Diversity and Distributions A Journal of Conservation Biogeography

BIODIVERSITY RESEARCH

Phylogenetic trees do not reliably predict feature diversity

Steven Kelly[1], Richard Grenyer[2] and Robert W. Scotland[1]*

[1]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK, [2]School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

*Correspondence: Robert W. Scotland, Department of Plant Sciences, South Parks Rd, University of Oxford, Oxford, OX1 3RB, UK.
E-mail: robert.scotland@plants.ox.ac.uk

ABSTRACT

Aim Phylogenetic trees provide a framework for understanding the evolution of features (properties, characters or traits) of species, where closely related species share many common or similar features. This property of phylogenetic trees has practical use in applications such as bio-prospecting, where an optimal strategy exploits phylogenetic information to target closely related species to search for shared features of interest. The implicit corollary of this is that distantly related species share few features in common. This property of phylogenetic trees is thought to be useful for conservation evaluation in choosing sets of species that maximize the present utilitarian benefits of extant feature diversity (such as biologically active compounds or source systems for genetic engineering) as well as maximizing the range of evolutionary trajectories into the future.

Location Global.

Methods Here, we investigate the relationship between phylogenetic trees and biological features through both simulation and meta-analysis of 223 publicly available feature matrices.

Results We demonstrate that phylogenetic tree distance, both in real and simulated datasets, is correlated with feature similarity only for a short relative distance along the tree, such that there is no relationship for the majority of the length of most phylogenetic trees. In other words, close relatives share more features than distant relatives but beyond a certain threshold increasingly more distant relatives are not more divergent in phenotype.

Main conclusions Measures of phylogenetic diversity based upon maximizing phylogenetic distance may not maximize feature diversity.

Keywords

Feature diversity, homoplasy, phenotype, phylogeny, predictivity, traits.

INTRODUCTION

Using the association between phylogenetic relationships and the distribution of features to inform conservation decisions has a formal history going back more than 20 years (May, 1990; Vane-Wright *et al.*, 1991; Faith, 1992a). The underlying assumption is that biological feature diversity is best sampled by the largest possible phylogenetic distance amongst species (Faith, 1992a,b): conservation strategies that select sets of species separated by maximal phylogenetic distances should result in a concomitant maximal diversity of features (Faith, 1992b). Maintaining this feature diversity should not only result in the greatest utilitarian benefits of biological diversity, but also contemporary evolutionary novelty (Isaac *et al.*, 2007) and the range of potential evolutionary trajectories in the future (Forest *et al.*, 2007). The use of phylogenetic measures in conservation has not been uncontroversial, but criticisms have focused upon the eventual goal of conservation strategies (Erwin, 1991; Krajewski, 1994) and the spatial distribution of phylogenetic diversity (Erwin, 1991; Krajewski, 1994; Rodrigues *et al.*, 2005, 2011) rather than on the assumption that maximizing phylogenetic diversity results in the selection of the most disparate phenotypes. In fact, the expectation that phylogenetic diversity identifies sets of taxa that maximize the accumulation of feature diversity is widespread (Faith, 1992a, 2013; Barker, 2002; Soutullo

DOI: 10.1111/ddi.12188
http://wileyonlinelibrary.com/journal/ddi

et al., 2005; Forest *et al.*, 2007; Cadotte & Davies, 2010; Collen *et al.*, 2011) (Text S1), although some authors are more cautious about this expectation (Winter *et al.*, 2013).

Discussion of the relationship between biological features and hierarchies of organisms is also much discussed in other scientific disciplines in addition to conservation. Research areas concerned with bio-prospecting (Cracraft, 2002), character evolution (Wagner, 1961) and phylogeny reconstruction (Hennig, 1966) have all explored the distribution of biological features in the context of phylogeny. In bio-prospecting, if a feature of interest is observed in a particular taxon but is not readily exploitable, then the optimal strategy would be to sample closely related taxa on the understanding that close relatives are more likely to contain the feature of interest than a random search of all species (Gower, 1974; Correll, 1977; Farris, 1979, 1982; Bottjer, 1980; Colless, 1981; Archie, 1984; Sneath & Hansell, 1985; Sneath, 1989; Sober & Steel, 2002; Maddison & Schulz, 2007; Mishler, 2009; Saslis-Lagoudakis *et al.*, 2011, 2012; Ronsted *et al.*, 2012). One such search of relatives occurred after the isolation of the anticancer agent taxol from bark of the Pacific Yew (*Taxus brevifolia*) (Goodman & Walsh, 2001). The quantities of *T. brevifolia* bark required for commercial taxol extraction were large enough to raise pragmatic and ecological concern, as bark removal kills individual trees. Bio-prospecting in con-generic species was, in this case, successful as a suitable precursor of taxol was discovered to be abundant in the leaves of *Taxus baccata,* the European Yew, which can be harvested commercially (Goodman & Walsh, 2001). Further active toxoids have been subsequently isolated from several con-generics (Vander Velde *et al.*, 1994). In another illustrative example of the predictive ability of phylogenies, a new large and robust phylogeny of fishes increased the number of species predicted to have a venom apparatus (active venom glands and a delivery mechanism) from 200 species to 1200 species (Smith & Wheeler, 2006). This prediction was subsequently verified by anatomical studies in a subset of taxa predicted to have venom. Other recent literature (Saslis-Lagoudakis *et al.*, 2011, 2012; Ronsted *et al.*, 2012) has explored the relationship between phylogeny, bio-prospecting, medicinal plants and chemical diversity with varying degrees of success. Taken together, these implicit properties of phylogenies were elegantly summarized by Cracraft (Cracraft, 2002) 'The expectation that closely related taxa share similarities not shared with more distant taxa is the foundation for comparative biology'.

