Earth and Environmental Science Transactions of the Royal Society of Edinburgh, **109**, 301–323, 2019 (for 2018)

Early Vertebrate Evolution

The challenges and potential utility of phenotypic specimen-level phylogeny based on maximum parsimony

Emanuel TSCHOPP[1,2,3*] and Paul UPCHURCH[4]

[1] Division of Paleontology, American Museum of Natural History, Central Park West @ 79th Street,
New York, NY 10024, USA.
Email: etschopp@amnh.org
[2] Dipartimento di Scienze della Terra, Università di Torino, Via Valperga Caluso 35, 10125 Torino, Italy.
[3] Museu da Lourinhã, Rua João Luís de Moura 95, 2530-157 Lourinhã, Portugal.
[4] Department of Earth Sciences, University College London, Gower Street, London, WC1E 6BT, UK.
*Corresponding author

ABSTRACT: Specimen-level phylogenetic approaches are widely used in molecular biology for taxonomic and systematic purposes. However, they have been largely ignored in analyses based on morphological traits, where phylogeneticists mostly resort to species-level analyses. Recently, a number of specimen-level studies have been published in vertebrate palaeontology. These studies indicate that specimen-level phylogeny may be a very useful tool for systematic reassessments at low taxonomic levels. Herein, we review the challenges when working with individual organisms as operational taxonomic units in a palaeontological context, and propose guidelines of how best to perform a specimen-level phylogenetic analysis using the maximum parsimony criterion. Given that no single methodology appears to be perfectly suited to resolve relationships among individuals, and that different taxa probably require different approaches to assess their systematics, we advocate the use of a number of methodologies. In particular, we recommend the inclusion of as many specimens and characters as feasible, and the analysis of relationships using an extended implied weighting approach with different downweighting functions. Resulting polytomies should be explored using *a posteriori* pruning of unstable specimens, and conflicting tree topologies between different iterations of the analysis should be evaluated by a combination of support values such as jackknifing and symmetric resampling. Species delimitation should be consistent among the ingroup and based on a reproducible approach. Although time-consuming and methodologically challenging, specimen-level phylogenetic analysis is a highly useful tool to assess intraspecific variability and provide the basis for a more informed and accurate creation of species-level operational taxonomic units in large-scale systematic studies. It also has the potential to inform us about past speciation processes, morphological trait evolution, and their potential intrinsic and extrinsic drivers in pre-eminent detail.

KEY WORDS: character weighting, cladistics, species delimitation, variability, vertebrate morphology.

Specimen-level phylogenetic analysis is becoming increasingly popular in vertebrate palaeontology, in particular (but not only) in dinosaur systematics (Yates 2003; Upchurch *et al.* 2004; Boyd *et al.* 2009; Makovicky 2010; Morschhauser *et al.* 2014; Scannella *et al.* 2014; Longrich 2015; Mounier & Caparros 2015; Tschopp *et al.* 2015; Campbell *et al.* 2016; Cau 2017). This kind of phylogenetic analysis includes single specimens instead of species or genera as operational taxonomic units (OTUs), and thus ignores earlier species- and/or genus-level identifications based on comparative studies. This approach was first advocated by Vrana & Wheeler (1992), and is widely used in molecular phylogenetic studies (e.g., Dettman *et al.* 2003; Godinho *et al.* 2005; Mayer & Pavlicev 2007; Bacon *et al.* 2012; Ahmadzadeh *et al.* 2013; Marzahn *et al.* 2016), but rarely by morphologists.

Specimen-level phylogenetic analyses can be considered a bottom-up approach to establish the monophyly of a species (Vrana & Wheeler 1992), and to reassess the referral of a particular specimen to a species (Longrich 2015; Campbell *et al.* 2016). Using specimens instead of species avoids the risk of including potentially chimeric species-level OTUs resulting from erroneous species identifications in earlier studies (Tschopp *et al.* 2015). Given these advantages over species-level analyses, specimen-level phylogenetic analysis has indeed predominantly been used for taxonomic and systematic purposes, mostly at low taxonomic levels (Yates 2003; Upchurch *et al.* 2004; Boyd *et al.* 2009; Scannella *et al.* 2014; Longrich 2015; Mounier & Caparros 2015; Tschopp *et al.* 2015; Campbell *et al.* 2016).

Longrich (2015) and Tschopp *et al.* (2015) specifically highlighted the ability of specimen-level phylogenetic analyses to act as a test for the homology of particular morphological features, and, thus, to assess a trait's phylogenetic informativeness *versus* its status as intraspecific variation. This issue

doi:10.1017/S1755691018000877

is particularly important in vertebrate palaeontology, where many species are represented by a single, incomplete specimen. The holotype of the sauropod dinosaur *Diplodocus longus* serves as an example here: it solely comprises caudal vertebrae and a chevron (McIntosh & Carpenter 1998; Tschopp & Mateus 2016), but these caudal vertebrae bear a peculiar ridge connecting the prezygapophyses, which appears to be otherwise shared only with one other specimen (Tschopp *et al.* 2015, 2018a; Tschopp & Mateus 2016). Whereas Carpenter (2017) interprets this ridge as homologous in the two specimens, and accepts it as a potential autapomorphy of the species *D. longus*, the specimen-level analysis of Tschopp *et al.* (2015) did not find that these two specimens formed a unique clade, suggesting that the occurrence of this ridge results from individual variation (Tschopp *et al.* 2015, 2018a).

Whereas these taxonomic issues are certainly important, the potential of specimen-level studies is far greater. Such a phylogenetic analysis not only provides information about relationships between individuals, but also on the importance and variability of certain traits in the evolution of the taxon under study. When correlated with a well-dated stratigraphy, first occurrences of diagnostic traits can theoretically be pinpointed to a particular time and place, and in some cases, speciation modes can be identified (Cau 2017). Further correlations with palaeoclimatic, palaeoenvironmental or molecular data could then yield information on evolution in pre-eminent detail. Moreover, key information on macroevolutionary patterns and processes (e.g., diversity, biogeography) can be determined from the fossil record (e.g., Alroy *et al.* 2008; Benson *et al.* 2014, 2016; Mannion *et al.* 2014, 2015; Tennant *et al.* 2016a, b; Close *et al.* 2017), but this ultimately depends on accurate counts of how many species or genera were present in a given temporal and/or spatial bin. The taxonomic identifications that underpin such studies have mostly been made on partially subjective grounds (especially when dealing with fossils), such as a systematist's personal view that a given autapomorphy does, or does not, warrant the erection of a new species or genus. Some recent specimen-level phylogenetic analyses (e.g., Tschopp *et al.* 2015) have introduced methods for imposing more explicit, quantified and consistent means for separating clusters of specimens into higher taxonomic units. The application of such approaches offers the prospect of producing more objective taxonomic units that can be counted in diversity and other macroevolutionary studies.

Palaeontological data sets, however, present a number of methodological challenges that researchers must deal with when setting up a specimen-level phylogenetic analysis. Herein, we review these issues, with a particular focus on methodologies using the maximum parsimony criterion, and propose a number of approaches to address these problems accurately, while also highlighting the potential for future applications of this methodology in palaeontology.

The institutional abbreviations used in this paper are as follows: American Museum of Natural History, New York, USA (AMNH); Museum of Paleontology, Brigham Young University, Provo, USA (BYU); Carnegie Museum of Natural History, Pittsburgh, USA (CM); Gunma Museum of Natural History, Gunma, Japan (GMNH-PV); Sauriermuseum Aathal, Switzerland (SMA); National Museum of Natural History, Smithsonian Institution, Washington DC, USA (USNM); Yale Peabody Museum, New Haven, USA (YPM).

1. Methodological challenges

Challenges for phenotypic specimen-level phylogenetic analysis can be grouped into three specific steps: (1) matrix construction; (2) phylogenetic methodology and interpretation of tree topology; and (3) species delimitation.

1.1. Matrix construction

1.1.1. Taxon sampling. Taxon sampling is a paramount factor affecting the accuracy of phylogenetic analysis (e.g., Bergsten 2005; Puslednik & Serb 2008; Brusatte 2010). In general, taxon (and in this case also specimen) sampling should be as extensive as possible. Molecular case studies have indicated that undersampling of specimens per species can lead to taxonomic over-splitting, and thus inflation of the number of recognised species (Bacon *et al.* 2012). In theory, we can be confident of sampling the most meaningful genetic variation in a species if we include a minimum of ten specimens per species (Saunders *et al.* 1984; Carstens *et al.* 2013). Although we do not know of any empirical study assessing the minimum numbers of specimens in phenotypic matrices, similar numbers might apply to morphological variation. However, there are obvious pragmatic constraints on both scoring a large number of OTUs and on performing phylogenetic analysis on larger datasets. In vertebrate palaeontology, many species are known from less than ten specimens per species. For example, the maximum number of specimens attributed to a single species in the analysis by Tschopp *et al.* (2015) was four (referred to *Diplodocus hallorum*), whereas Campbell *et al.* (2016) identified nine specimens as belonging to *Chasmosaurus russelli*. We believe, however, that these issues should not be seen as prohibitive: although we need to be aware of the methodological shortcomings, we have to work with the data we have at hand and address challenges with the necessary attention.

Within a dataset, different sampling strategies apply for ingroup and outgroup. Taxon selection for the ingroup, in part, depends on the scope of the analysis. In most specimen-level analyses, the main scope is a taxonomic revision (e.g., Yates 2003; Upchurch *et al.* 2004; Boyd *et al.* 2009; Makovicky 2010; Scannella *et al.* 2014; Longrich 2015; Mounier & Caparros 2015; Tschopp *et al.* 2015; Campbell *et al.* 2016). In this case, it is necessary to include all the available type specimens of the clade to be revised, because these are the 'name-bearing' specimens that will help to determine the identification of referred specimens during the post-phylogenetic-analysis phase of the study. Even if incomplete, adding OTUs generally has a positive impact on tree accuracy (Wilkinson 2003; Wiens 2006; Wiens & Tiu 2012; see Section 1.1.3). In order to best exploit this positive impact, it is of crucial importance to add as many reasonably complete non-type specimens as are available, which can facilitate indirect comparisons between more fragmentary specimens that do not have any anatomical overlap (Tschopp *et al.* 2015, 2018b). In the case of the sauropod *Camarasaurus*, type specimens of all the species that were at some point considered to belong to the genus are highly incomplete, and are often represented by non-overlapping parts of the skeleton (Table 1). In order to analyse their relationships correctly, it is, therefore, necessary to add more complete specimens like CM 11338 or GMNH-PV 101, which show anatomical overlap with nearly all the type specimens (Table 1), and can, therefore, serve as a link between non-overlapping ones.

When analysing character distribution and trait evolution rather than systematics, the inclusion of incomplete type specimens is not of crucial importance. However, because they might still bear unique, phylogenetically informative combinations of character states, the *a priori* exclusion of these incomplete taxa should follow certain guidelines (like, e.g., the ones outlined for the 'safe taxonomic reduction' process proposed by Wilkinson 1995; see also Norell & Gao 1997; Kearney & Clark 2003; Butler & Upchurch 2007).

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Table 1 Anatomical overlap in single specimens of the sauropod dinosaur *Camarasaurus*. Note that only by adding the two relatively complete non-type specimens, can most of the types be indirectly compared with each other. Type specimens are marked with an asterisk. Coloured cells mark which parts of the skeleton are represented. The specimens CM 11338 and GMNH-PV 101 have been described in literature, and assigned to *Camarasaurus lentus* (Gilmore 1925) and *Camarasaurus grandis* (McIntosh *et al*. 1996), respectively. Abbreviations: Cd = caudal vertebrae; Ch = chevrons; CV = cervical vertebrae; DV = dorsal vertebrae; Fl = forelimb; Hl = hindlimb; PcG = pectoral girdle; PvG = pelvic girdle; Sk = skull; SV = sacral vertebrae; T = teeth.

Taxon	Specimen(s)	Sk	T	CV	DV	SV	Cd	Ch	PcG	Fl	PvG	Hl
*Camarasaurus supremus**	**AMNH FARB 5760, X-c-1**			■								
*'Apatosaurus' grandis**	**YPM VP.001901, and parts of YPM VP.001902, VP.001905**	■	■	■	■		■	■	■	■	■	■
*'Caulodon' diversidens**	**AMNH FARB 5768**		■									
*'Amphicoelias' latus**	**AMNH FARB 5765**						■					■
*'Caulodon' leptoganus**	**AMNH FARB 5769**		■									
*'Morosaurus' impar**	**YPM VP.001900, VP.001903, VP.007680**					■			■			■
*'Morosaurus' robustus**	**in parts: YPM VP.001905**								■		■	■
*Camarasaurus leptodirus**	**AMNH FARB 5763**			■								
*Camarasaurus lentus**	**YPM VP.001910**	■	■	■	■	■	■		■	■	■	■
*'Morosaurus' agilis**	**USNM 5384**	■		■								
*'Uintasaurus' douglassi**	**CM 11069**			■	■							
*Camarasaurus annae**	**CM 8942**				■							
*'Cathetosaurus' lewisi**	**BYU 9047**			■	■	■	■	■	■	■	■	
Camarasaurus sp.	**CM 11338**	■	■	■	■	■	■	■	■	■	■	■
Camarasaurus sp.	**GMNH-PV 101**	■	■	■	■	■	■	■	■	■	■	■

In any phylogenetic analysis, outgroup selection is paramount for the correct optimisation of character states along the tree. Increased outgroup sampling is likely to have benefits in terms of phylogenetic accuracy (Nixon & Carpenter 1993; Bergsten 2005; Brusatte 2010) – if one includes only a single outgroup taxon, the analysis will find the ingroup as a monophyletic clade by default, excluding any possibility of testing this hypothesis *a priori* (Puslednik & Serb 2008). Outgroups should, therefore, cover a range of taxa from species closely related to the ingroup to more distantly related taxa (Bergsten 2005), with a relatively plesiomorphic taxon as the outgroup to all others (see Whitlock 2011).

For a systematic review, it can be necessary to include type specimens that are currently thought not to belong to the clade being revised, but have been attributed to it at some point in the past (see Tschopp *et al*. 2015). These should, therefore, be recovered in the outgroup by the analysis. In order to test these more recent identifications accurately, it is important to include at least one additional OTU from the taxon to which the type specimen is currently thought to belong. However, given that these OTUs were previously referred to the ingroup, it is probable that their actual higher-level taxon exhibits a number of convergently acquired features. Therefore, it is particularly important to add additional OTUs from intermediate phylogenetic positions, as outlined above. The more complete these additional outgroup OTUs, the lower the probability that convergences could outnumber phylogenetically informative characters, and thus the risk of an erroneous interpretation of homoplastic traits as homologies. Thus, the completeness of outgroup terminals becomes more important than the risk of creating chimeric OTUs by combining data from various individuals. Also, testing the monophyly of outgroup taxa is generally not the scope of a particular study. Therefore, if no complete specimen is available, species-level OTUs may be a good compromise for a particular outgroup. Indeed, completeness has often been put forward as one of the main criteria for the selection of a specific taxon in the outgroup (e.g., Whitlock 2011), and has often also led researchers to use higher-level taxa as outgroups, especially if the ingroup is composed of single specimens (e.g., Upchurch *et al*. 2004; Tschopp *et al*. 2015). However, the more inclusive these outgroup OTUs are, the more they are likely to be polymorphic, creating problems in scoring variable taxa (see Section 1.1.5). This problem is why various researchers have advocated the use of multiple species-level OTUs instead of higher-level taxa (see Prendini 2001; Brusatte 2010, and references therein). Thus, adding several species-level OTUs of a particular clade in the outgroup appears to be the best compromise between OTU completeness and scoring accuracy. By doing so, the specimen-level OTUs of the ingroup can be expected to fit into a strongly supported backbone topology defined by relatively complete outgroup OTUs. In those cases where one or more outgroup species or higher taxa are themselves considered to be problematic (e.g., chimaeric), then, ultimately, they should also be investigated via specimen-level phylogenetic analysis. This could lead to research programmes based on iterative studies that 'reciprocally illuminate' the taxonomic content of a series of closely related taxa.

Juvenile specimens can create problems for phylogenetic analyses, because some of the traits change throughout ontogeny, such that only adult individuals display the derived state necessary for a correct identification (Woodruff *et al.* 2017). Indeed, in some analyses, juveniles were found in a more 'basal' position compared to their respective species, because some of their apomorphic features had not developed yet (e.g., Campione *et al*. 2013; Carballido & Sander 2014). However, this is not always the case. In Upchurch *et al.* (2004), Tschopp *et al.* (2015) and Campbell *et al.* (2016), juvenile specimens were actually recovered in disparate, and often relatively derived, positions within the ingroup, and in sister-taxon relationships with adult specimens. Therefore, it appears that under certain circumstances, phylogenetic analysis is minimally (or not at all) influenced by ontogenetically variable features. Indeed, Carballido & Sander (2014) found that although early juvenile ontogenetic stages of the macronarian sauropod *Europasaurus* were recovered more 'basally' compared to adult

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Table 2 Missing data ratios of selected phylogenetic analyses. Tschopp & Mateus (2017) used an updated version of Tschopp *et al.* (2015), and collapsed the operational taxonomic unit (OTU) sampling to species based on the taxonomic interpretations of Tschopp *et al.* (2015).

			OTUs		Scores		Missing data	
Taxonomic level	**Analysis**	**Characters**	**Total**	**Ingroup**	**Total**	**Ingroup**	**Total (%)**	**Ingroup (%)**
Specimen	Upchurch *et al.* (2004)	32	16	11	319	196	38	44
	Scannella *et al.* (2014)	33	30	28	408	372	59	60
	Tschopp *et al.* (2015)	477	81	49	13404	7026	65	70
	Campbell *et al.* (2016)	155	40	19	3743	1617	40	45
Species	Arbour & Currie (2016)	177	44	41	3128	2659	60	63
	Mannion *et al.* (2017)	416	77	65	11124	7637	65	72
	Tschopp & Mateus (2017)	489	35	16	8806	3673	49	53

specimens, older juveniles and subadults grouped with the adult specimens.

In taxa, where derived clades experienced heterochronic evolutionary processes resulting in the retention of juvenile features into adulthood (as, e.g., during the theropod–bird transition; Bhullar *et al.* 2012), juvenile specimens of less derived taxa could resemble the more derived, neotenic forms. These juvenile specimens could, therefore, theoretically be recovered in more derived positions than the adults, but we do not know of any empirical study where such a result has been reported. However, both stem- and crown-slippage should be assessed and discussed as potential errors when including juvenile specimens in specimen-level phylogenetic analyses.

