Development and Assessment of Fully Automated and Globally Transitive Geometric Morphometric Methods, With Application to a Biological Comparative Dataset With High Interspecific Variation

TINGRAN GAO,[1]* GABRIEL S. YAPUNCICH ,[2,3]* INGRID DAUBECHIES,[1] SAYAN MUKHERJEE,[4] AND DOUG M. BOYER[3]

[1]Department of Mathematics, Duke University, Durham, North Carolina
[2]Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, North Carolina
[3]Department of Evolutionary Anthropology, Duke University, Durham, North Carolina
[4]Departments of Statistical Science, Mathematics, and Computer Science, Duke University, Durham, North Carolina

ABSTRACT

Automated geometric morphometric methods are promising tools for shape analysis in comparative biology, improving researchers' abilities to quantify variation extensively (by permitting more specimens to be analyzed) and intensively (by characterizing shapes with greater fidelity). Although use of these methods has increased, published automated methods have some notable limitations: pairwise correspondences are frequently inaccurate and pairwise mappings are not globally consistent (i.e., they lack transitivity across the full sample). Here, we reassess the accuracy of published automated methods—cPDist (Boyer et al. Proc Nat Acad Sci 108 (2011) 18221–18226) and auto3Dgm (Boyer et al.: Anat Rec 298 (2015a) 249–276)—and evaluate several modifications to these methods. We show that a substantial percentage of alignments and pairwise maps between specimens of dissimilar geometries were inaccurate in the study of Boyer et al. (Proc Nat Acad Sci 108 (2011) 18221–18226), despite a taxonomically partitioned variance structure of continuous Procrustes distances. We show these inaccuracies are remedied using a globally informed methodology within a collection of shapes, rather than relying on pairwise comparisons (c.f. Boyer et al.: Anat Rec 298 (2015a) 249–276). Unfortunately, while global information generally enhances maps between dissimilar objects, it can degrade the quality of correspondences between similar objects due to the accumulation of numerical error. We explore a number of approaches to mitigate this degradation, quantify their performance, and compare the generated pairwise maps (and the

Tingran Gao and Gabriel S. Yapuncich contributed equally to this work.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Division of Graduate Education; Grant number: 0991660; Grant sponsor: NSF; Grant numbers: BCS 1304045, BCS 1317525, BCS 1440742, BCS 1440588, and DGE 0991660.

*Correspondence to: Tingran Gao, Department of Mathematics, Duke University, 107A Biological Sciences Building, Campus Box Box 90383, Durham, NC 27708 E-mail: trgao10@math.duke.edu Or Gabriel S. Yapuncich, Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, NC E-mail: gabrielyapuncich@gmail.com

Received 12 May 2017; Revised 15 July 2017; Accepted 7 August 2017.

DOI 10.1002/ar.23700

Published online 00 Month 2017 in Wiley Online Library (wileyonlinelibrary.com).

shape space characterized by these maps) to a "ground truth" obtained from landmarks manually collected by geometric morphometricians. Novel methods both improve the quality of the pairwise correspondences relative to cPDist and achieve a taxonomic distinctiveness comparable to auto3Dgm. Anat Rec, 00:000–000, 2017. © 2017 Wiley Periodicals, Inc.

Key words: procrustes; transformational homology; morphological disparity; shape analysis; phenomics

Quantifying and comparing complex shapes is a key component of fields as diverse as evolutionary morphology, molecular biochemistry, computer vision, and computational anatomy. A variety of analytical methods have been developed to achieve this goal, including landmark-based geometric morphometrics (Gower, 1975; Dryden and Mardia, 1998; Rohlf and Slice, 1990; Slice, 2005), voxel-based morphometry (Ashburner and Friston, 2000), and spherical harmonics (Gerig et al., 2001; Styner et al., 2006). Of these methods, three-dimensional geometric morphometrics (3DGM) based on the alignment of spatial coordinates through Procrustes superimposition is particularly widespread in evolutionary morphological studies (for reviews, see Slice, 2007; Lawing and Polly, 2010; Cooke and Terhune, 2015). Though popular, 3DGM is nonetheless a time-consuming and labor-intensive process, requiring a substantial number of landmarks to be placed on each specimen by the researcher (Polly and MacLeod, 2008; Gunz et al., 2005; Wiley et al., 2005). The simulation study of Watanabe (2015) suggested shape characterization is unstable without at least 30 landmarks, a number which may not be feasible in samples spanning multiple genera. For researchers utilizing 3DGM, the reliance on user-determined landmarks generates a tradeoff between sample size and detail of morphological representation for a given time spent collecting data, and thus limits the explanatory power of morphological data. Without significant methodological advances, future geometric morphometric morphological studies are likely to remain limited.

To more thoroughly characterize shape variation and decrease processing time, geometric morphometric approaches have become increasingly automated, including both semiautomated (based on semilandmarks [Bookstein, 1997; Bookstein et al., 1999, 2002; Perez et al., 2006; Harcourt-Smith et al., 2008; Mitteroecker and Gunz, 2009], or eigensurfaces [Polly and Macleod, 2008; Sievwright and Macleod, 2012]) and fully automated [Boyer et al., 2011, 2012, 2015a; Koehl and Hass, 2015] shape characterization methods). Automated 3DGM methods improve researchers' ability to sample phenotypes intensively (by increasing the resolution of shape characterization) and extensively (by permitting the inclusion of more specimens). These outcomes neatly align with Houle et al.'s (2010) recommendations for advancing phenomics, the study of high-dimensional phenotypic data.

Initial evaluations of automated 3DGM methods recover performance similar to or better than user-determined landmarks for species discrimination (Boyer et al., 2011, 2015a) and shape characterization of certain types of shapes (Gonzalez et al., 2016). Still, as with user-based approaches, current automated 3DGM methods suffer from limitations. The first limitation concerns sample availability: a significant investment of time is required to convert specimens into 3D digital models. This limitation will be reduced as researchers continue to contribute data to online repositories such as MorphoSource (Boyer et al., 2016). The second limitation concerns the analytical workflow: while the fully automated method published in Boyer et al., (2011), the continuous Procrustes distance method (cPDist), successfully classifies specimens by species, the output of the method cannot be analyzed in a way analogous to user-determined landmarks due to the lack of *transitivity* of the resulting pairwise maps (i.e., the direct map from A to C is not the same as the map from A to B to C). Because the automated landmarks of cPDist are defined only on a pairwise basis, rather than from the entire collection of shapes, the algorithm does not produce a set of globally consistent landmarks (i.e., exhibit transitivity across the whole collection). This limitation prevents distance matrices produced by cPDist from being utilized in additional downstream 3DGM analyses. The third limitation stems from the computational intensity of automated 3DGM methods: the method published in Boyer et al. (2015a), auto3Dgm, produces a transitive set of "pseudolandmarks" that is applicable to the entire collection of shapes, but the method does not begin to yield consistent results unless over 1,000 pseudolandmarks are identified on each specimen (at least when each specimen is initially discretized as a mesh of no more than ~5,000 vertices) (Vitek et al., 2017). Even with a relatively powerful computer, auto3Dgm may take weeks to analyze a dataset of 200 specimens without access to parallel computational resources.

Despite these limitations, initial results and applications of automated 3DGM methods are encouraging (Boyer et al., 2011, 2012, 2015a, 2015b; Seiffert et al., 2015; Gonzalez et al., 2016). Still, the ability of fully automated methods to achieve certain goals in biological research has not been thoroughly explored. This study addresses the key issues concerning previously published automated 3DGM methods (Boyer et al., 2011, 2015a), including the following:

1. Assessing error rates of cPDist and identifying general properties of shapes which may indicate whether automated mappings are likely to be accurate.
2. Evaluating the ability of a Minimum Spanning Tree-based approach (cPMST, a variant of cPDist inspired by auto3Dgm) to avoid bad alignments.

3. Describing quantitative and qualitative differences in ordinated shape spaces recovered by automated 3DGM methods.

In addition to examining previously published automated methods, we also develop and evaluate several approaches for joining dissimilar shapes through intermediate shape sequences to insure landmark transitivity. As these approaches utilize the geometric information of the entire collection of shapes, we refer to these novel methods as "globally informed methods." To increase ease of application of the methods introduced in this study, MATLAB code is provided for each method in Supporting Information.

Background on Previous and Related Automated 3DGM Methods

The cPDist method (Boyer et al., 2011, 2012) begins by flattening disc-type shapes (such as a tooth crown without roots) into planar discs using a conformal (angle-preserving) projection. The algorithm then exhaustively searches the space of all conformal maps between the flattened shapes for an "optimal conformal map" minimizing energy functional. Conformal maps between discs are characterized by the choice of a pair of correspondence points and an in-plane rotation (thus only three degrees of freedom), which makes the exhaustive search highly efficient. The resulting conformal map then serves as an initialization for a final thin plate spline (TPS) procedure, which strives to stretch the two flattened shapes so that regions of "high curvature" (e.g., cusp tips or tuberosities) align. The value of the energy functional on the final map (composition of TPS and the optimal conformal map) defines a distance between the pair of shapes, which we refer to as the continuous Procrustes (cP) distance (Al-Aifari et al., 2013). When the cP distance is small, the final map is usually of high quality and can be leveraged to reveal interspecific variation (c.f. Boyer et al., 2011).

As an initial assessment of the biological relevance of these correspondence maps, Boyer et al., (2011) compared the classification success rates of cPDist and Procrustes distances computed from user-determined landmarks (a more traditional morphometric approach) on a mammalian molar dataset. The comparison showed that cPDist was able to taxonomically classify specimens at a rate better than or equal to the method based on user-determined landmarks. Since then, Boyer et al., (2012) used cPDist to confirm the attribution of a newly discovered fossil to a species previously believed to be much younger than indicated by dating of the fossil. However, due to lack of transitivity, it is not clear how the resulting correspondence maps of cPDist pipeline should be incorporated in a more traditional geometric morphometric workflow. Furthermore, when inspecting pairwise correspondence maps between specimens in detail, Boyer et al. (2012) observed anomalies in some of the maps (e.g., reversed alignments of the buccolingual axis). Though the distance matrices and ordinations based on them produced intelligible results, these errors raised questions about possible inaccuracies lurking in the analysis. We examine these inaccuracies in more detail here.

Boyer et al., (2015a) reported a different automated method, auto3Dgm, which guarantees transitivity, thereby permitting more familiar modes of downstream analysis. First, auto3Dgm computes all pairwise alignments and distances with a modified version of the Iterative Closest Points (ICP) algorithm (Besl and McKay, 1992). Transitivity is then imposed with the following procedure based on a Minimum Spanning Tree (MST) for the entire collection: (1) view the collection of shapes as a complete weighted graph (in which edge weights are defined by the pairwise distances), and extract an MST for this graph; 2) for any pair of shapes, the alignment between them is obtained by identifying the unique shortest path connecting them in the MST and composing the pairwise alignments along the edges constituting this path. By the nature of the MST, only alignments that yield small pairwise distances are involved in the final alignments. Auto3Dgm outputs a "pseudolandmark file" that can be analyzed with standard geometric morphometric software such as *morphologika*[2] (O'Higgins et al., 2006) or MorphoJ (Klingenberg, 2011). Boyer et al., (2015a) argued that this procedure should generally reduce the mapping errors, and verified this claim with three osteological datasets. MATLAB implementation of auto3Dgm is available on github (Gao, 2017) and the R code developed for Boyer et al. (2015a) is also available online (Boyer, 2017; Mukherjee, 2017).

Recently, Koehl and Hass (2015) suggested minimizing a novel metric, the symmetric deformation energy, when searching for a globally optimal conformal map between two closed surfaces. Based on their analysis, the program MatchSurface outperforms cPDist in approximating the ground truth pairwise distances computed from user-determined landmarks, in correctly classifying specimens to taxonomic groups, and in generating phenetic trees that more closely resemble trees generated from user-determined landmarks. However, MatchSurface may suffer from similar problems we have noted with cPDist (i.e., potential anomalies in pairwise maps, lack of transitivity, inability to utilize the method in a traditional morphometric workflow). In addition, because user-based approaches such as 3DGM have their own methodological challenges, it is unclear if automated 3DGM methods should be evaluated primarily by their ability to replicate pairwise distances computed from user-determined landmarks (although this is also the approach we use in this study). An in-depth comparative analysis of MatchSurface and the methods introduced here is beyond the scope of this article, but will be important for further development of automated geometric morphometric methods.

In a broader context, the development of automated 3DGM methods resonates with the emerging interest in the analysis of collections of shapes in the computer graphics community (e.g., Nguyen et al., 2011; Huang et al., 2012; Huang and Guibas, 2013; Chen et al, 2014). The starting point is the observation that pairwise shape registration often yields more accurate results between pairs of similar shapes than dissimilar ones: the resulting pairwise correspondence maps are much more meaningful when the shapes are near-isometric. Therefore, when working with a large collection of shapes, one can usually find a sequence of (pairwise similar) intermediate shapes between any pair of dissimilar shapes, and build the correspondence map between them by composing the more accurate pairwise correspondences along this sequence. This composition strategy leads to more accurate maps than direct pairwise comparisons. The

novel methods presented in this paper are all derived from this general idea, leveraging the size of the dataset and high-quality maps between similar shapes to improve the accuracy of correspondence maps between dissimilar shapes.

METHODS AND MATERIALS

We assess the qualities of automated correspondence analyses using the dataset originally published in Boyer et al., (2011), and available through various sources (http://arxiv.org/abs/1110.3649; Lipman, 2017). The sample consists of 116 mandibular second molars of living and fossil primates and their close relatives. Further details (such as included species and specimen information) can be found in the supplementary data of Boyer et al., (2011). Our main approach is to compare automated results of anatomical correspondence and geometric similarity among shapes in this dataset to a "ground truth"—a set of 16 user-determined landmarks placed on each specimen by experienced geometric morphometricians. As noted above, we have some reservations about regarding user-determined landmarks as the performance standard for automated correspondence analyses, but this framework facilitates comparisons between user-based and automated approaches. Abbreviations used throughout this paper can be found in Table 1, and we provide a summary of the key differences between the automated methods we evaluate here in Table 2.

Quantifying Errors of cPDist

The accuracy of pairwise correspondences can be quantified using Mean Square Error (MSE). MSE is calculated by first mapping user-determined landmarks from one tooth to another using the automated pairwise correspondence (creating a set of "propagated landmarks" on the second tooth), and then taking the average of the squared Euclidean distances between the user-determined and propagated landmarks. Larger MSEs indicate greater deviations between the two sets of landmarks (and a greater discordance between the user-determined and automated assessments). While there are other approaches to assess accuracy relative to user-determined landmarks (including directly comparing distance matrices generated by each method or the ordinations resulting from those matrices), MSE benefits by focusing on the "local" inaccuracy of correspondence maps at the level of individual landmarks.