In contrast to the assumed relationship between phylogeny and features discussed above, some authors examining the distribution of susceptibility to white mould in potatoes (Jansky *et al.*, 2006) and the distribution of web forms in theridiid spiders (Eberhard *et al.*, 2008), report no predictive relationship between the feature of interest and phylogeny. Other studies directly concerned with bio-prospecting propose that feature diversity is clumped at the tips of a phylogeny, but not at deeper nodes (Saslis-Lagoudakis *et al.*, 2011), while others report that the relationship between feature diversity and phylogeny was significant but not strong (Ronsted *et al.*, 2012).

The extent of the relationship between biological features and phylogenetic distance has also been much discussed in the context of the 'morphology and molecules' debate (Hillis, 1987; Patterson *et al.*, 1993; Scotland *et al.*, 2003; Wiens, 2004) as well as the extent of convergence and homoplasy across the tree of life (Sanderson & Donoghue, 1996; Donoghue & Ree, 2000; Wake *et al.*, 2011). Although these discussions have led to divided opinions, it is most widely accepted to use DNA sequence data to infer phylogenetic trees (Felsenstein, 1978, 2004; Hillis *et al.*, 1994; Hillis, 1995; Posada & Crandall, 1998). For molecular sequence data and morphological features, data saturation (when rates of substitution are sufficiently high to erase the signal of shared ancestry) and extensive homoplasy increase the need for corrective models to infer a correct tree (Huelsenbeck & Crandall, 1997; Posada & Crandall, 1998). For biological features (generally morphology, but also discreet biochemical or molecular secondary structures), the tendency in recent years has been either to combine these data (Kluge & Wolfe, 1993; Nixon & Carpenter, 1996) with molecular sequences in a simultaneous analysis or analyse them separately and look for agreement between data partitions (Huelsenbeck *et al.*, 1996). Nevertheless, although morphology has been eclipsed as the main source of evidence for inferring phylogeny and the relationship between biological features and phylogenetic distance is often problematic, there remains a substantial body of contemporary literature claiming a predictive relationship between biological features and phylogeny (Systematics Agenda 2000, 1994; Judd *et al.*, 1999; Barker, 2002; Cracraft, 2002; Soutullo *et al.*, 2005; Simpson, 2006; Forest *et al.*, 2007; Isaac *et al.*, 2007; Maddison & Schulz, 2007; Mishler, 2009; Cadotte & Davies, 2010; Collen *et al.*, 2011; Faith, 2013).

An important issue at the heart of this debate is the mode of feature evolution: how the probability, magnitude and direction of change in each feature (and thus all features) are related to the time over which that change has occurred. There are now a large number of models of feature evolution, the primary axes of variation being: whether the features are treated as continuously valued (often described under Brownian motion frameworks in which trait variance along a continuous dimension accumulates as a function of time/distance; Felsenstein, 1985) or discrete traits (often modelled under a Markov framework in which features occupy one of a range of states, transitioning over time between them with given probability; Pagel, 1994); whether rates of feature change can be heterogeneous across characters or over time [e.g. the 'early burst' model of feature change (Simpson, 1953) in which rates of change decline over time; (Harmon *et al.*, 2010)], whether change in one feature is conditional upon the state of others (e.g. Pagel & Meade, 2006) and whether the likelihood and magnitude of change are conditional on the state of the feature at the time (e.g. Ornstein–Uhlenbeck models for continuously varied characters; Felsenstein, 1988). Given that the topology and branch lengths of the phylogeny are usually specified in advance from molecular data (see above), and the set of feature states at the tips of the tree are known, by far, the most common procedure is to choose amongst models of feature evolution based upon the relative likelihood of observing the tip features under those models and to infer evolutionary events and ecological processes from the results, such as co-evolution of particular features (e.g. Harvey & Pagel, 1991), phylogenetic niche conservatism (e.g. Losos, 2008), community assembly rules (e.g. Webb, 2000) or adaptive radiations (e.g. Yoder *et al.*, 2010).

In other words, a model of feature evolution directly relates overall phenotypic diversity to phylogenetic distance, and there are lots of models of feature evolution in use. This of course does not mean that any model universally describes the evolution of all features, nor when one model does so that it does so particularly well: as we can see from examples above, the strength of phylogenetic signal in feature diversity is variable. Demonstration of phylogenetic signal in a feature, and therefore the resultant expectation of overall feature diversity, is only meaningful with reference to a defined model of feature evolution. As a consequence, when phylogenetic distance is used as a surrogate for unknown feature diversity, an implicit but strong assumption of some model of feature evolution is made under which phylogenetic signal of unmeasured features is present to a significant degree. The importance of validating this assumption has, we suggest, been lost in most uses of phylogenetic information in conservation to a greater extent than in other fields in which it is used. Although there remain numerous studies (see Mouquet *et al.*, 2012) in which distant relatives are assumed to have dissimilar phenotypes, theorists of community phylogenetics (Webb, 2000; Losos, 2008) have been at pains to point out that relative proximity on a phylogeny need not imply relative similarity in phenotype and that different implications must be drawn about ecological processes if feature diversity has strong or weak phylogenetic signal (Kraft *et al.*, 2007; Graham *et al.*, 2012). This sort of caution must become standard procedure too in conservation uses of phylogenies and suggests that a meta-analysis of the ability of phylogenetic distance to predict feature diversity is very timely. However, opportunities to validate this relationship are rare: the contemporary dominance of molecular phylogenetics, coupled with the tendency to evaluate models of feature evolution on only a small number of features, means that few datasets exist that possess both a molecular alignment and a reasonably sized matrix of binary phenotypic features.

We consider here that the assumption of a predictive relationship between branch length and feature diversity can be stringently tested without a molecular phylogeny, and without the prior assumption of a particular time-dependent model of feature evolution, by inferring a phylogeny directly from the features themselves. A phylogeny inferred directly from the features alone will be the maximum-likelihood fit between feature diversity and a phylogenetic tree, and therefore, it is the most stringent test of the assumption of whether features are correlated with tree distance. Even if the 'true' phylogenetic tree and branch lengths inferred from molecular data were radically different from the phylogeny inferred directly from the features for the same taxa, it would not alter the fact that the features will have a stronger correlation with tree distance on the 'feature phylogeny' than on any other possible phylogeny. In this sense, our methodology is a rigorous test of the assumption that tree distance is correlated with feature diversity.