The most straightforward approach to avoid potentially misleading information from juvenile specimens would be their exclusion from the dataset (Mounier & Caparros 2015). However, juveniles of extinct taxa are not always easily recognisable as such, and it remains unclear where in the ontogenetic trajectory to set a potential threshold for exclusion. Whereas early juveniles often exhibit clear features of immaturity, and should be excluded, sexual maturity could only be established with certainty in few fossil vertebrates (e.g., Sato *et al.* 2005; Ji *et al.* 2010; Sander 2012; Hastings & Hellmund 2015). Skeletal maturity, on the other hand, can be identified with histological studies (e.g., Cormack 1987; Chinsamy-Turan 2005; Klein & Sander 2008), but is rarely reached, and corresponds almost never with sexual maturity, also because many vertebrates continue to grow as adults (Klein & Sander 2008; Scheyer *et al.* 2010). Indeed, the vast majority of fossil vertebrate specimens were probably still actively growing at the time of death, but do not have morphological features that would identify them as young juveniles. The case study with the sauropod *Europasaurus holgeri* (Carballido & Sander 2014) has shown that phylogenetically informative features may develop late in ontogeny in sauropods, but also that autapomorphic features of the species were present in specimens that were not skeletally mature, based on the incomplete fusion of the neurocentral synchondrosis in the vertebrae (Carballido & Sander 2014; see also Section 1.1.6). Thus, whereas early juveniles can be identified and excluded, subadult to sexually mature individuals cannot be distinguished in most analyses because of a lack of data. Using the more easily recognisable skeletal maturity as a threshold for exclusion might be misleading, however, and even result in very low numbers of available specimens, given that most fossil vertebrate specimens were still growing at their point of death. The inclusion of actively growing individuals is, thus, a necessity, but also not necessarily misleading. However, more case studies, such as the one by Carballido & Sander (2014), should be performed in a variety of taxa to assess the timing of the development of synapomorphic and autapomorphic features during ontogeny in various subclades.

As with the fragmentary individuals, exclusion cannot be advised if the juvenile specimen is the type of an ingroup species (as occurs, for example, in diplodocid sauropods; Tschopp *et al.* 2015). Also, in some data sets, it might be the case that juveniles are the only (or one of a few) relatively complete specimens, and are thus important for indirect comparisons among ingroup specimens (e.g., in the sauropod *Camarasaurus*; Gilmore 1925; Table 2), or that they represent rare finds in specific geographical areas or time epochs (e.g., Early Pleistocene hominins; Mounier & Caparros 2015). A number of possible approaches for minimising the negative influence of ontogeny on phylogeny during character scoring, analysis and species delimitation are discussed at relevant points later in this paper.

1.1.2. Character selection and construction. Character selection is rarely explained in phylogenetic studies, but can significantly impact the outcomes of an analysis (Poe & Wiens 2000). In general, the inclusion of as many characters as possible is recommended, even if they are variable among and within species (Poe & Wiens 2000). Specimen-level phylogenetic analysis presents a special case, because it allows for independent assessments of trait variability (Longrich 2015; Tschopp *et al.* 2015), especially when using maximum parsimony approaches, which are designed to minimise the number of homoplasies (Wiley & Lieberman 2011). Homoplastic characters are generally regarded as evolving faster than phylogenetically highly informative traits (Sites *et al.* 1996), which often only produce a single character-state change within a phylogenetic analysis, and are thereby recovered as unambiguous synapomorphies for that particular clade. Homoplastic characters add ambiguous information to the data matrix, which has led many researchers to exclude them *a priori* (see Poe & Wiens 2000, and references therein). However, a combination of information from slow- and fast-evolving characters might actually be advantageous to resolve the tree at different taxonomic levels (Wiens 2006). Indeed, both simulations and real case studies have shown that the *a priori* exclusion of homoplastic characters decreases accuracy and resolution of the resulting phylogenetic tree (Chippendale & Wiens 1994; Sites *et al.* 1996; Wiens 1998; Prevosti & Chemisquy 2010), at least as long as they do not include a large amount of missing data (Wiens 2006; see discussion in Section 1.1.3).

Homoplastic characters in specimen-level phylogenetic analyses have a high probability of describing features that are intraspecifically variable (Tschopp *et al.* 2015). As such, they add noise, and could possibly obscure the phylogenetic signal of other characters (Sites *et al.* 1996; Pisani *et al.* 2012; Townsend *et al.* 2012). However, case studies yield ambiguous results: whereas in some instances, deletion of the most homoplastic characters appears to increase general support and accuracy (Sites *et al.* 1996), the opposite appears to be the case when deleting all homoplastic characters (Sites *et al.* 1996; Wiens 1998). In fact, the exclusion of homoplastic

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

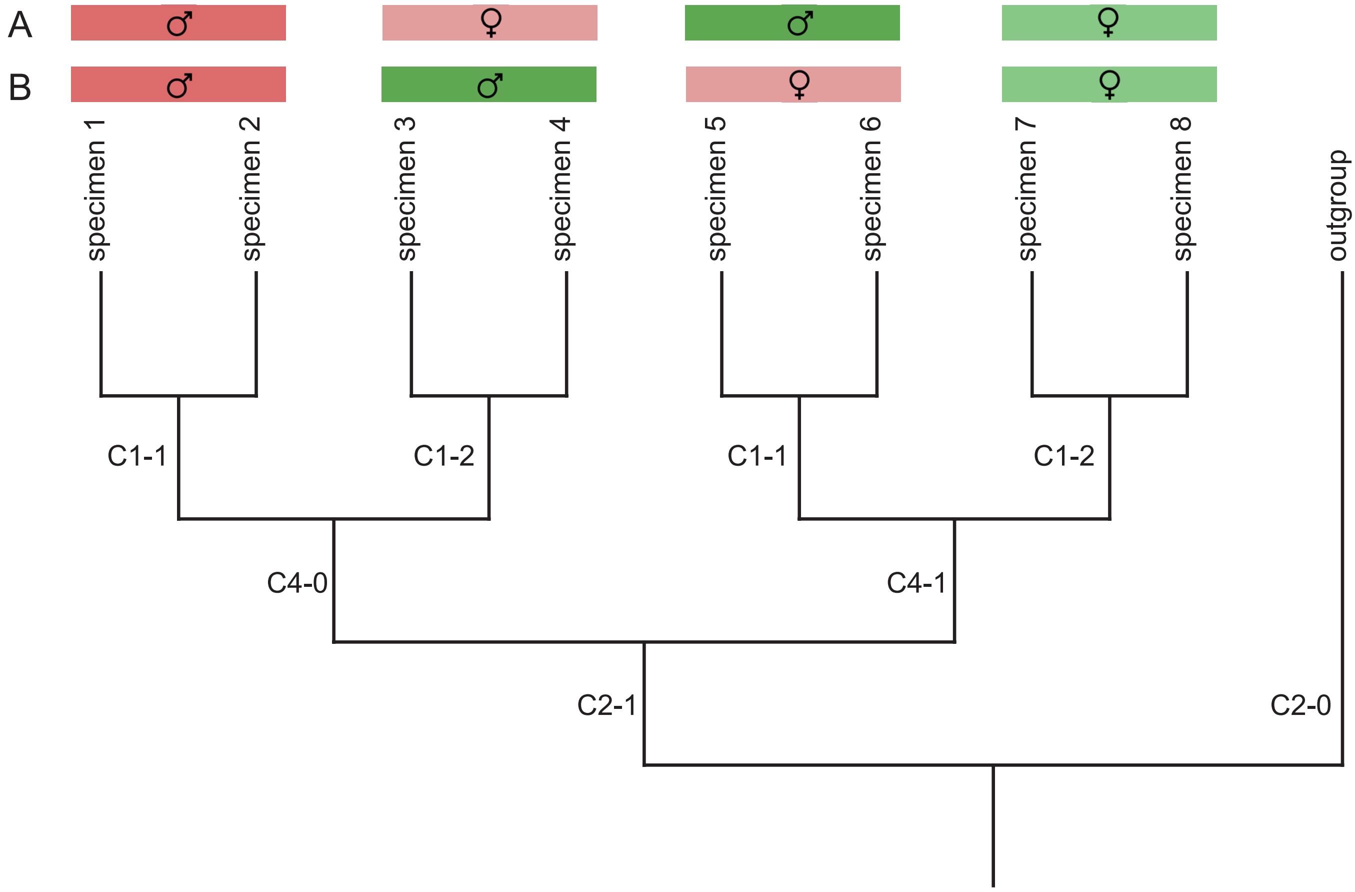

Figure 1 Potential influence of directed noise on tree topology. Directed noise because of sexual dimorphism, for example, can lead to misleading topologies. In this hypothetical tree, colours indicate the 'true' species (which we usually do not know in palaeontological datasets), tones and symbols the sexes. (A) The result when character 1 codes for a sexually dimorphic trait that equally occurs in both species. (B) The result if character 4 coded a sexually dimorphic feature. In case (B), character 1 codes for the true distinguishing feature between the species, but is overprinted by character 4, which codes for features shared among males or females across the two species. Mapping the character states diagnosing the subclades might help detect these phenomena: if distantly related subclades show the same diagnostic features (the different states of character 1 in the present example), they should be checked for potentially being sexually dimorphic.

characters might obscure potential phylogenetic information at a low taxonomic level (given that they evolve faster than other characters). The deleterious effects of increasing homoplasy resulting from adding more characters are outnumbered by positive effects on the accuracy of the phylogenetic analysis because of the additional information available (Prevosti & Chemisquy 2010). Also, it could be that certain traits are highly variable in one taxon, but less so in another clade (Farris 1969; Tschopp *et al.* 2015). Finally, the probability that the added noise created by homoplastic characters could produce a random signal that would be stronger than the one produced by highly phylogenetically significant characters, and that could thus overwhelm the latter, appears low (Farris 1969; De Laet 1997). In large datasets, we would expect it to be much more probable that the random support for different tree topologies within the noise would tend to be mutually contradictory instead of combining to obscure the true phylogenetic signals. Although this does not always appear to be the case when the number of character statements is small (Townsend *et al.* 2012; but see Prevosti & Chemisquy 2010), extensive taxon (or specimen, for that matter) sampling appears to reduce the negative impact of noise (Townsend *et al.* 2012).

Specimen-level phylogenetic analyses are potentially more prone to the effects of what can be termed 'directed' or 'coherent' noise (i.e., secondary non-phylogenetic signals in the data) that might overwhelm the true phylogenetic signal. Potential sources of such directed noise are shared ontogenetic or sexually dimorphic features, and ecologically controlled traits. These sources can result in the recovery of clusters of specimens in the most-parsimonious trees (MPTs), which represent juveniles (see Campione *et al.* 2013), males or females, or similar ecological adaptations instead of true phylogenetic relationships and/or species (Fig. 1). Whereas ontogenetic features can sometimes be recognised in fossil material, and sexually immature specimens could be excluded *a priori* (see Section 1.1.1), a similar approach is difficult for sexually dimorphic features. Osteological indicators for sex are rarely known in extinct taxa, but similar sex differences can occur across closely related taxa (e.g., in lacertid lizards; Arnold *et al.* 2007). In the worst-case scenario, individual female specimens from several taxa could, therefore, be grouped together, and form the sister-clade to a group of male specimens from the same taxa (see case B in Fig. 1). Indeed, Donoghue (pers. comm. in Vrana & Wheeler 1992) mentioned this as the main reason why he changed his mind after initially promoting specimen-level phylogenetics (see Donoghue 1985; de Queiroz & Donoghue 1990a, b; Vrana & Wheeler 1992). However, directed noise caused by sexual dimorphisms could potentially be identified in morphological datasets by character mapping: if a similar set of convergently acquired apomorphic features diagnoses subclades in equivalent phylogenetic positions in the sister clades at higher levels (Fig. 1), one should

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

give serious consideration to the potential confounding effects of sexually dimorphic features.

Ecological or functional convergences can occur differently in subsets of characters, resulting in an uneven distribution of homoplasy among the available characters. Such an uneven distribution has been shown to occur in mammals, where dental characters are more homoplastic than other osteological ones, and produce trees that are less compatible with molecular trees than the ones recovered using only non-dental osteological characters (Sansom *et al.* 2017). Such a different phylogenetic signal might indicate that teeth carry a largely functional signal instead of a phylogenetic one, and that, in extreme cases, the phylogenetic signal is overprinted by a functional and/or ecological signal. In order to assess if a dataset is affected by such an overprinting, it might be advisable to check if different subsets of characters carry different signals. This can be done by using Partitioned Bremer Support (see Parker 2016), or by dissecting the dataset into smaller sets including only the group of characters in question (e.g., dental *versus* cranial *versus* postcranial), and comparing the outcomes with a series of tests, as described in detail by Sansom *et al.* (2017).

Whereas the exclusion of homoplastic characters appears counter-productive, and negative effects can best be avoided by adding OTUs, this does not mean that highly homoplastic characters should have the same weight as highly parsimony-informative ones (Farris 1983; Goloboff 1993, 1995; Chippendale & Wiens 1994). This has led several workers (e.g., Farris 1969; Goloboff *et al.* 2008a, Goloboff 2014) to propose methods for identifying and downweighting homoplasies (see Section 1.2.1 for discussion of the different strategies), but these have not previously been considered in detail with respect to their utility in specimen-level phylogenetic analyses. In short, the most justified approach in character selection would be to use as many character statements as possible, including highly variable ones, as long as the latter do not include a high percentage of missing data.

1.1.3. Missing data. Missing entries can stem from both incompletely preserved specimens (particularly in vertebrate palaeontology, and in analyses at specimen level) and incompletely scored characters (Kearney & Clark 2003; Pol & Escapa 2009; Mannion & Upchurch 2010; Tschopp *et al.* 2018b). There is an expectation that, all things being equal, missing data are a particular problem for specimen-level analyses because greater completeness of OTUs in a conventional analysis is often achieved by combining multiple specimens into a single OTU. Whereas the use of individual specimens as OTUs reduces the risk of having chimeric higher-level OTUs, it will also tend to increase the relative amount of missing data per OTU. This would especially be the case when palaeontological species-level datasets are simply converted into specimen-level matrices. However, the challenge is not necessarily the missing data *per se*, but the amount of anatomical overlap between the included OTUs (see Section 1.1.1). Also, the relative amount of missing data in a palaeontological specimen-level analysis is not always higher when compared to species-level matrices (Table 2). It is, therefore, important to consider the real contents of a species-level OTU – if it only comprises an individual specimen, this should be stated clearly in the matrix. In fact, given that many fossil vertebrate species are only represented by a single specimen, phylogenetic analyses, even when formally run at species-level, are effectively often partial specimen-level analyses. Exceptions are analyses using similar matrices at different taxonomic levels, as, for instance, was done by Tschopp & Mateus (2017), who used a species-level matrix based on the specimen-level matrix of Tschopp *et al.* (2015). In their case, the amount of missing data was considerably reduced from 65 % (complete taxon sampling) or 70 % (only ingroup) in the specimen-level matrix to 49 % (complete) and 53 % (ingroup) in the species-level matrix (Table 2).

This reduction can also be quantified using the Character Completeness Metrics proposed by Mannion & Upchurch (2010), which considers the percentage of phylogenetic characters that can be scored for a specimen or species. The Chinese sauropod *Euhelopus zdanskyi*, for instance, is known from two incomplete specimens (Wiman 1929; Wilson & Upchurch 2009). The more complete one (PMU 24705) scores 47 % in character completeness, whereas, at the level of species, combining information from both specimens, character completeness increases to 68 % (Mannion & Upchurch 2010).

The metrics of Mannion & Upchurch (2010) are particularly low in sauropodomorph type specimens, which, on average, are only slightly more than half as complete as the species they typify, reaching 25.65 % of individual skeletal completeness. The situation is considerably better in ichthyosaurs, where holotype specimens have an average skeletal completeness of 45.49 % (ranging from 1 to 90.5 %), and reach 66 % of the completeness of the entire species (Table 3; based on data from Cleary *et al.* 2015). Whereas the completeness of sauropodomorph type specimens increased through time of description (Mannion & Upchurch 2010), there seems to be no such correlation in ichthyosaurs (Fig. 2). In any case, because species-level OTUs can always draw on one or more specimens, they are logically always equally or more complete than a specimen-level OTU.

The inclusion of highly incomplete specimens results in an extensive lack of anatomical overlap among the specimen-level OTUs in the matrix, and is likely to decrease resolution in the consensus trees (Huelsenbeck 1991; Kearney & Clark 2003; Wiens 2006; Butler & Upchurch 2007; Prevosti & Chemisquy 2010; Tschopp *et al.* 2015, 2018b). Both simulations and real case studies have shown that an increase in the relative amount of missing data lowers accuracy and increases errors (Wiens 2006; Prevosti & Chemisquy 2010; Sansom 2015). However, these case studies deleted information from already existing matrices, so that the result is not really about the impact of missing data in general, but about not including available data *a priori*, and thus the negative impact might be expected. When adding taxa or characters, even if they include a substantial amount of missing entries, accuracy increases in most cases, or at least remains similar to that achieved by the original matrix (Wiens 2006). Because missing data is no data, it cannot logically be added when adding incompletely scored characters or taxa – what we add is the amount of actual data scored in them. Therefore, even if the addition of more taxa and/or characters results in a relative increase of missing data in the entire dataset, we still increase the absolute amount of data that can be analysed, so that the positive results obtained by Wiens (2006) are to be expected.

Another concern is that character statements with a large number of missing entries may simulate the problem of long-branch attraction (Wiens 2006). This problem arises from the presence of two OTUs or characters, for which few data are available, but the information that is available might be convergent, as can be the case in highly homoplastic characters (see Section 1.1.2). Without the information on the true character-state distribution across the tree (because of too many missing entries), the two convergent taxa might be wrongly grouped together to the exclusion of others (Bergsten 2005; Wiens 2006; Tschopp *et al.* 2018b). However, even though adding new OTUs or characters might decrease the overall anatomical overlap in the dataset (Tschopp *et al.* 2018b), the addition of data is always recommended (Kearney & Clark 2003; Wiens 2006; Goloboff 2014). The relative

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Table 3 Skeletal completeness of holotype specimens of ichthyosaurs, and the species they typify. Data from Cleary *et al.* (2015).