Boyer et al. (2011) assessed cPDist's error rate by computing a variant of MSE between user-determined and propagated landmarks (Supporting Information, Table 8 of Boyer et al., 2011), which realigned propagated and user-determined landmarks with an additional Procrustes superimposition. Unfortunately, this realignment could potentially mask several types of mapping errors (Fig. 1). Most notably, the cPDist algorithm may result in an axial inversion in which incorrect sides are matched to one another (buccal–lingual [Fig. 1c] or anterior–posterior [Fig. 1d] inversions). To assess the prevalence of mapping errors in the analysis of Boyer et al., (2011), we perform an extensive (though not exhaustive) visual check of 16 propagated landmarks in 588 pairwise mappings. We observed five types of errors; the four primary errors are shown in Figure 1 (the fifth error type, a 90° rotation, occurred in only three of 161 instances of error). To test whether errors occur between dissimilar teeth, we compare the pairwise cP distances observed in the good and bad maps of this test sample.

We also test the hypothesis that more dissimilar teeth generate erroneous maps by comparing their Dirichlet Normal Energy (DNE), a measure of the bending energy of a surface that has been shown to partition species by diet reliably (Bunn et al., 2011; Ledogar et al., 2013; Winchester et al. 2014). For a reduced sample of comparisons between specimens from extant species (N = 138), species mean DNE values are taken from Bunn et al., (2011), and the absolute value of the difference between the source and target teeth calculated. Poor maps are expected to occur between teeth that exhibit a large difference in DNE.

Finally, we examine the skewness of the distribution of candidate maps for good and poor maps. To find an optimal map with minimum energy, cPDist searches among a large number (typically tens of thousands) of candidate conformal maps between two surfaces. In

TABLE 1. List of abbreviations used in this study

3DGM	Three-dimensional geometric morphometrics
ANOVA	Analysis of variance
cP	Continuous Procrustes
DNE	Dirichlet normal energy
GLAM	Globally informed automated method
ICP	Iterative closest points algorithm
LAST	Light approximate shortest path tree
MRPP	Multiple response permutation procedure
MSE	Mean square error
MST	Minimum spanning tree
PCA	Principal component analysis
SPT	Shortest path tree
TPS	Thin plate spline

TABLE 2. Globally informed methods developed for and analyzed in this study

Method type	Variations		Root shape	Feature-fixing	Pseudolandmark resolution	Source
	Alpha	Angle weight				
cPDist	-	-	Random; minimum distance to all others	-	64; 256; 1024	Boyer et al., 2011
cPMST	-	-		Off; on		Boyer et al., 2015a; this study
LAST	Mean; balance	-				This study
Composed LAST	Mean; median; balance; 1	-				This study
Viterbi	-	0%; 25%; 50%				This study

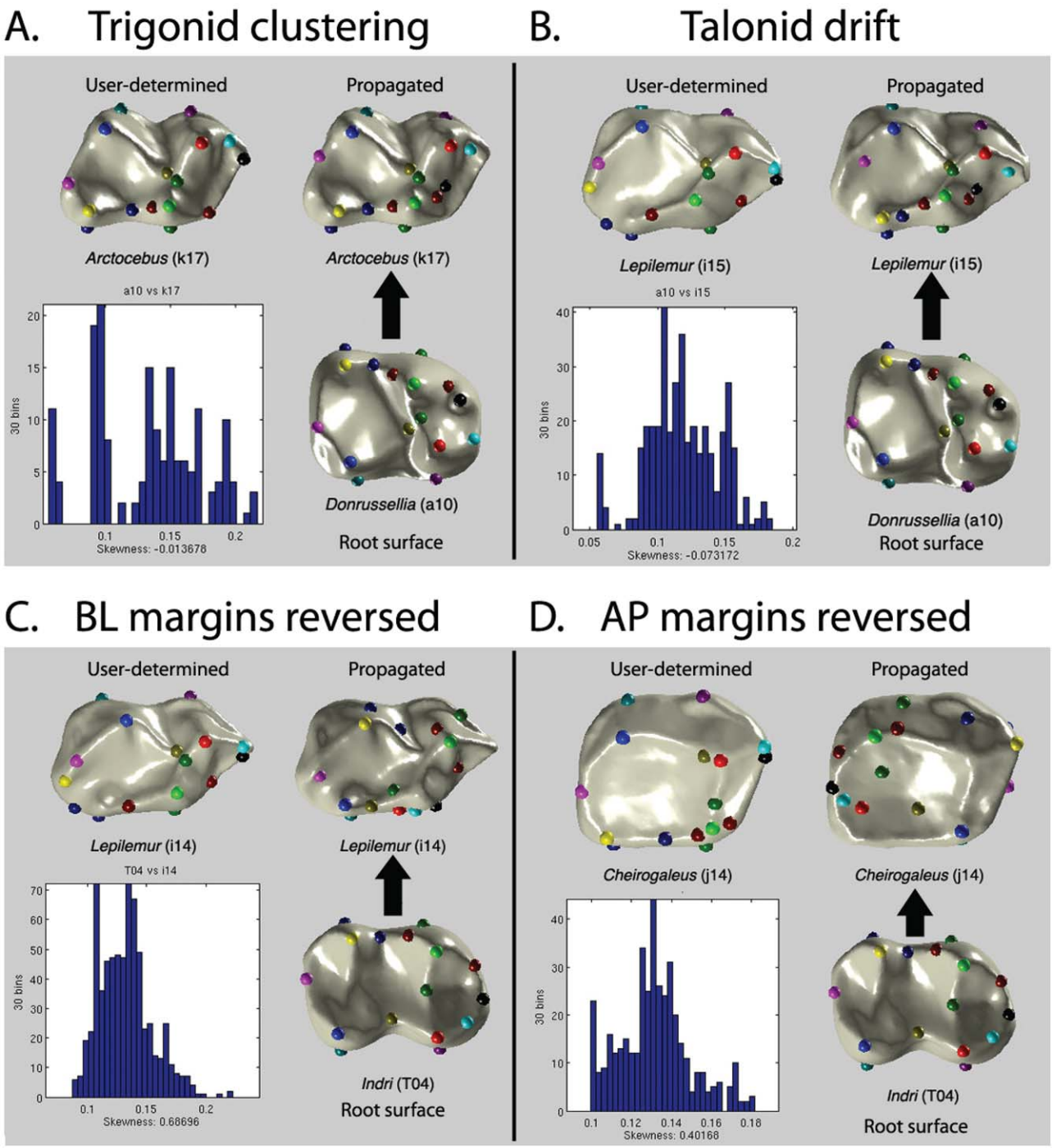

Fig. 1. Classification of errors made by cPDist using MATLAB interface. Each panel shows the root surface with user-determined landmarks (lower right), user-determined landmarks on target surface (upper left), and propagated landmarks on target surface (upper right). The distribution of the values of the continuous Procrustes functional of candidate maps is shown in the lower left of each panel. The four main error types are pictured here and include (A) Trigonid clustering: landmarks bunched in anterior portion of tooth; (B) Talonid drift: landmarks spread over posterior portion of tooth; (C) buccal-lingual (BL) inversion: landmarks are reversed from side-to-side; and (D) anterior–posterior (AP) inversion: landmarks are reversed from front-to-back. A fifth error type, 90° rotation, is not shown, but was only observed three times.

general, we expect good maps to be more strongly distinguished from the population of candidate maps than poor maps, as there are often many candidate maps with comparably low energy when a poor map forms the minimum in the exhaustive search. By measuring the skewness of the energy distribution of candidate maps, we assess the distinctiveness of the optimal map: if the optimal map is less distinct, the histogram of candidate

maps should be skewed to the right (positive skew), forcing cPDist to select from several candidate maps with similar cP distances.

Evaluating Accuracy of a Minimum-Spanning Tree Approach

As described above, auto3Dgm (Boyer et al., 2015a) improves alignments between dissimilar shapes and imposes transitivity for the entire collection using MST. However, while MST improves pairwise cP maps, the quality of the resulting maps is not guaranteed. The visually inspected test sample permits the identification of a threshold cP distance below which all maps are good maps. To gauge the accuracy of the MST approach, we evaluate if any edge lengths in the MST exceed this threshold.

We also compare the landmark MSEs of cPDist to cPMST, another MST-based approach inspired by auto3Dgm (Boyer et al., 2015a). While auto3Dgm's representation of surfaces through pseudolandmarks is analogous to user-determined landmarks (increasing the method's utility to comparative morphologists), these pseudolandmarks are not globally consistent in the same manner as user-determined landmarks. In auto3Dgm, pseudolandmarks are randomly sampled on each surface in a collection, without knowledge of the exact sampling procedure for other surfaces. The full set of user-determined landmarks is not likely to be included in the set of pseudolandmarks, making it impossible to calculate landmark MSEs for auto3Dgm. Here, we evaluate the accuracy of cPMST, an upgraded version of cPDist that generates globally consistent maps between all pairs of shapes within a collection, motivated by the MST approach first adopted in auto3Dgm. With cPMST, pairwise maps are defined for the pair of surfaces in their entirety (and necessarily including user-determined landmarks), so that landmark MSEs can be computed.

Development and Refinement of Globally Informed Methods

Though using an MST improves pairwise correspondences between dissimilar shapes, this approach could potentially degrade the quality of maps between similar shapes. Directly aligning similar teeth avoids the accumulation of random errors in pairwise comparisons, while composing alignments along a path through the MST involves intermediate shapes and can amplify random error. MSTs minimize the *total* sum of edge lengths in a subgraph connecting all the vertices; this global minimization permits large distortion of local distances. Many intermediate vertices may separate a pair of reasonably close vertices, so that the sum of the length of all intermediate hops may be larger than the direct pairwise distance.

The potential for map degradation with MST-based approaches and the results of our accuracy analyses (Results) made it clear that there was ample room for improving previously published automated 3DGM methods. We propose several approaches that maintain global transitivity but also attempt to reduce the potential for local distance distortion. These approaches are based on the research of tree-based metric space approximation and dynamic programming in computer science. We examine the effects of several different methodological variations: (1) alternative methods to avoid map degradation, (2) alternative root shapes for landmark propagation when transforming maps into pseudolandmarks, (3) alternative methods for postprocessing maps, and (4) the effects of pseudolandmark sampling resolution. All examined methods are summarized in Table 2 and described in detail below.

Alternative Methods to Avoid Map Degradation

Our first attempt to avoid potential degradation when mapping through intermediate shapes is to restrict the depth of the minimum spanning tree, and impose that any pair of shapes in the MST be separated by a controlled number (no greater than twice the tree depth) of intermediate shapes. This type of graph theoretic construction, known as a bounded-hop MST problem, is nondeterministic polynomial-time hard and has no practical polynomial-time and constant-factor approximation in general (Clementi et al., 2007). However, rather than globally minimizing the sum of edge lengths, it is almost trivial to minimize the number of hops along the tree that separates any pair of vertices: the best strategy is to designate one vertex as the root and connect any other vertices to it. Generalizing from counting the number of hops to measuring the length of the path, researchers study shortest-path trees (SPT) of a graph G, which are trees that span the graph with the property that any path connecting a vertex to the root is also the shortest path in the graph G between that particular vertex and the root. SPTs strive to minimize local distance distortions without controlling the total edge length, while MSTs minimize the latter but sacrifice the former. Leveraging the advantages of both MSTs and SPTs, the concept of Light Approximate Shortest-Path Tree (LAST) was developed to balance local distance distortion and total edge length (Khuller et al., 1995). We use LAST to alleviate the quality degradation of correspondences between similar teeth in the cP distance framework. Because LAST is also a tree, global transitivity is achieved.

Unlike MST and SPT, which are calculated directly from the data, LAST depends on a parameter, α (≥ 1), which controls the trade-off between the advantages of MST and SPT. When α is close to 1, the LAST becomes more like an SPT; as α approaches infinity, the LAST becomes more similar to an MST. We evaluate two candidates for α: the mean of the local distance distortions on an MST and the special value $=1+\sqrt{2}$, which generates an LAST "balanced" between the SPT and MST.

By altering how local distance distortion is computed, we develop a second set of LAST methods. Normally, computing the distance distortion of a tree between two vertices involves finding the shortest path on the tree that connects the pair, and then dividing the path length by the direct distance between the vertices. However, as the cP distance is defined as the minimum of an energy functional, a potentially more meaningful definition of local distance distortion uses the value of the energy functional obtained by composing pairwise correspondences along the shortest path (rather than the cumulative path length) in the numerator. We evaluate "composed" LAST methods with four candidates for α: the mean and median of the composed local distance

distortions, the special value $\alpha=1+\sqrt{2}$, and $\alpha=1$ (which does not generate an SPT with composed local distance distortion[1]).

We also attempted to fix map degradation with an approach that does not rely on trees (though this method is not strictly transitive). For any pair of shapes, the goal is to find the best path between them such that the composition of pairwise correspondences along the hops gives the most meaningful map. In the spirit of cP distance, which produces a map that minimizes an energy functional, we use the value of the same energy functional (called the "cP value") as an indicator for the quality of pairwise correspondences. Though the number of shapes in a collection is finite, the set of possible paths connecting a pair of shapes is exponentially large, making it difficult to search exhaustively for the optimum; computational complexity is even larger when looking for optimal paths between *all* pairs of shapes. This computational difficulty is tackled with a dynamical programming algorithm motivated by the Viterbi algorithm in the context of hidden Markov models. Each shape, viewed as a vertex in the complete distance graph, is treated as a hidden state of a Markov chain, and the transition probability from one state to another is determined by the cP distance between the two shapes, with smaller distances indicating greater probability. In this setup, the optimal path of composition between a pair of shapes can be interpreted as the most likely path connecting the two states. When the length of the path is fixed, optimal path searching can be done using the Viterbi algorithm (assuming the hidden state is identical to the observed state).

Owing to the efficiency of the Viterbi algorithm, it is easy to compute the optimal paths of all possible numbers of hops (ranging from 1 to one less than the total number of shapes in the collection), and to choose the path along which the composition of pairwise correspondences leads to the lowest possible cP value. This method is denoted as "Viterbi." As direct links between two shapes are the only paths with 1 hop, the Viterbi method is guaranteed to keep direct pairwise maps if they produce the lowest cP values among all paths of different number of hops, thus avoiding the accumulation of random errors through propagation along long paths in an MST. Though global transitivity is not maintained, the determination of the unique optimal path connecting any two shapes leads to good correspondence maps.

