Sampling diverse characters improves phylogenies: Craniodental and postcranial characters of vertebrates often imply different trees

Ross C. P. Mounce¹, Robert Sansom² and Matthew A. Wills¹*

¹The Milner Centre for Evolution, Department of Biology and Biochemistry, The University of Bath, The Avenue, Claverton Down, Bath BA2 7AY, UK; E-mail: m.a.wills@bath.ac.uk

²Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT

Words in main text: 6588 (including abstract)
Figures: 5
Tables: 2

*Correspondence: Matthew A. Wills
Department of Biology and Biochemistry
The University of Bath
The Avenue
Claverton Down
Bath BA2 7AY
United Kingdom

Telephone: 01225 383504
Fax: 01225 386779
E-mail: m.a.wills@bath.ac.uk
Sampling diverse characters improves phylogenies: Craniodental and postcranial characters of vertebrates often imply different trees

Ross C. P. Mounce\textsuperscript{1}, Robert Sansom\textsuperscript{2} and Matthew A. Wills\textsuperscript{1}

\textsuperscript{1}The Milner Centre for Evolution, Department of Biology and Biochemistry, The University of Bath, The Avenue, Claverton Down, Bath BA2 7AY, UK; E-mail: m.a.wills@bath.ac.uk

\textsuperscript{2}Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT

\begin{itemize}
\item Morphological cladograms of vertebrates are often inferred from greater numbers of characters describing the skull and teeth than from postcranial characters. This is either because the skull is believed to yield characters with a stronger phylogenetic signal (i.e., contain less homoplasy), because morphological variation therein is more readily atomized, or because craniodental material is more widely available (particularly in the palaeontological case). An analysis of 85 vertebrate datasets published between 2000 and 2013 confirms that craniodental characters are significantly more numerous than postcranial characters, but finds no evidence that levels of homoplasy differ in the two partitions. However, a new partition test based on tree-to-tree distances (as measured by Robinson Foulds metric) rather than tree length reveals that relationships inferred from the partitions are significantly different about one time in three, much more often than expected. Such differences may reflect divergent selective pressures in different body regions, resulting in different localized patterns of homoplasy. Most systematists attempt to sample characters broadly across body regions, but this is not always possible. We conclude that trees inferred largely from either craniodental or postcranial characters in isolation may differ significantly from those that would result from a more holistic approach. We urge the latter.
\end{itemize}
Despite the increasing importance of molecular and genomic data over the last two decades, morphology still makes an invaluable contribution to vertebrate phylogenetics, and is the only suitable source of data for palaeontological phylogenies (Asher and Müller, 2012; Hillis and Wiens, 2000). Levels of morphological homoplasy in vertebrate groups are generally lower than those amongst their invertebrate counterparts (Hoyal Cuthill et al. 2010), suggesting that the signal quality for vertebrates is relatively high. Morphological systematists usually seek to code as many valid characters from as wide a selection of organs and body regions as possible, and to analyse these simultaneously, typically citing the principle of total evidence (Kluge, 1989). Heterogeneity in the performance of characters is often quantified in retrospect, and can also be utilised formally to increase stability in various post hoc weighting schemes (Farris 1969; Goloboff 2014; Goloboff et al. 2008a). However, coded matrices are usually analysed holistically rather than in partitions defined a priori. This is not least because hidden support (the presence of weak signals within putative partitions that become reciprocally reinforcing, and therefore the dominant signal when all characters are considered) is most readily identified in this way (Gatesy et al. 1999; Gatesy and Arctander 2000). It is therefore conceded that different body regions can be subject to different levels and patterns of homoplasy (Gaubert et al. 2005), such that anatomical subsets of characters can yield significantly different trees. This is especially true when there is strong functional selection in particular organ systems, leading to convergence (Ji et al. 1999; Kivell et al. 2013; Tseng et al., 2011).

Few cladistic studies have analysed the performance of morphological characters within partitions explicitly (e.g., Sánchez-Villagra and Williams, 1998; Rae, 1999; Gould, 2001; Song and Bucheli, 2010). While analyses may concentrate upon characters of particular types or from particular organ systems, it is still rare that trees are inferred explicitly from subsets of these data (but see Rae, 1999; Vermeij and Carlson, 2000; O’Leary et al., 2003; Poyato-Ariza, 2003; Diogo, 2004; Young, 2005; Clarke and Middleton 2008; Smith, 2010; Farke et al., 2011). Notable exceptions are taphonomic studies that investigate the effects of omitting volatile, soft-part
characters with a low fossilisation potential (Sansom et al. 2010, 2011; Sansom and Wills, 2013; Pattinson et al., 2014; Sansom 2015).

Where levels of homoplasy across morphological character partitions are considered at all, these are usually investigated by comparing distributions of the consistency index (ci: Kluge and Farris, 1969) (Fig. 1). The ci is given simply as the minimum possible number of state changes (states minus characters) divided by the most parsimonious tree length. More appropriate functions underpin most post hoc weighting schemes (Goloboff 2014; Goloboff et al. 2008a). In the case of insects, Song and Bucheli (2010) found male genital characters to be less homoplastic than characters coding other aspects of form. In brachiopods, Leighton and Maples (2002) revealed that shell characters are significantly more homoplastic than those describing internal anatomy (a disconcerting finding in a group whose fossil record consists largely of hard parts, and in which phylogeny is inferred almost exclusively with recourse to such characters). Similarly, in hedgehogs, Gould (2001) reported significantly higher consistency indices for dental characters compared with those describing other aspects of anatomy. However, she also noted that the optimal trees inferred from the dental characters alone were seriously at odds with those inferred from the other characters. Kangas et al. (2004) noted that many morphological characters of mammalian teeth have exceptionally strong developmental interdependence, all being under the correlated control of a small number of genes. Most ambitiously, Sanchez-Villagra and Williams (1998) compared the consistency indices between dental, cranial and postcranial character partitions of eight mammalian datasets, but reported no significant differences. We note that the relative performance of particular characters and character partitions may be contingent, especially with respect to the taxonomic hierarchical level of study. Hence, characters entailing little or no homoplasy for shallow branches may exhibit greater homoplasy when considered at deeper levels. Goloboff et al. (2010) have proposed analytical approaches that take account of this variability in the evolutionary lability of characters along branches.