Species	Year	Holotype	Species (total)	%
Acamptonectes densus	2012	15.75	35.25	45
Arthropterygius chrisorum	1993	28.5	35	81
Brachypterygius cantabridgiensis	1888	4	17.5	23
Brachypterygius extremus	1904	9.5	12	79
Brachypterygius mordax	1976	27.5	59	47
Brachypterygius zhuravlevi	1998	11	30	37
Californosaurus perrini	1902	56	79.5	70
Callawayia neoscapularis	1994	49.5	78.5	63
Caypullisaurus bonapartei	1997	68	75.5	90
Chaohusaurus geishanensis	1972	75.75	100	76
Cymbospondylus petrinus	1868	1	86	1
Cymbospondylus piscosus	1868	1	3.5	29
Eurhinosaurus longirostris	1851	61.5	97	63
Excalibosaurus costini	1999	50	94	53
Grippia longirostris	1929	13	63.75	20
Guanlingsaurus liangae	2000	90.5	100	91
Guizhouichthyosaurus tangae	2000	70.5	100	71
Guizhouichthyosaurus wolonggangensis	2007	44	44	100
Hudsonelpidia brevirostris	1995	57.5	64	90
Ichthyosaurus breviceps	1881	89.5	96	93
Ichthyosaurus conybeari	1888	53	98	54
Maiaspondylus lindoei	2006	34	34	100
Mixosaurus kuhnschnyderi	1998	53.5	94	57
Mixosaurus panxianensis	2006	34.5	90	38
Nannopterygius enthekiodon	1871	74.5	80	93
Ophthalmosaurus icenicus	1874	47	98	48
Ophthalmosaurus yasykovi	1999	48.5	63	77
Phalarodon fraasi	1910	10	16.5	61
Platypterygius americanus	1939	31.5	44.5	71
Platypterygius hercynicus	1946	65.5	65.5	100
Platypterygius kiprijanoffi	1968	36	38.5	94
Qianichthyosaurus zhoui	1999	90.5	100	91
Shastasaurus alexandrae	1902	21.5	46	47
Shastasaurus pacificus	1895	5.5	41	13
Shonisaurus popularis	1976	65.5	81.5	80
Stenopterygius triscissus	1856	85	98	87
Stenopterygius uniter	1931	81.5	98.5	83
Undorosaurus gorodischensis	1999	53.5	55	97

amount of missing data should, thus, not be reduced by omitting taxa or characters; rather, its deleterious effects should be addressed using approaches such as differential weighting and 'reduced consensus', as will be discussed in Sections 1.2.1 and 1.2.3. Moreover, one way in which choice of character construction can reduce missing data is to convert multistate characters (coded within a single column in the data matrix) into their equivalent additive binary form. Although this is only appropriate for those multistate characters that capture a morphological transition series (i.e., ordered; see Section 1.2.2), the use of additive binary coding has the benefit of reducing the amount of missing data. For example, a single multistate character scoring the number of vertebrae in the neck would have to be scored as '?' whenever the neck of a specimen was incompletely preserved, but can be scored for at least some of the states for the equivalent additive binary character (e.g., a combination of 0s, 1s and ?s scores would inform the analysis that the specimen had at least a given number of neck vertebrae, even though the exact number remains unknown – see Upchurch (1998) for elaboration of this point).

1.1.4. Character-state scoring. Characters can be coded either in a discrete way or as continuous characters. These continuous characters are a type of quantitative character that use the specific ratios, ranges of measurements or specific numbers in meristic features as states (Goloboff *et al.* 2006). As such, this approach further develops the idea of gap-weighting (Thiele 1993), in which large differences in quantitative traits between OTUs are upweighted compared to minute ones, but avoids discretisation of the actual values obtained from the OTUs (Goloboff *et al.* 2006). Advocates of such an approach mostly highlight the fact that state boundaries in discrete, quantitative character statements are often arbitrary, and their choice rarely explained and justified by the researchers (see Rae 1998, and references therein). Thus, the risk of influencing the analysis by choosing state boundaries that favour the recognition of a pre-conceived clade is relatively high (Mannion *et al.* 2013).

The implementation of continuous characters in the phylogenetics software TNT treats them by default as ordered (Goloboff *et al.* 2006). Thus, given that every single score forms its own character state, the sum of steps in a single continuous character is much higher than any discrete binary character. As already pointed out by Goloboff *et al.* (2006), there are weighting strategies that can be applied to address this issue, which will be discussed in Section 1.2.1.

General issues with this approach concern the choice of exact values or ranges as character scores, the use of mean or maximum or minimum values, and how to address incompleteness and deformation in fossils. Although these issues apply to any kind of phylogenetic analysis, they are particularly common when working at the specimen level, mostly because the sample size on which ratios and other values can be based is much lower than when working with species or higher-level taxa (e.g., some ranges, means, etc. will be based on a maximum sample size of two, like, for instance, the tibia:femur ratio in a single individual, or cannot be obtained from individual specimens, because of incompleteness).

Rae (1998) argued for the use of means or medians in the scoring of continuous characters, because variation could occur randomly or due to measurement errors, rendering 'central tendencies' (as he termed them) more appropriate estimations of the actual distribution of values within an OTU. When working with fossils, taphonomic deformation can add to the variation of numerical values, and even lead to differential character scoring (Tschopp *et al.* 2013), in particular when using continuous data. As exemplified in Figure 3, two cervical vertebrae of a single sauropod individual (SMA 0011, the holotype of *Galeamopus pabsti*, in this case) can be compressed transversely (Fig. 3a) or dorsoventrally (Fig. 3b), which leads to highly diverging shapes and ratios. Furthermore, specimen incompleteness might skew the analysis towards an extreme when only a statistical outlier can be sampled. If only a single, incomplete element is preserved from a specimen, it could even be that the incompleteness renders it impossible to obtain precise measurements and ratios (and thus precludes scoring as continuous characters), although they might be scorable in a discrete version of the character (Mannion *et al.* 2013). For instance, no exact ratios concerning tibial robustness are obtainable from a tibia lacking its distal end, but the preserved length might still result in a robustness ratio that exceeds the defined boundary of a

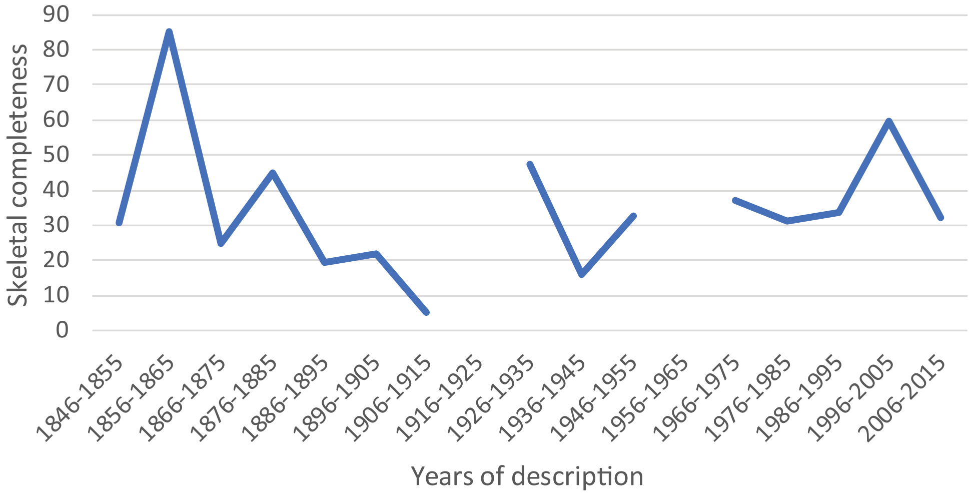

Figure 2 Skeletal completeness of holotype specimens of ichthyosaurs. Average completeness of species erected within ten-year bins from 1846 to 2015 are plotted. No species was erected between 1916 and 1925, and 1956 and 1965. Data from Cleary *et al.* (2015).

discretised state (e.g., the proximal-width-to-proximodistal-length ratio might be '0.15 or lower', showing that it lies below the state boundary of 0.2). In this case, a continuous character could not be scored when using central tendencies, but one could argue that a range could be included. This range could span from the ratio using the preserved length as minimum value to the highest value exhibited by any other OTU. However, such a range would exaggerate the actual variability and overlap with a large number of more precise ranges from other individuals, effectively hiding phylogenetic information (Giovanardi 2017). Taphonomically increased ranges due to deformation processes pose the same problem.

Given that it is statistically more probable that a single element found from a vertebral column, for instance, is closer to the central tendency than to any minimum or maximum value displayed along the column, and given that ranges pose their own risks, especially when working with fossils, mean or median values should be preferred over ranges, or minimum or maximum values. Discretisation of a quantitative character can be useful in ratios that are more prone to deformational processes (Arbour & Currie 2012; Tschopp *et al.* 2013), effectively hiding potentially misleading information. However, state boundaries in discrete characters should be defined based on statistical analyses rather than on preconceived taxonomic or phylogenetic interpretations. There is a large number of papers concerning discretisation of continuous data in statistics (e.g., Jiang & Sui 2015; Cano *et al.* 2016, and references therein), and some methods are also implemented in the usual office packages for computers. To our knowledge, a study on which kind of discretisation would work best in phylogenetics has not yet been made.

1.1.5. Polymorphisms. Polymorphic traits are traits that are variable within species (Wiens 1995, 2000). At the species level, they can be treated differently, and several theoretical approaches have been compared by Wiens (1995, 2000), who suggested the use of a frequency approach, meaning that species should be scored for the character state that occurs with the highest frequency within the species. By splitting a species-level OTU into single specimens, some polymorphisms can be avoided, because they derive from intraspecific variability.

Although reducing polymorphisms deriving from intraspecific variability, a specimen-level approach can still be affected by polymorphisms. In single specimens, these can be created by serial variation throughout the vertebral column (e.g., Barbadillo & Sanz 1983; Wilson 2012; Chamero *et al.* 2014; Böhmer *et al.* 2015; Tschopp 2016), bilateral asymmetry (e.g., Palmer 1996; Hoso *et al.* 2007) or pathologic processes (e.g., Rothschild & Martin 2006; Foth *et al.* 2015; Tschopp *et al.* 2016). Whereas an exclusion of pathologic data is advisable for obvious reasons, serial variation and bilateral asymmetry can still provide important phylogenetic data (Palmer *et al.* 1994; Böhmer *et al.* 2015). Even though polymorphisms in a single specimen-level OTU might indicate that the trait is individually variable and has no phylogenetic/taxonomic significance, this is difficult to establish *a priori* and should be evaluated in light of specimen-level relationships – exclusion is, therefore, not an appropriate option (Wiens 1998; Poe & Wiens 2000). In serially variable traits, frequency-based approaches could work in a similar way as in the studies reported by Wiens (1995, 2000). In vertebral columns with distinct regionalisation (see Müller *et al.* 2010 for a review in tetrapods), it can also make sense to subdivide the column into separate morphological areas, as is often done in sauropod dinosaurs and squamates (see, e.g., the descriptions and characters for anterior cervical, or posterior caudal vertebrae in Carballido *et al.* 2012; D'Emic 2012; Gauthier *et al.* 2012; Mannion *et al.* 2013; Otero *et al.* 2014; Tschopp *et al.* 2015, 2018c). Often, such subdivisions are made numerically, because clear-cut morphological boundaries are difficult to identify (Mannion *et al.* 2013; Tschopp *et al.* 2015), but increasing information is now available on serial variation in a number of vertebrate animals based on geometric morphometrics, so that more detailed and less arbitrary morphological subdivisions can be made (e.g., Müller *et al.* 2010; Burnell *et al.* 2012; Böhmer *et al.* 2015). Splitting vertebral columns into subregions is an analogous approach to subdividing taxa into lower-level taxonomic units in order to minimise the number of polymorphisms. A combination of character splitting and frequency-based scoring approaches, therefore, seems the best option in this case, even though this would also increase the relative amount of missing data.

Bilaterally asymmetric traits can occur due to developmental plasticity or as a result of abnormal developmental processes. Whereas the latter should be treated as pathology and excluded, the first could still be phylogenetically informative because it may indicate a trend to acquiring a new feature that may become fixed by natural selection (Palmer 1996). Distinguishing between the two may be difficult in fossils, but in systems where asymmetry is ubiquitous, as, for instance, in the lamination pattern of vertebrae of saurischian dinosaurs (Wilson 1999, 2012), it is probably safe to assume they derive from plasticity instead of widespread pathology.

Generally, only a small number (usually two) of bilaterally occurring elements are present in a vertebrate skeleton.

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

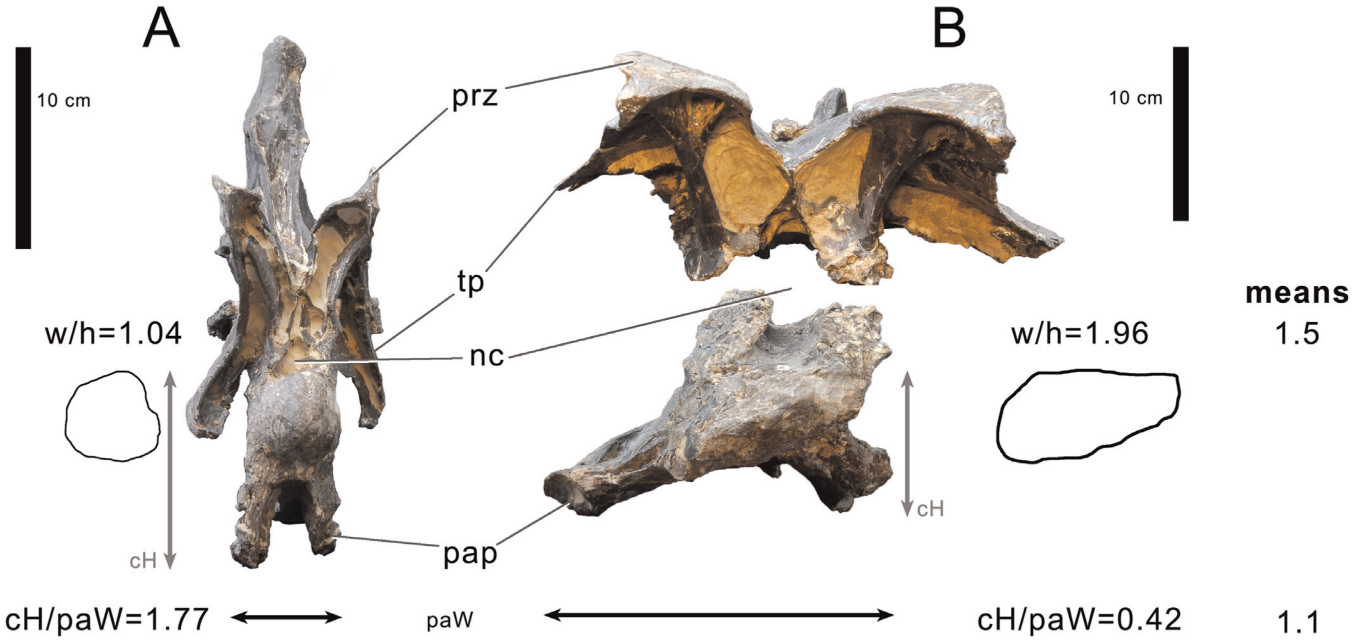

Figure 3 Taphonomic deformation impacts measurements and ratios. The cervical vertebrae 5 (A) and 8 (B) of *Galeamopus pabsti* SMA 0011 were compressed transversely and dorsoventrally, respectively. Anterior condyle outline shape and ratios such as centrum height/width across parapophyses are examples of affected measurements and ratios potentially useful for phylogenetic analysis. Minimum and maximum values or ranges can, therefore, yield misleading data for continuous character scores, and central tendencies such as means should be preferred. Abbreviations: cH = centrum height; h = height; nc = neural canal; pap = parapophysis; paW = width across parapophyses; prz = prezygapophyses; tp = transverse process; w = width. Vertebrae figured in anterior view (modified from Tschopp & Mateus 2017), and scaled to the same anterior condyle height.

Frequency-based, or majority approaches, therefore, cannot be applied. Possible treatments of such characters outlined by Wiens (1995) include: (1) the 'any-instance' method, where the sheer occurrence of a trait (even if only on one of several equivalent elements) is treated as if the character state was invariably present; (2) the 'missing' method, where asymmetric traits are scored as missing data; (3) the 'polymorphic' method includes polymorphic scores; (4) the 'scaled', 'unordered' or 'unscaled' methods, where binary characters are coded such that they include a third polymorphic character state as state 1. The character can then be treated as ordered ('scaled') or unordered, and binary characters, where no asymmetry was observed can be coded as normal binary character statements, without a polymorphic intermediate state ('unscaled', see Wiens 1995 for more details). The 'any-instance' method would be the most straight forward approach in a specimen-level analysis, but ignores a potential phylogenetic signal in the occurring asymmetry. Also, it remains unclear how to score an asymmetrical individual in a multistate character, following this method (Wiens 1995). Scoring a specimen as '?' in the trait in which it shows bilateral asymmetry results in loss of information, and the same happens when using the polymorphic approach if the character is binary, because the analysis treats a polymorphic score in binary character statements as '?' (Wiens 1995, 1998; Brazeau 2011). Of the two latter treatments, a score as polymorphic at least provides information to a researcher who inspects the data matrix, because it clearly indicates the presence of two or more states, whereas a score as 'missing' completely hides any information. The treatments that include the most potential phylogenetic information are those where a separate polymorphic character state is included (in the present case, this state might be called 'bilaterally asymmetric'). When applying this approach to a real dataset, the scaled method yielded the highest accuracy, although without large differences compared to the unscaled method (Wiens 1998).

Bilateral asymmetries can be an issue in continuous characters, in particular in meristic features. For instance, tooth counts in lizard dentaries and maxillae often vary in left and right elements (Arnold *et al.* 2007). Given that these variations are usually small, and counts generally precise, this might be a case where scoring ranges could actually be helpful in order to include as much morphological information as possible, without risking widely overlapping ranges among large numbers of individuals in the dataset.

1.1.6. Ontogenetic traits. As previously mentioned, ontogenetically variable traits can introduce problems into specimen-level analyses. However, there are several approaches one can adopt during scoring and subsequent steps in the analysis, if it is necessary to include a juvenile specimen. In sauropod dinosaurs, the number and prominence of vertebral laminae, and vertebral pneumatisation, strongly increases during ontogeny (Wilson 1999; Wedel *et al.* 2000; Wedel 2003; Bonnan 2007; Schwarz *et al.* 2007; Tschopp & Mateus 2017), which led Carballido & Sander (2014) to propose four Morphological Ontogenetic Stages (MOS) applicable to sauropod vertebrae. In the case of *Europasaurus holgeri*, Carballido & Sander (2014) found that when scoring all the different MOS as distinct OTUs in the phylogenetic analysis, the juvenile MOS 1 and 2 occurred in a more 'basal' position compared to MOS 3 and 4. This is probably due to the fact that a large number of vertebral character statements used in sauropod dinosaur phylogenetics code for variation in these traits, and that well-developed lamination and pneumatisation is both an adult and a phylogenetically derived feature among sauropods (Wilson 2012). Many other ontogenetically variable features are known in the vertebrate skeleton, so that the most straightforward approach would simply be to avoid scores of ontogenetically variable traits in obviously juvenile specimens. If scored, these characters can be downweighted during the analysis, and not considered for species delimitation (see

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Sections 1.2.1 and 1.3.1), but the exclusion of these scores altogether would probably still be more methodologically sound.

1.2. Phylogenetic methodology

1.2.1. Character weighting. Specimen-level analyses provide an opportunity to include characters coding for minute differences in morphology, and check whether or not they might be informative at some taxonomic level. However, such characters might not have a genetic basis, but could represent individual variation caused by plasticity, ecophenotypic effects or any other non-genetic cause (Tschopp *et al.* 2015), which manifests as homoplasy in the phylogenetic analysis (see Section 1.1.2). In such cases, equal weighting is not advisable, in particular when working with large-scale specimen-level analyses. Indeed, Goloboff *et al.* (2008a, 2018) have shown that weighting against homoplasy increased the reliability and stability of tree topologies in morphological datasets.

