYSpC: Numerical taxonomy system, ver. 2.1. Exeter Publishing, Ltd, Setauket, New York.
- Rohlf, F., 2015. The TPS Series. Hystrix, in press.
- Rohlf, F.J., Slice, D., 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zoo.* 39, 40–59.
- Shea, B., 1983. Allometry and heterochrony in the African apes. *Am. J. Phys. Anthropol.* 62, 275-289.
- Shea, B. 1986. Ontogenetic approaches to sexual dimorphism in anthropoids. *J. Hum. Evol.* 1(2), 97-110.
- Shea, B., 1988. Heterochrony in primates. In: McKinney, M.L. (Ed.), *Heterochrony in Evolution: A Multidisciplinary Approach.*, Plenum Press, New York, pp. 237-263.
- Sholts, S.B., Flores, L., Walker, P.L., Warmlander, S.K.T.S., 2011. Comparison of coordinate measurement precision of different landmark types on human crania using a 3D laser scanner and a 3D digitiser: implications for applications of digital morphometrics. *Int. J. Osteoarchaeol.* 21, 535–543.
- Strait, D.S., Grine, F.E, Fleagle, J.G., 2007. Analyzing hominid phylogeny. In: Henke, W.,

- Tattersall, I. (Eds.) Handbook of Paleoanthropology. Springer, New York, pp. 1781-1806.
- Strait, D.S., Grine, F.E., Moniz, M.A., 1997. A reappraisal of early hominid phylogeny. *J. Hum. Evol.* 32, 17-82.
- Thalmann O., Wegmann, D., Spitzner M., Arandjelovi, M., Guschanski, K., Leuenberger, C., Bergl, R.A., Vigilant, L., 2011. Historical sampling reveals dramatic demographic changes in western gorilla populations. *BMC Evol. Biol.* 11(85), 1-10.
- Viscosi V., Cardini, A., 2011. Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. *PLoS ONE* 6: e25630.
- Wall, J.D. 2013., Great ape genomics. *ILAR J.* 54:82-90, doi: 10.1093/ilar/ilt048
- Wood, B. 1992., Early hominid species and speciation. *J. Hum. Evol.* 22, 351-365.
- Wood, B, Constantino, P., 2007. *Paranthropus boisei*: Fifty years of evidence and analysis. *Yearb. Phys. Anthropol.* 50, 106–132.
- Zelditch, M., Swiderski, D., Sheets, D., Fink, W., 2004. Geometric Morphometrics for Biologists: A Primer: Elsevier Academic Press. Waltham, MA.
- Zelditch, M.L, Swiderski, D.L, Sheets, H.D., 2012. Geometric morphometrics for biologists: a primer. Academic Press, London.
- Zollikofer, C.P., Ponce de Leon, M.S. 2005. Virtual Reconstruction: A Primer in Computer-Assisted Paleontology and Biomedicine. Wiley-Blackwell, New York.

Figure Legends

Fig. 1. Landmark configuration seen on the inferior (left) and lateral (right) views of the temporal bone of *Homo sapiens*. Please refer to Table 2 for the definitions of each landmark.

Fig. 2. Molecular and morphological species (7SP) trees with *Homo* and *Pan* lineages emphasized in grey. (a) 10KT consensus molecular time tree (estimated time of divergence in millions of years is shown on the horizontal axis). Neighbor joining trees of average females (left) and males (right) based on (b) size (cm), (c) shape (units of Procrustes shape distance - uPRD), (d) size-corrected shape (uPRD). For (b) and (c) percentages of node repeatability are shown by the corresponding node; these are not provided for (d) as *Homo* is never found any closer to members of its sister genus *Pan*.

Fig. 3. Molecular and morphological species/subspecies (11SSP) trees (see previous Figure for a more detailed legend). (a) 10KT tree with estimated evolutionary ages. Neighbor joining trees of average females (left) and males (right) based on (b) size, (c) shape, (d) size-corrected shape; nodes are shown with percentages of node repeatability only in (b) and (c), where *Homo* cluster with or close to species/subspecies of its sister genus, *Pan*.