The dominant practice of total evidence analysis for morphology contrasts with the more qualified approach that was sometimes adopted with molecular data (Bull et al., 1993; Felsenstein, 1988; Pamilo and Nei, 1988; Maddison, 1997; Nichols, 2001; Degnan and Rosenberg, 2009).
Molecular systematists have debated the relative merits of partitioned versus combined analyses, although the consensus has emerged in favour of the latter (Gatesy et al., 1999; Gatesy and Baker, 2005; Kjer and Honeycutt, 2007; Thompson et al., 2012). Historically, the issue of morphological versus molecular incongruence has been more to the fore (Gatesy and Arctander, 2000), motivated by striking examples of conflict between molecular and morphological cladograms in some groups (Mickevich and Farris, 1981; Bledsoe and Raikow, 1990; Bremer, 1996; Poe, 1996; Baker et al., 1998; Hillis and Wiens, 2000; Wiens and Hollingsworth, 2000; Jenner, 2004; Draper et al., 2007; Springer et al., 2007; Pisani et al., 2007; Near 2003; Mayr, 2011), although see Lee and Camens (2009).

**WHY EXAMINE THE CONGRUENCE OF CRANIODENTAL AND POSTCRANIAL PARTITIONS?**

Some systematists hold that craniodental and postcranial characters convey signals of differing quality (Ward, 1997; Collard et al., 2001; Naylor and Adams, 2001; Finarelli and Clyde, 2004). However, the evidence for this is piecemeal and largely anecdotal. Many practitioners take a more holistic approach to sampling characters (Sánchez-Villagra and Williams, 1998), sampling densely from as many anatomical regions as possible. However, even where potential characters are reasonably homogenously distributed throughout the body, “certain body regions and organs still hold a considerable mystique for taxonomists as classificatory tools, while others are neglected” (Sokal and Sneath, 1963; page 85). Arratia (2009) noted that actinopterygian systematists focus on cranial characters, despite rich seams of underexploited data within the fin rays and fulcra. Murray and Vickers-Rich (2004) suggested that the crania and mandibles of birds often provide the most informative characters because of their structural complexity. Similarly, Cardini and Elton (2008) demonstrated that characters of the chondrocranium were most informative in studies of *Cercopithecus* monkeys, and suggested that this might apply across primates and perhaps across all mammals. Lastly, Ruta and Bolt (2008) found that characters of the lower jaws of temnospondyls recovered many of the same relationships as those inferred from a more holistic data set.
Studies of extinct organisms *necessarily* focus on those characters capable of fossilisation (typically shells and bones). In vertebrates, disproportionate numbers of characters are often coded from the most recalcitrant skeletal elements, notably teeth in mammals (Billet, 2011; Alejandra Abello, 2013). More generally, there is a tendency for soft part characters to resolve as more derived apomorphies than those characters with a higher preservation potential. Removal of soft characters therefore results in preferential 'stemward slippage' across animal groups (Sansom *et al.*, 2010, 2011; Sansom and Wills, 2013); the phenomenon whereby taxa resolve closer to the root of the tree than they otherwise would.

In this study, we apply a variety of methods to explore differences in the strength and nature of phylogenetic signals in craniodental and postcranial partitions of 85 published vertebrate datasets. We address the following questions: 1. Do levels of homoplasy in craniodental character partitions differ from those in postcranial character partitions (Sanchez-Villagra and Williams, 1998), and are any observed differences more than simply a function of differing numbers of characters within these partitions? We quantify this using conventional indices of homoplasy (Kluge and Farris, 1969; Archie, 1996) modelled with respect to data set parameters. 2. Is there more conflict between craniodental and postcranial characters than we would expect for random partitions, and do craniodental and postcranial characters support significantly different trees as a result? We investigate this using the incongruence length difference test (Mickevich and Farris, 1981; Farris *et al.*, 1995a,b) and a new partition homogeneity test based upon tree distance metrics rather than differences in tree lengths.

### Materials and Methods

#### THE DATA SETS

Phylogenetic datasets published between 2000 and 2013 were sourced from the literature. We restricted our focus to discrete character morphological matrices composed entirely of vertebrate taxa, and analysed using equal weights maximum parsimony. Although morphological data can be
analysed with model-based likelihood (Lewis, 2000; Lee and Worthy, 2012) and Bayesian
(Nylander et al., 2003; Pollitt et al., 2005; Clarke and Middleton, 2008; Tsuj and Mueller, 2009;
Bouchard-Cote et al., 2012) methods, the considerable majority of published morphological trees
are generated utilising maximum parsimony. Matrices were also garnered from Brian O'Meara's
TreeBASE mirror (O'Meara, 2009), Graeme Lloyd's collection of dinosaur matrices (Lloyd, 2009)
and MorphoBank (O'Leary and Kaufman, 2011). We excluded matrices with fewer then eight taxa
or partitions with fewer than eight parsimony-informative characters (for reasons of statistical
power). We interpret craniodental characters here as those pertaining to the skull (cranium plus
mandible and dentition).