Downweighting can be implemented *a priori*, or during the tree search, or iteratively after each tree search (Farris 1969; Goloboff 1993, 2014; De Laet 1997; Goloboff *et al.* 2008a). The most intuitively correct, and least subjective way, to downweight potential homoplasies is to use a method called 'implied weighting' (Goloboff 1993), which is implemented in the phylogenetic software package TNT (Goloboff *et al.* 2008b). This approach downweights characters with widespread homoplasy as part of the tree search function (Goloboff 1993, 2014; Goloboff *et al.* 2008a, 2018). The equation is as follows:

$$\text{weight} = k/(k + [\text{observed steps} - \text{minimum steps}])$$

where k is the 'concavity value'.

This equation shows that implied weighting can be performed with different concavity values ('k-values'; Goloboff 1993, 1995, 2014). These values describe the slope of the curve, defining how strongly characters with different homoplastic rates are downweighted. The lower the k-value, the more strongly a highly homoplastic character is downweighted during the phylogenetic analysis compared to a less variable character. A k-value approaching zero would, therefore, effectively exclude homoplastic characters, whereas one approaching infinity would weight them all equally. However, other than avoiding extreme values, there seems to be little biological or methodological basis for selecting any specific k-value (Goloboff 1995; Turner & Zandee 1995). Recent studies showed that a k-value of around 12 produced the most accurate results in a series of morphological datasets (Goloboff *et al.* 2018), but it is possible that this value varies slightly in different taxa, or even in different phylogenetic analyses of a single taxon. However, this cannot be used as an argument to dismiss implied weighting *a priori*, it just means that one should perform different analyses with varying k-values, and compare the results (Goloboff *et al.* 2008a) by using statistical or stratigraphic measurements, as will be discussed in Section 1.2.3. Ultimately, implied weighting might provide a simple solution to the problem found by Sites *et al.* (1996): that is, the exclusion of all homoplastic characters reduced accuracy, whereas the exclusion of only the most homoplastic ones increased it.

Implied weighting as initially proposed by Goloboff (1993) can be negatively influenced by missing data, because characters with a large amount of the latter have a higher probability of showing fewer homoplasies, and would thus tend to be upweighted relative to more completely scored characters (Goloboff 2014). In a worst-case scenario, where the data set includes very incompletely scored characters, the weaker downweighting could effectively lead to a strengthening of the long-branch attraction phenomenon simulated by the missing data (Wiens 2006; Tschopp *et al.* 2018b). Nonetheless, real case studies using matrices with missing data showed that implied weighting approaches performed better than equal weighting (Prevosti & Chemisquy 2010). Moreover, Goloboff (2014) implemented the so-called 'extended implied weighting' approach in the software TNT, which not only downweights the characters based on their homoplastic rate and the chosen k-value, but also adapts the k-value for every character individually based on its proportion of missing entries. Congreve & Lamsdell (2016) dismissed this methodology, in part, because polymorphic or inapplicable characters are often treated as missing data and could, therefore, be wrongly penalised by an extended implied weighting approach. However, the proposed methodology actually just enables the use of different k-values for every single character (Goloboff 2014), so that these issues could also be addressed manually instead of applying the default, automated script (Goloboff *et al.* 2018). Moreover, at least inapplicable character states can be recognised by the latest versions of TNT and, thus, be excluded from the algorithms for extended implied weighting (Goloboff *et al.* 2018).

Simulations using modelled phylogenies have recently shown that traditional implied weighting performs worse than equal weighting and probability-based approaches such as Bayesian (Congreve & Lamsdell 2016; O'Reilly *et al.* 2016). On the other hand, case studies using real morphological matrices appear to show the contrary (Prevosti & Chemisquy 2010; Brinkman *et al.* 2017), and also extended implied weighting seemed to work well under certain circumstances when analysing specimen-level data in lizards (Villa *et al.* 2017). One reason for these discrepant conclusions could be that the modelled phylogenies did not accurately represent a real distribution of homoplasy within a morphological dataset (Goloboff *et al.* 2018). By analysing the actual distribution of homoplasies in numerous morphological data sets, Goloboff *et al.* (2018) showed that earlier simulations (Congreve & Lamsdell 2016; O'Reilly *et al.* 2016) did, indeed, represent this distribution incorrectly. Comparisons of the methodologies with newly simulated trees based on the distribution of homoplasies found in real data sets resulted in extended implied weighting being the strategy that recovered the most accurate trees, followed by the traditional implied weighting approach (Goloboff *et al.* 2018). Even though implied weighting retrieved a proportionally larger number of both correct and incorrect groupings in data sets with more homoplasy, compared to equal weights (Congreve & Lamsdell 2016; Goloboff *et al.* 2018), the relative amount of added correct groups exceeded the relative increase of incorrect groups, thereby increasing overall accuracy, especially when using extended implied weighting (Goloboff *et al.* 2018). Collapsing branches with low support was shown by Goloboff *et al.* (2018) to reduce the number of incorrect groups, but this also reduces the number of weakly supported, correct groups (Goloboff *et al.* 2018), and generally lowers the information content of the recovered trees by increasing the number of polytomies.

These issues become especially important if the matrix was specifically constructed to test assumptions of homology at the level of single individuals, which likely results in more homoplasies in the data set. At present, it is not yet clear whether a stronger downweighting function might help to reduce the number of incorrect retrieved groups in data sets with a larger amount of homoplasies, and if these incorrect

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

groups might be identified somehow if we do not know the correct tree. Moreover, only Goloboff *et al.* (2018) also included the extended implied weighting approach in their simulations, and most other studies used rather strong downweighting functions (e.g., $k = 1$, 3, 5 and 10 in Congreve & Lamsdell 2016). Additional tests with real data sets (such as that of Villa *et al.* 2017) and a higher range of downweighting functions will be needed to compare the performance of different weighting methods, including extended implied weighting, in order to resolve this debate.

Tschopp *et al.* (2015) noted that using an implied weighting strategy was useful to address the potentially misleading ontogenetically variable characters, because the ontogenetic changes add variability to these characters, which, therefore, have a higher homoplastic rate, and so are downweighted more strongly than less variable ones. However, if the characters are highly parsimony-informative among adult specimens, the variability introduced by juvenile specimens would partly obscure this information, and, combined with implied weighting, even reduce its impact on the calculation of the most-parsimonious trees. Omitting scores for ontogenetically variable traits in obviously juvenile specimens, therefore, appears more appropriate than applying implied weighting to reduce their deleterious effects.

1.2.2. Character ordering. Phylogenetic characters can have multiple states that describe different relative sizes or shapes of a single feature. Multistate characters can be treated as ordered or unordered, or with step-matrices (Hauser & Presch 1991; Wilkinson 1992; Wilson 2002; Brazeau 2011). Ordering and step matrices impose different degrees of directional morphological state transformations onto the character concerned, whereas a treatment as unordered accepts all possible changes between character states as equally probable (Wilkinson 1992; Brazeau 2011). For instance, in an ordered character with three states (0, 1, 2), a morphological change from state 0 to state 2 would need two evolutionary steps, and thus also increase the length of the most-parsimonious tree relative to a treatment of the same character as unordered. By using a step-matrix, a researcher can define the possible direct evolutionary steps even more precisely, and can allow for a so-called 'easy loss character', in which the evolution from character state 0 to 2 costs more than from 2 to 0, implying that it is more likely that the character will pass through state 1 on its evolutionary way to 2, whereas the reversal could be direct (Wilson 2002). The differences and rationales of why, how and if multistate characters should be ordered have been reviewed recently by Brazeau (2011), and apply to phylogenetic analyses at any taxonomic level equally, so that there is no need to discuss it in detail here. Brazeau (2011) concluded that multistate characters should be ordered if they code for quantitative characters, or if they describe an obvious morphological transformational series. We follow this recommendation here. The use of step-matrices, even if theoretically adding methodological soundness, probably has little influence on the result in most cases, but needs additional time investment to prepare the file for the analysis. The implementation of which characters should be ordered, on the other hand, is uncomplicated and fast.

1.2.3. Tree searches and consensus trees. Whereas no specific requirements apply to the methodology of tree searches when using specimen-level matrices, several points have to be addressed once a set of trees has been obtained, and before proceeding to species delimitation. The basic tree topology can be influenced by ontogenetically variable characters, consensus methods can hide phylogenetic structure and analyses under differential weighting (as recommended in Section 1.2.1) can produce conflicting tree topologies.

Ontogenetically variable characters can influence tree topology, and thus also taxonomic interpretations. If one prefers downweighting over exclusion of ontogenetic character states (as in Tschopp *et al.* 2015), the position of juvenile specimens in the phylogenetic trees, on which species delimitation will be based, will be influenced by these characters. Another approach was followed by Campbell *et al.* (2016), who conducted a specific test to assess the influence of ontogeny on tree topology. They followed the principles of a 'ontogenetic analysis' as initially proposed by Brochu (1996), and ran it in parallel to the phylogenetic analysis. In an ontogenetic analysis, only traits known to be ontogenetically variable are used as character statements (Brochu 1996; Carr & Williamson 2004; Campbell *et al.* 2016). Character states are adapted to follow supposed ontogenetic changes, and multistate characters are treated as ordered during the parsimony analysis.

Campbell *et al.* (2016) used their ontogenetic analysis to check if small-to-large-sized skulls of two different species of *Chasmosaurus* fell on two distinct ontogenetic trajectories, which could be used to distinguish the two species. Although the result of their study was negative, such an approach could also be used to verify if the topology of the tree recovered by the ontogenetic analysis reproduces the findings of the phylogenetic one. If this is the case, and if the ontogenetic trajectory also correlates with an increase in body size, one should expect that the topology found by both analyses was strongly influenced by ontogenetically variable characters. Such an integrative approach of ontogenetic and phylogenetic analysis is probably more appropriate than simply reducing the weight and, thus, impact of ontogenetic character states as done by Tschopp *et al.* (2015). A combination of an ontogenetic analysis and the exclusion of obviously juvenile character states during scoring for the phylogenetic analysis under implied weighting approaches would likely provide the most accurate results.

Most specimen-level analyses of fossil taxa have had to cope with the problem of a very high number of most-parsimonious trees, and, therefore, large polytomies in the strict consensus tree (Yates 2003; Scannella *et al.* 2014; Tschopp *et al.* 2015; Campbell *et al.* 2016). Polytomies can derive from both the genuine absence of a branching pattern (so-called 'hard polytomies'), and insufficient data in the phylogenetic matrix to recover an entirely resolved tree ('soft polytomies'; Maddison 1989; Purvis & Garland 1993). Thus, in the context of specimen-level analyses, hard polytomies would represent the lack of phylogenetic structure below the level of species, and could be used as an indication for the delimitation of species (see Section 1.3.1 for a detailed assessment). However, complete strict consensus trees do not always report the entirety of phylogenetic signal present in the matrix (Wilkinson 1995), so a distinction of hard and soft polytomies is crucial before making positive inferences based on an apparent lack of hierarchical structure. It is possible that a few, highly unstable, taxa (specimens in this case) might produce large soft polytomies, even though the rest of the included OTUs remain stable (Wilkinson 1995). Often, the main reason for this instability is missing data in fragmentary specimens lacking anatomical overlap (see Section 1.1.3). One approach to ameliorate such a problem is to prune the unstable OTUs from the trees *a posteriori*, and then apply tests to identify their most-parsimonious phylogenetic positions (e.g., see Tschopp *et al.* 2015). The underlying tree structure hidden in 'soft polytomies' in the complete strict consensus tree can, thus, be revealed by reduced strict consensus approaches, or *a posteriori* pruning of the most unstable taxa.

Multiple conflicting tree topologies can be generated by the presence of unstable taxa, as discussed previously, but can

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

also occur because of the application of an array of different starting assumptions or analytical protocols to the same data set. Thus, alternative positions of specimens have to be tested with a number of approaches. There are several support measures to evaluate tree accuracy. Given that these are not specific to specimen-level phylogeny, we will only discuss them briefly herein. Following our recommendation to use different weighting strategies during phylogenetic analysis, we will specifically focus on the impact of weighting on the various support measures.

The most widely used support metrics are resampling measures such as bootstrapping and jackknifing. Källersjö *et al.* (1999) and Goloboff *et al.* (2008a) used jackknife frequencies to calculate and compare group support between analyses with different character sets or weighting strategies. Källersjö *et al.* (1999) compared analyses based on molecular data under equal weighting, with and without the highly homoplastic third-codon positions, whereas Goloboff *et al.* (2008a) compared different *k*-values in implied weighting. The two approaches are equivalent, because equal weighting and the exclusion of characters basically represent the two extremes of *k*-values in implied weighting (Goloboff 1993; De Laet 1997). However, jackknife frequencies and bootstrapping have been reported to produce distorted support values under certain circumstances, when the analyses to be compared use different weighting strategies (Goloboff *et al.* 2003). Instead, Goloboff *et al.* (2003) proposed the use of symmetric resampling, which normalises the impact of the up- and downweighting of characters based on a probability constant ('*P*'); however, even here, absolute values of support can be hard to interpret, and might even support groupings that are not found in the optimal trees (so-called 'spurious groups'; Goloboff *et al.* 2003; Kopuchian & Ramirez 2010). Thus, rather than absolute support values from resampling, Goloboff *et al.* (2003) suggested the use of the frequency differences between contradictory groups, frequency slopes derived from curves formed by the use of different values of *P* or a sample of the values at a particular threshold of *P*. For further details, we refer the reader to Goloboff *et al.* (2003). All of these support measures have their own problems (Goloboff *et al.* 2003; Kopuchian & Ramirez 2010), and, to our knowledge, frequency differences have rarely been used to calculate group support in vertebrate palaeontology (e.g., Marx 2011; Mannion *et al.* 2013). Frequency differences can actually support spurious groups, just like absolute values (Goloboff *et al.* 2003; Kopuchian & Ramirez 2010). Frequency slopes can be misleading, because they can change drastically along the curve (Goloboff *et al.* 2003; Kopuchian & Ramirez 2010). Finally, the threshold for the specific sample (i.e., where to compare group support) depends on the dataset (Goloboff *et al.* 2003). In their case studies using real phylogenetic matrices and varying weighting and resampling strengths, Kopuchian & Ramirez (2010) found that Jackknife resampling methods generally performed better than bootstrapping, but that symmetric resampling did not uniformly perform better than traditional jackknifing. Although symmetric resampling is more consistent than the traditional method in which groups are supported, it also finds more spurious groups (Kopuchian & Ramirez 2010). Perhaps unexpectedly, Kopuchian & Ramirez (2010) also found a tendency that the absolute values still performed better than the frequency differences. Thus, it remains somewhat unclear which of these statistical support measures is actually the most reliable, so that a pluralistic approach is probably warranted at this stage.

Bremer supports (initially proposed as decay analysis; Bremer 1988, 1994; Donoghue *et al.* 1992) depend on the calculation of suboptimal topologies to test which clades are also found in trees that are longer than the most-parsimonious trees. In analyses with substantial amounts of missing data, this can become a computing problem, because it is likely that the number of MPTs is already very large (e.g., >60,000 in Tschopp *et al.* 2015). Moreover, Bremer supports can be strongly influenced by single, very unstable, OTUs (Wilkinson *et al.* 2000), as occurs relatively often in palaeontological specimen-level analysis. An alternative might be the so-called Double Decay Analysis developed by Wilkinson *et al.* (2000), but this approach has rarely been used in vertebrate palaeontology, or has been found to be unfeasible even in only moderately large data sets with around 50 OTUs and up to 221 characters (Butler *et al.* 2008; Brusatte *et al.* 2010). Finally, it remains unclear how to interpret the fractional tree lengths resulting from the use of continuous characters and/or implied weighting approaches (Goloboff & Farris 2001). When using TNT, tree length under implied weights is reported to four decimal places, such that increases can occur by as little as 0.0001. Given that these fractional tree lengths, and thus also the Bremer support values, change with the applied *k*-value, it remains uncertain how different Bremer support values should be compared between conflicting tree topologies resulting from analyses with different *k*-values. This could potentially be addressed by using the Relative Fit Difference (RFD) developed by Goloboff & Farris (2001). RFD calculates the difference of how often a certain node is supported *versus* contradicted by the data, providing a percentage. Therefore, the tree length itself does not impact the RFD, and topologies from different weighting strategies could be compared (Goloboff & Farris 2001). The RFD was used to calculate support for specific nodes in Mannion *et al.* (2013), but limitations on the number of trees that can be stored using TNT resulted in the highest detectable support values being 44 %. Nevertheless, RFD might be the most useful and most easily applicable derivative of Bremer supports to compare conflicting topologies resulting from differing weighting strategies.

A similar approach to Bremer support, based on differential tree lengths, was used by Tschopp *et al.* (2015). In that study, specimens recovered in conflicting positions in the analyses under equal and implied weighting were subjected to constrained tree searches, in which the questionable specimens were forced to lie in the position found by the other analysis. Because the absolute values of tree lengths using differential weighting are hard to compare, Tschopp *et al.* (2015) compared relative increases in tree length between the constrained tree searches to infer the most-parsimonious phylogenetic position of critical specimens. However, relative length increase in the tests of Tschopp *et al.* (2015) were nearly always very low (below 1 %), and it remains unclear if the observed differences really are statistically significant.

Low support for specific groups within a tree might generally result from implied weighting approaches, if the synapomorphies uniting a group are highly homoplastic, and, therefore, downweighted. In the case of specimen-level analyses, where highly homoplastic characters might represent individual variation, low support could indicate that these OTU clusters represent spurious groups within a species, instead of potentially distinct subpopulations. Collapsing weakly supported nodes based on relative fit differences, as initially proposed by Goloboff & Farris (2001), could be used to circumvent this issue, but has never been applied in any specimen-level analysis to date. As mentioned in the discussion concerning the use of implied weighting, weak group support can also occur in correct groups, so that a collapse of these nodes always runs the risk of obscuring potentially useful information (see also Goloboff *et al.* 2018). However, collapsing groups with very low relative fit differences might be a promising approach to avoid spurious within-species tree resolution.

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Whereas all the analyses discussed above concern data intrinsic to the phylogenetic matrix and analysis, stratigraphic indices might provide an alternative to test for the support of specific clades using extrinsic data, in particular in palaeontological datasets. Stratigraphic data of the single OTUs can be implemented directly using some approaches of Bayesian Inference (Cau 2017), but no convincing strategy has yet been proposed for adding this data in parsimony analyses. Instead, stratigraphic data and phylogenetic topology can be treated as separate data sources and compared using an array of indices that capture aspects of how well a branching topology matches the stratigraphic order of the appearance of taxa. A number of such stratigraphic indices have been proposed, reviewed in detail by Bell & Lloyd (2015), who also presented an easily usable script for the statistics software R (Bell & Lloyd 2014). One limiting factor is that, in many cases when working with specimen-level phylogeny, specimens come from similar strata or the strata are not dated with enough precision to be able to apply stratigraphic indices in a significant way. Nevertheless, in cases where finely resolved stratigraphic data are available for all or most specimens, very detailed analyses of character evolution through time can be attempted, as discussed in Section 4.3. Of course, including stratigraphic data in the analysis, or using it to decide on a more ‘accurate’ tree topology will render subsequent biostratigraphic studies based on these trees circular, just as in palaeobiogeographic studies of taxa, where fossil material is attributed to extant species based on their geographical occurrence (Bell *et al.* 2010).