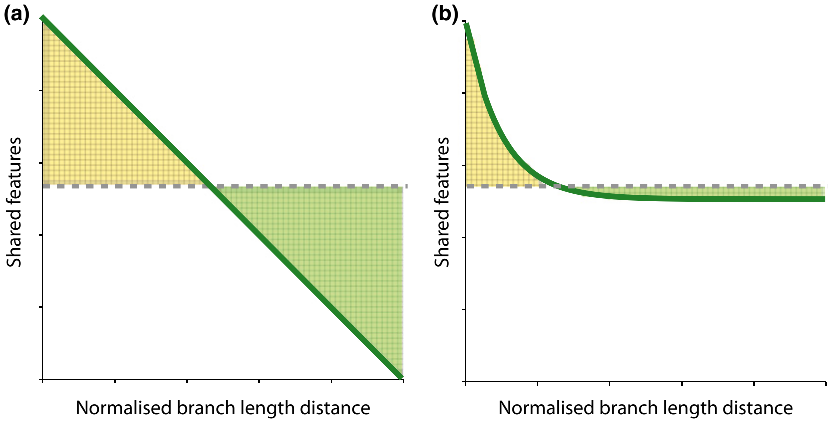

Figure 1 Summary figure of relationship between feature diversity, branch length distance, bio-prospecting and measures of phylogenetic diversity. (a) For a perfectly congruent dataset, the yellow hashed area indicates the proportion of the tree length where the phylogeny predicts shared features better than a random search. This is the sort of reasoning employed in bio-prospecting. The green hatched area indicates increasing tree distance and dissimilarity which is the property maximized in phylogenetic conservation measures. (b) In many datasets, the proportion of tree distance that is predictive of shared feature diversity is small and the proportion of tree distance that correlates with dissimilarity may not be distinguishable from random.

Figure 1a illustrates the relationship between shared feature diversity and branch length distance on a hypothetical phylogenetic tree in which all features evolve once on the tree (no homoplasy). The slope indicates that closely related species share more features in common than distantly related species. In contrast, plotting shared feature diversity between the species using a random sampling strategy for the same number of non-homoplastic features, but ignoring any information about how closely related the species are, gives the horizontal dashed line (Fig. 1a) which reflects the average shared feature diversity between taxa for the entire matrix of features. The ability of a phylogenetic tree to predict the distribution of features better than a random sampling strategy – as used in bio-prospecting – is illustrated by yellow hatch shading (Fig. 1a). The ability of a phylogenetic tree to predict the distribution of dissimilar features (maximizing feature diversity) – as used in conservation – better than a random sampling strategy is illustrated by the green hatch shading. Both bio-prospecting and conservation strategies use different ends of an assumed relationship between feature diversity and phylogenetic distance. An interesting question therefore is what happens to the framework depicted in Fig. 1a, when real datasets of biological traits are examined. Here, using both simulated and real data, we ask the question how predictive are phylogenetic trees of feature diversity?

METHODS

Source data

In July 2011, we downloaded 110 binary data matrices and 113 multistate data matrices from TreeBASE (Piel *et al.*, 2010) covering plant, animal and fungal groups and comprising a wide range of taxonomic ranks as terminal taxa. We excluded 68 super-tree matrices in which the characters do not represent features of organisms and eight matrices that comprised taxa with only missing data or identical scores for the majority of terminals. We also excluded two matrices of RNA secondary structure and two matrices of protein architecture because there were insufficient numbers to provide statistics and also because our focus was mainly on traits of general conservation value. We also excluded any matrices with combined datasets or DNA sequence datasets which were mis-classified in TreeBASE as containing phenotypic data. After filtering, the binary matrices comprised 26 matrices of morphological data, 71 of restriction site data and 13 containing insertion/deletion or gene presence/absence data. The 113 multistate data matrices were all of morphological data. Matrix identifiers and references as in TreeBASE for all binary and multistate matrices, along with matrix information including number of taxa, number of characters, number of character states and number of autapomorphies are available in (Appendix S1 and Appendix S2 in Supporting Information). The matrices are readily accessed by cutting and pasting matrix identifiers from column A in Supporting Information Appendices 1 & 2 into TreeBASE (http://treebase.org/treebase-web/urlAPI.html).

Phylogenetic tree inference

For each matrix, we inferred a phylogenetic tree directly from each discrete dataset, using maximum likelihood. We used RAxML v7.2.8 (Stamatakis, 2006) to estimate both the topology and branch lengths. In each case, either a binary or multistate model was selected with a GTR model of character evolution, optimization of substitution rates and a Gamma model of rate heterogeneity with an estimated alpha parameter and 25 distinct rate categories. Data matrices are available from TreeBASE as described above, maximum-likelihood trees and plots of feature diversity against normalized branch length distance for all matrices used in this study are available from the authors on request.

As a result of our choice of tree inference method (i.e. maximizing the fit between biological features and the phylogenetic tree), it is implicit that all other tree topologies, even those inferred from molecular sequence data, will have equal or weaker relationships between feature disparity and branch length than the topology we infer. To provide an illustration of this point, we compared a tree inferred from a 10 chloroplast gene concatenated multiple sequence alignment with a tree inferred from a 22 morphological character matrix sampled from the same taxa (Cohen, 2011). We inferred the phylogenetic tree from the morphological character matrix (feature phylogeny) using the method described above. The nucleotide sequence phylogeny (molecular phylogeny) was inferred from the concatenated sequence data using maximum likelihood implemented by RAxML v7.2.8 (Stamatakis, 2006) and using MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003); in both cases, a GTR+I+Gamma model was employed. For MrBayes two runs, each of 4 chains was initiated and allowed to run for 200,000 generations sampling every 500 generations. Convergence was assessed through visual inspection of log-likelihood traces and through analysis of the standard deviation of split frequencies. The analysis had reached stationary phase after 50,000 generations, and these first 50,000 generations were discarded as burn-in prior to inferring the consensus tree. We examined the pairwise relationship between feature diversity and branch length distance between taxa on both the feature phylogeny (Fig. 2a) and the molecular phylogenies (Fig. 2b). Figure 2a,b show that the squared Spearman's ranked correlation coefficient (SSRCC) value for the tree inferred directly from the feature phylogeny is higher than that of the tree inferred from the sequence data (here, Spearman's Rho is a nonparametric measure of correlation between branch length distance and feature diversity). Our intent, therefore, in this analysis and in our analyses of all matrices in this study is to use a model of phylogenetic inference that allows the closest possible relationship between feature diversity and a phylogenetic tree to be recovered because this correlation will be the strongest of all potential phylogenetic trees.