Table 1. Sample composition and size variation (mean and standard deviation of temporal bone centroid size).

Subspecies	M ^a	F ^a	N	M size (mm)	F size (mm)
<i>Pongo pygmaeus</i>	8	9	17	110.0 (10.0)	92.0 (3.8)
<i>P. abelii</i>	4	6	10	104.3 (8.3)	92.5 (4.7)
<i>Gorilla beringei beringei</i>	7	8	15	135.1 (5.7)	113.7 (5.3)
<i>G. b. graueri</i>	6	6	12	125.2 (6.2)	109.9 (3.1)
<i>G. gorilla gorilla</i>	14	4	18	128.9 (5.5)	112.2 (7.7)
<i>G. g. diehli</i>	5	7	12	133.6 (4.9)	108.5 (4.9)
<i>Pan paniscus</i>	5	7	12	86.6 (2.0)	83.9 (3.6)
<i>Pan troglodytes verus</i>	2	3	5	90.4 (0.8)	88.5 (4.1)
<i>P. t. schweinfurthii</i>	9	8	17	94.0 (3.2)	90.5 (2.9)
<i>P. t. troglodytes</i>	6	5	11	90.9 (2.4)	88.1 (4.6)
<i>Homo sapiens</i>	20	14	34	93.7 (6.5)	89.9 (5.8)
Total	8	77	163	-	-
	6				

^a Males are abbreviated using M and females using F.

Table 2. Anatomical landmarks (please also refer to Fig. 1).

Landmark	Description
1	Intersection of the infratemporal crest and sphenosquamosal suture.
2	Most lateral on the margin of foramen ovale (not present on temporal bone).
3	Most anterior on the articular surface of the articular eminence.
4	Most inferior on entoglenoid process.
5	Most inferior on the medial margin of the articular surface of the articular eminence.
6	Center of the lateral margin of the articular surface of the articular eminence.
7	Center of articular eminence.
8	Deepest within mandibular fossa.
9	Most inferior on the postglenoid process.
10	Anterior margin of tympanic element that is closest to carotid canal.
11	Apex of the petrous part of the temporal bone.
12	Most posterolateral on the margin of the carotid canal entrance.
13	Most lateral on the vagina of the styloid process (whether process is present or not).
14	Most lateral on the margin of the stylomastoid foramen.
15	Most lateral on the jugular fossa.
16	Centre of the inferior tip of the mastoid process.
17	Most inferior on the external acoustic meatus.
18	Most inferolateral on the tympanic element of the temporal.
19	Point of inflection where the braincase curves laterally into the supraglenoid gutter, in coronal plane of mandibular fossa.
20	Lateral margin of zygomatic process of the temporal bone at the position of the postglenoid process.

21 Auricular.

22 Porion.

Table 3. Evolutionary allometry: multivariate regressions of shape on size using non-comparative (‘standard’) and comparative methods (weighted or unweighted, depending on whether branch lengths are used or not). Analyses are replicated both using MorphoJ PIC approach and Geomorph approach based on Euclidean distances.

data	sex	model	Morph		Geomorph		
			oJ		ph		
			R ²	<i>p</i> ^a	R ²	F _{1,5/9}	<i>p</i>
7SP	F ^b	standard	21.8%	0.2304	21.8%	1.396	0.2312
		weighted	15.4%	0.5424	15.4%	0.913	0.0798
		unweighted	13.4%	0.5625	13.4%	0.773	0.0964
	M ^b	standard	24.6%	0.1173	24.6%	1.627	0.1191
		weighted	15.4%	0.4401	15.4%	0.909	<i>0.0041</i>
		unweighted	16.3%	0.3210	16.3%	0.971	<i>0.0028</i>
	11SP	standard	22.7%	<u><i>0.0009</i></u>	22.7%	2.648	<u><i>0.0006</i></u>
		weighted	55.7%	<u><i>0.0005</i></u>	55.7%	11.334	0.4584
		unweighted	9.5%	0.2981	9.5%	0.941	<i>0.0047</i>
	M	standard	23.9%	<i>0.0023</i>	23.9%	2.822	<u><i>0.0019</i></u>
		weighted	39.7%	<i>0.0096</i>	39.7%	5.918	0.5195
		unweighted	12.5%	0.1930	12.5%	1.285	<i>0.0027</i>

^aUnless specified otherwise, significant ($p < 0.05$) and highly significant ($p < 0.0021$, Bonferroni corrected for the total of 24 tests) p values are emphasized respectively in italics and underscored italics. Results that are in agreement using different tests are emphasized using a grey (non significant) or black (significant) background.