Our resulting sample comprised 85 matrices, spanning all major vertebrate groups. A small
number of these datasets contained characters that were not strictly morphological (e.g., character
618 relating to habitat choice in the matrix of Spaulding et al., 2009). These were removed prior to
any further analysis. We also removed phylogenetically uninformative taxa within partitions using
the principles of safe taxonomic reduction (Wilkinson, 1995). We additionally removed taxa with
large amounts of missing data that were demonstrated empirically to obfuscate resolution within
partitions, usually because they inflated search times and numbers of optimal trees to impractical
levels. The number and percentage of craniodental and postcranial characters in each matrix were
recorded, as well as the fraction of missing entries within these partitions. Wilcoxon signed-rank
tests were used to assess the significance of differences between these medians in different
groups.

Simply recording percentages of missing cells has limitations, as these can be distributed
randomly (usually less problematic analytically) or can be concentrated within particular taxa. Such
concentrations are often observed in real data sets, and particularly in matrices including fossil
taxa (Cobbett et al., 2007). Our pruning of taxa removed the worst of these effects. Simulations
have demonstrated that it is the signal within the characters that are coded that is critical in
determining the placement of particular taxa (Wiens, 2003ab) rather than numbers of missing cells
per se.
All phylogenetic analyses were performed using TNT (Goloboff et al., 2008b), using equally weighted parsimony. We also reproduced any assumptions regarding character order, polarity and rooting. Empirically, we determined that comprehensive searches involving 200 parsimony ratchet iterations (Nixon, 1999) and 100 drift iterations per replication, with 10 rounds of tree fusion (Goloboff, 1999) were effective at recovering the set of MPTs reported by the original authors in each case. These settings were used throughout.

TESTING WHETHER CRANIODENTAL AND POSTCRANIAL CHARACTERS EXHIBIT DIFFERENT LEVELS OF HOMOPLASY

Consistency and retention indices. – There are two intuitive ways to calculate differences in mean/median consistency indices (ci; Kluge and Farris, 1969) and retention indices (ri; Farris, 1989) for characters in partitions of a dataset (ci and ri in lower case pertain to individual characters). The usual approach is to find the optimal tree or trees for all characters analysed simultaneously (the global MPT(s)) and to take mean values for characters reconstructed on this/these (Sánchez-Villagra and Williams, 1998; Song and Bucheli, 2010). However, there are theoretical partition size effects, even in the absence of differences in the levels of character conflict within partitions. All other things being equal, the characters within the larger partition are likely to have higher ci values on average (Fig. 1). The other approach is to report metrics for characters within each partition analysed independently. However, the ensemble CI and ensemble RI (and therefore ci and ri for individual characters) are influenced by data set dimensions (Archie, 1989; Sanderson and Donoghue, 1989; Faith and Cranston, 1991; Klassen et al., 1991): there is a strong, negative correlation between CI and the number of taxa and a weaker, negative relationship between CI and the number of characters (Archie, 1989, 1996; Archie and Felsenstein, 1993).

There are two ways in which differences between indices can be tested. For individual data matrices, Mann-Whitney or t-tests can be applied to character ci and ri values, with the null that these have the same median or mean in the two partitions. For the more general comparison across all 85 matrices simultaneously, Wilcoxon signed ranks or paired t-tests can be used to test...
the nulls that (either) the median/mean ci or ri in craniodental and postcranial partitions were similar, or that the median/mean CI and RI indices for the two partitions were similar.

Homoplasy excess ratio (HER). – The homoplasy excess ratio (HER; Archie and Felsenstein, 1993) was proposed as an adjunct to the ensemble consistency index (CI), and argued to be relatively immune to its worst shortcomings. HER is given by \( \frac{\text{MEANNS} - L}{\text{MEANNS} - \text{MINL}} \), where MEANNS is the mean length of the most parsimonious trees resulting from a large sample of matrices (here 999) in which the state assignments within each character have been randomised. L is then the optimal length of the original dataset, and MINL is the minimum possible length of the dataset (number of states minus number of characters). HER was calculated for craniodental and postcranial partitions in isolation, and we then tested for differences in partitions across all 85 datasets using the Wilcoxon signed ranks test.

TESTING INCONGRUENCE BETWEEN CRANIODENTAL AND POSTCRANIAL CHARACTER PARTITIONS

Incongruence length difference (ILD) test. – To assess the significance of congruence between whole character partitions as measured by optimal tree length, the ILD test (Mickevich and Farris, 1981; Farris et al., 1995ab; Barker and Lutzoni, 2002) was applied to the matrices in TNT using 999 replicates (Allard et al., 1999a,b). The ILD score is given by \( L_{AB} - \frac{(L_A + L_B)}{L_{AB}} \), where \( L_{AB} \) is the optimal tree length (in steps) of the simultaneous analysis of both partitions together (the total evidence analysis), and \( L_A \) and \( L_B \) are the optimal tree lengths for partitions A and B analysed independently. To determine the significance of the observed ILD score, random partitions of the same size (number of characters) as the specified partitions are also generated to yield a distribution of randomized ILD scores. Given the nature of phylogenetic data, the suitability of this test has been questioned on a variety of grounds (Dolphin et al. 2000; Hipp et al. 2004; Ramirez 2006; Planet 2005, 2006). Despite this, the ILD test remains commonly used to compare the congruence of data partitions. We did not apply the arcsine transformation of Quicke et al. (2007) because they justified their correction on the basis of empirical and simulated molecular data (morphological data have different statistical properties).
TESTING WHETHER PARTITIONS SUPPORT DIFFERENT TREES

The incongruence relationship difference (IRD) test: a new test of the congruence of relationships.