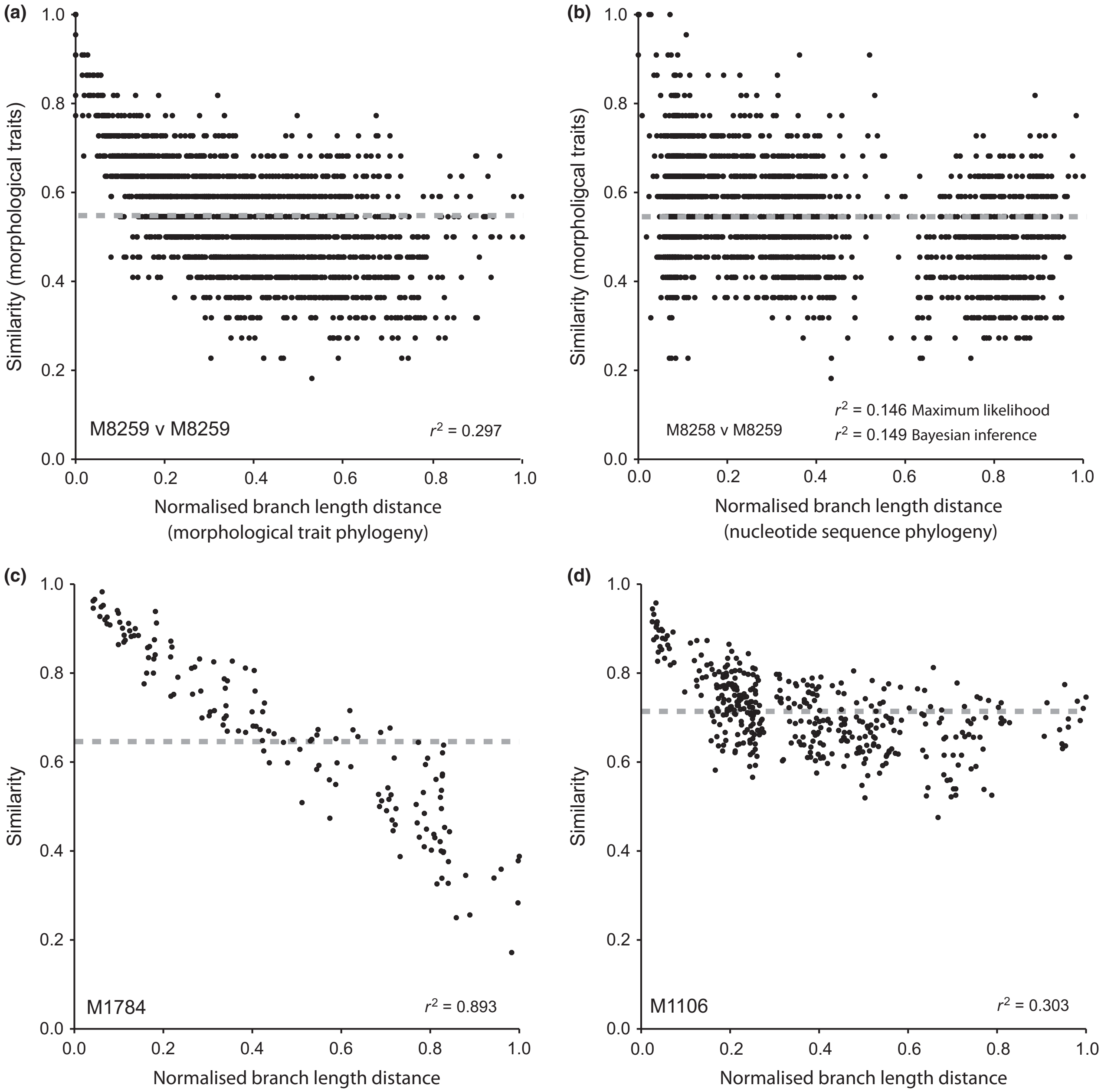

Figure 2 Examples of real data matrices from TreeBASE. (a) Relationship between character trait similarity and normalized branch length distance for a maximum-likelihood tree inferred directly from character trait matrix published by Cohen (2011) TreeBASE M8259. (b) Relationship between character trait similarity and normalized branch length distance for a maximum-likelihood tree inferred from a 10 chloroplast gene concatenated sequence matrix Cohen (2011). Also provided is the SSRCC from a Bayesian inference tree inferred from the same molecular dataset. (c) As in (a) but inferred using M1784 which exhibits a strong correlation between feature diversity and branch length distance. (d) As in (a) but inferred using M1106 which exhibits a weak relationship between feature diversity and branch length distance. The dashed grey line indicates the probability that two randomly selected taxa would share the same character state. In all cases, Spearman r^2 values describing the correlation between character trait similarity and normalized branch length distance are given.