1.3. Post-phylogenetic analysis

1.3.1. Species delimitation. Specimen-level cladistic analyses allow the reassessment of taxonomic assignments and nomenclature without having to accept previous identifications or referrals (Tschopp *et al.* 2015; Cau 2017). However, it does not provide direct evidence for the delimitation of taxonomic levels such as species or genera, and there seems to be no single objective criterion to do so, be it based on morphology or molecular data (Sites & Marshall 2004; Carstens *et al.* 2013; Satler *et al.* 2013; Kimura *et al.* 2016). Disagreements over species delimitation can stem from the use of different data, from variable evolutionary processes acting on different sources of data and from different methodological approaches (Wiens & Penkrot 2002; Dettman *et al.* 2003; Sites & Marshall 2004; Carstens *et al.* 2013; Satler *et al.* 2013; Kimura *et al.* 2016). Whereas many approaches exist for molecular data (see reviews in Sites & Marshall 2004; Carstens *et al.* 2013), only a small proportion of them are applicable to morphological data, and only a few approaches have been proposed to address the problem of species and genus distinctions based on morphology specifically (Wiens & Penkrot 2002; Sites & Marshall 2004; Benson *et al.* 2012; Tschopp *et al.* 2015; Kimura *et al.* 2016).

Species-delimitation methods can be tree-based or character-based (Wiens & Penkrot 2002). Although all these approaches have to be guided by tree topology, monophyly (the most straightforward criterion for the definition of species and genera) cannot be used as the sole criterion for recognising species in a specimen-level analysis. In the case of anagenetic speciation, some but not all members of a species become ancestors of a descendent species (Wiens & Penkrot 2002), which renders the ancestral species as a whole necessarily paraphyletic (Brummit 2002; Longrich 2015). Such a pattern could theoretically be detectable in a phylogenetic tree resulting from a specimen-level analysis. Because of this, some researchers advocated the entire abandonment of the species-level taxon in phylogenetic nomenclature (e.g., Pleijel & Rouse 2000), but, by doing so, some individual organisms might not be referable to a ‘least-inclusive taxonomic unit’ (*sensu* Pleijel & Rouse 2000; see also Baum 1998). Some species-delimitation approaches used in molecular studies allow for paraphyletic species (Carstens *et al.* 2013), but they have not yet been further developed for application to morphological data. Carr *et al.* (2017) presented a species-level phylogenetic analysis of tyrannosaurid dinosaurs, and inferred anagenetic speciation based on sister-taxon relationships and differential stratigraphic but overlapping geographic ranges. An adaptation of such an approach to specimen-level analyses holds promise but has not yet been attempted. Proposed approaches for morphological data by various researchers are explained and discussed below.

Wiens & Penkrot (2002) proposed a tree-based method combining information from bootstrap supports and geographic distribution of the included OTUs (populations in their case, but this could equally be applied to specimens). Following this approach, species delimitation depends on how weakly or strongly supported a specific clade is, and how much tree topology follows geographical segregation between populations (Wiens & Penkrot 2002). Additionally, Wiens & Penkrot (2002) proposed a character-based approach, which uses the occurrence of fixed and exclusive diagnostic features as cut-off points to define species boundaries. However, these two approaches did not lead to the same conclusions in their study case of the iguanian *Sceloporus*, and yielded discordant results compared to approaches based on molecular data (Wiens & Penkrot 2002). That the two approaches almost necessarily lead to discordant results should be expected, given that they are based on fundamentally different ideas of character evolution: as shown by Sites & Marshall (2004), tree-based methods are often based on recognising phylogenetic splits or nodes, which do not necessarily have to be diagnosable by distinct apomorphic features. Indeed, Wiens & Penkrot (2002) noted that some species, as recognised by their character-based approach, actually just represented groupings of OTUs that did not exhibit any of the diagnostic features used to define other species, and that no diagnostic feature could be statistically proven to be fixed in any of these clades. Additionally, Kimura *et al.* (2016) demonstrated that the appearance of diagnostic features is delayed in respect to lineage splitting in murid mammals. High intraspecific variability among osteological features has also been shown in the lacertid lizard *Lacerta* (Villa *et al.* 2017), where no single trait could be identified as a unique, unambiguous autapomorphy of a species; rather, only combinations of traits were found to be species-specific.

The tree-based approach of Wiens & Penkrot (2002) relies on bootstrap support measures. In specimen-level phylogenetic analysis, bootstrap values rarely reach 70 % – a value proposed to indicate high support by Hillis & Bull (1993) and used as a cut-off value by Wiens & Penkrot (2002) – or even 88 % (as proposed by Zander 2004). Nonetheless, the type of support value could be changed to one less prone to the negative impacts of morphological data and missing entries (see discussion in Section 1.2.3), and a stratigraphic criterion could be added to the geographic one when analysing fossil OTUs. In general, integrating different types of data to test interpretations of species delimitations is expected to lead to more accurate results, and is being applied increasingly frequently in extant organisms (see Carstens *et al.* 2013, and references therein, for examples).

The proposed species delimitation methods of Benson *et al.* (2012) and Tschopp *et al.* (2015) can be regarded as adaptations of approaches used in molecular specimen-level studies based on genetic distances. Benson *et al.* (2012) calculated

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

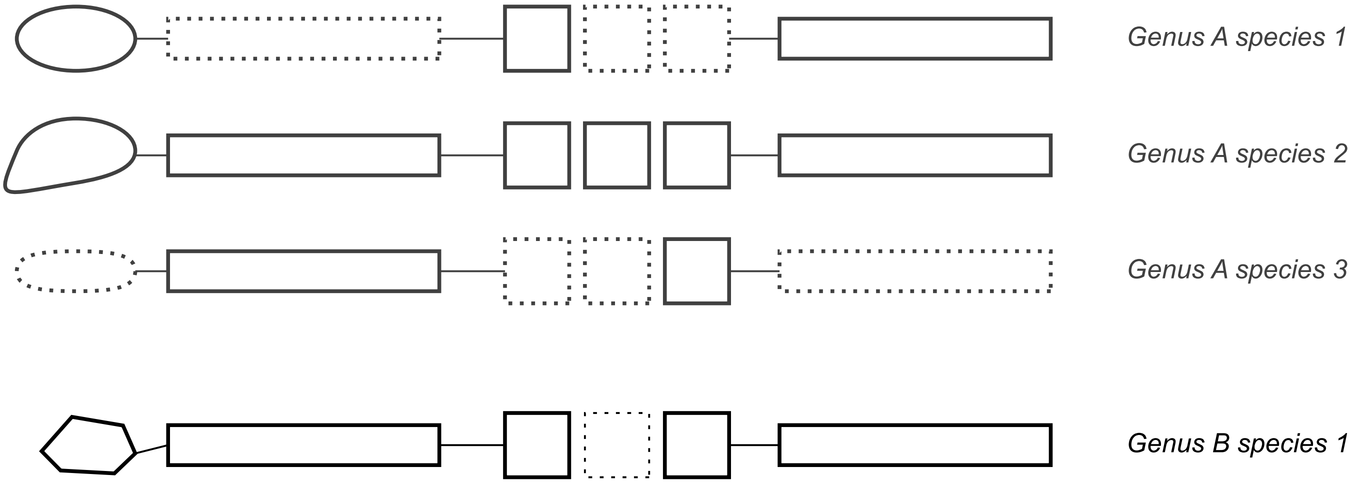

Figure 4 Missing data can reduce pairwise dissimilarity scores to 0 %. Four hypothetical skeletons, where only skull shape (to the left) changes. Rounded skulls are an autapomorphy of genus A, and angled ones an autapomorphy for genus B. Different skull shapes distinguishing species within genus A. Hypothetical, not preserved, elements are marked with dashed lines. In such a simplified case, a skeleton not preserving postcranial elements can still be identified at species level (e.g., *Genus A species 1*), whereas the incomplete fossil actually belonging to *Genus A species 3* does not show any dissimilarities with any species of genera A and B, and can only be referred to a higher-level taxon. Pairwise dissimilarity between this fragmentary specimen and the specimens of the other species would be 0 %.

morphological dissimilarity between species of different genera of plesiosaurs. They identified the comparable character states between the various OTUs within the genera, and calculated how many of them are scored differently. By doing so, Benson *et al.* (2012) included a value of completeness of the sampled species and specimens. However, highly fragmentary specimens might simply not preserve characters coding for variation at species level, but only at genus or even higher systematic levels. If this is the case, dissimilarity scores between these fragmentary specimens and more complete ones of potentially different species within the same genus will approach 0 %; this would obviously not represent the true extent of differences that would be recognisable if a complete skeleton were available (Fig. 4).

The distance method was applied by Tschopp *et al.* (2015), who also developed an additional approach, which they termed ‘apomorphy count’. Recovered apomorphies are qualitatively assessed based on their variability within the clade they define, and among the other OTUs. At the level of specimens, recovered ‘autapomorphies’ of single specimens are not necessarily species autapomorphies, whereas recovered ‘synapomorphies’ of specific clades might actually represent autapomorphic features of a particular species. Single-specimen ‘autapomorphies’ are, therefore, especially prone to simply code for intraspecific variability. Consequently, Tschopp *et al.* (2015) excluded recovered ‘autapomorphies’ from their counts, if they were shared with other specimens of closely related species (i.e., shown to be homoplastic; Fig. 5). Additionally, Tschopp *et al.* (2015) excluded ‘synapomorphies’ from their apomorphy counts, if they were variable within the clade they define, shared with specimens of other clades and found solely by one of the two analyses they performed. Apomorphies considered valid after this step (which could be both ‘autapomorphies’ and ‘synapomorphies’) are then counted for two branches of a dichotomy, and summed in order to determine the number of major morphological changes between the two. As such, only characters deemed significant enough by the software TNT to be considered apomorphies, and which are not too variable among the ingroup are counted. These apomorphies can also be distributed unequally: in an extreme case, they could all occur on one branch of the dichotomy only, with the sister-group having not a single apomorphic feature. The apomorphy count, therefore, also partially accounts for unequal rates of morphological evolution.

Based on earlier taxonomic interpretations of specific and generic distinctions, for which sister-taxon relationships have been confirmed by their specimen-level analysis, Tschopp *et al.* (2015) then defined thresholds for how many significant morphological changes were historically accepted within a species and within a genus, and applied these consistently across their ingroup taxon Diplodocidae. In the latter study, two traditionally recognised species clusters were confirmed by the analysis (*Apatosaurus ajax* and *A. louisae*, and *Diplodocus carnegii* and *D. hallorum*), and changes between these sister-groups amounted to a maximum of 12, leading Tschopp *et al.* (2015) to use 13 changes as a minimum threshold to justify generic separation. At the species level, a number of specimens historically referred to a single species were found as sister-OTUs by Tschopp *et al.* (2015) as well. Differences between these specimens summed to maximally five, so that six changes were considered as sufficient for justifying specific distinctions (Tschopp *et al.* 2015). However, it is important to note that the absolute number of changes depends on the dataset, and can thus not be uniformly applied to any specimen-level phylogenetic analysis. Concerns about this method are the fact that the resulting absolute numbers vary between any single phylogenetic analysis performed, and that highly incomplete specimens are likely to show fewer apomorphic features. Both methods (pairwise dissimilarity and apomorphy counts), in part, take earlier, and well-accepted, interpretations of species and genera as a basis for the definition of the taxonomic thresholds, and thus also include the taxonomical history of a given clade to some extent.

Kimura *et al.* (2016) proposed a combination of phenetic, ecological and diagnosability criteria to study lineage sorting in murid mammals, based on morphometric and carbon isotope analyses. Although their study was not based on a phylogenetic analysis, these criteria could be easily adapted for use with a cladogram. Interestingly, and thanks to their extensive and stratigraphically well dated data set, Kimura *et al.* (2016) found that the different species-delimitation thresholds did not occur simultaneously, but that, based on the phenetic criterion, new species could be recognised earlier in geological time than based on the other criteria. This finding correlates

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

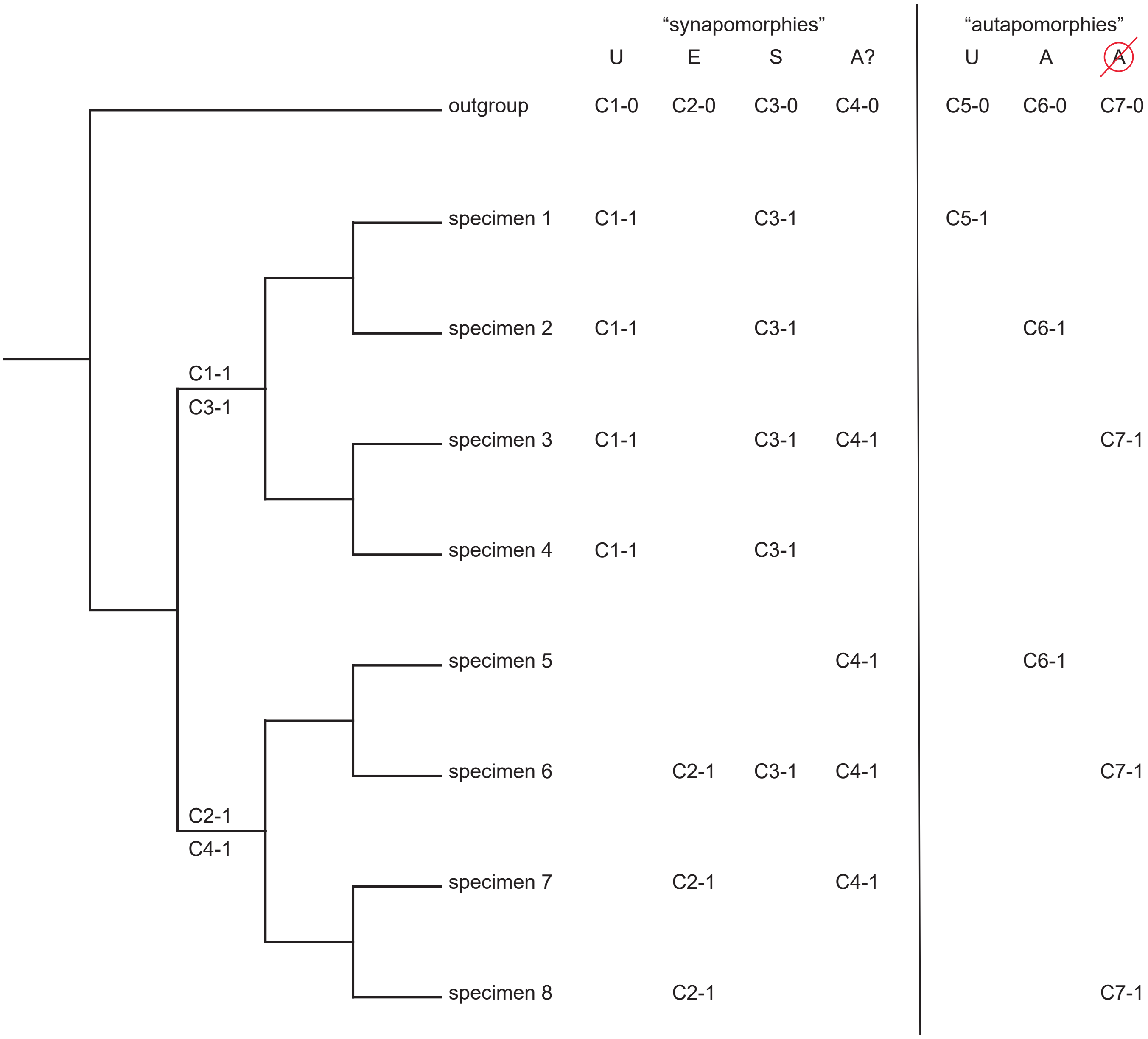
"synapomorphies"
U
E
S
A?
"autapomorphies"
U
A
A
outgroup C1-0 C2-0 C3-0 C4-0 C5-0 C6-0 C7-0
specimen 1 C1-1 C3-1 C5-1
specimen 2 C1-1 C3-1 C6-1
specimen 3 C1-1 C3-1 C4-1 C7-1
specimen 4 C1-1 C3-1
specimen 5 C4-1 C6-1
specimen 6 C2-1 C3-1 C4-1 C7-1
specimen 7 C2-1 C4-1
specimen 8 C2-1 C7-1
C1-1
C3-1
C2-1
C4-1

Figure 5 Qualitative assessment of 'synapomorphies' and 'autapomorphies' within a specimen-level context, following Tschopp *et al.* (2015). Acronyms with numbers indicate the character states that diagnose particular clades (in the tree), and the hypothetical distribution of these derived states among the ingroup. 'Synapomorphies' can be unambiguous (U, shared among all members of the clade they diagnose, and only among them), exclusive (E, occur only in specimens belonging to the clade they diagnose, but not in all of the specimens), shared (S, shared among all members of the clade they diagnose, but not only) and ambiguous (A, shared by most members of the clade they diagnose, and also by specimens belonging to other groups). The latter are the most dubious 'synapomorphies', and probably not all of them should be considered valid. Tschopp *et al.* (2015) did not consider ambiguous 'synapomorphies' found only by one of their two analyses for the apomorphy count. 'Autapomorphies' can be unambiguous (U) and ambiguous (A). Ambiguous 'autapomorphies' shared with specimens in a closely related clade were not counted for the apomorphy count as implemented by Tschopp *et al.* (2015).

well with the interpretation of a species as a lineage, as is the case in the General Lineage Concept (de Queiroz 1998). Based on the assumption that species lineages diverge gradually during the process of speciation, and that they gradually accumulate distinguishing features along the way, different operational criteria (such as the ones used by Kimura *et al.* 2016), can be plotted onto diverging lineages, and evaluated in light of the General Lineage Concept.

Although the studies and approaches mentioned above yielded promising results concerning species delimitation, it remains unclear if the outcomes represent accurate identifications of the boundaries between true biological species. Indeed, populations exist today that are only reproductively isolated due to behavioural incompatibility (e.g., Nanda & Singh 2012). Although this can obviously not be detected in extinct species, behavioural incompatibility can be a first step during cladogenesis in the context of the General Lineage Concept, followed by morphological distinctiveness due to diverging evolution. While morphologically indistinct, 'biological species' might be an issue when comparing extinct with extant forms, it is not necessarily a problem when working with fossil taxa alone. What we need to develop are consistent and reproducible studies for taxonomic clustering at the lowest possible level. In palaeontological datasets, this can only be done based on morphological differences. Even if these clusters do not represent exactly true biological species, a use of distance

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

measures or apomorphy counts will produce consistent and objective units that can be counted in diversity studies.