– Much like the ILD test, this is a randomisation-based test. However, partitions are compared via the distances between the optimal trees that result from them, rather than via tree length (ILD) or a matrix-representation of topology (TILD; Wheeler, 1999). There are many possible tree-to-tree distance measures including symmetric difference (RF; Bourque, 1978; Robinson and Foulds, 1981; Pattengale et al. 2007) quartets distance (QD; Estabrook et al., 1985), nearest neighbour interchange distance (NNID; Waterman and Smith, 1978), nodal distance (Bluis and Shin, 2003), maximum agreement subtree distance (Goddard et al., 1994; de Vienne et al., 2007), transposition distance (Rossello and Valiente, 2006), subtree prune and regraft distance (SPR; Goloboff, 2008), and path-length difference (PLD; Zaretskii, 1965; Williams and Clifford, 1971). For reasons of familiarity (it is among the most well characterised; e.g. Steel and Penny, 1993) and ease of computation, we chose to use the symmetric difference or Robinson and Foulds distance (RF) as our measure of tree-to-tree distance. We note that all other implementations are possible. We illustrate the approach for two examples: firstly the theropod data of Ezcurra and Cuny (2007) (Fig. 3a) and secondly the mammalian data of Beck (2008) (Fig. 3b). The results from each partition are illustrated as 50% majority rule consensus trees for ease of visualization. The left hand tree in both cases is that derived from the analysis of craniodental characters alone, while the right hand tree is inferred from just the postcranial characters. The open circles indicate nodes present in one partition tree that are absent from the other, and the total number of such nodes gives the measure of RF between the trees. This value corresponds to the incongruence relationship difference (IRD_{MR} in the case of the majority rule trees illustrated).

All MPTs from the analysis of each partition were saved and then compared to each other in two different ways. (1) ‘Nearest neighbours’ (IRD_{NND}) for up to 10,000 trees in each partition: the mean of the minimum distance between each tree in one set, compared with the trees in the other (and vice versa) (Cobbett et al., 2007). (2) The distance between the 50% majority-rule consensus trees (from up to 10,000 fundamentals) for each partition (IRD_{MR}) (Figs. 3 and 4).
generated random partitions of the original data in the original proportions, and repeated the above exercises in order to yield a distribution of randomized partition tree-to-tree distances. Distances for the original partitions were deemed significantly different from this distribution when they lay in its 5% tail. Our p-values were derived from 999 replicates. All but three (98%) of our 170 partitions yielded less than 10,000 trees; the imposition of a 10,000 bound for the remainder was a necessary limitation to restrict prohibitively long searches in poorly resolved partitions.

Tests not implemented. – The topological incongruence length difference (TILD) test (Wheeler, 1999) is analogous to the ILD test, but is applied to a matrix representation (GIC; Farris, 1973; also known as MRP coding, Baum, 1992; Ragan, 1992) of the branching structure of a consensus of the optimal trees from the data partitions. The test appears to have limited discriminatory power and high type I error rate (Wills et al. 2009).

Rodrigo et al. (1993) proposed three interrelated tests to investigate differences in relationships directly. The first of these determines whether the symmetrical difference distance (RF: Robinson and Foulds, 1981) between sets of MPTs from independent analyses of the two data set partitions is distinguishable from the distribution of RF distances between a large sample of pairs of random trees. Only weak congruence between partitions is needed to pass this test. The second test of Rodrigo et al. (1993) compares the partitions directly, and determines if there is any overlap between the MPTs derived from the two partitions upon bootstrap resampling. This is problematic because the probability of encountering common trees changes with the bootstrap parameters (Lutzoni, 1997), especially the number of replicates (Page, 1996). The third test compares RF distances between partitions with those between trees bootstrapped from within partitions. Although a useful test, it may have limitations, particularly where the partitions of the data set are of very different sizes, and especially where the number of characters in the smaller partition is also small relative to the number of terminals. In such cases, bootstraps of the smaller partition may consistently yield poor resolution and low RF distances between trees within this partition (Page, 1996). The IRD\textsubscript{NND} test proposed above controls this partition size difference.
Results

HOMOPLASY IN CRANIODENTAL AND POSTCRANIAL DATA PARTITIONS

Across our sample of 85 datasets, craniodental partitions had more characters (median = 58) than postcranial partitions (median = 50) (Wilcoxon signed ranks; \( V = 2288.5, p = 0.044 \)) (Table 1). The percentage of missing data cells was comparable in craniodental (median = 12.8%) and postcranial (median = 16.9%) partitions, although this difference was significant (paired Wilcoxon; \( V = 868.5, p = 0.006 \)). The mean ensemble consistency indices (CI) for craniodental and postcranial characters across all 85 data sets were not significantly different (paired \( t = 1.184, p = 0.240 \)), with postcranial partitions (\( X = 0.564 \)) having slightly higher values than craniodental partitions (\( X = 0.550 \)). Using the mean partition (per character) CI index across all matrices (characters optimised onto the globally optimal tree(s) for the entire matrix) revealed a non-significant difference (\( X = 0.632 \) and 0.627 for craniodental and postcranial partitions respectively; paired \( t = 0.450, p = 0.654 \)). Mann-Whitney tests of craniodental and postcranial CI values within the 85 data sets yielded 40 significant (\( p < 0.05 \)) results (four or five might be expected). Twenty-one of these 40 had higher means (less homoplasy) for cranial characters, despite their larger partition size. A simple linear model was used to express partition CI in terms of the log of the number of taxa, the log of the number of characters and the log of the percentage of missing data (+1) across all 170 partitions. The term for missing data was not significant, but both the log of the number of taxa (\( p < 0.001 \)) and the log of the number of characters (\( p < 0.014 \)) were highly so (multiple \( R^2 = 0.458, p = < 0.001 \)). A subsequent paired t-test of the residual CI values from this model revealed no significant difference (\( t = 0.917, p = 0.362 \)) between craniodental and postcranial partitions. We note that other variables have been demonstrated empirically to influence CI (Donoghue and Ree, 2000; Hoyal Cuthill et al., 2010).