Measures of phylogenetic distance and character dissimilarity

Given that our tree inference method optimizes the relationship between feature diversity and tree topology, we sought to determine whether under these ideal conditions that the branch lengths of the tree can be used to select the most divergent phenotypes. Branch length distances and measures of character similarity between taxa were calculated (from the phylogenetic trees and the data matrices, respectively) using custom Perl scripts available from the authors on request. Similarity in characters between taxa was calculated as the proportion of all characters exhibiting the same state in both taxa. Branch length distance was calculated as the shortest sum-total branch length distance between any two terminal taxa and was normalized to the largest pairwise branch length distance on that tree. Correlations between character similarity and normalized branch length distance measures were calculated for each data matrix using SSRCC to avoid the requirement for additional model assumptions. To determine whether our results were influenced by the hierarchical level of the terminal taxa for each matrix, we partitioned all matrices into one of three classes dependent on whether the terminal taxa comprised mainly species, genera or family/higher taxonomic units. We also determined whether there were statistically significant correlations between SSRCC and a number of different matrix characteristics. The characteristics comprised the number of characters, the number of character states, the number of autapomorphies and the number of taxa. We also wanted to check whether developments in data assembly methods over time influenced our findings. We therefore checked whether our results were influenced by publication date, so correlated publication date with SSRCC for each matrix.

Measures of phylogenetic diversity (PD)

Choices of taxa to maximize phylogenetic diversity (PD) were made using the PDA algorithm (Minh *et al.*, 2006).

Simulation of character matrices

Character matrices were simulated by adding increasing amounts of homoplasy to a fully congruent data matrix. For each simulated character matrix, the tree was inferred as described above. Homoplasy was added stochastically to the matrix in increments of 20% starting with a data matrix with all congruent characters states in which all character states evolved once with no homoplasy (Fig. 3a) and ending with a data matrix with all characters states having been randomly re-assigned (Fig. 3f).

Null expectations and randomizations

Where necessary, a null expectation of the probability of any two taxa sharing the same character state was obtained by randomly sampling 10,000 character states and taxon pairs within a data matrix. This null expectation is shown as a red and grey dashed line in Fig. 3 a–f and Fig. 2c,d, respectively.

To provide the null expectation distributions for Fig. 4, we randomly generated 1000 binary data matrices with character states

$$S = \{0, 1\}$$

Each matrix was assigned a randomly selected matrix population probability $p(M)$ where

$$p(M) = \{x \mid x \in R,\ 0 < x < 1\}$$

which defined the proportion of the matrix that is composed of the first element from set S. Thus, for each taxon within a

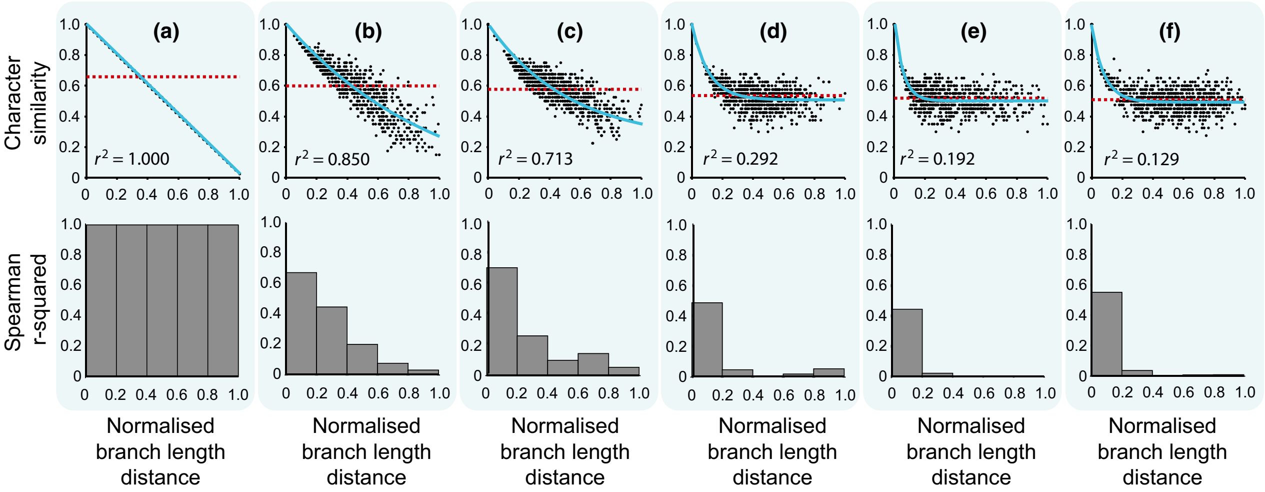

Figure 3 The strength of relationship between character trait similarity and normalized branch length distance for data matrices containing increasing amounts of homoplasy. (a) A fully congruent data matrix containing no homoplasy. (b) A data matrix containing 80% congruent character states and 20% stochastically simulated homoplastic character states. (c) 60% congruent and 40% homplastic character states. (d) 40% congruent and 60% homoplastic character states. (e) 20% congruent and 80% homoplastic character states. (f) Entirely stochastic matrix. In all cases, the blue line indicates the fitted exponential decay model, and the dashed red line indicates the probability that two randomly selected taxa would share the same character state. SSRCC values describing the correlation between character trait similarity and normalized branch length distance are given. Below each scatter plot, the binned SSRCC for increasing normalized tree distance are provided.

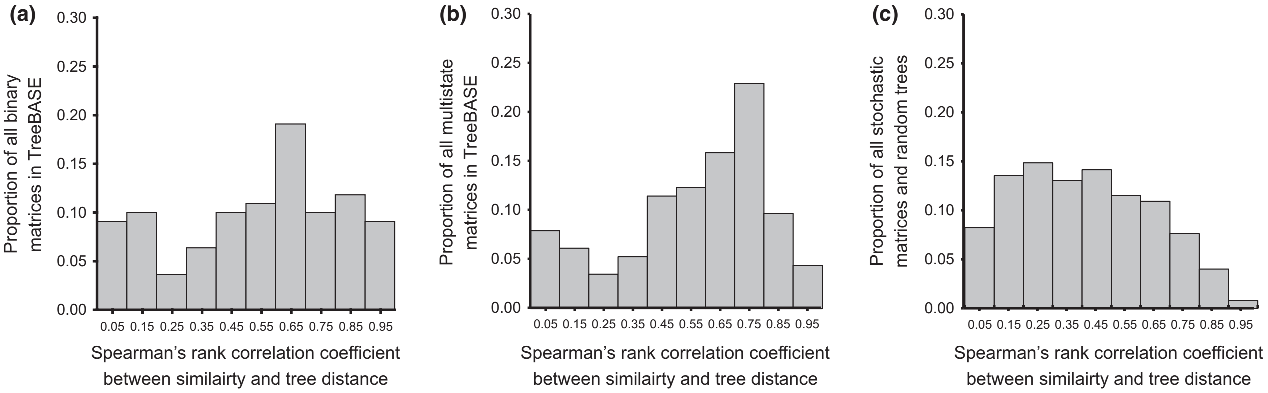

Figure 4 The distribution of Spearman's rank correlation coefficients for all data matrices in TreeBASE (a) All binary character trait matrices. (b) All multistate character trait matrices. (c) The null expectation distribution for stochastic matrices paired with random trees.