For the Viterbi method, the transition probability between any two states can depend on metric geometric information other than path distance, such as the angle between consecutive hops in a path. More obtuse angles (i.e., closer to 180 degrees) between consecutive hops are likely preferable, as more acute angles lead to more torturous paths, which may increase random error accumulation. After proper renormalization, distances and angles can be combined with convex weights to determine the transition probability. In this study, we compare the effect of angles weighted at 0%, 25%, and 50% in the computation of transition probability.

Alternative Root Shapes for Pseudolandmark Propagation

For these novel methods, globally consistent pseudolandmarks can be generated by randomly sampling a number of vertices and then propagating these pseudolandmarks to all remaining shapes after pairwise correspondences are established. These pseudolandmarks depend on the distribution of points sampled on the initial shape, but are independent of the choice of the initial shape (although as Viterbi methods are not strictly globally transitive, the initial shape does affect pseudolandmarks). To evaluate the effect of the choice of root shape on shape characterization, we generate pseudolandmarks from two different root shapes (a randomly chosen tooth [*Chronolestes simul* IVPP V10696-2] and the tooth with minimum average distance from other teeth in the collection) for all methods.

Alternative Methods of Postprocessing

The final novel development examined here is a postprocessing step that performs an additional TPS procedure to align geometric characteristics (e.g., vertices of locally maximum conformal factors, locally maximum/minimum Gaussian curvatures) of two shapes. As these geometric characteristics may be generally understood as "features," we call this step "Feature-Fix." Feature-fixing is similar to the last step in the computation of cP distances in Boyer et al. (2011), which helps to correct the random errors accumulated through correspondence compositions but sometimes introduces artificiality in regions without geometric characteristics. We implement all novel methods with and without feature-fixing.

Sampling Density of Pseudolandmarks

Several previous studies have highlighted the importance of landmark sampling density for shape characterization (Watanabe, 2015; Vitek et al., 2017). Vitek et al. (2017) suggested that auto3Dgm does not produce consistent results unless 1000 pseudolandmarks are generated on each specimen. To examine the influence of sampling density on methods developed in this study, we implement all methods with 64, 256, and 1024 pseudolandmarks sampled.

Comparing Effects of Globally Informed Methods on the Characterization of Geometric Affinities

The different types of globally informed methods (N = 8: MST, LAST, Viterbi and their variations), the root shape used to generate pseudolandmarks (N = 2), the use of feature-fixing in postprocessing (N = 2), and the sampling density of pseudolandmarks (N = 3) can be understood as "parameters" of the cP distance improvement framework. To compare variance patterns within

[1]Note that when local distance distortion is measured by the ratio between energy of the composed map and the direct distance (i.e., energy of the direct map), it is possible for $\alpha = 1$ even when the two maps are different. This is in contrast with measuring local distance distortion by the ratio between the length of the minimum path and the direct distance, in which case $\alpha = 1$ requires the minimum path to equal the direct link and thus forcing the entire tree to be an SPT.

and among these parameters, we first generate pseudolandmarks for the sample using all 120 distinct parameter combinations. Each combination can be represented by a scatterplot of 116 points (one for each tooth in the dataset) and compared using Procrustes analysis. To accomplish this, we reshape the *x–y–z* spatial coordinates of all pseudolandmarks on each tooth into a vector, and run principal component analysis (PCA) on the 116 vectors. The PCA generates 116 principal component scores for each of the 116 teeth in the dataset. Using the first three principal component scores as coordinates of the teeth creates a scatterplot in three-dimensional Euclidean space. This procedure essentially embeds an abstract metric structure into a Euclidean space of reduced dimensionality, to which a traditional landmark-based Procrustes analysis can be applied. Finally, we run generalized Procrustes analysis on these scatterplots in *morphologika*[2.5] (O'Higgins et al., 2006), and take the first 26 principal scores (the minimum number accounting for at least 95% of the variance) as a 26-dimensional feature vector encoding a parameter combination (Supporting Information, S1).

We implemented vector equivalents of one-way ANOVA, two-way ANOVA, and a linear mixed model in MATLAB (for further detail, see Supporting Information, S2). One-way ANOVA is used to evaluate the significance of each parameter. Two-way ANOVA reveals pairwise interaction effects among all parameters; we focus on interactions between method and each of the other parameters. The linear mixed model is used to determine which factors lead to shape characterizations that are most similar to the user-determined ground truth of the dataset.

Evaluating Accuracy of Globally Informed Methods

We gauge the accuracy of the novel globally informed methods in two ways. First, we compare the landmark MSEs of all new methods to the landmark MSE generated by cPDist. The performance of cPMST (without feature-fixing) relative to cPDist was chosen as a baseline for evaluating the accuracy of the globally informed methods presented here. A novel method is deemed preferable to the cPMST baseline if (1) there are fewer pairwise mappings in which the landmark MSE of the novel method is greater than landmark MSE of cPDist (i.e., there are fewer positive residuals when the landmark MSEs of a novel method are plotted against the landmark MSEs of cPDist) or (2) the mean or maximum positive residual of a novel method is less than the mean or maximum positive residual of cPMST. Either of these patterns may indicate that the novel method does not accumulate random error through intermediate correspondences.

Comparing the distribution of landmark MSEs under each method provides another way to evaluate the accuracy of different methods. For inaccurate methods, the distribution of landmark MSEs may have a higher mean or display greater variance (or both) than the landmark MSE distribution of a more accurate method. Here we compare the landmark MSE distance matrices of all main methods using a Multiple Response Permutation Procedure (MRPP), which tests for significant differences between sampling units (method + feature-fixing in this study) (Zimmerman et al., 1985). As MRPP compares within and between group dissimilarities, it is similar to multivariate analysis of variance, but does not require data to exhibit multivariate normality (Zimmerman et al., 1985), making the method appropriate for comparing MSE matrices. MRPP evaluates observed within-group distances relative to average between-group distances of two random groups; the groups may be regarded as significantly different if average within-group distances are less than average between-group distances. MRPP was performed using the *mrpp* function in the R package *vegan* (Oksanen et al., 2017). The analysis returns several values of note: observed delta (δ), the overall mean of group mean distances, weighted by the number of groups; expected delta E (δ), expected delta under the null hypothesis of no group structure; A, a chance-corrected estimate of the proportion of variance explained by group membership; and P, the significance of the test. Each MRPP analysis was run with 999 permutations. The number of permuted between-group distances that are less than the observed within-group δ determines the test's significance. Additionally, as significant between-group differences may be the result of a difference in means (location) or a difference of variance (dispersion) (Warton et al., 2012), the homogeneity of variance of each method was also compared. Analysis of multivariate homogeneity of variance was performed with the *betadisper* function in *vegan* (Oksanen et al., 2017); significance was evaluated with Tukey's Honest Significant Differences.

Owing to the positive correlation between MSE and cP distances, methods that reduce cP distances between shapes may reduce the MSE of propagated landmarks. However, this phenomenon is undesirable since reduced cP distances may imply that the method is less sensitive to shape differences. To examine the interaction of MSE and cP distance, we generated matrices of MSE scaled to cP distances ($\sqrt{}$(MSE)/cP distance), and performed MRPP and homogeneity of variance analyses on this set of matrices as well. Matrix heat maps of MSE, cP distances, and scaled MSE for each method are in Supporting Information, S2.

RESULTS

Quantifying Errors of cPDist

Visual inspection of many pairwise mappings reveals that alignment errors in cPDist occur with undesirable frequency. Of the 583 pairwise mappings inspected, 161 (27.6%) exhibit some type of error. The majority of these errors are relatively subtle and represent some landmark distortion at either the anterior or posterior end of the tooth (trigonid clustering [N = 57, 9.8%], talonid drift [41, 7.0%]). The remaining 63 mappings display major errors, either buccal–lingual inversions (33, 5.7%), anterior–posterior inversions (27, 4.6%), or 90° rotations (3, <1%).

Errors generally arose when landmarks were propagated between dissimilar teeth. To our advantage, dissimilarity can be quantified using cP distances. The cP distances between teeth with good maps averaged 0.058 (max = 0.116, min = 0.025, sd = 0.014), while cP distances between teeth with bad maps averaged 0.073 (max = 0.114, min = 0.046, sd = 0.013). The difference between these distributions is highly significant (Tukey's $Q = 17.44$, $P < 0.001$). There are also significant

TABLE 3. Mann–Whitney *U*-test results for cP distances, skewness, and Dirichlet Normal Energy (DNE) of good and bad maps

	Good maps	Trigonid clustering	Talonid drift	BL reversed	AP reversed	Rotated 90°
cP distances						
Good maps		5790	4200	1730	1690	59
Trigonid clustering	***		1099	674	570	38
Talonid drift	***	1.0000		419	383	18
BL reversed	***	0.3890	0.0779		429	29
AP reversed	***	0.8498	0.4966	1.0000		24
Rotated 90	0.1024	1.0000	0.6790	1.0000	1.0000	
	Good maps	Trigonid cluster	Talonid drift	BL reversed	AP reversed	Rotated 90°
Skewness						
Good		9319	4867	5130	4223	95
Trigonid	0.0866		907	940	720	15
Talonid	***	0.9023		509	464	19
BL reversed	0.1765	1.0000	1.0000		428	9
AP reversed	0.3627	1.0000	1.0000	1.0000		10
Rotated 90	0.1685	0.2638	0.7577	0.3308	0.5713	
	Good	Trigonid clustering	Talonid drift	BL reversed	AP reversed	
Dirichlet Normal Energy						
Good		240.5	114.5	263	347.5	
Trigonid	1.0000		8	16	14.5	
Talonid	0.3151	1.0000		12	30	
BL reversed	0.2433	1.0000	1.0000		52.5	
AP reversed	0.0941	1.0000	1.0000	1.0000		

Mann–Whitney *U* statistic in upper triangle; Bonferroni corrected *P* values in lower triangle.
Abbreviations: BL, buccolingual; AP, anteroposterior.
Significance codes: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

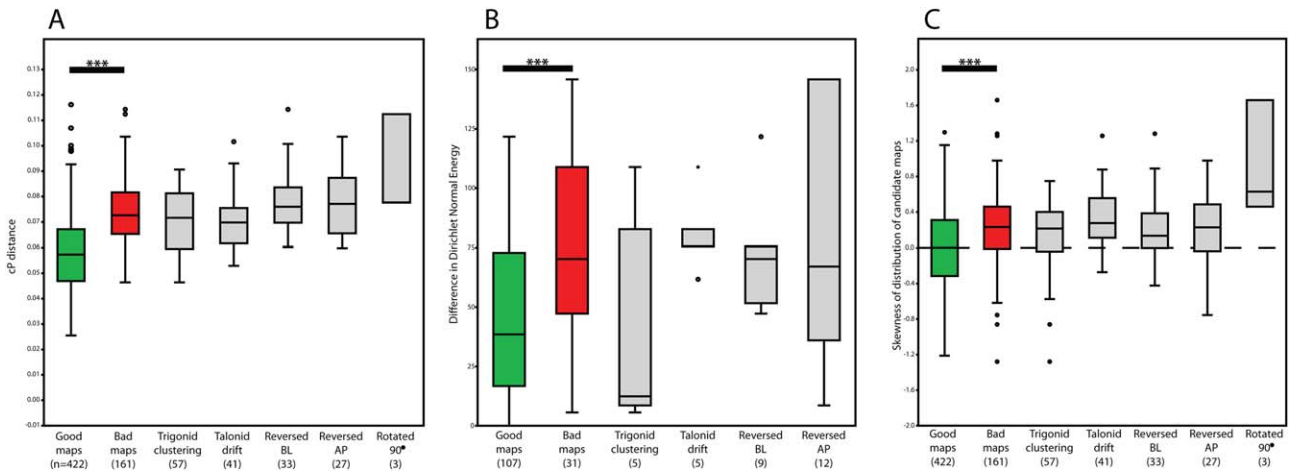

Fig. 2. Boxplots of cP distances (A), Dirichlet Normal Energy (B), and skewness of the distribution of candidate maps (C) for good maps, bad maps, and maps of each error type. Asterisks denote significant differences ($P < 0.001$) between group means. Boxes include 25%–75% quartiles; whiskers extend to furthest points <1.5 times the interquartile range. Circles indicate outliers.

differences between the cP distances of good maps and all particular error types except 90° rotation (Table 3). Although there are no significant differences in the cP distances between any two error classes, the subtler errors of trigonid clustering and talonid drift tend to occur in pairwise comparisons involving smaller cP distances than the side-to-side inversions or 90° rotation (Fig. 2a).

The difference in DNE was significantly lower (Tukey's $Q = 5.44$, $P < 0.001$) for good mappings (mean = 46.2, max = 121.7, min = 0, sd = 34.5) than for bad maps (mean = 74.6, min = 5.6, max = 145.8, sd = 41.5). However, there were no significant differences between DNE contrasts of good mappings and any particular error type (Table 3 and Fig. 2b). These results confirm our hypothesis that bad mappings occur when landmarks are propagated between highly dissimilar teeth.

Finally, we calculated the skewness of the distributions of the candidate maps for all 583 pairwise

mappings. For 422 good mappings, map distributions had a mean skewness of 0.007 (min = −1.21, max = 1.30, sd = 0.42), which is not significantly different from 0 ($P = 0.84$). The 161 bad maps had a mean skewness of 0.22 (min = −1.28, max = 1.66, sd = 0.39), which is significantly different from 0 ($P < 0.001$), indicating that the distribution of candidate maps exhibits significant positive skew when cPDist selects a bad map. Significant differences were recovered in the skewness of good and bad maps (Tukey's $Q = 7.88$, $P < 0.001$), and between good maps and those with talonid drift, trigonid clustering, and buccal–lingual inversions (Table 3 and Fig. 2c). These results confirm the prediction that good maps are much more distinct from other candidate maps than bad maps. When candidate maps are normally distributed, the map with the smallest cP distance is likely to be an accurate map. When candidate maps are positively skewed, cPDist chooses between several maps with similar cP distances. Given the plurality of possibilities and numerical error, the algorithm is more likely to select a bad map.