Homoplasy excess ratio (HER) values (Archie 1989, 1996; Archie and Felsenstein, 1993) were similar in the craniodental (\( X = 0.582 \)) and postcranial skeleton (\( X = 0.571 \)) (paired \( t = 0.621, p = 0.537 \)). A linear model of HER in terms of the logs of numbers of characters and taxa and the percentage of missing data (+1) revealed no significant independent variables. Finally,
retention indices were significantly higher for postcranial than craniodental partitions when measured across all characters in a partition (RI; paired $t = 2.654, p = 0.009$) but not as the average of per character values within a partition ($r_i$; paired $t = 1.538, p = 0.128$). Linear modelling of the partition RI in terms of the logs of numbers of characters and taxa and the percentage of missing data (+1) revealed no significant independent variables.

**CONGRUENCE BETWEEN CRANIODENTAL AND POSTCRANIAL SIGNALS (ILD TESTS)**

When originally described, the ILD test was used with a standard significance level of 5% (0.05). At this level, 31 of our 85 dataset partitions had significant character incongruence (Table 1). Some have advocated more stringent levels (e.g., Cunningham, 1997): with $p < 0.010$, we still rejected the null for 23 of our datasets. We note that the correlation between ILD $p$-values and the percentage of missing data within a data set was not significant ($r = -0.115, p = 0.125$), although there was a significant correlation between ILD $p$-values and the difference in the percentage of missing data cells in the two partitions ($r = -0.161, p = 0.031$). When culling our matrices to those with a difference of just 5% or less in the fraction of missing cells in the two partitions ($n = 40$), we still observed 11 data sets with an ILD significant at $p < 0.050$ (a similar rate of rejection: $G = 2.653, p = 0.103$).

Logistic regression was used to model the binary outcome of the ILD test (significant or not at $p < 0.05$) as a function of log of the number of taxa, log of the number of characters, the imbalance in number of characters between partitions (as a percentage of the total number), the percentage of missing entries in the data set, and the log of the imbalance in the percentage of missing entries between partitions. Terms for the number of characters ($p < 0.001$) and the imbalance in character numbers ($p = 0.021$) were retained in the minimum adequate model selected by the progressive deletion of non-significant terms ($p > 0.05$).

**THE SIMILARITY OF RELATIONSHIPS IMPLIED BY PARTITIONS**

Across all 85 data sets, 27 had significantly incongruent relationships (IRD$_{NND}$) at $p < 0.05$, of which 14 were also significant at $p < 0.01$ (Table 1). Correlation between $p$-values and the
difference in the percentage of missing data between partitions was not significant ($\tau = -0.012, p = 0.876$). Moreover, the rate of rejection of the null at $p < 0.05$ was similar for the culled ($n = 40$) data set and those matrices ($n = 45$) with more than 5% difference in missing data between partitions (eleven and sixteen significant results respectively) ($G = 0.637, p = 0.425$). Logistic regression of the binary outcome of the IRD$_{nnd}$ test ($p < 0.05$ or otherwise) as above yielded no significant terms in the minimum adequate model.

The correlation between $p$-values for the nearest neighbour and majority rule variants of the IRD test was highly significant ($p < 0.001$) but not particularly tight ($\tau = 0.387$). The latter offers a very imprecise proxy for the distances measured by nearest neighbours, and we do not advocate its use.

**HIGHER TAXONOMIC DIFFERENCES IN CONGRUENCE**

Partitioning datasets into six broad and inclusive taxonomic groups (Aves, other Ornithodira, Synapsida, other reptiles, amphibians (including early tetrapods), and fishes) revealed some significant differences (Fig. 5). In particular, synapsids and amphibians were less likely to have congruent partitions according to the ILD test than other groups ($G = 11.808, p = 0.038$) (Table 2). However, there were significant differences in the log of the number of characters across groups ($F = 3.095, p = 0.013$), largely accounted for by the contrast between synapsids and fishes (Tukey HSD, $p = 0.003$). Taxonomic group was not significant as a factor in logistic regression models, and the residuals from models omitting taxonomic group membership retained no differences between groups (Kruskal-Wallis $\chi$-squared $= 2.976, p = 0.704$). A broadly similar taxonomic pattern was observed for the IRD$_{nnd}$ test, although differences in frequencies of null rejection across higher taxa were not significant ($G = 4.283, p = 0.510$).

**Discussion**

**CRANIODENTAL AND POSTCRANIAL PARTITIONS CONTAIN SIMILAR LEVELS OF HOMOPLASY**
It is well known that the CI is negatively correlated with the number of taxa in a matrix (Archie, 1989; Sanderson and Donoghue, 1989; Faith and Cranston, 1991; Klassen et al., 1991), but it also has a weaker, negative relationship with the number of characters (Archie, 1989, 1996; Archie and Felsenstein, 1993). Despite the greater number of craniodental characters than postcranial characters across our sample of data sets, we found no significant difference in CI (paired t = 1.184, p = 0.240). Unsurprisingly, when modelling out both data matrix dimensions (numbers of taxa and characters) and the percentage of missing data (+1) across all 170 partitions, residual CI values were even more similar (for residuals: t = 0.917, p = 0.362). Homoplasy excess ratio (HER) values were also similar in the craniodental (\(\bar{X} = 0.582\)) and postcranial skeleton (\(\bar{X} = 0.571\)) (paired t = 0.621, p = 0.537).

We note that the absence of a clear difference between craniodental and postcranial levels of homoplasy does not necessarily imply that additional characters of equivalent phylogenetic informativeness can be garnered from the two partitions with comparable ease. One partition may have been exhausted with considerable care, the other not. Our conclusions therefore necessarily relate to the coded data. We also note that a high CI within a partition could be the result of a strong phylogenetic signal, or the developmental non-independence of characters. Distinguishing between these causes requires detailed developmental and underpinning genetic knowledge, which are often unavailable.