None of the proposed species-delimitation approaches is without problems. In fact, the various competing species-delimitation methods are based on different species concepts (Adams 2001; Sites & Marshall 2004; Kimura *et al.* 2016), and effectively represent the operational ways of how to apply these concepts to recognise species in nature (Adams 2001). Given that the numerous species concepts (both theoretical and operational) just define species at different steps of the speciation process (and can, indeed, be united in the General Lineage Concept for species, as proposed by de Queiroz 1998), it is paramount to apply a number of operational criteria to assess species delimitation (Sites & Marshall 2004; Bacon *et al.* 2012; Satler *et al.* 2013). Conflicting outcomes can then be evaluated qualitatively in light of speciation processes, as has been successfully achieved with palaeontological material by Kimura *et al.* (2016). Such a need for an integrative approach to species delimitation has been confirmed by the results of a case study of fungi by Dettman *et al.* (2003), where the phylogenetic species recognition approach (based on genetic distance) identified an additional species, which was still able to produce viable offspring with the sister-group. Similarly, generally accepted species of plants exhibited only some of the criteria applied in case studies of palms and *Primula*, implying that speciation has not yet led to complete lineage sorting in these taxa (Bacon *et al.* 2012; Schmidt-Lebuhn *et al.* 2012). These examples of molecular studies and the case study of fossil murids by Kimura *et al.* (2016) show that by applying different operational concepts to taxa with a good fossil record, it is possible to trace morphological speciation along a phylogenetic tree.

2. Ceratopsian case study

In order to illustrate some of the challenges outlined above, we conducted a case study based on the analysis of Campbell *et al.* (2016) on chasmosaurine ceratopsians, which used a modified version of the matrix of Sampson *et al.* (2010). Reanalysis of this study is informative, because Campbell *et al.* (2016) did not apply several of the methodological steps outlined herein to address specific challenges. For instance, Campbell *et al.* (2016) treated all multistate characters as unordered and performed the analysis under equal weights (J. Campbell, pers. comm. 2018). They pruned OTUs only *a posteriori*, as recommended herein, but the deleted taxa were selected based on their amount of missing data rather than their instability in the MPTs. Finally, Campbell *et al.* (2016) delimited species based on a morphometric study of a character of the frill (the variable angle of an embayment on the posterodorsal bar) rather than either the distance measure or apomorphy count approaches outlined previously.

2.1. Methodology

Herein, we treated the multistate characters that appeared to describe clear morphological transitions as ordered (characters 40, 41, 50, 60, 68, 70, 80, 89). Some of these characters had to be rescored to bring the states into the right order to describe a linear transition (characters 40, 41, 50, 70, 80; see supplementary material, available at https://doi.org/10.1017/S1755691018000877). During the analysis with TNT v. 1.1 (Goloboff *et al.* 2008b), we applied an extended implied weighting strategy, with a *k*-value of 5, and otherwise followed the search strategies of Campbell *et al.* (2016). A second analysis was performed with the original matrix under equal weights, applying only the character ordering, as outlined in Section 1.2.2, and agreement subtree and pruned tree options in TNT in order to assess possible hidden phylogenetic structure in the large polytomy found by Campbell *et al.* (2016, fig. 5a).

Before applying species delimitation methods, we collapsed the nodes with low supports by using tree bisection and reconnection (TBR), as suggested by Goloboff *et al.* (2018). We tentatively applied the apomorphy count to the resulting tree as a means of delimiting species. Given that the ingroup just includes two genera, we excluded all ambiguous 'synapomorphies', and all ambiguous 'autapomorphies' shared with any other member of the ingroup during the qualitative assessment of the apomorphies found by TNT (the final counts are given in the supplementary material). The apomorphy count had to be slightly adapted because the TBR collapsing resulted in a partly unresolved tree, so that the sums of apomorphies could not always be counted between two branches of a dichotomous node. Therefore, we calculated the average count across all possible sister-group relationships within a polytomy.

2.2. Results

2.2.1. Analysis under extended implied weights. The analysis with ordered multistate characters and under extended implied weights yielded a single, completely resolved phylogenetic tree with a length of 15.69203 (Fig. 6). The only clade of the ingroup recovered by Campbell *et al.* (2016), including the two specimens referred to *Vagaceratops irvinensis*, is also found here, as part of a larger clade, which also includes the type specimen of *Chasmosaurus russelli* (CMN 8800; Fig. 6). This entire clade forms the sister group to a clade including the type specimen of *C. belli* (CMN 0491; Fig. 6). Three specimens are found as successively more basal OTUs to these two clades: ROM 839, CMN 1254 and AMNH FARB 5401, which are the type specimens for *C. brevirostris*, *C. canadensis* and *C. kaiseni*, respectively. All the specimens referred to *C. russelli* by Campbell *et al.* (2016) are found in the clade with the type specimen of *C. belli*, whereas the type specimen of *C. russelli* is found in a clade with two specimens previously referred to *C. belli* (Fig. 6).

2.2.2. Analysis under equal weights. The reanalysis of the original matrix provided as supplementary material by Campbell *et al.* (2016) under equal weights, and with ordering of some multistate characters (see list in Section 2.1), yielded more than 30,000 most-parsimonious trees (we only allowed TNT to store 30,000 trees for this preliminary analysis) with a length of 297 steps – four more than reported by Campbell *et al.* (2016) – which is probably a result of the ordering of some of the multistate characters in our analysis.

Our reanalysis found the same large polytomy within Chasmosaurinae as did Campbell *et al.* (2016). Neither the *a posteriori* pruning processes, as implemented in TNT, nor an agreement subtree revealed more underlying phylogenetic structure.

2.2.3. Apomorphy count. The sums of changes between two branches of a node ranged from zero to three, which is very low compared to those reported by Tschopp *et al.* (2015). However, as pointed out previously, these absolute numbers depend on how a matrix is constructed. As a guideline to subdivide species following historical taxonomic practice, we took the sums of changes between the clades, including the holotypes of the two generally accepted species *Chasmosaurus belli* and *C. russelli*, which amounts to two (Fig. 6). For the necessary number to define a genus, we checked the sum of changes between the entire clade attributed to *Chasmosaurus* and its closest outgroup, *Agujaceratops*, which corresponds to three. Based on these counts, *Vagaceratops irvinensis* would only be considered a different species within a paraphyletic

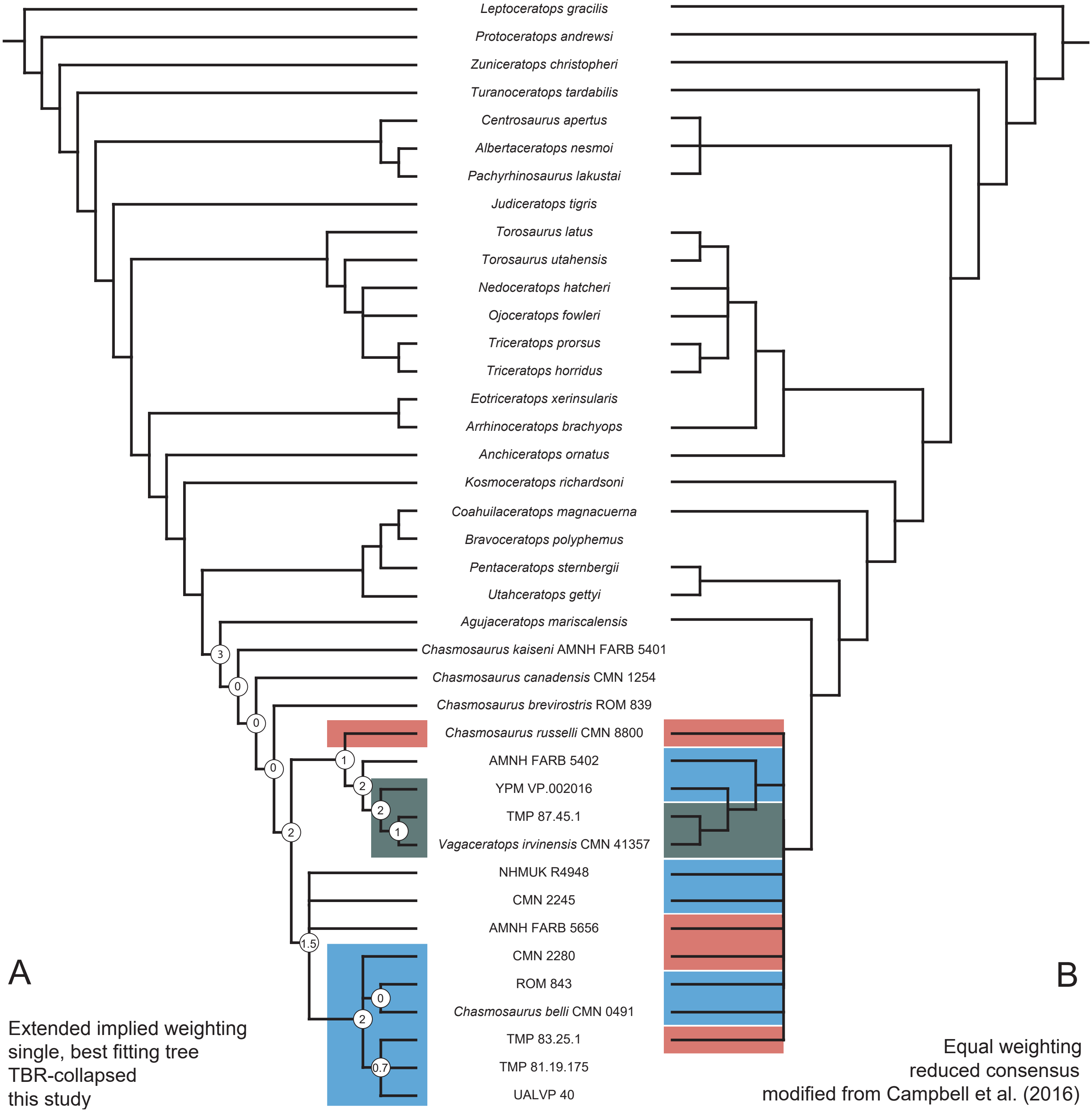
Leptoceratops gracilis
Protoceratops andrewsi
Zuniceratops christopheri
Turanoceratops tardabilis
Centrosaurus apertus
Albertaceratops nesmoi
Pachyrhinosaurus lakustai
Judiceratops tigris
Torosaurus latus
Torosaurus utahensis
Nedoceratops hatcheri
Ojoceratops fowleri
Triceratops prorsus
Triceratops horridus
Eotriceratops xerinsularis
Arrhinoceratops brachyops
Anchiceratops ornatus
Kosmoceratops richardsoni
Coahuilaceratops magnacuerna
Bravoceratops polyphemus
Pentaceratops sternbergii
Utahceratops gettyi
Agujaceratops mariscalensis
Chasmosaurus kaiseni AMNH FARB 5401
Chasmosaurus canadensis CMN 1254
Chasmosaurus brevirostris ROM 839
Chasmosaurus russelli CMN 8800
AMNH FARB 5402
YPM VP.002016
TMP 87.45.1
Vagaceratops irvinensis CMN 41357
NHMUK R4948
CMN 2245
AMNH FARB 5656
CMN 2280
ROM 843
Chasmosaurus belli CMN 0491
TMP 83.25.1
TMP 81.19.175
UALVP 40
3
0
0
0
1
2
2
1
2
1.5
0
2
0.7
A
Extended implied weighting
single, best fitting tree
TBR-collapsed
this study
B
Equal weighting
reduced consensus
modified from Campbell et al. (2016)

Figure 6 Different weighting strategies lead to conflicting tree topologies in ceratopsian dinosaurs. The tree obtained under extended implied weighting (A) is better resolved than the one under equal weighting (B), modified from Campbell *et al*. 2016, even after TBR-collapsing. The systematic referrals of Campbell *et al*. (2016) are contradicted by the apomorphy count applied to the tree obtained using the extended implied weighting approach (see numbers in circles in A). The specimens referred to *Chasmosaurus russelli* are highlighted in red; specimens referred to *C. belli* are highlighted in blue; and specimens referred to *Vagaceratops irvinensis* are highlighted in dark green. Non-highlighted specimens in (A) are specimens with unclear taxonomic assignments (see Sections 2.2 to 2.3).

C. russelli, and not a distinct genus. However, both nodes along the lineage from *C. russelli* to *V. irvinensis* have an apomorphy count of two. A similar condition occurs along the stem of *C. belli* (Fig. 6). Accepting the General Lineage Concept, these continued, elevated counts might be an indication of gradual morphological change during the speciation process. However, stratigraphic tests would be needed to sustain such a claim. The apomorphy count thus supports the validity of three species within *Chasmosaurus*, but no distinct genus *Vagaceratops*.

2.3. Discussion

As shown in Section 2.2.1, the analysis under extended implied weighting recovered a much more resolved tree than the one under equal weights, even after TBR-collapsing. Moreover, most of the referrals by Campbell *et al*. (2016) could not be confirmed based on this tree topology, indicating that the single character proposed as distinguishing the two species *Chasmosaurus russelli* and *C. belli* by Campbell *et al*. (2016) might not be taxonomically informative. According to these authors, the two species can be distinguished by the embayment

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

of the posterior parietal bar, which is deep in *C. russelli* and shallow in *C. belli*. Although we cannot know the correct phylogenetic tree, our study implies that this character should be assessed in more detail, in particular concerning alternative interpretations such as sexual dimorphism. The latter has already been tentatively suggested by Lehman (1990), and might have to be reconsidered given our analysis.

Our results highlight the importance of using different weighting strategies, and a combination of methodological approaches that suit the specific challenges of a specimen-level phylogenetic analysis. However, given that our tests are only preliminary, the underlying causes of the potentially conflicting taxonomic interpretations based on Campbell *et al.* (2016), and the tree recovered herein, are better addressed by experts in chasmosaurine anatomy.

3. Recommendations

Recommendations for the various steps of a specimen-level phylogenetic analysis are collated and summarised in this section. For detailed rationales and case studies, see the discussion in Section 1.

3.1. Matrix construction

Phylogenetic matrices for specimen-level analyses should generally include as many data points as possible. Neither character selection nor OTU sampling should be guided by the amount of missing data. An inclusion of all holotype specimens in the analysis is necessary for systematic reviews. The only justification for *a priori* exclusion of certain specimens is when they are incomplete, juvenile non-type specimens, which could mislead the analysis because of the typically higher number of plesiomorphic traits in individuals of an early ontogenetic stage.

Character scoring should include approaches to address polymorphisms along the vertebral column and bilateral asymmetry. The most straightforward and promising approaches are frequency or majority scoring for serially variable characters, and the inclusion of an intermediate character state for bilaterally asymmetric traits. Continuous characters can be used, but should be scored with a value representing a central tendency instead of ranges or minimum or maximum values. If juvenile specimens have to be included, they should not be scored for reportedly ontogenetically variable traits. Ordered multistate characters should be represented by their additive binary equivalents in order to reduce the impact of missing data.

3.2. Phylogenetic methodology

Characters should be weighted differentially, using an extended implied weighting approach as implemented in the software TNT with variable *k*-values. Multistate characters should be treated as ordered if they are quantitative (including continuous characters), or if they describe clear transitions in morphology.

Polytomies in the resulting consensus trees cannot be taken as evidence for species-level clades, but have to be analysed for possible hidden phylogenetic structure by using reduced consensus approaches. At the same time, weakly supported nodes should be collapsed to avoid the recovery of spurious groups. Conflicting topologies recovered after performing the analysis with different weighting constants are best evaluated using a combination of methods (e.g., jackknifing, relative length increases in constrained searches). Additional tests might be based on data extrinsic to the analysis itself, such as stratigraphic or geographic ranges, but this must be stated clearly to avoid circularity in subsequent biostratigraphic or palaeobiogeographic studies.

3.3. Species delimitation

Species delimitation should be carried out based on several approaches, and the differing results assessed from a cautious taxonomic perspective. A combination of tree-based approaches with measures of morphological distance and possibly additional, extrinsic data are expected to provide the most accurate results. However, when using extrinsic data, the same concerns apply here as when testing for accuracy in tree topology (see Section 3.2).

4. Future research

4.1. Validation of the method

As has happened frequently with many other biological and palaeobiological techniques, the development and application of specimen-level morphological phylogenetic methods have proceeded prior to any attempt to validate its accuracy. Validation of the methodologies of morphological specimen-level phylogenetic analyses, using extant taxa, is the first step that should be undertaken. This has been proposed for species delimitation methods by Sites & Marshall (2004), and has been carried out using molecular approaches in some fungi and plants (Dettman *et al.* 2003; Bacon *et al.* 2012). Without such tests, any follow-up study addressing the further potential of specimen-level analyses based on morphology (see Section 4.3) will be flawed and lack a firm methodological base. Extant taxa have to be chosen carefully, and should represent species and genera, where several recent phylogenetic studies based on multiple molecular sequence data confirm at least monophyly of the ingroup. Validation studies should be undertaken for a number of disparate and distantly related clades, in order to assess if the methodology that works best is the same across clades, or has to be adapted for each group of organisms. The studies of Wiens & Penkrot (2002) on lizards and Dettman *et al.* (2003) on fungi would suggest the latter: whereas different methodologies led to discordant results in lizards (Wiens & Penkrot 2002), the opposite was the case in fungi (Dettman *et al.* 2003). Thus, a wide survey seems to be necessary to detect significant patterns.

Aside from a general test of whether specimen-level morphological phylogenetic analyses are capable of accurately identifying species among extant taxa, validation and testing is also needed for each of the alternative steps and assumptions available to the researcher. For example, it would be interesting to examine whether the morphological distance approach of Benson *et al.* (2012) or the apomorphy-based approach of Tschopp *et al.* (2015) yields the most accurate assessments of species delimitations among extant taxa where the 'correct' answer is already known based on molecular phylogenies or direct field observations of reproductive isolation. Again, it might be that different protocols are variably successful with particular clades or types of organisms, but this has yet to be investigated in any detail.

Simulations are an additional tool to assess methodological issues, but their utility and applicability to a wide taxonomic range depend strongly on study design (Carstens *et al.* 2013). Therefore, validation studies with real morphological data preferably gathered first-hand should be expected to provide more meaningful results. Nevertheless, simulations could prove to be highly useful to model and address the impact of missing data and of the treatment of ontogenetic features on tree topology (see Wiens 2003 and Carballido & Sander 2014

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

for examples simulating missing data and the influence of ontogeny, respectively).

4.2. Beyond parsimony

In addition to validation in different taxa, it will also be important to analyse and compare the accuracy and performance of phylogenetic methodologies other than parsimony, such as Bayesian inference, Maximum Likelihood and Network analysis. Bayesian inference has been shown to be a promising tool for specimen-level phylogeny, because it is possible to allow for the recognition of ancestor-descendent pairs (Cau 2017). However, there is an ongoing debate on the accuracy of maximum parsimony *versus* probability-based approaches, in particular regarding the applicable models of character evolution in probability-based approaches when analysing morphological data (e.g., Wright & Hillis 2014; O'Reilly *et al*. 2016; Goloboff *et al*. 2018; Sansom *et al*. 2018). Network analysis might represent a promising approach because it is able to recognise patterns of reticulate evolution and horizontal gene or trait transfer (Morrison 2005), which should be expected to be ubiquitous when using individual organisms as OTUs. Comparisons of these different approaches are rare in vertebrate palaeontology, however, so it remains unclear to what extent these methodologies can fulfil their promise. Therefore, we herein concentrated on parsimony approaches, but we note that the entire discussion concerning the interpretation of phylogenetic topology equally applies to trees recovered by means of other methodologies.