Evaluating Accuracy of a Minimum Spanning Tree Approach

Visual inspection of propagated landmark errors reveals that bad maps are generated when two shapes are quite dissimilar from one another. If any of the branches of the MST utilized by cPMST connect dissimilar shapes, the global map could also do a poor job propagating landmarks. Given the distribution of cP distances of all 161 observed bad maps, cP distances of <0.047 are outside the 95% confidence interval (mean = 0.073, sd = 0.013). Only one inspected bad map had a cP distance <0.047; this map exhibited trigonid clustering, a relatively minor propagation error. Of the 115 branches in the MST, only 3 branches (2.6%) had cP distances >0.047. Additionally, only one MST branch had a cP distance greater than the mean of all observed good mappings (0.062, *Galago senegalensis* K05 – *Cynocephalus volans* u16). Thus, it seems likely that propagating landmarks through the MST greatly reduces the likelihood of serious errors such as inversions or rotations. Figure 3 compares the landmark MSEs of cPDist with cPMST (without feature-fixing), and demonstrates that the latter has a much lower MSE than the former. However, as discussed above, the MST approach also has the undesirable property of increasing landmark MSE between shapes that are similar; we evaluate the accuracy of the novel globally informed methods relative to cPDist below.

Comparing Effects of Globally Informed Methods on the Characterization of Geometric Affinities

Of the five one-way ANOVAs run (Table 4), only sampling resolution was non-significant. The most significant factor was feature-fixing ($P < 0.0001$). The method of sequential comparison (MST v. LAST or Viterbi, etc.) was also highly significant ($P = 0.0005$), followed by the propagation root of pseudolandmarks ($P = 0.005$), and composedness for LAST trees ($P = 0.01$).

Two-way ANOVAs were used to determine whether interaction effects existed between any pairs of factors. Because our assessment of error in propagated user-determined landmarks was explained largely by the choice of different methods (see results of linear mixed model analysis below), we checked for interaction effects between method and each of the other factors. Significant interaction effects were found between method and feature-fixing, and method and composedness (Table 5). Neither propagation root nor sampling resolution interacted significantly with method (Table 5). To include composedness in the two-way ANOVA with method, we were limited to using only the 48 data points that included balanced distribution of method types for each of the composed groups.

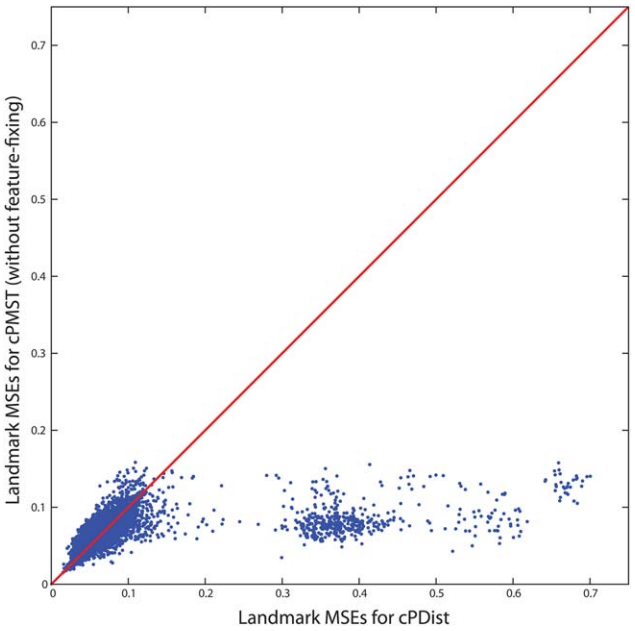

Fig. 3. Pairwise comparisons of landmark MSE generated by cPDist and cPMST (without feature-fixing). For many pairwise comparisons, cPMST (without feature-fixing) has much lower landmark MSE than cPDist. In particular, large (>0.1583) MSEs are reduced with cPMST (without feature-fixing) compare to cPDist. Red line indicates line of equivalence ($y = x$).

A linear mixed model including all factors and interaction effects (Table 6) shows a pattern of relative significance comparable to ANOVA results, although method explains more variance than feature-fixing does. Sequentially dropping nonsignificant terms and rerunning the analysis leads to a final model with three terms, included method, feature-fixing, and an interaction term between the two (Table 6). This result is strongly consistent with ANOVA results (Tables 4 and 5).

The ANOVA and linear mixed model results highlight the importance of method, feature-fixing, and their interaction. To further assess how feature-fixing influences method effects, we split the dataset into two subsets. The first subset contained analyses that used feature-fixing, while the second subset contained analyses that did *not* use feature-fixing. We then ran one-way ANOVA with method as the factor. Results of this one-way ANOVA indicate that the method type is much more significant when feature-fixing is *not* used (Table 7).

Owing to unequal sample sizes, we could not test the effect of composedness in our linear model. However,

TABLE 4. Results of one-way ANOVAs assessing parameter effects on shape space

Source	Sum sq.	DF	Mean sum sq.	*F* value	*P* value
Method					
Groups	9.0277	7	1.2897	4.0731	***
Error	35.4626	112	0.3166		
Total	44.4902	119			
Feature-Fix					
Groups	6.9913	1	6.9913	21.9998	***
Error	37.499	118	0.3178		
Total	44.4902	119			
Root					
Groups	2.867	1	2.867	8.1278	**
Error	41.6232	118	0.3527		
Total	44.4902	119			
Resolution					
Groups	0.1671	2	0.0836	0.22057	0.80239
Error	44.3231	117	0.3788		
Total	44.4902	119			
Composedness (48 methods only)					
Groups	2.1184	1	2.1184	7.097	*
Error	13.7304	46	0.2985		
Total	15.8487	47			

Significance codes: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

TABLE 5. Results of two-way ANOVAs assessing parameter effects on shape space

Source	Sum sq.	DF	Mean sum sq.	*F* value	*P* value
Method vs feature-fix					
Method	9.0277	7	1.2897	5.8952	***
Feature-fix	6.9913	1	6.9913	31.9575	***
Method:feature-fix	5.7195	7	0.8171	3.7349	***
Error	22.7518	104	0.2188		
Total	44.4902	119			
Method vs root					
Method	9.0277	7	1.2897	4.6674	***
Root	2.8670	1	2.8670	10.3758	**
Method:root	3.8587	7	0.5512	1.9949	0.0627
Error	28.7369	104	0.2763		
Total	44.4902	119			
Method vs resolution					
Method	9.0277	7	1.2897	3.5524	**
Resolution	0.1671	2	0.0836	0.2302	0.7948
Method:resolution	0.4435	14	0.0317	0.0873	1.0000
Error	34.8519	96	0.3630		
Total	44.4902	119			
Method vs composedness (48 methods only)					
Method	1.5290	1	1.5290	6.3377	*
Composedness	2.1184	1	2.1184	8.7806	**
Method:composedness	1.5862	1	1.5862	6.5746	*
Error	10.6152	44	0.2413		
Total	15.8487	47			

Significance codes: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

composedness was a significant parameter in one-way ANOVA and exhibited significant interaction effects with method. Splitting the dataset into composed and uncomposed subsets and running one-way ANOVAs with method as a factor indicates that method is more significant on uncomposed maps (Table 7).

Evaluating Accuracy of Globally Informed Methods

With the exceptions of LAST ($\alpha =$ balance) and the Viterbi methods (with or without feature-fixing), all of the proposed globally informed methods have much lower maximum landmark MSEs than cPDist, suggesting that these globally informed methods successfully avoid large-scale misalignments and correspondingly elevated landmark MSE between user-determined and propogated landmarks (Table 8). As LAST ($\alpha =$ balance) and the Viterbi methods (with or without feature-fixing) have maximum landmark MSEs that are comparable to the maximum landmark MSE of cPDist, these methods may be susceptible to misalignments of similar magnitude as cPDist. Among the novel methods with much lower maximum landmark MSEs than cPDist, only

TABLE 6. Linear mixed model for assessment of parameters explain variance from user-based landmark approach

Factor	Sum sq.	DF	Mean sum sq.	*F* value	*P* value
Initial model					
Method	0.4086	7	0.0584	4.7011	***
Feature-fix	0.0966	1	0.0966	7.7771	**
Method:feature-fix	0.2452	7	0.0350	2.8211	*
Root	0.0402	1	0.0402	3.2365	0.0758
Resolution	0.0349	2	0.0174	1.4040	0.2516
Method:root	0.0704	7	0.0101	0.8102	0.5814
Method:resolution	0.0541	14	0.0039	0.3113	0.9911
Error	0.9934	80	0.0124		
Final model					
Method	0.4086	7	0.0584	5.0890	***
Feature-fix	0.0966	1	0.0966	8.4187	**
Method:feature-fix	0.2452	7	0.0350	3.0538	**
Error	1.1930	104	0.0115		

The initial model included all four factors (method, feature-fixing, root, and resolution) and interactions between method and the other three factors. Nonsignificant factors were sequentially dropped to arrive at the final model.
Significance codes: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

TABLE 7. Conditional ANOVAs for assessment of interaction effects between method and feature-fixing, and method and composedness

Source	Sum sq.	DF	Mean sum sq.	*F* value	*P* value
Methods conditioned with feature-fixing					
Groups	7.3487	7	1.0498	3.3596	**
Error	16.2492	52	0.3125		
Total	23.598	59			
Methods conditioned without feature-fixing					
Groups	7.3984	7	1.0569	8.4520	***
Error	6.5026	52	0.1250		
Total	13.901	59			
Methods conditioned on composed maps					
Groups	1.5079	1	1.5079	5.4329	*
Error	6.1061	22	0.2775		
Total	7.614	23			
Methods conditioned on uncomposed maps					
Groups	1.6073	1	1.6073	7.8418	***
Error	4.5092	22	0.2050		
Total	6.1164	23			

Significance codes: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

composed LAST (α = median; no feature-fixing) has a lower maximum landmark MSE value than the performance baseline established by cPMST (without feature-fixing) (maximum landmark MSEs = 0.1507 and 0.1583, respectively). Still, many other methods that do not utilize feature-fixing (e.g., LAST [α = mean], composed LAST [α = balance], composed LAST [α = mean]) have similarly low maximum landmark MSEs (Table 8).

Two trends are apparent when feature-fixing is implemented: (1) methods with feature-fixing have greater maximum landmark MSE than the same method without feature-fixing, and (2) methods with feature-fixing are appear to be less susceptible to the accumulation of numerical error during landmark propagation, as indicated by the percentage of points above the line $y = x$ (Table 8). These positive residuals indicate pairwise mappings in which landmark MSE has increased for the novel method relative to the landmark MSE observed for cPDist (Fig. 3 provides an example for cPMST). While no single method has the lowest mean, median, or maximum positive residual, the set of cPMST, LAST (α = mean), and three composed LAST (α = balance, mean, or median) methods share low positive residuals with low dispersion (Table 8). When feature-fixing is implemented with these methods, both the magnitude and dispersion of these residuals increase but the number of positive residuals decreases. The decrease in the number of positive residuals is observed for all methods except the Viterbi methods.

Compared to cPMST, Viterbi methods (with or without feature-fixing) have smaller median positive residuals and larger mean positive residuals. Based on the percentage of points above $y = x$, Viterbi methods (with or without feature-fixing) experience less inflation of landmark MSE during landmark propagation (Table 8). However, all Viterbi methods have large maximum positive residuals and high variance in the distribution of these residuals. Furthermore, while the Viterbi methods have the greatest number of points below $y = x$ (indicating that cPDist has greater landmark MSE), the means and

TABLE 8. Summary statistics of differences in landmark MSE under each globally informed method relative to cPDist

Method	Variant	Feature-fixing	Maximum landmark MSE	% of points above $y = x$	Positive residuals				% of points below $y = x$	Negative residuals			
					Mean	Median	Maximum	SD		Mean	Median	Maximum	SD
cPDist	-	N	0.7003	na	na	na	na	na	na	na	na	na	na
cPMST	-	N	0.1583	56.0	**0.0092**	**0.0070**	0.0645	0.0082	42.6	0.0629	0.0082	0.5785	0.1239
LAST	Mean	N	0.1617	55.2	0.0115	0.0082	0.0787	0.0106	43.4	0.0623	0.0086	0.5785	0.1230
LAST	Balance	N	0.6901	55.2	0.0170	0.0059	0.6048	0.0641	43.3	0.0526	0.0060	0.5500	0.1099
Composed LAST	1	N	0.4488	**66.3**	0.0313	0.0106	0.3844	0.0721	**32.4**	0.0628	0.0093	0.6033	0.1272
Composed LAST	Balance	N	0.1583	56.0	**0.0092**	**0.0070**	0.0645	0.0082	42.6	0.0629	0.0082	0.5785	0.1239
Composed LAST	Mean	N	0.1627	65.3	0.0164	0.0145	0.0652	0.0123	33.3	**0.0773**	**0.0103**	0.5785	**0.1350**
Composed LAST	Median	N	**0.1507**	57.6	0.0098	0.0078	**0.0573**	**0.0080**	41.0	0.0653	0.0085	0.5785	0.1261
Viterbi	0.0	N	0.6801	35.1	0.0169	0.0066	0.4930	0.0455	41.8	0.0479	0.0072	0.5972	0.1101
Viterbi	0.25	N	0.6679	35.4	0.0131	0.0060	0.4658	0.0366	42.9	0.0461	0.0072	0.6026	0.1071
Viterbi	0.5	N	0.6801	35.9	0.0131	0.0060	0.4658	0.0362	42.7	0.0461	0.0071	**0.6165**	0.1080
cPMST	-	Y	0.1854	45.8	0.0099	0.0076	0.0752	0.0088	54.2	0.0528	0.0096	0.5886	0.1125
LAST	Mean	Y	0.1968	47.9	0.0109	0.0082	0.0810	0.0100	52.1	0.0548	0.0099	0.5886	0.1141
LAST	Balance	Y	0.6737	43.5	0.0192	0.0072	0.5869	0.0655	56.5	0.0443	0.0091	0.5464	0.0975
Composed LAST	1	Y	0.4750	55.5	0.0341	0.0102	0.3825	0.0762	44.5	0.0488	0.0091	0.6085	0.1110
Composed LAST	Balance	Y	0.1854	45.8	0.0099	0.0076	0.0752	0.0088	54.2	0.0528	0.0096	0.5886	0.1125
Composed LAST	Mean	Y	0.1887	52.1	0.0118	0.0092	0.0726	0.0102	47.9	0.0578	0.0101	0.5967	0.1174
Composed LAST	Median	Y	0.1796	48.3	0.0105	0.0083	0.0809	0.0091	51.7	0.0546	0.0095	0.5886	0.1146
Viterbi	0.0	Y	0.6844	44.2	0.0147	0.0068	0.4872	0.0396	55.8	0.0385	0.0087	0.6129	0.0964
Viterbi	0.25	Y	0.6163	43.5	0.0124	0.0066	0.4876	0.0325	56.5	0.0378	0.0088	0.5949	0.0944
Viterbi	0.5	Y	0.6844	43.7	0.0124	0.0065	0.4876	0.0321	56.3	0.0378	0.0087	0.6117	0.0950

Positive residuals indicate the globally informed method has greater landmark MSE than cPDist, whereas negative residuals indicate globally informed method has smaller landmark MSE than cPDist. "Maximum landmark MSE" indicates the inflection point above which the globally informed method always decreases MSE relative to cPDist. Note that the % of points above or below the line $y = x$ may not sum to 100, as some points may be on the line. Bold text represents maximum or minimum values as appropriate. SD, standard deviation.