**PARTITIONS HAVE INCONGRUENT SIGNALS MORE OFTEN THAN WE EXPECT**

Our ILD test results demonstrate that our partitions are incongruent about one time in three: 31 from 85 datasets. Assuming a significance level (false positive rate) of 5%, we would expect four or five datasets to be significantly incongruent by chance. Significant incongruence is therefore detected across our sample of datasets (binomial test \(p < 0.001\), assuming a 5% false positive error rate), although this is partly accounted for by differences in partition parameters. We make no inferences concerning the overall quality of individual data sets on the strength of these results,
and note that partitions were imposed by us in each case (rather than reflecting distinctions made by the original authors).

CRANIODENTAL AND POSTCRANIAL PARTITIONS OFTEN IMPLY SIGNIFICANTLY DIFFERENT RELATIONSHIPS

Results from the incongruence relationship difference (IRD_{NND}) tests were broadly similar to those from the ILD test: 32% of data sets yielded significantly different trees from the two partitions. As with the ILD test, we would expect just three or four data sets (5%) to reject the null by chance (a highly significant difference: binomial test $p < 0.001$). Our empirical sample suggests that the IRD_{NND} test using the Robinson Foulds distance is less likely to yield a significant result than the ILD, and is less susceptible to differences in partition size and differences in levels of missing data. The IRD_{NND} therefore has certain advantages over the ILD, and also offers a more intuitive index of congruence (i.e., one based directly on differences in topological branching structure rather than differences in tree length). We note, however, that the IRD_{NND} is insensitive to differences in branch lengths. Many authors have identified difficulties with the incongruence length difference test (ILD) (Dolphin et al. 2000; Hipp et al. 2004; Ramirez 2006; Planet 2005, 2006) and we do not repeat these here.

As with all similar metrics, the IRD is to some extent arbitrary (Wheeler, 1999). Different indices of tree-to-tree distances will yield different distances and $p$-values, and the IRD can be implemented with many such measures. For example, the Robinson Foulds distance penalises distant and shallow branch transpositions much more strongly than close and shallow transpositions, whereas the maximum agreement subtree distance (Finden and Gordon, 1985), for example, makes only a marginal distinction between these two cases. Similarly, the maximum agreement subtree distance is less sensitive to the depth of the transposition (Cobbett et al., 2007). Alternative metrics will have other, and perhaps more desirable, properties (Lin et al., 2011). Our choice of the Robinson Foulds distance here was partly pragmatic, as it is computationally less demanding than most alternatives.
WHAT DOES PARTITION INCONGRUENCE IMPLY?

In studying the evolution of form, it is now relatively common to recognize anatomical modules (Mitteroecker and Bookstein, 2007, 2008; Klingenberg, 2008; Cardini and Elton, 2008; Lü et al. 2010; Hopkins and Lidgard, 2012; Cardini and Polly, 2013; Goswami et al., 2013, 2015). These are regions of the body (or suites of landmarks) within which morphological changes are strongly correlated through evolutionary time, but between which there is significantly less coordination. Different selective forces may operate on these modules or components of the mosaic (Gould, 1977; Maynard Smith, 1993; Kemp, 2005; Lü et al., 2010), and they may therefore exhibit different evolutionary rates and trends (Mitteroecker and Bookstein, 2007, 2008; Klingenberg, 2008). In the context of phylogenetic characters, differing pressures on modules may favour particular patterns of convergence and homoplasy, and therefore suites of characters that imply different trees (Clarke and Middleton, 2008). The skull of many tetrapod groups has often been regarded as biomechanically and functionally somewhat independent of the rest of the skeleton (Ji et al., 1999; Koski, 2007; Mitteroecker and Bookstein, 2008) hence the difficulty of making many inferences about the one from the other. We note that fishes have the lowest overall incongruence between partitions and the most similar levels of per character consistency index (ci) (Fig. 5). This may result from greater functional and biomechanical integration between the head and trunk in fishes compared with other vertebrate groups (i.e., fishes lack a functional neck) (Klingenberg, 2008; Larouche et al., 2015).

We note that while anatomical modules are usually envisaged as physically proximate suites of landmarks or characters, it is possible for characters to evolve in a coordinated manner across the body as the result of particular selective pressures (Kemp, 2007). For example, adaptations for swimming, digging or flying (Gardiner et al., 2011; Abourachid and Hofling, 2012; Allen et al., 2013) might entail correlated suites of change across the body in a manner that would not be apparent from studying straightforward divisions into body regions (e.g., head, body and limbs). Developmental regulatory processes may also entail counterintuitive suites of coordinated character change across the body (Kharlamova et al. 2007; Chase et al. 2011), but our objective was not to test for such suites of correlations here.
In the most general terms, character selection and coding clearly has an impact upon inferred phylogeny. Most straightforwardly, alternative data sets for identical sets of taxa can yield different trees (Freitas and Brown, 2004; Munoz-Duran, 2011; Penz et al. 2013), and this argues strongly for the synthesis of all characters. More generally, systematists rightly exercise their judgement in deciding which aspects of morphology to codify as putative homologies. Wings in birds, bats and pterosaurs are not considered homologous as wings *a priori* (although they are as limbs) because the weight of evidence unambiguously rules this out (and in the absence of any formal analysis). In many cases, however, the decision is less straightforward, and the *a priori* omission of characters believed to be analogous or strongly homoplastic may unintentionally overlook useful signal at some level in the tree. Finally, it is often observed empirically that the tree(s) derived from a given matrix can alter markedly with the omission, reweighting or ordering of characters (Wills, 1998), such that even modest perturbations to the data yield large changes to the resulting trees.