4.3. Potential of phenotypic specimen-level phylogeny

Detailed phylogenetic trees of species known from well-dated stratigraphic successions provide the basis for the study of physical drivers of evolution. Where phylogeny is analysed at the level of individual specimens, external factors do not have to be applied to a species as a whole, but can be applied to single individuals or populations, and specific morphological traits. Thus, once validated with extant taxa, specimen-level phylogeny, combined with fine-scale stratigraphic field work and geological studies revealing palaeoenvironmental and palaeoclimatic factors, could yield information concerning morphological trait evolution within (and possibly across) evolutionary lineages through deep time in pre-eminent detail. Such an approach would allow highly localised and detailed correlations with data on environment and climate in the locale where a diagnostic trait first occurred, and can even help to track speciation processes through the accumulation of new morphological traits.

5. Conclusions

Phenotypic specimen-level phylogenetic analysis has a high potential for significant advances in the study of morphological variability, trait evolution and speciation in deep time. However, certain steps during matrix construction, phylogenetic analysis and interpretation of tree topology have to be followed in order to obtain accurate results. These mostly concern the inclusion of as much data as possible to obtain statistical significance, the application of appropriate weighting strategies to reduce the impact of characters possibly simply describing individual variation and the use of a number of complementing approaches to species delimitation, evaluating potentially conflicting results in light of the General Lineage Concept for species. We also highlight the need for validation studies with extant taxa, where the attribution of specimens to a particular species is known *a priori*, and can be used to infer the best-fitting methodology in a specific taxon.

6. Supplementary material

Supplementary material is available online at https://doi.org/10.1017/S1755691018000877.

7. Acknowledgements

ET has received funding from the European Union's Seventh Framework programme for research and innovation under the Marie Skłodowska-Curie grant agreement no. 609402-2020 researchers: Train to Move (T2M). ET is currently supported by the Theodore Roosevelt Memorial Fund, Division of Paleontology, and Richard Gilder Graduate School, American Museum of Natural History. We thank the Willi Hennig Society for providing the software TNT for free, and Campbell and colleagues for providing their data online as freely accessible supplementary material. Finally, we would like to acknowledge the inspirational role played by Professor Jenny Clack, whose work on tetrapod evolution and the invasion of the land stands as a model of how palaeobiology can revolutionise our understanding of evolutionary patterns and processes. PU is particularly grateful to Jenny for all of her support during his PhD and later career.

8. References

Adams, B. J. 2001. The species delimitation uncertainty principle. *Journal of Nematology* **33**, 153–60.

Ahmadzadeh, F., Flecks, M., Rödder, D., Böhme, W., Ilgaz, Ç., Harris, D. J., Engler, J. O., Üzüm, N. & Carretero, M. A. 2013. Multiple dispersal out of Anatolia: biogeography and evolution of oriental green lizards. *Biological Journal of the Linnean Society* **110**, 398–408.

Alroy, J., Aberhan, M., Bottjer, D. J., Foote, M., Fürsich, F. T., Harries, P. J., Hendy, A. J. W., Holland, S. M., Ivany, L. C., Kiessling, W., Kosnik, M. A., Marshall, C. R., McGowan, A. J., Miller, A. I., Olszewski, T. D., Patzkowsky, M. E., Peters, S. E., Villier, L., Wagner, P. J., Bonuso, N., Borkow, P. S., Brenneis, B., Clapham, M. E., Fall, L. M., Ferguson, C. A., Hanson, V. L., Krug, A. Z., Layou, K. M., Leckey, E. H., Nürnberg, S., Powers, C. M., Sessa, J. A., Simpson, C., Tomašových, A. & Visaggi, C. C. 2008. Phanerozoic trends in the global diversity of marine invertebrates. *Science* **321**, 97–100.

Arbour, V. M. & Currie, P. J. 2012. Analyzing taphonomic deformation of ankylosaur skulls using retrodeformation and Finite Element Analysis. *PLOS ONE* **7**, e39323.

Arbour, V. M. & Currie, P. J. 2016. Systematics, phylogeny and palaeobiogeography of the ankylosaurid dinosaurs. *Journal of Systematic Palaeontology* **14**, 385–444.

Arnold, E. N., Arribas, O. & Carranza, S. 2007. Systematics of the Palaearctic and Oriental lizard tribe Lacertini (Squamata: Lacertidae: Lacertinae), with descriptions of eight new genera. *Zootaxa* **1430**, 3–86.

Bacon, C. D., McKenna, M. J., Simmons, M. P. & Wagner, W. L. 2012. Evaluating multiple criteria for species delimitation: an empirical example using Hawaiian palms (Arecaceae: Pritchardia). *BMC Evolutionary Biology* **12**, 23.

Barbadillo, L. J. & Sanz, J. L. 1983. Análisis osteométrico de las regiones sacra y presacra de la columna vertebral en los lagartos Ibéricos *Lacerta viridis* Laurenti, *Lacerta lepida* Daudin y *Lacerta schreiberi* Bedriaga. *Amphibia-Reptilia* **4**, 215–39.

Baum, D. A. 1998. Individuality and the existence of species through time. *Systematic Biology* **47**, 641–53.

Bell, C. J., Gauthier, J. A. & Bever, G. S. 2010. Covert biases, circularity, and apomorphies: a critical look at the North American Quaternary Herpetofaunal Stability Hypothesis. *Quaternary International* **217**, 30–36.

Bell, M. A. & Lloyd, G. T. 2014. Strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence: a tutorial. *Dryad Digital Repository* 1–14. DOI: 10.5061/dryad.4k078.

Bell, M. A. & Lloyd, G. T. 2015. Strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence. *Palaeontology* **58**, 379–89.

Benson, R. B. J., Evans, M. & Druckenmiller, P. S. 2012. High diversity, low disparity and small body size in plesiosaurs (Reptilia, Sauropterygia) from the Triassic–Jurassic boundary. *PLOS ONE* **7**, e31838.

Benson, R. B. J., Campione, N. E., Carrano, M. T., Mannion, P. D., Sullivan, C., Upchurch, P. & Evans, D. C. 2014. Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. *PLOS Biology* **12**, e1001853.

Benson, R. B. J., Butler, R. J., Alroy, J., Mannion, P. D., Carrano, M. T. & Lloyd, G. T. 2016. Near-stasis in the long-term diversification of Mesozoic tetrapods. *PLOS Biology* **14**, e1002359.

Bergsten, J. 2005. A review of long-branch attraction. *Cladistics* **21**, 163–93.

Bhullar, B.-A. S., Marugán-Lobón, J., Racimo, F., Bever, G. S., Rowe, T. B., Norell, M. A. & Abzhanov, A. 2012. Birds have paedomorphic dinosaur skulls. *Nature* **487**, 223–26.

Böhmer, C., Rauhut, O. W. M & Wörheide, G. 2015. Correlation between Hox code and vertebral morphology in archosaurs. *Proceedings of the Royal Society B* **282**, 20150077.

Bonnan, M. F. 2007. Linear and geometric morphometric analysis of long bone scaling patterns in Jurassic neosauropod dinosaurs: their functional and paleobiological implications. *The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology* **290**, 1089–111.

Boyd, C. A., Brown, C. M., Scheetz, R. D. & Clarke, J. A. 2009. Taxonomic revision of the basal neornithischian taxa *Thescelosaurus* and *Bugenasaura*. *Journal of Vertebrate Paleontology* **29**, 758–70.

Brazeau, M. D. 2011. Problematic character coding methods in morphology and their effects. *Biological Journal of the Linnean Society* **104**, 489–98.

Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**, 795–803.

Bremer, K. 1994. Branch support and tree stability. *Cladistics* **10**, 295–304.

Brinkman, D., Rabi, M. & Zhao, L. 2017. Lower Cretaceous fossils from China shed light on the ancestral body plan of crown softshell turtles (Trionychidae, Cryptodira). *Scientific Reports* **7**, 6719.

Brochu, C. A. 1996. Closure of neurocentral sutures during crocodilian ontogeny: implications for maturity assessment in fossil archosaurs. *Journal of Vertebrate Paleontology* **16**, 49–62.

Brummitt, R. K. 2002. How to chop up a tree. *Taxon* **51**, 31–41.

Brusatte, S. L. 2010. Representing supraspecific taxa in higher-level phylogenetic analyses: guidelines for palaeontologists. *Palaeontology* **53**, 1–9.

Brusatte, S. L., Benton, M. J., Desojo, J. B. & Langer, M. C. 2010. The higher-level phylogeny of Archosauria (Tetrapoda: Diapsida). *Journal of Systematic Palaeontology* **8**, 3–47.

Burnell, A., Collins, S. & Young, B. A. 2012. Vertebral morphometrics in *Varanus*. *Bulletin de la Societe Geologique de France* **183**, 151–58.

Butler, R. J., Upchurch, P. & Norman, D. B. 2008. The phylogeny of the ornithischian dinosaurs. *Journal of Systematic Palaeontology* **6**, 1–40.

Butler, R. J. & Upchurch, P. 2007. Highly incomplete taxa and the phylogenetic relationships of the theropod dinosaur *Juravenator starki*. *Journal of Vertebrate Paleontology* **27**, 253–56.

Campbell, J. A., Ryan, M. J., Holmes, R. B. & Schröder-Adams, C. J. 2016. A re-evaluation of the chasmosaurine ceratopsid genus *Chasmosaurus* (Dinosauria: Ornithischia) from the Upper Cretaceous (Campanian) Dinosaur Park Formation of Western Canada. *PLOS ONE* **11**, e0145805.

Campione, N. E., Brink, K. S., Freedman, E. A., McGarrity, C. T. & Evans, D. C. 2013. '*Glishades ericksoni*', an indeterminate juvenile hadrosaurid from the Two Medicine Formation of Montana: implications for hadrosauroid diversity in the latest Cretaceous (Campanian-Maastrichtian) of western North America. *Senckenbergiana Lethaea* **93**, 65–75.

Cano, A., Nguyen, D. T., Ventura, S. & Cios, K. J. 2016. ur-CAIM: improved CAIM discretization for unbalanced and balanced data. *Soft Computing* **20**, 173–88.

Carballido, J. L., Salgado, L., Pol, D., Canudo, J. I. & Garrido, A. 2012. A new basal rebbachisaurid (Sauropoda, Diplodocoidea) from the Early Cretaceous of the Neuquén Basin; evolution and biogeography of the group. *Historical Biology* **24**, 631–54.

Carballido, J. L. & Sander, P. M. 2014. Postcranial axial skeleton of *Europasaurus holgeri* (Dinosauria, Sauropoda) from the Upper Jurassic of Germany: implications for sauropod ontogeny and phylogenetic relationships of basal Macronaria. *Journal of Systematic Palaeontology* **12**, 335–87.

Carpenter, K. 2017. Comment (Case 3700) – opposition against the proposed designation of *Diplodocus carnegii* Hatcher, 1901 as the type species of *Diplodocus* Marsh, 1878 (Dinosauria, Sauropoda). *The Bulletin of Zoological Nomenclature* **74**, 47–49.

Carr, T. D., Varricchio, D. J., Sedlmayr, J. C., Roberts, E. M. & Moore, J. R. 2017. A new tyrannosaur with evidence for anagenesis and crocodile-like facial sensory system. *Scientific Reports* **7**, 44942.

Carr, T. D. & Williamson, T. E. 2004. Diversity of late Maastrichtian Tyrannosauridae (Dinosauria: Theropoda) from western North America. *Zoological Journal of the Linnean Society* **142**, 479–523.

Carstens, B. C., Pelletier, T. A., Reid, N. M. & Satler, J. D. 2013. How to fail at species delimitation. *Molecular Ecology* **22**, 4369–83.

Cau, A. 2017. Specimen-level phylogenetics in paleontology using the Fossilized Birth-Death model with sampled ancestors. *PeerJ* **5**, e3055.

Chamero, B., Buscalioni, Á. D., Marugán-Lobón, J. & Sarris, I. 2014. 3D geometry and quantitative variation of the cervico-thoracic region in Crocodylia. *The Anatomical Record* **297**, 1278–91.

Chinsamy-Turan, A. 2005. *The microstructure of dinosaur bone*. Baltimore, MD: Johns Hopkins University Press, 216 pp.

Chippindale, P. T. & Wiens, J. J. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology* **43**, 278–87.

Cleary, T. J., Moon, B. C., Dunhill, A. M. & Benton, M. J. 2015. The fossil record of ichthyosaurs, completeness metrics and sampling biases. *Palaeontology* **58**, 521–36.

Close, R. A., Benson, R. B. J., Upchurch, P. & Butler, R. J. 2017. Controlling for the species-area effect supports constrained long-term Mesozoic terrestrial vertebrate diversification. *Nature Communications* **8**, 1–11.

Congreve, C. R. & Lamsdell, J. C. 2016. Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology* **59**, 447–62.

Cormack, D. H. 1987. *Ham's histology*. 9th edn. Philadelphia, PA: Lippincott Williams & Wilkins, 732 pp.

De Laet, J. 1997. A reconsideration of three-item analysis, the use of implied weights in cladistics, and a practical application in Gentianaceae. PhD Dissertation, Catholic University of Leuven, Belgium. 214 pp.

D'Emic, M. D. 2012. The early evolution of titanosauriform sauropod dinosaurs. *Zoological Journal of the Linnean Society* **166**, 624–71.

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation. *In* Howard, D. J. & Berlocher, S. H. (eds) *Endless forms: species and speciation*, 57–75. Oxford: Oxford University Press.

de Queiroz, K. & Donoghue, M. J. 1990a. Phylogenetic systematics and species revisited. *Cladistics* **6**, 83–90.

de Queiroz, K. & Donoghue, M. J. 1990b. Phylogenetic systematics or Nelson's version of cladistics? *Cladistics* **6**, 61–75.

Dettman, J. R., Jacobson, D. J., Turner, E., Pringle, A. & Taylor, J. W. 2003. Reproductive isolation and phylogenetic divergence in neurospora: comparing methods of species recognition in a model eukaryote. *Evolution* **57**, 2721–41.

Donoghue, M. J. 1985. A critique of the biological species concept and recommendations for a phylogenetic alternative. *The Bryologist* **88**, 172–81.

Donoghue, M. J., Olmstead, R. G., Smith, J. F. & Palmer, J. D. 1992. Phylogenetic relationships of dipsacales based on rbcL sequences. *Annals of the Missouri Botanical Garden* **79**, 333–45.

Farris, J. S. 1969. A successive approximations approach to character weighting. *Systematic Biology* **18**, 374–85.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. *In* Platnick, N. & Funk, V. A. (eds) *Advances in cladistics Vol. 2, proceedings of the second meeting of the Willi Hennig Society*, 7–36. New York: Columbia University Press.

Foth, C., Evers, S. W., Pabst, B., Mateus, O., Flisch, A., Patthey, M & Rauhut, O. W. M. 2015. New insights into the lifestyle of *Allosaurus* (dinosauria: Theropoda) based on another specimen with multiple pathologies. *PeerJ* **3**, e940.

Gauthier, J. A., Kearney, M., Maisano, J. A., Rieppel, O. & Behlke, A. D. B. 2012. Assembling the squamate tree of life: perspectives from the phenotype and the fossil record. *Bulletin of the Peabody Museum of Natural History* **53**, 3–308.

Gilmore, C. W. 1925. A nearly complete articulated skeleton of *Camarasaurus*, a saurischian dinosaur from the Dinosaur National Monument, Utah. *Memoirs of the Carnegie Museum* **10**, 347–84.

Giovanardi, S. 2017. Evaluation of Several cladistic methodologies and their impact on a paleontological dataset: the case of

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Diplodocidae (dinosauria: Sauropoda). Master's Thesis, Università di Torino, Italy. 56 pp.

Godinho, R., Crespo, E. G., Ferrand, N. & Harris, D. J. 2005. Phylogeny and evolution of the green lizards, *Lacerta* spp. (Squamata: Lacertidae) based on mitochondrial and nuclear DNA sequences. *Amphibia-Reptilia* **26**, 271–85.

Goloboff, P. A. 1993. Estimating character weights during tree search. *Cladistics* **9**, 83–91.

Goloboff, P. A. 1995. Parsimony and weighting: a reply to Turner and Zandee. *Cladistics* **11**, 91–104.

Goloboff, P. A. 2014. Extended implied weighting. *Cladistics* **30**, 260–72.

Goloboff, P. A., Farris, J. S., Källersjö, M., Oxelman, B., Ramírez, M. J. & Szumik, C. A. 2003. Improvements to resampling measures of group support. *Cladistics* **19**, 324–32.

Goloboff, P. A., Mattoni, C. I. & Quinteros, A. S. 2006. Continuous characters analyzed as such. *Cladistics* **22**, 589–601.

Goloboff, P. A., Farris, J. S. & Nixon, K. C. 2008a. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–86.

Goloboff, P. A., Carpenter, J. M., Arias, J. S. & Esquivel, D. R. M. 2008b. Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–73.

Goloboff, P. A., Torres, A. & Arias, J. S. 2018. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics* **34**, 407–37.

Goloboff, P. A. & Farris, J. S. 2001. Methods for quick consensus estimation. *Cladistics* **17**, S26–34.

Hastings, A. K. & Hellmund, M. 2015. Rare *in situ* preservation of adult crocodylian with eggs from the Middle Eocene of Geiseltal, Germany. *PALAIOS* **30**, 446–61.

Hauser, D. L. & Presch, W. 1991. The effect of ordered characters on phylogenetic reconstruction. *Cladistics* **7**, 243–65.

Hillis, D. M. & Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**, 182–92.

Hoso, M., Asami, T. & Hori, M. 2007. Right-handed snakes: convergent evolution of asymmetry for functional specialization. *Biology Letters* **3**, 169–73.

Huelsenbeck, J. P. 1991. When are fossils better than extant taxa in phylogenetic analysis? *Systematic Zoology* **40**, 458–69.

Ji, Q., Wu, X. & Cheng, Y. 2010. Cretaceous choristoderan reptiles gave birth to live young. *Naturwissenschaften* **97**, 423–28.

Jiang, F. & Sui, Y. 2015. A novel approach for discretization of continuous attributes in rough set theory. *Knowledge-Based Systems* **73**, 324–34.

Källersjö, M., Albert, V. A. & Farris, J. S. 1999. Homoplasy increases phylogenetic structure. *Cladistics* **15**, 91–93.

Kearney, M. & Clark, J. M. 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. *Journal of Vertebrate Paleontology* **23**, 263–74.

Kimura, Y., Flynn, L. J. & Jacobs, L. L. 2016. A palaeontological case study for species delimitation in diverging fossil lineages. *Historical Biology* **28**, 189–98.