TABLE 9. Summary statistics of pairwise cPdistances, landmark MSE, and scaled MSE by method

Method	Variant	Feature-fixing	cP distance			Landmark MSE			Scaled MSE		
			Mean	Maximum	SD	Mean	Maximum	SD	Mean	Maximum	SD
cPDist	-	-	0.0665	0.1599	0.0202	0.0859	0.7003	0.0962	4.1559	13.1445	1.3491
cPMST	-	N	0.0646	0.1360	0.0177	0.0645	0.1583	0.0215	3.9989	11.8016	0.9091
LAST	Mean	N	0.0702	0.1373	0.0211	0.0654	0.1617	0.0223	3.7563	11.8016	0.9511
LAST	Balance	N	0.0645	0.1470	0.0186	0.0727	0.6901	0.0764	4.0782	11.8016	1.0101
Composed LAST	1	N	0.0699	0.1402	0.0186	0.0863	0.4488	0.0753	4.0415	11.8016	1.1308
Composed LAST	Balance	N	0.0646	0.1360	0.0177	0.0645	0.1583	0.0215	3.9989	11.8016	0.9091
Composed LAST	Mean	N	0.0682	0.1422	0.0179	0.0710	0.1638	0.0222	3.9690	11.8016	0.8617
Composed LAST	Median	N	0.0652	0.1360	0.0174	0.0650	**0.1555**	**0.0214**	3.9757	11.8016	0.8951
Viterbi	0.0	N	0.0581	0.1205	0.0145	0.0719	0.6800	0.0616	4.4542	13.1445	1.1375
Viterbi	0.25	N	0.0582	0.1207	0.0150	0.0709	0.6679	0.0609	4.4364	13.1445	1.1354
Viterbi	0.5	N	0.0581	0.1222	0.0149	0.0710	0.6800	0.0612	4.4432	13.1070	1.1312
cPMST	-	Y	0.0771	0.1788	0.0216	**0.0620**	0.1854	0.0230	3.2714	8.4501	0.7432
LAST	Mean	Y	**0.0805**	0.1727	**0.0235**	0.0628	0.1968	0.0239	**3.1649**	8.4501	0.7548
LAST	Balance	Y	0.0764	0.1696	0.0214	0.0694	0.6737	0.0722	3.3382	**8.3625**	0.8365
Composed LAST	1	Y	0.0797	**0.1872**	0.0214	0.0832	0.4750	0.0759	3.4331	8.7927	0.9775
Composed LAST	Balance	Y	0.0771	0.1788	0.0216	**0.0620**	0.1854	0.0230	3.2714	8.4501	0.7432
Composed LAST	Mean	Y	0.0795	0.1777	0.0213	0.0646	0.1887	0.0235	3.2285	8.4501	**0.7276**
Composed LAST	Median	Y	0.0777	0.1745	0.0213	0.0630	0.1796	0.0232	3.2636	8.4501	0.7404
Viterbi	0.0	Y	0.0742	0.1755	0.0203	0.0711	0.6844	0.0612	3.4847	12.1532	0.9508
Viterbi	0.25	Y	0.0742	0.1595	0.0202	0.0701	0.6163	0.0608	3.4681	12.1532	0.9571
Viterbi	0.5	Y	0.0743	0.1644	0.0202	0.0702	0.6844	0.0612	3.4679	10.6384	0.9497

Bold text represents maximum or minimum values as appropriate.
Abbreviations: MSE, mean square error; SD, standard deviation.

medians of these negative residuals are smaller and the standard deviations are lower than the values recovered for cPMST. So while Viterbi methods experience less inflation of landmark MSE due to the accumulation of numerical error, these methods are likely to be more prone to large-scale misalignments (similar to cPDist). Essentially, compared to all other methods, landmark MSEs of the Viterbi methods have higher correlations with the landmark MSE of cPDist, which can be seen in bivariate plots of the landmark MSEs of these methods (Supporting Information, S2 Appendix).

Results from multivariate homogeneity of variance tests confirm significant differences in the dispersion of landmark MSE under different methods (Supporting Information, S1 and S2). cPDist has the highest variance (measured as the Euclidean distance from each matrix entry to the centroid of the matrix), and all Viterbi methods exhibit similarly high dispersion. The variance of landmark MSEs for cPDist and all Viterbi methods is significantly greater than any other method, and cPDist has significantly greater variance than all Viterbi methods (Supporting Information, S1 and S2). A slightly different pattern of variance emerges when MSEs are scaled to pairwise cP distances (Supporting Information, S1 and S2). cPDist maintains the greatest dispersion, but all methods that utilize feature-fixing have reduced variance relative to their non-feature-fixing counterparts (Supporting Information, S1 and S2). In most cases, pairwise comparisons of the same method with and without feature-fixing produce significant differences in terms of variance; methods with feature-fixing always have lower variance than methods without. Viterbi methods without feature-fixing also have relatively high dispersion, similar to the unscaled MSE results.

The observed heterogeneity of variance renders any test of significant differences in method means suspect. With this caveat, MRPP indicates there are significant differences between methods, as gauged by both MSE and scaled MSE comparisons (MSE: $\delta = 0.47$, $E(\delta) = 0.60$, $A = 0.21$, $P = 0.001$; scaled MSE: $\delta = 12.8$, $E(\delta) = 14.64$, $A = 0.13$, $P = 0.001$). Both within- and between-method MSE and scaled MSE are detailed in Supporting Information, S1. To compensate for the observed heterogeneity of variance, we also performed MRPP analysis for a reduced sample of MSE matrices, excluding all methods with significantly higher dispersion (cPDist, all six Viterbi methods). In this analysis, significant differences between methods were still recovered ($\delta = 0.31$, $E(\delta) = 0.47$, $A = 0.34$, $P = 0.001$). The observed patterns of heterogeneity in scaled MSE comparisons were too diffuse to permit a similarly restricted analysis.

Table 9 provides the summary statistics for cP distances, landmark MSE, and scaled MSE for all methods. cPMST (with feature-fixing) has the lowest mean landmark MSE (0.062), while LAST (α = mean, feature-fixing) has the highest mean cP distance and the lowest mean scaled MSE (3.165). When considering only those methods without feature-fixing, the set of cPMST, LAST (α = mean), and three composed LAST (α = balance, mean, or median) methods share similar characteristics: moderate mean and maximum cP distances, low mean and maximum landmark MSEs, and the lowest mean and maximum scaled MSEs (Table 9). Compared to other methods without feature-fixing, this set of methods also exhibits lower standard deviations for landmark and scaled MSE.

There are several trends across methods when feature-fixing is implemented. Relative to the same method without feature-fixing, the mean, maximum, and standard deviation of cP distances increase for all methods with feature-fixing (Table 9). For scaled MSE, the

mean, maximum, and standard deviation all decrease with feature-fixing. For landmark MSE, not all methods exhibit the same trends when feature-fixing is implemented. While mean landmark MSE decreases with feature-fixing for all methods, maximum landmark MSE increases for all methods except LAST (α = balance) and Viterbi (angle weight = 25%). The variance of landmark MSE increases for all methods except LAST (α = balance) and all three Viterbi methods (Table 9). Finally, while the Viterbi methods (with or without feature-fixing) have the lowest mean scaled MSE, they share the highest maximum scaled MSE with cPDist, and have high variance relative to other methods (Table 9).

DISCUSSION

Quantifying Errors of cPDist

When comparing highly dissimilar shapes, the pairwise correspondence found by cPDist is particularly error-prone in terms of landmark MSE. This substantial error rate presents a minor paradox: despite the errors lurking in the dataset, Boyer et al. (2011) still recovered a strong correlation between user-determined and automatically determined distances and had a nearly equivalent success rate in taxonomic classification. We reconcile this paradox by noting that errors of cPDist occur most often between geometrically dissimilar shapes, resulting in large cP distances due to the restricted search space (conformal maps plus TPS) as opposed to faithfully capturing the geometric dissimilarity. Consequently, cPDist gets the "right answer" for the wrong reasons when comparing very different shapes. Furthermore, the classification method used in Boyer et al. (2011) assessed only whether small cP distances are taxonomically reliable. As we have confirmed that only large cP distances are unreliable, there is no real contradiction between our results and Boyer et al.'s (2011) assessment. However, our findings highlight that the approach of Boyer et al. (2011) does not provide sufficiently informative pairwise correspondences, which are essential for geometric morphological studies.

Evaluating Accuracy of a Minimum Spanning Tree Approach

As all edges of the MST have smaller cP distances than the minimum cP distance of the observed bad maps, it is quite likely that using an MST sufficiently addresses the issue of aligning structures with very different morphologies. In our test case, branches within the MST were small enough to alleviate concerns of major misalignments. However, more work should be done on a wider variety of data sets to determine more precisely the sample properties (e.g., the morphological gaps between objects or the range of morphologies) that increase risk for misalignments even with MST. Additionally, cPMST has the unfortunate side effect of increasing landmark propagation error (reflected in the cP distances) when two similar shapes are not directly connected by an edge of the MST. This issue is partially addressed by subsequent tree-based methods and is discussed further below.

Comparing Effects of Globally Informed Methods on the Characterization of Geometric Affinities

Using ANOVA and linear mixed models, we ascertained the effects of five different factors on the characterization of shape affinities by the globally informed approaches developed for this study (Tables 4–6). The strongest effects were produced by the choice of method and whether or not feature-fixing was used. These two factors had an interaction effect such that including feature-fixing in the protocol reduced the impact of method. From this information alone, it is unclear if feature-fixing is beneficial. Reduction of the method effect through feature-fixing may be beneficial if the results became both more consistent and biologically meaningful. On the other hand, the observed statistical effect may be a consequence of feature-fixing increasing random error, and thus leading to more variable and less biologically meaningful results. To determine which is more likely for this particular dataset, we assess the shape space characterization of a few example methods below.

The root for pseudolandmark propagation had little obvious effect on the results (Tables 4–6). Our analyses suggest that the best approach is to use the specimen with the minimum average difference from all other specimens in the collection as the starting point for landmark propagation. Thought the root shape chosen this way is *not* the same among all 120 analyses, no significant interaction between the root and method was recovered (Tables 5 and 6). If root had a strong direct effect, we should observe a significant interaction between root and method (as the root was different under each method). Though the effect of the root was minimal in this analysis, it could still be an important parameter for other datasets, particularly those that include highly dissimilar shapes (e.g., if the combined astragalus and calcaneus dataset of Boyer et al., 2015a was analyzed with a conformal method).

For our dataset, the impact of composed versus uncomposed maps seems more important than the root, as the composedness has a significant interaction with method. However, as stated above, it could not be included in the linear model due to unequal sample sizes (and unbalanced experiment design) relative to other factors. One-way ANOVA run on method using a composed/uncomposed split sample suggests that composed maps tend to decrease differences between methods. Therefore, it is probably more desirable to account for cP distances measured for composed maps when sequentially concatenating pairwise correspondences between similar shapes (unless there is a reason to prefer properties of an individual method).

Finally, pseudolandmark sampling density had virtually no effect. This is highly encouraging, and suggests that relatively fewer pseudolandmarks may be used with conformal approaches to decrease the computational intensity of downstream analyses.

Evaluating Accuracy of Globally Informed Methods

The alternative tree-based approaches presented here were developed primarily to minimize the accumulation

of numerical error as landmarks are propagated through a tree. Based on the analysis of landmark MSE residuals above, a set of methods perform similar to cPMST, including LAST (α = mean) and three composed LAST (α = balance, mean, or median) methods. While none of these approaches has substantially better performance than cPMST, composed LAST (α = median) does exhibit fewer positive landmark MSE residuals, a lower maximum positive residual, and lower variance of these residuals (Table 8), indicating that it may be somewhat preferable to cPMST. In addition, with or without feature-fixing, composed LAST (α = median) exhibits a higher mean cP distance and lower scaled MSE than cPMST (Table 9).

Because feature-fixing aims to match geometric characteristics on the shapes (e.g., cusp tips or basins) and many user-determined landmarks lie near these positions, we expected feature-fixing to reduce landmark MSE. However, based on the analysis of MSE residuals, feature-fixing appears to have a mixed effect on landmark MSE. The procedure tends to increase maximum landmark MSE, but decrease the number of positive MSE residuals (for all except the Viterbi methods) (Table 8). Furthermore, because feature-fixing increases cP distances, scaled MSEs are lower when feature-fixing is implemented (Table 9). Thus, based on landmark MSE, the potential benefits of feature-fixing are ambiguous.

Multivariate homogeneity of variance tests performed on MSE matrices (S1 Appendix) reveal that cPDist and all six Viterbi methods have significantly greater variance than any other method. This result is not surprising, as visual inspection of pairwise cPDist mappings revealed multiple instances of propagation error (reflected by high landmark MSEs). Reduced variance in the MST-based methods suggests that erroneous mappings have been largely eliminated, which was a primary goal in developing these subsequent tree-based methods. The Viterbi methods are different from other tree-based methods as they aim to minimize the distance functional but sacrifice global transitivity.

Homogeneity of variance tests performed on the scaled MSE matrices reveal a pattern not seen in the MSE matrices (Supporting Information, S1). With scaled MSE, all methods utilizing feature-fixing have lower variance than their counterparts without feature-fixing. The variance of Viterbi methods utilizing feature-fixing is similar to the variance of non-Viterbi methods without feature-fixing. In general, scaled MSE distances are substantially reduced since feature-fixing tends to increase pairwise cP distances but not landmark MSEs. This effect can also be seen in the matrix heat maps (Supporting Information, S2 Appendix), and supports the significance of feature-fixing as an important parameter for the performance of the tree-based improvement methods.