given data matrix, the probability of observing each character state was independent of both taxon and character within the matrix. For each randomly generated data matrix, an independent randomly generated tree topology was also produced. Thus, the topology of the tree and the data matrix are completely independent. Pairwise maximum-likelihood branch length distances between taxa were then calculated using RAxML v7.2.8 (Stamatakis, 2006). Here, the topology of the tree was constrained but again the maximum-likelihood branch length distances between taxa were estimated using a GTR model of character evolution, optimization of substitution rates and a Gamma model of rate heterogeneity with an estimated alpha parameter and 25 distinct rate categories. For each matrix, normalized branch length distances and measures of character similarity between taxa were calculated as previously described and compared via SSRCC. The distribution of correlation coefficients is provided in Fig. 4c and thus represents the null expectation that there is no relationship between phylogenetic trees and their underlying data matrices.

RESULTS

Figure 2a,b show an example of feature similarity amongst a set of organisms compared with phylogenetic distance over both a phylogenetic tree inferred directly from the features themselves (Fig. 2a) and a phylogenetic tree inferred from molecular sequence data using both Bayesian and maximum-likelihood methods (Fig. 2b).

Figure 2c,d show two examples of the real data matrices from TreeBASE (see Methods) in which our methodology has been applied. These were chosen to illustrate informative and typical examples of real biological data. Neither matrix represents an extremum, but instead resides near the lower and upper quartile boundaries of our dataset, respectively. In Fig. 2c, branch length distance is a good predictor of biological feature diversity over half of the length of the tree, whereas in Fig. 2d, branch length distance is a very poor indicator of feature diversity, except over very short distances, that is, amongst close relatives near the tips of the phylogenetic tree.

To determine the effect of homoplasy on the relationship between features and trees, we simulated a data matrix in which all features evolved uniquely (once) on the tree with no homoplasy (Fig. 3a) and stochastically added increasing levels of homoplasy (Fig. 3b–f). The proportion of homoplasious to non-homoplasious characters can also be interpreted as the distribution of rates where characters evolve under a Markov process. The results show that with increasing levels of homoplasy, the relationship between tree distance and biological similarity decays such that increased phylogenetic distance does not result in decreased biological similarity (Fig. 3). In all cases, even with entirely stochastic data (i.e. Fig. 3f), there is some relationship between character similarity and tree distance over very short distances in the phylogenetic tree (i.e. normalized branch length distances of less than 0.2, Fig. 3). However, only in the case of data with no homoplasy (Fig. 3a) do large distances on a phylogenetic tree correlate with large differences in character similarity. Moreover, for all levels of stochastically introduced homoplasy, normalized branch length distances exceeding 0.4 provide little or no information about the character similarity between taxa. Thus, selecting the taxa with the largest branch length distances in phylogenetic trees, unless the data are perfectly congruent, does not ensure that the taxa in question are phenotypically the most dissimilar.

Figure 4 shows the frequency distribution of SSRCC between feature similarity and tree distance for all of the 110 binary and 113 multistate datasets we sampled from TreeBASE. A broad spectrum of correlation values is observed, from little or no correlation (left-hand side of Fig. 4a,b; matrices similar to the example in Fig. 2d) to a very strong correlation (right-hand side of Fig. 4a,b; matrices similar to Fig. 2c). The null expectation distribution of global SSRCC, that is, where there is no relationship between tree and features is shown in Fig. 4c (also see methods). Comparison of real feature matrices with the null expectation reveals that published phylogenetic datasets contain more structured data than entirely stochastic matrices (Fig. 4). However, in all cases, the correlation between feature similarity and normalized branch length distances decreases with increasing normalized branch length distance (data not shown). This reveals that for all matrices, beyond some level of phylogenetic distance (usually a normalized branch length distance of ~0.4), there is no further decrease in feature similarity with increasing phylogenetic distance.

Lower global correlations are more common in the binary datasets than in the multistate datasets. This is probably due to the increased likelihood of observing homoplasy in binary characters for a given rate of state change (Donoghue & Ree, 2000). However, even for multistate characters, a global correlation of less than 0.5 is observed in 34% of all matrices (for illustrative purposes, a global correlation of 0.5 is midway between Fig. 3c,d). The mean SSRCC for all matrices is 0.57, and for species level matrices, it is 0.651, genus 0.567 and family and above 0.398 (Fig. S1). This is consistent with our finding that phylogenetic trees are more informative over shorter distances as species trees comprise shorter evolutionary distances than genus or family level trees. We found no significant correlation between SSRCC and number of taxa, number of characters, number of character states and number of autapomorphies (Table S1). We found no significant relationship between publication date of the matrix and the SSRCC for that matrix (Fig. S2), which we consider rules out technological or philosophical changes in feature identification as a potential source of bias.