Klein, N. & Sander, M. 2008. Ontogenetic stages in the long bone histology of sauropod dinosaurs. *Paleobiology* **34**, 247–63.

Kopuchian, C. & Ramírez, M. J. 2010. Behaviour of resampling methods under different weighting schemes, measures and variable resampling strengths. *Cladistics* **26**, 86–97.

Lehman, T. M. 1990. The ceratopsian subfamily Chasmosaurinae: sexual dimorphism and systematics. *In* Carpenter, K. & Currie, P. J. (eds) *Dinosaur systematics: approaches and perspectives*, 211–29. Cambridge: Cambridge University Press.

Longrich, N. 2015. Systematics of *Chasmosaurus* – new information from the Peabody Museum skull, and the use of phylogenetic analysis for dinosaur alpha taxonomy. *F1000Research* **4**, 1468.

Maddison, W. 1989. Reconstructing character evolution on polytomous cladograms. *Cladistics* **5**, 365–77.

Makovicky, P. J. 2010. A redescription of the *Montanoceratops cerorhynchus* holotype, with a review of referred material. *In* Ryan, M. J., Chinnery-Allgeier, B. J. & Eberth, D. A. (eds) *New perspectives on Horner dinosaurs: the Royal Tyrrell Museum Ceratopsian Symposium*, 68–82. Bloomington and Indianapolis, IN: Indiana University Press.

Mannion, P. D., Upchurch, P., Barnes, R. N. & Mateus, O. 2013. Osteology of the Late Jurassic Portuguese sauropod dinosaur *Lusotitan atalaiensis* (Macronaria) and the evolutionary history of basal titanosauriforms. *Zoological Journal of the Linnean Society* **168**, 98–206.

Mannion, P. D., Upchurch, P., Benson, R. B. J. & Goswami, A. 2014. The latitudinal biodiversity gradient through deep time. *Trends in Ecology & Evolution* **29**, 42–50.

Mannion, P. D., Benson, R. B. J., Carrano, M. T., Tennant, J. P., Judd, J. & Butler, R. J. 2015. Climate constrains the evolutionary history and biodiversity of crocodylians. *Nature Communications* **6**, 9438.

Mannion, P. D., Allain, R. & Moine, O. 2017. The earliest known titanosauriform sauropod dinosaur and the evolution of Brachiosauridae. *PeerJ* **5**, e3217.

Mannion, P. D. & Upchurch, P. 2010. Completeness metrics and the quality of the sauropodomorph fossil record through geological and historical time. *Paleobiology* **36**, 283–302.

Marx, F. G. 2011. The more the merrier? A large cladistic analysis of Mysticetes, and comments on the transition from teeth to baleen. *Journal of Mammalian Evolution* **18**, 77–100.

Marzahn, E., Mayer, W., Joger, U., Ilgaz, Ç., Jablonski, D., Kindler, C., Kumlutaş, Y., Nistri, A., Schneeweiss, N., Vamberger, M., Žagar, A. & Fritz, U. 2016. Phylogeography of the *Lacerta viridis* complex: mitochondrial and nuclear markers provide taxonomic insights. *Journal of Zoological Systematics and Evolutionary Research* **54**, 85–105.

Mayer, W. & Pavlicev, M. 2007. The phylogeny of the family Lacertidae (Reptilia) based on nuclear DNA sequences: convergent adaptations to arid habitats within the subfamily Eremiadinae. *Molecular Phylogenetics and Evolution* **44**, 1155–63.

McIntosh, J. S., Miles, C. A., Cloward, K. A. & Parker, J. R. 1996. A new nearly complete skeleton of *Camarasaurus*. *Bulletin of the Gunma Museum of Natural History* **1**, 1–87.

McIntosh, J. S. & Carpenter, K. 1998. The holotype of *Diplodocus longus*, with comments on other specimens of the genus. *Modern Geology* **23**, 85–110.

Mounier, A. & Caparros, M. 2015. The phylogenetic status of *Homo heidelbergensis* – a cladistic study of Middle Pleistocene hominins. *BMSAP* **27**, 110–34.

Morrison, D. A. 2005. Networks in phylogenetic analysis: new tools for population biology. *International Journal for Parasitology* **35**, 567–82.

Morschhauser, E. M., You, H., Li, D. & Dodson, P. 2014. Juvenile cranial material of *Auroraceratops rugosus* (Ceratopsia: Ornithischia) and implications for the phylogenetic placement of juvenile specimens. *Journal of Vertebrate Paleontology, Program and Abstracts* **2014**, 192.

Müller, J., Scheyer, T. M., Head, J. J., Barrett, P. M., Werneburg, I., Ericson, P. G. P., Pol, D. & Sánchez-Villagra, M. R. 2010. Homeotic effects, somitogenesis and the evolution of vertebral numbers in recent and fossil amniotes. *Proceedings of the National Academy of Sciences* **107**, 2118–23.

Nanda, P. & Singh, B. N. 2012. Behavioural reproductive isolation and speciation in *Drosophila*. *Journal of Biosciences* **37**, 359–74.

Nixon, K. C. & Carpenter, J. M. 1993. On outgroups. *Cladistics* **9**, 413–26.

Norell, M. A. & Gao, K. 1997. Braincase and phylogenetic relationships of *Estesia mongoliensis* from the Late Cretaceous of the Gobi Desert and the recognition of a new clade of lizards. *American Museum Novitates* **3211**, 1–25.

O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., Pisani, D & Donoghue, P. C. J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters* **12**, 20160081.

Otero, R. A., Soto-Acuña, S., O'Keefe, F. R., O'Gorman, J. P., Stinnesbeck, W., Suárez, M. E., Rubilar-Rogers, D., Salazar, C. & Quinzio-Sinn, L. A. 2014. *Aristonectes quiriquinensis*, sp. nov., a new highly derived elasmosaurid from the upper Maastrichtian of central Chile. *Journal of Vertebrate Paleontology* **34**, 100–25.

Palmer, A. R. 1996. From symmetry to asymmetry: phylogenetic patterns of asymmetry variation in animals and their evolutionary significance. *Proceedings of the National Academy of Sciences* **93**, 14279–86.

Palmer, A. R., Strobeck, C. & Chippindale, A. K. 1994. Bilateral variation and the evolutionary origin of macroscopic asymmetries. *In* Markow, T. A. (ed.) *Developmental instability: its origins and evolutionary implications*, **2**, 203–20. Netherlands: Springer.

Parker, W. G. 2016. Revised phylogenetic analysis of the Aetosauria (Archosauria: Pseudosuchia); assessing the effects of incongruent morphological character sets. *PeerJ* **4**, e1583.

Pisani, D., Feuda, R., Peterson, K. J. & Smith, A. B. 2012. Resolving phylogenetic signal from noise when divergence is rapid: a new look at the old problem of echinoderm class relationships. *Molecular Phylogenetics and Evolution* **62**, 27–34.

Pleijel, F. & Rouse, G. W. 2000. Least-inclusive taxonomic unit: a new taxonomic concept for biology. *Proceedings of the Royal Society of London B: Biological Sciences* **267**, 627–30.

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press

Poe, S. & Wiens, J. J. 2000. Character selection and the methodology of morphological phylogenetics. *In* Wiens, J. J. (ed.) *Phylogenetic analysis of morphological data*, 20–36. Washington, DC: Smithsonian Institution Press.

Pol, D. & Escapa, I. H. 2009. Unstable taxa in cladistic analysis: identification and the assessment of relevant characters. *Cladistics* **25**, 515–27.

Prendini, L. 2001. Species or supraspecific taxa as terminals in cladistic analysis? Groundplans versus exemplars revisited. *Systematic Biology* **50**, 290–300.

Prevosti, F. J. & Chemisquy, M. A. 2010. The impact of missing data on real morphological phylogenies: influence of the number and distribution of missing entries. *Cladistics* **26**, 326–39.

Purvis, A. & Garland, T. 1993. Polytomies in comparative analyses of continuous characters. *Systematic Biology* **42**, 569–75.

Puslednik, L. & Serb, J. M. 2008. Molecular phylogenetics of the Pectinidae (Mollusca: Bivalvia) and effect of increased taxon sampling and outgroup selection on tree topology. *Molecular Phylogenetics and Evolution* **48**, 1178–88.

Rae, T. C. 1998. The logical basis for the use of continuous characters in phylogenetic systematics. *Cladistics* **14**, 221–28.

Rothschild, B. M. & Martin, L. D. 2006. Skeletal impact of disease. *New Mexico Museum of Natural History and Science Bulletin* **33**, 1–226.

Sampson, S. D., Loewen, M. A., Farke, A. A., Roberts, E. M., Forster, C. A., Smith, J. A & Titus, A. L. 2010. New horned dinosaurs from Utah provide evidence for intracontinental dinosaur endemism. *PLOS ONE* **5**, e12292.

Sander, P. M. 2012. Reproduction in early amniotes. *Science* **337**, 806–08.

Sansom, R. S. 2015. Bias and sensitivity in the placement of fossil taxa resulting from interpretations of missing data. *Systematic Biology* **64**, 256–66.

Sansom, R. S., Wills, M. A. & Williams, T. 2017. Dental data perform relatively poorly in reconstructing mammal phylogenies: morphological partitions evaluated with molecular benchmarks. *Systematic Biology* **66**, 813–22.

Sansom, R. S., Choate, P. G., Keating, J. N. & Randle, E. 2018. Parsimony, not Bayesian analysis, recovers more stratigraphically congruent phylogenetic trees. *Biology Letters* **14**, 20180263.

Satler, J. D., Carstens, B. C. & Hedin, M. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, *Aliatypus*). *Systematic Biology* **62**, 805–23.

Sato, T., Cheng, Y., Wu, X., Zelenitsky, D. K. & Hsiao, Y. 2005. A pair of shelled eggs inside a female dinosaur. *Science* **308**, 375.

Saunders, I. W., Tavaré, S. & Watterson, G. A. 1984. On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–91.

Scannella, J. B., Fowler, D. W., Goodwin, M. B. & Horner, J. R. 2014. Evolutionary trends in *Triceratops* from the Hell Creek Formation, Montana. *Proceedings of the National Academy of Sciences* **111**, 10245–50.

Scheyer, T. M., Klein, N. & Sander, P. M. 2010. Developmental palaeontology of Reptilia as revealed by histological studies. *Seminars in Cell & Developmental Biology* **21**, 462–70.

Schmidt-Lebuhn, A. N., de Vos, J. M., Keller, B. & Conti, E. 2012. Phylogenetic analysis of *Primula* section *Primula* reveals rampant non-monophyly among morphologically distinct species. *Molecular Phylogenetics and Evolution* **65**, 23–34.

Schwarz, D., Ikejiri, T., Breithaupt, B. H., Sander, P. M. & Klein, N. 2007. A nearly complete skeleton of an early juvenile diplodocid (Dinosauria: Sauropoda) from the Lower Morrison Formation (Late Jurassic) of north central Wyoming and its implications for early ontogeny and pneumaticity in sauropods. *Historical Biology* **19**, 225–53.

Sites, J. W., Davis, S. K., Guerra, T., Iverson, J. B. & Snell, H. L. 1996. Character congruence and phylogenetic signal in molecular and morphological data sets: a case study in the living iguanas (Squamata, Iguanidae). *Molecular Biology and Evolution* **13**, 1087–105.

Sites, J. W. & Marshall, J. C. 2004. Operational criteria for delimiting species. *Annual Review of Ecology, Evolution, and Systematics* **35**, 199–227.

Tennant, J. P., Mannion, P. D. & Upchurch, P. 2016a. Environmental drivers of crocodyliform extinction across the Jurassic/Cretaceous transition. *Proceedings of the Royal Society B* **283**, 20152840.

Tennant, J. P., Mannion, P. D. & Upchurch, P. 2016b. Sea level regulated tetrapod diversity dynamics through the Jurassic/Cretaceous interval. *Nature Communications* **7**, 12737.

Thiele, K. 1993. The Holy Grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* **9**, 275–304.

Townsend, J. P., Su, Z. & Tekle, Y. I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biology* **61**, 835–49.

Tschopp, E. 2016. Nomenclature of vertebral laminae in lizards, with comments on ontogenetic and serial variation in Lacertini (Squamata, Lacertidae). *PLOS ONE* **11**, e0149445.

Tschopp, E., Russo, J. & Dzemski, G. 2013. Retrodeformation as a test for the validity of phylogenetic characters: an example from diplodocid sauropod vertebrae. *Palaeontologia Electronica* **16**, 2T.

Tschopp, E., Mateus, O. & Benson, R. B. J. 2015. A specimen-level phylogenetic analysis and taxonomic revision of Diplodocidae (Dinosauria, Sauropoda). *PeerJ* **3**, e857.

Tschopp, E., Wings, O., Frauenfelder, T. & Rothschild, B. 2016. Pathological phalanges in a camarasaurid sauropod dinosaur and implications on behaviour. *Acta Palaeontologica Polonica* **61**, 125–34.

Tschopp, E., Brinkman, D., Henderson, J., Turner, M. A. & Mateus, O. 2018a. Considerations on the replacement of a type species in the case of the sauropod dinosaur Diplodocus Marsh, 1878. *Geology of the Intermountain West* **5**, 245–62.

Tschopp, E., Tschopp, F. A. & Mateus, O. 2018b. Overlap indices: tools to quantify the amount of anatomical overlap among groups of incomplete terminal taxa in phylogenetic analyses. *Acta Zoologica* **99**, 169–76.

Tschopp, E., Villa, A., Camaiti, M., Ferro, L., Tuveri, C., Rook, L., Arca, M. & Delfino, M. 2018c. The first fossils of Timon (Squamata: Lacertinae) from Sardinia (Italy) and potential causes for its local extinction in the Pleistocene. *Zoological Journal of the Linnean Society* **184**, 825–56. DOI: 10.1093/zoolinnean/zly003.

Tschopp, E. & Mateus, O. 2016. Case 3700 *Diplodocus* marsh, 1878 (Dinosauria, Sauropoda): proposed designation of *D. carnegii* Hatcher, 1901 as the type species. *Bulletin of Zoological Nomenclature* **73**, 17–24.

Tschopp, E. & Mateus, O. 2017. Osteology of *Galeamopus pabsti* sp. nov. (Sauropoda: Diplodocidae), with implications for neurocentral closure timing, and the cervico-dorsal transition in diplodocids. *PeerJ* **5**, e3179.

Turner, H. & Zandee, R. 1995. The behaviour of Goloboff's tree fitness measure F. *Cladistics* **11**, 57–72.

Upchurch, P. 1998. The phylogenetic relationships of sauropod dinosaurs. *Zoological Journal of the Linnean Society* **124**, 43–103.

Upchurch, P., Tomida, Y. & Barrett, P. M. 2004. A new specimen of *Apatosaurus ajax* (Sauropoda: Diplodocidae) from the Morrison Formation (Upper Jurassic) of Wyoming, USA. *National Science Museum Monographs* **26**, 1–118.

Villa, A., Tschopp, E., Georgalis, G. L. & Delfino, M. 2017. Osteology, fossil record and palaeodiversity of the European lizards. *Amphibia-Reptilia* **38**, 79–88.

Vrana, P. & Wheeler, W. 1992. Individual organisms as terminal entities: laying the species problem to rest. *Cladistics* **8**, 67–72.

Wedel, M. J. 2003. The evolution of vertebral pneumaticity in sauropod dinosaurs. *Journal of Vertebrate Paleontology* **23**, 344–57.

Wedel, M. J., Cifelli, R. L. & Sanders, R. K. 2000. Osteology, paleobiology, and relationships of the sauropod dinosaur *Sauroposeidon*. *Acta Palaeontologica Polonica* **45**, 343–88.

Whitlock, J. A. 2011. A phylogenetic analysis of Diplodocoidea (Saurischia: Sauropoda). *Zoological Journal of the Linnean Society* **161**, 872–915.

Wiens, J. J. 1995. Polymorphic characters in phylogenetic systematics. *Systematic Biology* **44**, 482–500.

Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology* **47**, 625–40.

Wiens, J. J. 2000. Coding morphological variation within species and higher taxa for Phylogenetic Analysis. *In* Wiens, J. J. (ed.) *Phylogenetic analysis of morphological data*, 115–45. Washington, DC: Smithsonian Institution Press.

Wiens, J. J. 2003. Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? *Journal of Vertebrate Paleontology* **23**, 297–310.

Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics* **39**, 34–42.

Wiens, J. J. & Penkrot, T. A. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology* **51**, 69–91.

Wiens, J. J. & Tiu, J. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLOS ONE* **7**, e42925.

Wiley, E. O. & Lieberman, B. S. 2011. *Phylogenetics: theory and practice of phylogenetic systematics*. Hoboken, NJ: John Wiley & Sons, 497 pp.

Wilkinson, M. 1992. Ordered versus unordered characters. *Cladistics* **8**, 375–85.

Wilkinson, M. 1995. More on reduced consensus methods. *Systematic Biology* **44**, 435–39.

Wilkinson, M. 2003. Missing entries and multiple trees: instability, relationships, and support in parsimony analysis. *Journal of Vertebrate Paleontology* **23**, 311–23.

Wilkinson, M., Thorley, J. L. & Upchurch, P. 2000. A chain is no stronger than its weakest link: double decay analysis of phylogenetic hypotheses. *Systematic Biology* **49**, 754–76.

Wilson, J. A. 1999. A nomenclature for vertebral laminae in sauropods and other saurischian dinosaurs. *Journal of Vertebrate Paleontology* **19**, 639–53.

Wilson, J. A. 2002. Sauropod dinosaur phylogeny: critique and cladistic analysis. *Zoological Journal of the Linnean Society* **136**, 215–75.

Wilson, J. A. 2012. New vertebral laminae and patterns of serial variation in vertebral laminae of sauropod dinosaurs. *Contributions from the Museum of Paleontology, University of Michigan* **32**, 91–110.

Wilson, J. A. & Upchurch, P. 2009. Redescription and reassessment of the phylogenetic affinities of *Euhelopus zdanskyi* (Dinosauria: Sauropoda) from the Early Cretaceous of China. *Journal of Systematic Palaeontology* **7**, 199–239.

Wiman, C. 1929. Die kreide-dinosaurier aus shantung. *Palaeontologia Sinica* **6**, 1–67.

Woodruff, D. C., Fowler, D. W. & Horner, J. R. 2017. A new multifaceted framework for deciphering diplodocid ontogeny. *Palaeontologia Electronica* **20**, 1–53.

Wright, A. M. & Hillis, D. M. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLOS ONE* **9**, e109210.

Yates, A. M. 2003. The species taxonomy of the sauropodomorph dinosaurs from the Löwenstein Formation (Norian, Late Triassic) of Germany. *Palaeontology* **46**, 317–37.

Zander, R. H. 2004. Minimal values for reliability of bootstrap and jackknife proportions, decay index, and Bayesian posterior probability. *Phyloinformatics* **2**, 1–13.

MS received 19 February 2018. Accepted for publication 27 September 2018

https://doi.org/10.1017/S1755691018000877 Published online by Cambridge University Press