Quantitative Comparison of Ordinations Generated by Globally Informed Methods

The previous analyses permit the identification of those methodological parameters with strong effects on shape space characterization, but they are not informative regarding certain aspects (e.g., is feature-fixing beneficial?). Our final analysis compares the ordinated shape spaces of those globally informed methods that

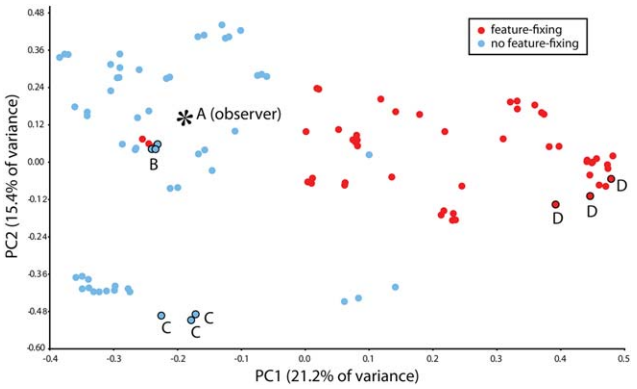

Fig. 4. Results of principal component analysis comparing different parameter combinations of automated alignment and mapping. Each point represents an automated analysis of 116 tooth surfaces from Boyer et al. (2011). Points that plot close together represent analytic protocols yielding similar representations of shape affinities for the surfaces in the test dataset. Note that only around 36% of the total variance is represented on these first two principal components. As confirmed by the statistical analyses detailed in the text, this plot indicates feature-fixing has the strongest effect on shape affinities among sampled teeth. To our surprise, analyses without feature-fixing characterize shape affinities in a way more similar to the user-based "ground truth." Three treatments are examined in more detail: (A) user-determined landmarks; (B) GLAM1 (LAST; α = balance; no feature-fixing), the approach most similar to the "ground-truth"; (C) GLAM2 (LAST; composed; α = balance; no feature-fixing); and (D) GLAM3 (LAST; composed; α = balance; feature-fixing). The separation between A and both C and D suggest that the latter two approaches characterize shape affinities in a distinct manner relative to A. As each of these treatments was run with three different pseudolandmark resolutions, each treatment is represented three times. The relatively minor variance of each treatment under differing resolutions demonstrates that pseudolandmark sampling has little effect on shape space characterization (also confirmed by statistical analyses in the text).

are most and least similar to the shape space characterized by user-determined landmarks. Principal components analysis of the vectors encoding method parameter combinations (Supporting Information, S1) reveals these methods (Fig. 4), which are subsequently referred to as GLobally-informed Automated Method 1 (GLAM1: LAST; α = balance; no feature-fixing), GLAM2 (LAST; composed; α = balance; no feature-fixing), and GLAM3 (LAST; composed; α = balance; feature-fixing). In addition to these three novel methods, we also compare the shape spaces generated by user-determined landmarks and auto3Dgm.

To compare the ordinated shape spaces of these five methods, we first identified 18 phylogenetically cohesive taxonomic groups (Supporting Information, S1) that were fairly distinctive when visualized on the first two principal components of the user-determined landmarks (Fig. 5a). We then ran one-way ANOVAs on a vector representing the first 46 principal component scores (Supporting Information, S1). As might be expected from the taxonomic separation apparent in Figure 5, we found all samples to have highly significant interspecific variance (Table 10), with the highest significance level for the user-determined landmarks. Treatments more similar to the user-determined landmarks in Figure 5 (GLAM1 and GLAM2) had higher significance levels than the treatment further removed from the ground truth

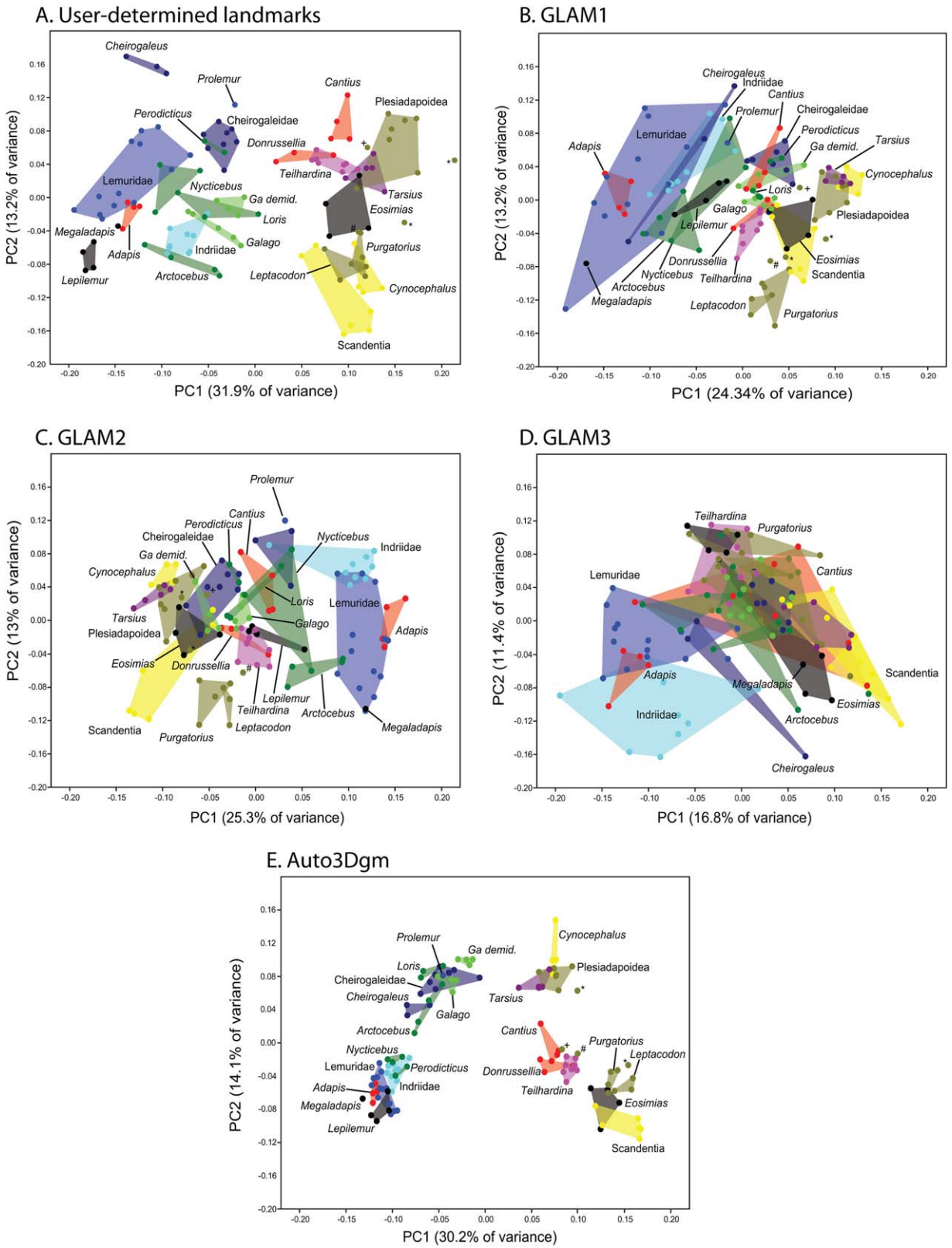

Fig. 5. Principal components analyses characterizing shape affinities in a sample of 116 teeth using five alternative approaches. Approaches include (A) user-determined landmarks; (B) GLAM1 (LAST; α = balance; no feature-fixing), the approach most similar to the "ground-truth" in Figure 4; (C) GLAM2 (LAST; composed; α = balance; no feature-fixing); (D) GLAM3 (LAST; composed; α = balance; feature-fixing); and (E) auto3Dgm. Neither C nor D was expected to look similar to A or B based on Figure 4. Minimum convex polygons include individual specimens of closely related species expected to be similar based on visual inspection and traditional comparative analyses. The degree to which each method successfully distinguished groups was evaluated with a series of vector ANOVAs in which taxonomic groups shown in these images were the treatment effects. From these analyses, it appears that A, B, and E do the best job of separating taxonomic groups (Table 10). The specific surfaces used in each group and the data for each analysis is provided (Supporting Information, S1). Exemplar teeth of each group are shown in Figure 7.

TABLE 10. One-way ANOVA on taxomonic groups to assess which parameter combinations result in shape spaces with greatest between-group distinctiveness

Source	Sum sq.	DF	Mean sum sq.	*F* value	*P* value
User-determined landmarks					
Groups	3.0454	17	0.1791	17.2013	***
Error	0.9894	95	0.0104		
Total	4.0347	112			
GLAM1 (LAST; α = balance; no feature-fixing)					
Groups	1.5673	17	0.0922	9.0052	***
Error	0.9726	95	0.0102		
Total	2.5399	112			
GLAM2 (LAST; composed; α = balance; no feature-fixing)					
Groups	1.7827	17	0.1049	10.4511	***
Error	0.9532	95	0.0100		
Total	2.7359	112			
GLAM3 (LAST; composed; α = balance; feature-fixing)					
Groups	2.1162	17	0.1245	5.8937	***
Error	2.0065	95	0.0211		
Total	4.1228	112			
Auto3Dgm					
Groups	2.5471	17	0.1498	15.5601	***
Error	0.9148	95	0.0096		
Total	3.4618	112			

Significance codes: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$.

(GLAM3). Though auto3Dgm does not specify explicit maps relating surfaces, its ordinated shape space appears to be most similar to the user-based result in this limited (though potentially representative) analysis.

User-determined landmarks recover substantially more between-taxon variance than globally informed methods (Table 10). The increase in between-group variance suggests that user-determined landmarks are better at capturing between-taxon differences. It is quite likely that user-determined landmarks focus on those features that exhibit a large amount of between-group variance, so that morphological expertise permits more variance to be captured by relatively fewer landmarks. This focus has been argued to be the primary advantage that semi-automated methods have over fully automated alternatives (Gunz and Mitteroecker, 2013). However, as the manual landmarks were collected with *a priori* taxonomic knowledge, it is also possible the researcher inadvertently biased their landmarking to maintain intraspecific consistency. Repeating manual landmark data collection would mitigate such bias and insure the absence of such effects. However, to avoid adding more sources of error, any additional data collection would have to be done by a researcher with equivalent anatomical expertise, which risks introducing a similar bias if they remember species-specific morphological patterns well. A possible approach would be to collect landmark data in random taxonomic order over a widely spaced time interval.

Finally, we were surprised to find that feature-fixing generally reduced the similarity between globally informed methods and user-determined landmarks (Fig. 4). As user-determined landmarks are often close to (but not necessarily exactly coincident with) positions of extreme geometric configuration, the discrepancy between user-determined and feature-fixed landmarks can be as large as several edges away on the discretized triangular mesh. Such a difference is comparable to the magnitude of some landmark MSEs between very similar shapes. Additionally, the TPS procedure involved in application of feature-fixing does not control for shape distortion at regions lacking anchor points, though some user-determined landmarks are positioned in such regions. Therefore, alternatives to TPS such as bounded distortion (Lipman, 2012) or quasi-conformal (Meng and Lui, 2015) maps may be preferable for geometric morphological analysis, as these methods guarantee low distortion on the regions of shapes even without geometric characteristics.

Qualitative Assessment and Biological Implications of Ordinations Generated by Globally Informed Methods

To understand how geometric affinities of particular taxonomic groups differ qualitatively across methods, we now focus on the details exhibited by each ordination. Figure 6 provides the terminology for some notable features of therian (marsupial and placental mammals) mandibular molars, and Figure 7 provides the examples of these teeth for the sample used in this study.

For all methods, the presence and prominence of the paraconid, the most anterior cusp of the tooth, drives variation on PC1 (Fig. 5). User-determined landmarks and auto3Dgm are most similar in this pattern: for these two methods, there is a slight gap between the distributions of extant strepsirrhines (all lack a paraconid) plus *Adapis* (which has a relatively small paraconid) and tarsiers, non-primates, and the remaining fossil taxa (all of which have prominent paraconids) (Fig. 7). This distinction blurs in the ordinations of globally informed methods as those strepsirrhines with relatively prominent trigonids invade the space of paraconid-bearing taxa. Specifically, lorisiforms (galagos, *Nycticebus, Perodicticus, Loris*, and *Arctocebus*), *Lepilemur*, and all cheirogaleids except *Cheirogaleus* (which has a strongly reduced trigonid) overlap with early fossil euprimates (*Teilhardina, Cantius*, and *Donrussellia*), certain plesiadapiforms (e.g., *Pronothodectes*), and the treeshrew *Ptilocercus* in the three globally informed methods. Among

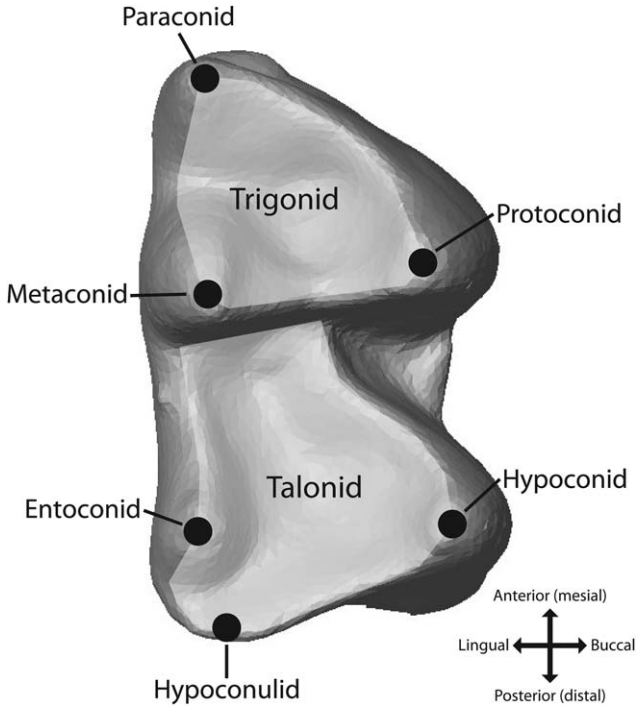

Fig. 6. Nomenclature of the primary features of a therian mandibular molar. Occlusal view.

these methods, the overlap is least pronounced in GLAM1, which also happens to be the globally informed method that is closest to user-determined landmarks in Figure 4.