In morphological phylogenetics, it is usual to combine all available data (Kluge *et al.*, 1989). Our results therefore reinforce the importance of this approach. While the patterns inferred from particular organ systems or suites of characters may be misleading (in the same way and for the same reasons that individual characters may merely introduce homoplasy and noise), combined analysis of all available characters often allows a globally strong phylogenetic signal to emerge from conflicting local homoplasy (Gatesy *et al.* 1999). We demonstrate that in vertebrate studies of this type, an exclusive focus upon characters of either the cranium or postcranium (at the expense of those of the other partition (e.g., Fitzgerald (2010) (craniodental only) and Mayr and Mourer-Chauvire (2004) (postcranial only))) will significantly influence the resultant optimal tree(s) about 30% of the time, irrespective of major group. We therefore strongly advocate garnering character data intensively from all anatomical regions whenever possible.

When analysing fossils, it is usually impossible to sample across the same suite of characters that would ideally be coded for extant species (Wiens, 2003a,b; Cobbett *et al.*, 2007). For example, in fossil crocodyliforms, the vast majority of characters are coded from the skull (e.g., O’Connor *et al.*, 2010; Turner and Sertich, 2010; Cau and Fanti, 2011; Hastings *et al.*, 2011;
Puertolas et al., 2011) and it is difficult to be confident that we are not merely inferring a 'craniodental' tree. The only (and indirect) way to test this would be to conduct parallel analyses upon the closest living representatives of the clade. However, the (quite possibly limited) utility of this approach depends upon the phylogenetic proximity of the extant exemplars, the presumed constancy of selective pressures on putative modules through time and across clades (a big assumption: Hunt, 2008; Frazzetta 2012), and the similarity of the available coded data. A related issue in the context of fossil vertebrates is the preferential preservation of hard part characters (bones rather than muscles or other more volatile tissues). An analogous concern, therefore, is whether skeletal and soft-part characters convey a consistent phylogenetic signal (Diogo, 2004). If they do not, then this has implications for the manner in which fossil vertebrates are interpreted and analyzed (Sansom et al., 2010; Sansom and Wills 2013; Pattinson et al., 2014) and is an area particularly needing detailed future work.

Conclusions

1. Systematists typically abstract significantly more characters from the skull than the rest of the skeleton. However, tests for levels of homoplasy in the craniodental and postcranial partitions of our sample of 85 matrices revealed no significant differences, irrespective of how homoplasy was measured. Systematists appear to be coding characters of similar internal consistency from both regions of the body. It is unclear to what extent the bias towards coding craniodental characters reflects a real bias in their distribution, the availability of material, or arises from the preconceptions of systematists.

2. Craniodental and postcranial character partitions exhibited significant incongruence (ILD tests) in 31 of our 85 sample data sets. Likewise, our new IRD_{NND} test found significantly different relationships in 27 from 85 cases.

3. Although vertebrate systematists sometimes code morphological characters preferentially from particular anatomical regions (through necessity or by choice), received wisdom is that a broad and unbiased sampling is usually preferable when possible. Here, we provide
empirical evidence for this view. Our results strongly support dense sampling of characters from across the vertebrate skeleton.

4. The IRD\textsubscript{NND} test appears to be less sensitive to differences in numbers of characters or amounts of missing data between partitions than the ILD test. We advocate its use as an ancillary test for addressing differences in implied relationships between data partitions.

5. There were significant differences in the distribution of significant partition inhomogeneity across higher taxonomic groups, as measured by the ILD test. Synapsids and amphibians were less likely to have congruent craniodental and postcranial partitions than other groups. However, these differences were no longer significant when the imbalance in partition size was modelled out. Neither were the relationships inferred from craniodental and postcranial characters more likely to conflict significantly in some higher taxa than in others (IRD\textsubscript{NND} test). Our findings therefore have some generality across vertebrates.

ACKNOWLEDGEMENTS

We would like to thank all authors who kindly supplied us with their datasets (knowingly or otherwise). We are grateful to Marcello Ruta and anonymous reviewers whose comments enabled us to substantially improve this paper. Ward Wheeler and Pablo Goloboff also made invaluable comments on an earlier version of this work as presented at Hennig XXIX. Some of our analyses were made possible by the generous free use of a high performance computing cluster; Bioportal, University of Oslo. RCPM gathered data and performed analyses. RS scripted IRD tests within \textit{TNT} and performed analyses. MAW conceived and designed the study, devised/programmed the original tests in \textit{PAUP}* and performed analyses. RCPM and MAW summarised results and wrote the paper. We thank Elizabeth Wills for drawing the silhouettes in Figure 3. This work was supported by a University of Bath URS award to RCPM, Leverhulme Trust Grant F/00351/Z and BBSRC grant BB/K015702/1 to MAW, JTF Grant 43915 to Mark Wilkinson and MAW, and NERC fellowship NE/I020253/1 to RSS.


Parrilla-Bel, J., M. T. Young, M. Moreno-Azanza, and J. I. Canudo. 2013. The first metriorhynchid crocodylomorph from the Middle Jurassic of Spain, with implications for evolution of the subclade Rhacheosaurini. *PLoS One* 8: e54275+


Sansom, R. S. 2015. Bias and sensitivity in the placement of fossil taxa resulting from interpretations of missing data. *Syst. Biol.* **64**: 256-266.