Our results indicate that feature diversity is a function of tree distance over short distances on a phylogenetic tree, that is, between close relatives. Similarity does not decline with further increasing phylogenetic distance, and thus, a pair of taxa separated over a phylogeny by the greatest distance are unlikely to be the most dissimilar. To determine whether this was true for all of the matrices in this study, we selected the taxon pair that maximized phylogenetic distance and compared trait similarity across them. Figure 5 shows the frequency, across all 110 binary and 113 multistate data matrices, with which a choice of the most distant pair of taxa on each tree results in also choosing the pair of taxa with the lowest total character similarity (i.e. the most dissimilar). In approximately two-thirds of data matrices that we sampled, the choice of the most distant pair of taxa on the tree misses at least 50 pairs of taxa with more dissimilar character states. Moreover, maximizing phylogenetic distance selects one of the ten most dissimilar taxon pairs in only 17% of all matrices. These findings are independent of data type and are observed whether the analysis is restricted to restriction site, morphological data or indel and gene presence/absence matrices (data not shown).

DISCUSSION

The main finding from our analyses is that the disparity in discrete biological features does not reliably accumulate as a monotonic function of phylogenetic divergence. Phenotypic state is strongly correlated only for a relatively short distance on a phylogenetic tree (i.e. most character states are shared only amongst close relatives), and that in a large proportion of cases, this relationship rapidly decays so that the relationship between shared features and tree distance is often not distinguishable from a random sampling strategy across much of the tree. In other words, close relatives share more features than distant relatives but beyond a threshold, increasingly more distant relatives are not reliably more divergent in phenotype. Our results show that as homoplasy increases beyond low levels, the relationship between shared features and tree length breaks down such that the resultant diversity of character states may often not be distinguishable from randomly sampling characters from the data matrix of taxa independent of how they are related (Fig. 1b). The mistaken assumption has been to extrapolate from close relatives sharing features in common to distant relatives sharing less features in common, and that this relationship is monotonically conserved across a tree. In contrast, for many empirical data matrices, it seems that the extent of homoplasy precludes paths across much of the tree being correlated with feature diversity. Consequently, there are implications for those fields that routinely make such assumptions.

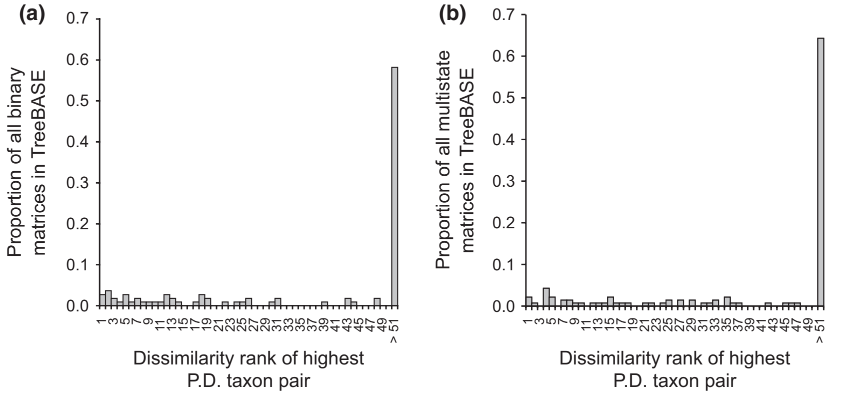

Figure 5 The distribution of feature diversity rank of the maximum PD pair of taxa. (a) All binary data matrices used in this study. (b) All multistate state data matrices examined in this study.

Our findings demonstrate that conservation strategies based on measures of phylogenetic diversity will quite often not achieve the aim of maximizing feature diversity (Fig. 5). In particular, choosing a small number of species from a larger set on the assumption that they will maximize phenotypic diversity is particularly problematic (Fig. 5). One of the seminal papers on PD (Faith, 1992a) discussed the pitfalls of increasing amounts of homoplasy in feature diversity for inferring a robust phylogeny, but considered 'molecular data, while not necessarily representing features of direct conservation interest, can provide phylogenies that are predictive, through PD calculations, of more general feature diversity patterns'. In this paper, we have tested this assumption, albeit indirectly, and shown that the most predictive phylogeny possible for general feature diversity for a substantial number of datasets decays quickly as to render PD calculations indistinguishable from a random sampling strategy across much of the tree. We consider that the significance of our study has been to ask the question, if feature diversity is problematic when it comes to inferring phylogeny – which is why time calibrated molecular phylogenies are generally used – then why would we expect a monotonic or even close relationship between feature diversity and phylogeny? We consider that the reason for this mistaken assumption stems from an over-extrapolation from the uncontroversial claims that close relatives are more similar and distant relatives less so (Fig. 1a), but not taking into account that this is not maintained across tree distance when homoplasy and a lack of unique derived characters are taken into account (Fig. 1b). Our contention is not that feature diversity evolves randomly, but solely that feature diversity across phylogenetic tree distance is not distinguishable from a random sampling strategy. The reason this is so, which we found surprising, is that very modest levels of homoplasy reduce the power of phylogeny to predict feature diversity better than random (Fig. 1a,b).

Our methodology, which involves inferring phylogeny directly from the features, means that our results are conservative and that feature diversity relative to molecular phylogenies must yield equal or less of a relationship between feature diversity and phylogeny. We therefore consider our results represent the best-case scenario for a correlation between feature diversity and phylogeny, whereas the actual level of correlation will be less than we report here. One potential way, in which the relationship between distance and disparity could be made weaker, would be for 'early burst' modes of feature evolution (Simpson, 1953) to be prevalent. In such modes, the majority of feature evolution happens very early on in the group's history, after which transition rates decline. This rate heterogeneity over time results in overall disparity being relatively low over very long distances, with only pairs of taxa joined via nodes quite close to root exhibiting the highest values. Evidence for early burst modes of feature evolution being common is not strong (Yoder *et al.*, 2010; Ingram *et al.*, 2012); however, early bursts are proposed to have occurred very deep in the tree of large clades such as mammals (e.g. Cooper & Purvis, 2010). From a conservation viewpoint, early bursts mean that sampling taxa across large phylogenetic distances would result in conservation of little feature diversity.