PC2 of the user-determined landmark plot (Fig. 5a) separates teeth with a hypoconulid that is close to the entoconid and substantially posterior to the hypoconid (e.g., *Lepilemur* and *Tupaia*) from teeth with a more mesially and buccally positioned hypoconulid (e.g., *Cheirogaleus*). Trends of variation in PC2 of some globally informed methods are similar to the user-determined landmarks in certain respects. However, none of the automated 3DGM methods reflect variation in the talonid cusps on PC2, as indicated by the different positions of *Cynocephalus*, galagos, and certain lorisids (these taxa all have strongly lingually positioned hypoconulids). The auto3Dgm plot, with the smallest within-group distributions, is most similar to the hypoconulid-driven trend of the user-determined landmarks, but differs in the recovered overlap of galagos and cheirogaleids. In contrast, the user-determined landmarks pull galagos toward lemurids and indriids. Similarly, the interspecific distribution of lorisids is inverted between the user-based approach and auto3Dgm. With user-determined landmarks, *Arctocebus* and *Loris* overlap more with indriids and lemurids, while *Perodicticus* and *Nycticebus* overlap with cheirogaleids. In auto3Dgm, the trend is reversed (though galagos overlap most extensively with *Arctocebus* and *Loris* in both methods). In fact, auto3Dgm shows tight clustering of these taxa, revealing strong affinities of *Nycticebus* and *Perodicticus* with indriids and lemurids, and affinities of *Loris* and *Arctocebus* with cheirogaleids and galagos. From a phylogenetic perspective, these alternative groupings are not intuitive. However, from a functional perspective, the user-determined landmarks group frugivorous lorisids (e.g., *Perodicticus*) and omnivorous cheirogaleids (e.g., *Microcebus*) in one region and insectivorous lorisiforms (e.g., *Arctocebus* and galagos) with folivorous indriids (e.g., *Avahi*) in another.

The globally informed method most similar to the user-based approach in Figure 4 (GLAM1, Fig. 5b) does not exhibit any obvious trends on PC2 for strepsirrhines, but shows a distribution of the four lorisids more consistent with a dietary interpretation (low relief frugivorous lorises overlapping with omnivorous cheirogaleids, and high-relief insectivorous lorisiforms overlap with insectivorous galagids and folivorous indriids). In other automated methods, taxonomic groups overlap too extensively for succinct description, including among lorisids. Nonetheless, all three of the globally informed methods (Fig. 5b–d) preserve separation between galagos and cheirogaleids (like user-determined landmarks but unlike auto3Dgm) while also maintaining a large separation between galagos and indriids (unlike user-determined landmarks). Therefore, from a phylogenetic perspective, the globally informed methods return more intuitive results than either user-determined landmarks or auto3Dgm.

The relative positions of nonprimate taxa and primitive fossil primates are similar in the five example plots (Fig. 5). Extreme values are typically tupaiid treeshrews (an extant nonprimate that is insectivorous), and tend to cluster close to *Leptacodon* (a fossil nonprimate) and *Purgatorius* (the oldest and most basal known stem-primate). *Ptilocercus*, a more omnivorous treeshrew, is usually separated from this cluster. The ambiguous fossil taxon *Altanius* (either euprimate or stem-primate) plots near this cluster but generally has less extreme PC2 values. Eosimiidae, a group of purported stem-anthropoid primates, also plots near the tupaiid-*Leptacodon*-*Purgatorius* cluster with at least one individual always plotting close to *Ptilocercus*. Early euprimates *Cantius*, *Teilhardina*, and *Donrussellia* plot near one another with slightly less extreme PC1 and PC2 scores. Stem-primates more derived than *Purgatorius* (*Paromomys, Plesiolestes, Pronothodectes, Chronolestes, Elphidotarsius*) tend to plot separately from the early euprimates (but also further from treeshrews, *Purgatorius*, Eosimiidae, *Altanius*, etc.) while overlapping *Tarsius* (an extant haplorrhine). The non-primate *Cynocephalus* plots with treeshrews in the user-based method, a result consistent with these taxa being outgroups to Primates. Similarities between *Cynocephalus* and *Tupaia* in the user-based result are driven by the strongly lingual and posterior position of the hypoconulid. In contrast, all of the globally informed methods show *Cynocephalus* overlapping primarily with *Tarsius*, linking this dermopteran to a crown haplorhine. The automated results seem to reflect gross similarities between *Cynocephalus* and *Tarsius* (e.g., both taxa have relatively square occlusal outlines), but they are not phylogenetically intuitive.

The most intriguing result generated by these ordinations is the position of the Eocene primate *Eosimias*. Although the family Eosimiidae is typically assigned to anthropoid primates (Beard et al., 1994, 1996; Gebo et al., 2000; Beard and Wang, 2004; Bajpai et al., 2008), this assessment has been questioned (Godinot and Mahboubi, 1994; Simons and Rasmussen, 1994; Simons, 1995; Gunnell and Miller, 2001; Miller et al., 2005). All

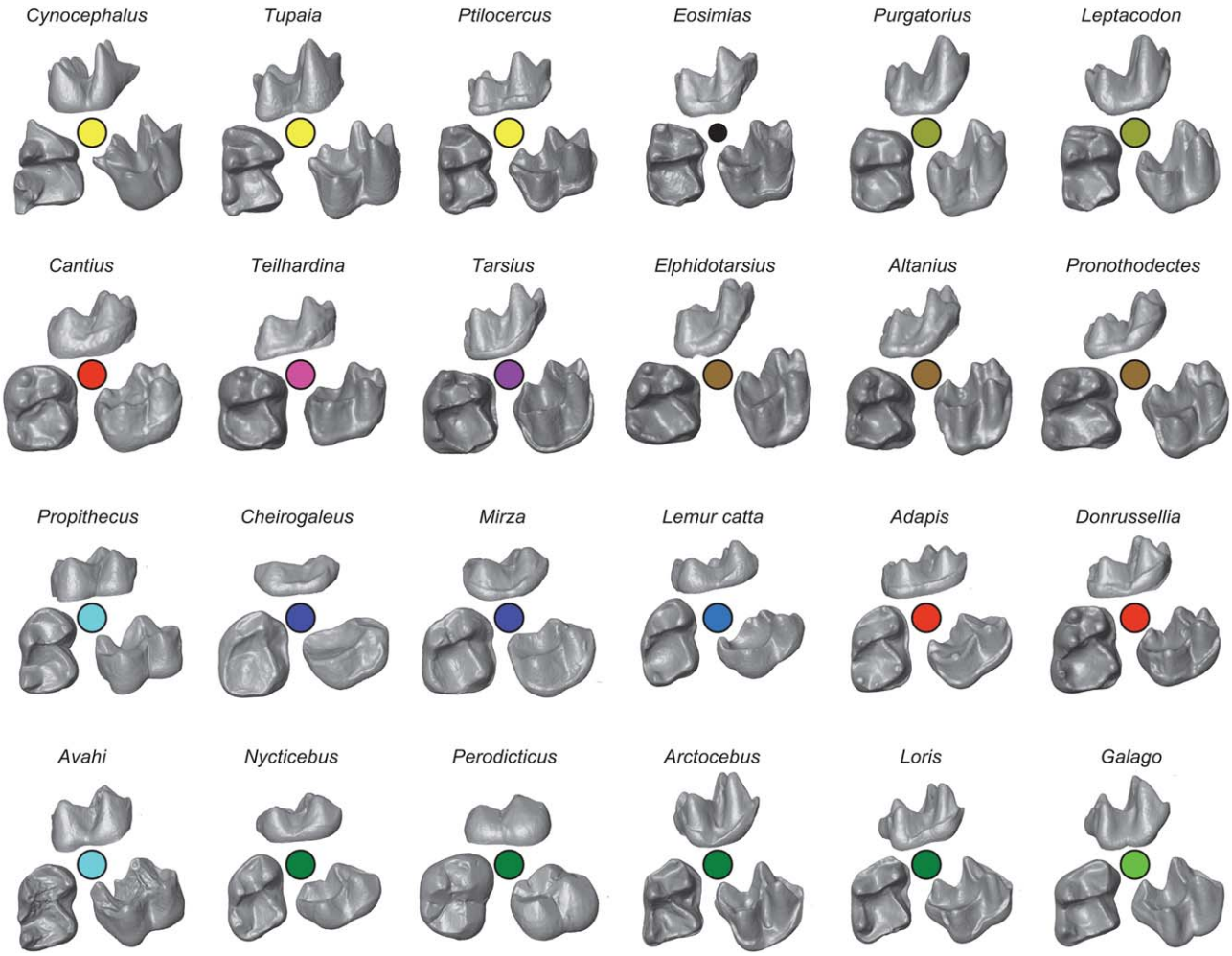

Fig. 7. Images of teeth of representative taxa in this study. For each taxon, views are buccal (upper), occlusal (lower left), and a three-quarter profile (lower right). Colors correspond to minimum convex polygons shown in Figure 5.

methods compared here, including user-determined landmarks, indicate that the second mandibular molars of *Eosimias* are more similar to non-primates and stem-primates of the sample than to early euprimates or tarsiers. These results are consistent with the perspective that Eosimiidae may be positioned more basally than stem Anthropoidea (*contra* Beard et al., 1994, 1996; Gebo et al., 2000), and questions purported similarities between Eosimiidae and Tarsiidae. Furthermore, the recovered dental affinities are concordant with recent results from analyses of primate ankle bones (Boyer et al., 2015c; Seiffert et al., 2015; Yapuncich et al., 2017) that suggest eosimiid ankle morphology is similar to that inferred for the common ancestor of primates of modern aspect. In our opinion, these findings point to the need for continued examination of eosimiid relationships, anthropoid relationships, and certain patterns of early primate evolution.

The Impact of Feature-Fixing on Shape Ordinations

Of the methods compared, GLAM3, the only method utilizing feature-fixing, shows the least taxonomic differentiation (Fig. 5d). It is also the only method that does not place the subfossil *Megaladapis* in the region occupied by *Adapis* and lemurids. The early euprimate *Cantius* is scattered across the entire plot area. Stem-primates, *Leptacodon*, *Purgatorius*, and *Teilhardina* are largely overlapping and oddly plot near *Lepilemur* (Fig. 5d). *Cheirogaleus* spans over half the range of PC2 values. This scatter is reflected in the relatively low value obtained by taxonomic ANOVA for this group as well (Table 10). Two observations suggest that GLAM1 and GLAM2 (which do not use feature-fixing) reflect sample geometry better than GLAM3 (which uses feature-fixing). First, in the plot of method by treatment type (Fig. 4), they were closer to the result from user-determined landmarks. Second, the one-way ANOVA on taxonomic groups of the vector distribution of ordinations produced by GLAM1 has a higher P value than GLAM3, suggesting greater taxonomic distinctiveness (Table 10).

The observation that feature-fixing degrades taxonomic signal is worth further investigation. TPS, the main technique involved in feature-fixing, strives to align specified corresponding anchor points and generate a smooth interpolation between the shapes without

guaranteed control for the distortion of the final map. The cP distance, in a certain sense, measures the minimum global average distortion of a class of candidate maps between two surfaces. Therefore, maintaining low global distortion may be more important for producing cP distances that faithfully reflect the geometric dissimilarity than precisely matching geometrically characteristic point features. Furthermore, the quality of the interpolated map depends heavily on the choice of anchor points: if two corresponding anchors appear to close to each other, TPS will face numerical stability issues. If the correspondences between anchor points are incomplete or wrongly specified (e.g., for teeth of low relief and probably fewer detectable extremal points such as the omnivorous *Cheirogaleus*), TPS may generate less biologically meaningful maps. It is therefore possible that the various sources of map and distance distortion in feature-fixing methods generate less taxonomically cohesive results. Rather than a major improvement for the automated 3DGM methods, feature-fixing may primarily facilitate map visualization compared to existing geometric morphological analysis.

Comparison of Novel Globally Informed Methods and auto3Dgm

It is surprising that auto3Dgm has a pattern and magnitude of taxonomic distinctiveness more similar to the user-determined landmarks (Fig. 5e and Table 10); in many ways, auto3Dgm is relatively naïve compared to other tree-based techniques presented here. However, we believe the novel globally informed methods represent improvements over auto3Dgm for three reasons: (1) pseudolandmark sampling density does not appear to affect the novel methods of this study, (2) the novel methods produce more intuitive ordinations, and (3) the novel methods evenly fill the ordination space. First, Vitek et al. (2017) found that ordinations produced by auto3Dgm are sensitive to pseudolandmark sampling density (i.e., number of points per tooth). For the methods presented in this study, downstream analyses were not sensitive to sampling density (Table 5). Second, several intertaxonomic affinities in the auto3Dgm result differ strongly from user-determined landmarks and are less functionally or geometrically intuitive than results from the novel methods of this study. In the case of the relative positions of galagos, cheirogaleids, and indriids, results from the novel methods are more intuitive, in both a functional and phylogenetic sense.

Finally, the novel methods evenly fill the ordination space in a manner more similar to the user-based approach (Fig. 5). The distribution of specimens in the auto3Dgm ordinations often form a Y pattern in which points cluster linearly through regions of space, giving a much different perspective on how filled the shape space is compared to user-based methods (Fig. 5). The Y pattern also seems to appear in interspecific PCA ordinations generated by Generalized Procrustes Surface Analysis (Pomidor et al., 2016), a recent shape analysis method of similarly high dimensionality. The diffuse distribution seems problematic, as it may reflect a highly skewed distribution of values in the correlation matrix of the PCA. Gonzalez et al., (2016) recommend converting pseudolandmarks to sliding landmarks to minimize surface bending energy or average Procrustes distance between specimens. In this way, they eliminate the Y pattern; however, this may potentially reflect the addition of random noise. Alternatively, replacing PCA with a dimension reduction method more suitable for high-dimensional data (such as *t*-distributed stochastic neighbor embedding [van der Maaten and Hinton, 2008]) may reduce the strength of the Y pattern. Whatever the remedy for auto3Dgm, the Y pattern does not manifest in any of the novel methods presented here, suggesting they are indeed improvements to existing automated methods.

CONCLUSIONS

This study has addressed and overcome limitations of previously published automated 3DGM methods, cPDist (Boyer et al., 2011), and auto3Dgm (Boyer et al., 2015a), and provided a detailed description of how the results of existing and novel automated methods compare to a user-determined landmark approach. Both the globally informed methods proposed here and auto3Dgm reflect the similar geometric patterns as user-based methods. Relative to cPDist, the dramatic reduction in MSE of propagated landmarks of the globally informed methods shows that global sample information is a critical component of automated analysis on samples with large shape differences. Other modifications did not definitively improve the similarity to a user-based approach. In particular, feature-fixing, or the automatic manipulation of maps to maintain type II landmark (Bookstein, 1991) representation, appears to add error and reduce taxonomic cohesiveness. Though composed LAST methods may potentially reduce the increased map inaccuracy between similar shapes that are not directly connected by an edge in the MST, no approach permits full retention of map quality of direct comparisons between similar shapes. This suggests that developing alternative approaches for analyzing collections of highly dissimilar shapes remains an interesting and challenging problem.