Table and figure legends
Table 1. Summary statistics for the 85 vertebrate morphological data sets analysed herein. Data set dimensions (numbers of taxa (ntax) and informative characters (nchar)) refer to the pre-processed matrices after applying safe taxonomic deletion rules (see text for details). ‘Cran. char.’ and ‘Post. char’ denote characters per partition. ‘Cran. miss %’ and ‘Post. miss %’ report the percentage of missing data cells for partitions. ‘ILD’ column reports the p-value resulting from an incongruence length difference test with 999 random partitions. IRD columns report the p-values resulting from incongruence relationship difference tests with 999 random partitions. ‘IRD\textsubscript{NND}’ denotes the results of the IRD test using the Robinson Foulds tree-to-tree distance for nearest neighbouring trees. CI columns give ensemble consistency indices for partitions of the data set (craniodental or postcranial). Similarly, ci columns report the mean per character consistency indices for partitions of the data set (craniodental or postcranial) when optimised onto the globally most parsimonious trees. HER gives the homoplasy excess ratio for partitions of the data set (craniodental or postcranial) derived from 999 randomized matrices. An expanded version of this table containing additional statistics is provided within the Supporting Information.
Table 2. Number of data matrices in higher taxonomic groups, with a tally of those with significant ($p < 0.05$) results for various partition homogeneity tests. G-test and $p$ values quoted for the null that significant results are equally likely in the five higher taxonomic groups in each case.
Figure 1. (a and b) Characters sampled from different anatomical regions can yield radically different most parsimonious trees (MPT) when analysed in isolation. In both cases, there is no homoplasy within either region (characters 1-3 or characters 4-6), and a single MPT results in each case. (c) Combining the data from both partitions (characters 1-6) yields four MPTs, the strict consensus of which (illustrated) is entirely unresolved. Character statistics have been averaged over the four trees. (d) Two additional characters (5’ and 6’) are sampled from the same region as ‘b’, and these have the same distribution as 5 and 6 respectively. Analysis of all characters now reveals a single MPT with relationships identical to those in ‘b’ (characters 4-6). Characters 4-6, 5’ and 6’ contain no homoplasy: all conflicts are resolved with a cost to characters 1-3. In this case, the MPT is identical to the result that would be obtained by a clique analysis (sensu Le Quesne 1969).
Figure 2. Calculation of two partition inhomogeneity metrics for ‘craniodental’ and ‘postcranial’ partitions of a hypothetical data set. In this example there are equal numbers of craniodental (1-9) and postcranial (10-18) characters, but this need not be the case. For the Incongruence Length Difference (ILD) measure, maximally parsimonious trees (MPTs) are inferred from the craniodental and postcranial partitions of the data independently. The summed lengths of these trees (11 steps + 12 steps) is the sum of partition lengths (23 steps). In parallel with this, an MPT is inferred from both partitions analysed simultaneously. This tree is longer (25 steps) than the sum of partition lengths (23 steps), and the difference between them is the ILD (25 – 23 = 2). The ILD represents the reduction in homoplasy afforded by the isolation of the two partitions (two extra steps are needed when the partitions are combined). For the Incongruence Relationship Difference (IRD) measure, the branching structure of the craniodental and postcranial partition MPTs are compared (rather than their lengths) using one of several possible tree-to-tree distance metrics. Here, we illustrate the symmetric difference distance (RF) of Robinson and Foulds (1991). Open circles mark branches in either the craniodental or postcranial MPT that are absent from the other. The tally of these unique branches on both trees is the RF (2 + 2 = 4). Some background level of ILD or IRD is anticipated wherever a data set contains homoplasy. In order to interpret these observed metrics, therefore, we need to know what values would be expected for partitions of similar data sets in similar proportions. Random character partitions are used to
generate null distributions for both the ILD and IRD, and observed values deemed significantly different from the null if they lie in some specified fraction of the tails.

**Figure 3.** Most parsimonious trees derived from craniodental and postcranial partitions can be significantly more different than we would expect. Tanglegrams computed using Dendroscope (Huson and Scornavacca,
The theropod data of Ezcurra and Cuny (2007) yielded one most parsimonious tree from the craniodental partition and six from the postcranial partition, the latter summarized as a majority rule tree merely for ease of visualization (we advocate the use of tests based upon mean nearest neighbours within sets of most parsimonious trees). Branches labelled with circles are unique to one or other tree (those unlabelled are common to both). The Robinson Foulds (RF) distance between the two is simply the sum of unique branches (6+6=12). While the ILD test returned a highly significant result ($p = 0.011$) and our new incongruence relationship difference test (IRD$_{NND}$) using RF did not ($p = 0.791$). (b) Six craniodental and four postcranial trees in the mammalian data of Beck (2008), again summarized as majority rule consensus trees for visualization. In this case, the incongruence length difference (ILD) test for partition homogeneity returned a highly significant result ($p = 0.004$) whereas our incongruence relationship difference (IRD$_{NND}$) test did not ($p = 0.259$). Indicative images are, from top to bottom on right hand side: *Ornithorhynchus*, *Tachyglossus*, *Didelphis*, *Monodelphis*, *Caenolestes*, *Dasyurus*, *Phascolarctos*, *Vombatus*, *Perameles*, *Echymipera*, *Macropus*, *Phalanger*, *Petaurus*, *Vincelestes*. Image of *Tachyglossus* courtesy of echidnasclub.com. See text for further explanation.

**Figure 4.** Summarising sets of most parsimonious trees (MPTs) for partitions prior to calculating tree-to-tree distances is computationally much faster than calculating distances between nearest neighbours. However, majority rule trees present modal or most frequent relationships, and may therefore plot eccentrically in tree
space. This is an undesirable property when attempting to summarise distances between sets of trees. Figure shows tree-to-tree distances for craniodental and postcranial partitions of the mammalian data of Pujos (2007). Distance matrices have been plotted in two dimensions using non-metric multidimensional scaling (NMDS), and rotated using principal components analysis (PCA). Circles indicate craniodental trees and squares indicate postcranial trees. Open symbols denote original MPTs, filled symbols (black) denote majority rule trees.

![Figure 5. Summary of results from three partition homogeneity tests, subdivided by higher taxonomic group.](http://creativecommons.org/licenses/by-sa/3.0/)

Black bars denote significant ($p < 0.05$) differences in $ci$ values between partitions (Mann-Whitney U tests). Gray bars denote significant ILD results, while light gray bars indicate significant $IRD_{NND}$ results. Images from phylopic.org (courtesy of Michael Scroggie, Oscar Sanisidro, B. Kimmel and Steven Coombs). Salamander and pterosaur courtesy of Matt Reinbold (modified by T. Michael Keesey) and Mark Witton.