^b Males are abbreviated using M and females using F.

Table 4. Permutation tests for the phylogenetic signal in size, shape and size-corrected (size-corr.) shape using matrix correlation (r) and a Mantel test, squared change parsimony (SQCP), and K_{mult} . As in the regressions using comparative methods, tests are performed either including branch lengths (weighted analysis) from the KT tree or setting them to the same unit lengths (unweighted analysis).

data	sex	Component	model	matrix		SQCP		K_{mult}	
				r	p^a	tree length	p^*	K	p
7SP	F ^b	Size	weighted	0.083	0.1502	224.860	0.0523	1.235	0.0794
			unweighted	0.304	0.1005	201.354	0.0497	1.367	0.0804
		Shape	weighted	0.061	0.3078	0.062	0.0043	0.614	0.0062
			unweighted	0.009	0.4878	0.057	0.0077	0.738	0.0088
		size-corr. shape	weighted	0.090	0.3097	0.053	0.2145	0.554	0.0460
			unweighted	-0.089	0.7106	0.050	0.1503	0.671	0.1389
M ^b		Size	weighted	0.181	0.1104	591.039	0.0128	1.280	0.0344
			unweighted	0.485	0.0646	519.680	0.0145	1.470	0.0441
		Shape	weighted	0.242	0.1841	0.066	0.0038	0.717	<u>0.0011</u>
			unweighted	0.150	0.1804	0.061	0.0023	0.810	<u>0.0015</u>
		size-corr. shape	weighted	0.283	0.1614	0.055	0.1026	0.637	0.0186
			unweighted	0.042	0.3810	0.051	0.0863	0.733	0.0722

11SP F	Size	weighted	0.271 0.0775	241.530 <u>0.0008</u>	0.226 <i>0.0362</i>
		unweighted	0.524 <i>0.0043</i>	201.262 <u>0.0005</u>	1.852 <i>0.0039</i>
	Shape	weighted	0.216 0.1760	0.080 <u>0.0001</u>	0.040 <u>0.0005</u>
		unweighted	0.020 0.4524	0.071 <u>0.0001</u>	0.681 <u>0.0001</u>
	size-corr. shape	weighted	0.126 0.1867	0.073 <i>0.0432</i>	0.031 <i>0.0052</i>
		unweighted	-0.206 0.9244	0.066 <i>0.0121</i>	0.586 <i>0.0305</i>
M	Size	weighted	0.333 <i>0.0237</i>	671.675 <u>0.0003</u>	0.229 <i>0.0194</i>
		unweighted	0.644 <u>0.0006</u>	557.749 <u>0.0006</u>	1.847 <u>0.0014</u>
	Shape	weighted	0.344 0.1368	0.091 <u>0.0001</u>	0.043 <i>0.0023</i>
		unweighted	0.078 0.2759	0.082 <u>0.0001</u>	0.698 <u>0.0001</u>
	size-corr. shape	weighted	0.282 0.1686	0.081 <i>0.0424</i>	0.035 <i>0.0155</i>
		unweighted	-0.177 0.9079	0.074 <i>0.0309</i>	0.617 <i>0.0159</i>

^a Unless specified otherwise, significant ($p < 0.05$) and highly significant ($p < 0.0021$, Bonferroni corrected for the total of 24 tests) p values are emphasized respectively in italics and underscored italics. The Bonferroni corrected p is again 0.0021 as there are 24 tests of a given aspect of temporal morphology – e.g., size – regardless of sex and species or subspecies taxonomic level. Results that are in agreement using different tests are emphasized using a grey (non significant) or black (significant) background.