Optimal Applications for Automated Approaches

Because the examined parameter combinations of globally informed methods and auto3Dgm produce similar shape ordinations that are largely consistent with results of a user-based approach, we recommend using automated approaches when (1) the primary questions concern patterns of overall variation in biological shapes, (2) a large number of type II landmarks (defined in Bookstein, 1991 as areas of high local curvature) are not consistently available, or (3) measurement/landmark selection may increase the potential for biased results. These recommendations are similar to those of Gonzalez et al., (2016), who evaluated the performance of auto3Dgm against a semi-automated landmark method. The first two conditions are related, since fewer type II landmarks limit the ability to assess overall variation. User-based approaches for assessing shape *disparity* (which refers to the overall structure of interest) become increasingly limited as sample diversity increases. Polly (2008) highlights the limitation of user-based approaches that require many biologically equivalent landmarks to represent patterns of shape variation and emphasizes the importance of methods that lack this requirement.

How can one represent the absence of a paraconid quantitatively if the feature does not exist in all specimens? The researcher could landmark the space where the paraconid "would be," but this is obviously subjective. Automated 3DGM methods do not make such explicit assumptions about feature equivalence and can therefore more comprehensively measure shape variation. The third condition, reducing the potential for user bias, is particularly salient when the anatomical structure of a fossil has been linked to a particular taxonomic identification or phylogenetic hypothesis, as in the case of *Eosimias* we have highlighted above.

In sum, this study affirms the utility and reliability of automated approaches through thorough comparisons of automated approaches and user-based approaches. The globally informed methods presented here have several important advantages over user-based approaches and previously published automated 3DGM methods. Future work will improve the use of global sample information in computing correspondence maps between objects and will hopefully recover even greater geometric fidelity in automated morphometric analysis.

ACKNOWLEDGEMENTS

The authors express their thanks to curators and staff at the American Museum of Natural History for providing access to dental specimens to be molded, cast, and μCT-scanned; to D. Krause and J. Groenke of Stony Brook University for access to facilities at the Vertebrate Paleontology Fossil Preparation Lab; to C. Rubin and S. Judex of Stony Brook University for access to μCT scanners; and to K. Christopher Beard for access to eosimiid dental casts.

REFERENCES

Al-Aifari R, Daubechies I, Lipman Y. 2013. Continuous Procrustes distance between two surfaces. Commun Pure Appl Math 66:934–964.

Ashburner J, Friston KJ. 2000. Voxel-based morphometry—the methods. NeuroImage 11:805–821.

Bajpai S, Kay RF, Williams BA, Das DP, Kapur VV, Tiwari BN. 2008. The oldest Asian record of Anthropoidea. Proc Natl Acad Sci USA 105:11093–11098.

Beard KC, Qi T, Dawson MR, Wang B, Li C. 1994. A diverse new primate fauna from middle Eocene fissure-filings in southeastern China. Nature 368:604–609.

Beard KC, Tong Y, Dawson MR, Wang J, Huang X. 1996. Earliest complete dentition of an anthropoid primate from the late middle Eocene of Shanxi Province, China. Science 272:82–85.

Beard KC, Wang J. 2004. The eosimiid primates (Anthropoidea) of the Heti Formation, Yuanqu Basin, Shanxi and Henan Provinces, People's Republic of China. J Hum Evol 46:401–432.

Besl PJ, McKay ND. 1992. Method for registration of 3D shapes. IEEE Trans Pattern Anal Mach Intell 14:239–256.

Bookstein FL. 1991. Morphometric tools for landmark data. Geometry and biology. Cambridge: Cambridge University Press.

Bookstein FL. 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med Image Anal 1:225–243.

Bookstein F, Schafer K, Prossinger H, Seidler H, Fieder M, Stringer C, Weber GW, Arsuaga J-L, Slice DE, Rohlf FJ, Recheis W. 1999. Comparing frontal cranial profiles in archaic and modern *Homo* by morphometric analysis. Anat Rec 257:217–224.

Bookstein FL, Streissguth AP, Sampson PD, Connor PD, Barr HM. 2002. Corpus callosum shape and neurophysical deficits in adult males with heavy fetal alcohol exposure. NeuroImage 15:233–251.

Boyer DM. 2017. Facilities, databases, and software. Department of Evolutionary Anthropology, Duke University [Internet]. Accessed http://www.dougmboyer.com/facilities-databases-software.html.

Boyer DM, Lipman Y, St. Clair E, Puente J, Patel BA, Funkhouser TA, Jernvall J, Daubechies I. 2011. Algorithms to automatically quantify the geometric similarity of anatomical surfaces. Proc Nat Acad Sci 108:18221–18226.

Boyer DM, Costeur L, Lipman Y. 2012. Earliest record of *Platychoerops* (Primates, Plesiadapidae), a new species from Mouras Quarry, Mont de Berru, France. Am J Phys Anthropol 149:329–346.

Boyer DM, Puente J, Gladman JT, Glynn C, Mukherjee S, Yapuncich GS, Daubechies I. 2015a. A new fully automated approach for aligning and comparing shapes. Anat Rec 298:249–276.

Boyer DM, Winchester JM, Glynn C, Puente J. 2015b. Detailed anatomical orientations for certain types of morphometric measurements can be determined automatically with geometric algorithms. Anat Rec 298:1816–1823.

Boyer DM, Yapuncich GS, Butler JE, Dunn RH, Seiffert ER. 2015c. Evolution of postural diversity in primates as reflected by the size and shape of the medial tibial facet of the talus. Am J Phys Anthropol 157:134–177.

Boyer DM, Gunnell GF, Kaufman S, McGeary TM. 2016. MorphoSource: archiving and sharing digital specimen data. Paleontol Soc Pap 22:157–181.

Bunn JM, Boyer DM, Lipman Y, St. Clair EM, Jernvall J, Daubechies I. 2011. Comparing Dirichlet normal surface energy of tooth crowns, a new technique of molar shape quantification for dietary inference, with previous methods in isolation and in combination. Am J Phys Anthropol 145:247–261.

Chen Y, Guibas L, Huang Q. 2014. Near-optimal joint object matching via convex relaxation. Preprint. Available: arXiv:1402.1473. Accessed: April 12, 2014.

Clementi AE, Di Ianni M, Lauri M, Monti A, Rossi G, Silvestri R. 2007. On the bounded-hop MST problem on random Euclidean instances. Theor Comput Sci 384:161–167.

Cooke SB, Terhune CE. 2015. Form, function, and geometric morphometrics. Anat Rec 298:5–28.

Dryden IL, Mardia KV. 1998. Statistical shape analysis. New York: Wiley.

Gao T. 2017. MST-based generalized dataset Procrustes distance by Jesús Puente. Database: GitHub [Internet]. Accessed: https://github.com/trgao10/PuenteAlignment.

Gebo DL, Dagosto M, Beard KC, Qi T, Wang J. 2000. The oldest known anthropoid postcranial fossils and the early evolution of higher primates. Nature 404:267–278.

Gerig G, Styner M, Gerig G, Styner M, Jones D, Weinberger D, Lieberman J. 2001. Shape analysis of brain ventricles using SPHARM. In: IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. p 171–178.

Gonzalez PN, Barbeito-Andrés J, D'Addona LA, Bernal V, Perez SI. 2016. Technical note: performance of semi and fully automated approaches for registration of 3D surface coordinates in geometric morphometric studies. Am J Phys Anthropol 160:169–178.

Godinot M, Mahboubi M. 1994. Les petits primates simiiformes de Glib Zegdou (Eocene inferier a moyen d'Algerie. C R Acad Sci II 319:357–364.

Gower JC. 1975. Generalized Procrustes analysis. Psychometrika 40:33–51.

Gunnell GF, Miller ER. 2001. Origin of Anthropoidea: dental evidence and recognition of early anthropoids in the fossil record, with comments on the Asian anthropoid radiation. Am J Phys Anthropol 114:177–191.

Gunz P, Mitteroecker P. 2013. Semilandmarks: a method for quantifying curves and surfaces. Hystrix Ital J Mammal 24:103–109.

Gunz P, Mitteroecker P, Bookstein FL. 2005. Semilandmarks in three dimensions. In: Slice DE, editor. Modern morphometrics in physical anthropology. New York: Kluwer Academic/Plenum Publishers. p 73–98.

Harcourt-Smith WEH, Tallman M, Frost SR, Wiley DF, Rohlf JF, Delson E. 2008. Analysis of selected hominoid joint surfaces using laser scanning and geometric morphometrics: a preliminary

report. In: Dagosto M, Sargis EJ, editors. Mammalian evolutionary morphology: a tribute to Frederick S. Szalay. New York: Springer Science. p 373–383.
Houle D, Govindaraju DR, Omholt S. 2010. Phenomics: the next challenge. Nat Rev Genet 11:855–866.
Huang Q-X, Zhang G-X, Gao L, Hu S-M, Butscher A, Guibas L. 2012. An optimization approach for extracting and encoding consistent maps in a shape collection. ACM Trans Graph 31:125.1–125.11.
Huang Q, Guibas L. 2013. Consistent shape maps via semidefinite programming. Comp Graph Forum 32:177–186.
Khuller S, Raghavachari B, Young N. 1995. Balancing minimum spanning trees and shortest-path trees. Algorithmica 14:305–321.
Klingenberg CP. 2011. MorphoJ: an integrated software package for geometric morphometrics. Mol Ecol Resources 11:353–357.
Koehl P, Hass J. 2015. Landmark-free geometric methods in biological shape analysis. J R Soc Interface 12:20150795.
Lawing AM, Polly PD. 2010. Geometric morphometrics: recent applications to the study of evolution and development. J Zool 280:1–7.
Ledogar JA, Winchester JM, St. Clair EM, Boyer DM. 2013. Diet and dental topography in pitheciine seed predators. Am J Phys Anthropol 150:107–121.
Lipman Y. 2012. Bounded distortion mapping spaces for triangular meshes. ACM Trans Graph 31:108.
Lipman Y. 2017. CPSurfComp. Database: Weizmann Institute [Internet]. Accessed: http://www.wisdom.weizmann.ac.il/~ylipman/CPsurfcomp/.
Meng TW, Lui LM. 2015. The theory of computational quasiconformal geometry of point clouds. Preprint. Available: arXiv: 1510.05104. Accessed: July 15, 2017.
Miller ER, Gunnell GF, Martin RD. 2005. Deep time and the search for anthropoid origins. Yearb Phys Anthropol 128:60–95.
Mitteroecker P, Gunz P. 2009. Advances in geometric morphometrics. Evol Biol 36:235–247.
Muhkerjee S. 2017. Automated 3D geometric morphometrics. Department of Statistical Science, Duke University (Durham, NC). Accessed: http://www2.stat.duke.edu/~sayan/auto3dgm.
Nguyen A, Ben-Chen M, Welnicka K, Ye Y, Guibas L. 2011. An optimization approach to improving collections of shape maps. Comp Graph Forum 30:1481–1491.
O'Higgins P, Bastir M, Kupczik K. 2006. Shaping the human face. Int Cong Ser 1296:55–73.
Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2017. vegan: community ecology package. R package version 2.4-3.
Perez SI, Bernal V, Gonzalez PN. 2006. Differences between sliding semilandmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. J Anat 208:769–784.
Polly PD. 2008. Developmental dynamics and G-matrices: can morphometric spaces be used to model phenotypic evolution?. Evol Biol 35:83–96.
Polly PD, MacLeod N. 2008. Locomotion in fossil Carnivora: an application of eigensurface analysis for morphometric comparison of 3D surfaces. Paleontol Electron 11:13.
Pomidor BJ, Makedonska J, Slice DE. 2016. A landmark-free method for three-dimensional shape analysis. PLoS One 11: e0150368.
Rohlf FJ, Slice DE. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. Syst Zool 39:40–59.
Seiffert ER, Costeur L, Boyer DM. 2015. Primate tarsal bones from Egerkingen, Switzerland, attributable to the middle Eocene adapiform *Caenopithecus lemuroides*. PeerJ 3:e1036.
Sievwright H, MacLeod N. 2012. Eigensurface analysis, ecology, and modelling of morphological adaptation in the falconiform humerus (Falconiformes: Aves). Zool J Linn Soc 165:390–419.
Simons EL. 1995. Skulls and anterior teeth of *Catopithecus* (Primates: Anthropoidea) from the Eocene and anthropoid origins. Science 268:1885–1888.
Simons EL, Rasmussen DT. 1994. A remarkable cranium of *Plesiopithecus teras* (Primates, Prosimii) from the Eocene of Egypt. Proc Natl Acad Sci USA 91:9946–9950.
Slice DE. 2005. Modern morphometrics in physical anthropology. New York: Kluwer Academic/Plenum Publishers.
Slice DE. 2007. Geometric morphometrics. Annu Rev Anthropol 36: 261–281.
Styner M, Oguz I, Xu S, Brechbuhler C, Pantazis D, Levitt JJ, Shenton ME, Gerig G. 2006. Framework for the statistical shape analysis of brain structures using SPHARM-PDM. Insight 1071:242–250.
van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605.
Vitek NS, Manz CL, Gao T, Bloch JI, Strait SG, Boyer DM. 2017. Semi-supervised determination of pseudocryptic morphotypes using observer-free characterizations of anatomical alignment and shape. Ecol Evol 7:5041. DOI: 10.1002/ece3.3058
Warton DI, Wright ST, Wang Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods Ecol Evol 3:89–101.
Watanabe A. 2015. How many landmarks are enough? Identifying adequate sampling of landmarks for capturing the shape of specimens. Society of Vertebrate Paleontology Annual Meeting, Dallas, TX. Available: http://vertpaleo.org/Annual-Meeting/Home.aspx.
Wiley DF, Amenta N, Alcantara DA, Ghosh D, Kil YJ, Delson E, Harcourt-Smith W, Rohlf FJ, St. John K, Hamann B. 2005. Evolutionary morphing. Proc IEEE Vis 2005:8.
Winchester JM, Boyer DM, St. Clair EM, Gosselin-Ildari AD, Cooke SB, Ledogar JA. 2014. Dental topography of platyrrhines and prosimians: convergence and contrasts. Am J Phys Anthropol 153: 29–44.
Yapuncich GS, Seiffert ER, Boyer DM. 2017. Quantification of the position and depth of the flexor hallucis longus groove in euarchontans, with implications for the evolution of primate positional behavior. Am J Phys Anthropol DOI: 10.1002/ajpa.23213.
Zimmerman GM, Goetz H, Mielke PW. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology 66:606–611.