^b Males are abbreviated using M and females using F.

Table 5. Summary of the results of the tests for the phylogenetic signal reported in the previous table: percentages of marginally significant ($p < 0.1$), significant ($p < 0.05$) and highly significant ($p < 0.0021$) tests are shown. K_{mult} , the most reliable test, and the summary percentages from all tests regardless of type (lower half of the table) are emphasized in italics.

	$p < 0.10$			$p < 0.05$			$p < 0.0021$		
	matrix r	SQCP	<i>K_{mult}</i>	matrix r	SQCP	<i>K_{mult}</i>	matrix r	SQCP	<i>K_{mult}</i>
size	62.5%	100.0%	<i>100.0%</i>	37.5%	87.5%	<i>75.0%</i>	12.5%	50.0%	<i>12.5%</i>
shape	0.0%	100.0%	<i>100.0%</i>	0.0%	100.0%	<i>100.0%</i>	0.0%	50.0%	<i>62.5%</i>
size-corr. shape	0.0%	62.5%	<i>87.5%</i>	0.0%	50.0%	<i>75.0%</i>	0.0%	0.0%	<i>0.0%</i>
	overall			overall			overall		
size	<i>87.5%</i>			<i>66.7%</i>			<i>25.0%</i>		
shape	<i>66.7%</i>			<i>66.7%</i>			<i>37.5%</i>		
size-corr. shape	<i>50.0%</i>			<i>41.7%</i>			<i>0.0%</i>		

Table 6. Numerical estimates of congruence between the KT molecular phylogeny and the morphological trees. Significant tests, which indicate more similarity than expected by chance, are emphasized as in previous tables: partition measure (PM); leaves to prune to get greatest agreement subtree (LP); nearest neighbor interchange (NNI).

method	Component	sex	7SP metric	p^a	11SP metric	p
PM	size	F	2	<i>0.010</i>	12	1.0000
		M	2	<i>0.010</i>	12	1.0000
	shape	F	0	<u><i>0.001</i></u>	6	<i>0.0068</i>
		M	0	<u><i>0.001</i></u>	10	0.2897
	size corr. shape	F	6	0.310	14	1.0000
		M	4	0.060	14	1.0000
LP	size	F	1	0.052	4	0.9696
		M	1	0.052	4	0.9696
	shape	F	0	<u><i>0.001</i></u>	2	<i>0.0240</i>
		M	0	<u><i>0.001</i></u>	2	<i>0.0240</i>
	size corr. shape	F	2	0.585	5	1.0000
		M	2	0.585	5	1.0000
NNI	size	F	1	<i>0.010</i>	7	0.2029
		M	1	<i>0.010</i>	7	0.2029
	shape	F	0	<u><i>0.001</i></u>	3	<i>0.0043</i>

	M	0	<u>0.001</u>	5	0.0471
size corr. shape	F	4	0.376	12	0.9756
	M	3	0.162	11	0.8997

^a Unless specified otherwise, significant ($p < 0.05$) and highly significant ($p < 0.004$, Bonferroni corrected for the total of 24 tests) p values are emphasized respectively in italics and underscored italics. The Bonferroni corrected p is now 0.004 as there are 12 tests for each morphological component.

Table 7. Summary of the results of tests for congruence between molecular and morphological trees: percentages of marginally ($p < 0.1$), significant ($p < 0.05$) and highly significant ($p < 0.004^a$) tests are shown.

component	$p < 0.10$			$p < 0.05$			$p < 0.004$		
	PM	LP	NNI	PM	LP	NNI	PM	LP	NNI
size	50.0%	50.0%	50.0%	50.0%	0.0%	50.0%	0.0%	0.0%	0.0%
shape	75.0%	100.0%	100.0%	75.0%	100.0%	100.0%	75.0%	50.0%	75.0%
size-corr. shape	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

^aBonferroni correction for 12 tests, which overall is the number of tests for all datasets regardless of sex and type of similarity metrics: i.e., two tests (males and females) for two datasets (7SP and 11SSP) by three types of dissimilarity metrics.

Fig. 1.

Fig. 2.

Fig. 3.