Although our results are at odds with the central assumption of phylogenetic diversity calculations, they are compatible with several widely accepted aspects of phylogenetic thinking including the following: homoplasy is prevalent for much of the tree of life (Lankester, 1870; Sanderson & Donoghue, 1996; Wake *et al.*, 2011; Tenaillon *et al.*, 2012); deep nodes in phylogenies are difficult to resolve with morphological data (Olmstead *et al.*, 2001; Bremer *et al.*, 2002; Wortley *et al.*, 2005); the relationship between feature diversity and hypotheses of common ancestry is not monotonic as otherwise phenetic and parsimony tree-building methods would be unproblematic (Lewis, 2001; Sober & Steel, 2002; Felsenstein, 2004).

There are two provisos we would make in interpreting our results. Firstly, our results are restricted to the phylogenetic breadth of taxa sampled in the publicly available data matrices we examined the majority of which were assembled for phylogenetic reconstruction. Secondly, the generality of our results depends on how well the characters in our datasets represent the universe of un-sampled phenotypic characters. We consider that if any filtering, conscious or unconscious, has occurred in the assembly of the original datasets, it would most likely be a favouring of characters with low homoplasy and high discriminant ability and would again bias against our primary finding that homoplasy is too pervasive for predictivity or a consistent distance–disparity relationship to be assumed.

The assumption that feature diversity is correlated or predicted by phylogeny, in part, stems from the widely held assumption that Linnaean classifications are predictive of features of organisms (Judd *et al.*, 1999; Goodman & Walsh, 2001; Cracraft, 2002; Simpson, 2006; Mishler, 2009). We wondered how can this be true at the same time as feature diversity having a poor correlation with phylogenetic distance. Classifications have three important ranks, species, genera and families and ranks above the family, for example, order and classes often suffer from a lack of diagnostic features due to a poor correlation between features and phylogeny (APG, 1998; Judd *et al.*, 1999; Simpson, 2006). Also, classifications are broad brush and lack resolution presumably because of a lack of suitable diagnostic features. Our results show that the mean SSRCC for all matrices is 0.57 with a small decrease in value from species to genus and family level matrices. Increasing taxonomic rank is therefore one variable but should be viewed in combination with the degree of phylogenetic resolution in determining the correlation between feature diversity and phylogeny.

CONCLUSIONS

There are very sound reasons for conservation to be based on, or at least incorporate phylogeny when choosing which species to prioritise. Conserving as much evolutionary history as possible is an important goal, and PD calculations on molecular phylogenies are one way this can be implemented objectively. However, our results demonstrate that conserving evolutionary history does not necessarily capture the most diverse set of biological features.

ACKNOWLEDGEMENTS

S.K. is a Leverhulme Trust Early Career Fellow. We thank Kevin Gaston, Mike Sanderson, Bob May and two anonymous referees for comments on the manuscript.

REFERENCES

APG. (1998) An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden*, **85**, 531–553.

Archie, J.W. (1984) A new look at the predictive value of numerical classifications. *Systematic Zoology*, **33**, 30–51.

Barker, G.M. (2002) Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society*, **76**, 165–194.

Bottjer, P.D. (1980) Farris information content and phylogenetic versus evolutionary classification – The philosophical differences remain. *Systematic Zoology*, **29**, 382–386.

Bremer, B., Bremer, K., Heidari, N., Erixon, P., Olmstead, R.G., Anderberg, A.A., Källersjö, M. & Barkhordarian, E. (2002) Phylogenetics of asterids based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. *Molecular Phylogenetics and Evolution*, **24**, 274–301.

Cadotte, M.W. & Davies, J. (2010) Rarest of the rare: advances in combining evolutionary distinctiveness and scarcity to inform conservation at biogeographical scales. *Diversity and Distributions*, **16**, 376–385.

Cohen, J.I. (2011) A phylogenetic analysis of morphological and molecular characters of *Lithospermum* L. (Boraginaceae) and related taxa: evolutionary relationships and character evolution. *Cladistics*, **27**, 559–580.

Collen, B., Turvey, S.T., Waterman, C., Meredith, H.M.R., Kuhn, T.S., Baillie, J.E.M. & Isaac, N.J.B. (2011) Investing in evolutionary history: implementing a phylogenetic approach for mammal conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2611–2622.

Colless, D.H. (1981) Predictivity and stability in classifications: some comments on recent studies. *Systematic Zoology*, **30**, 325–331.

Cooper, N. & Purvis, A. (2010) Body size evolution in mammals: complexity in tempo and mode. *The American Naturalist*, **175**, 727–738.

Correll, R.L. (1977) The application of maximal predictive classification to the Epacaridaceae. *Taxon*, **26**, 65–67.

Cracraft, J. (2002) The seven great questions of systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanical Garden*, **89**, 127–144.

Donoghue, M.J. & Ree, R.H. (2000) Homoplasy and developmental constraint: a model and an example from plants. *American Zoologist*, **40**, 759–769.

Eberhard, W.G., Agnarsson, L. & Levi, H.W. (2008) Web forms and the phylogeny of theridiid spiders (Araneae:Theridiidae): chaos from order. *Systematics and biodiversity*, **6**, 415–475.

Erwin, T.L. (1991) An evolutionary basis for conservation strategies. *Science*, **253**, 750–752.

Faith, D.P. (1992a) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.

Faith, D.P. (1992b) Systematics and conservation- on predicting the feature diversity of subsets of taxa. *Cladistics-the International Journal of the Willi Hennig Society*, **8**, 361–373.

Faith, D.P. (2013) Biodiversity and evolutionary history: useful extensions of the PD phylogenetic diversity assessment framework. *Annals of the New York Academy of Sciences*, **1289**, 69–89.

Farris, J.S. (1979) The information content of the phylogenetic system. *Systematic Zoology*, **28**, 483–519.

Farris, J.S. (1982) Simplicity and Informativeness in systematics and phylogeny. *Systematic Zoology*, **31**, 413–444.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–410.

Felsenstein, J. (1985) Phylogenies and the comparative method. *The American Naturalist*, **125**, 1–15.

Felsenstein, J. (1988) Phylogenies and quantitative characters. *Annual Review of Ecology & Systematics*, **19**, 445–471.

Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer Associates, Sunderland, MA.

Forest, F., Grenyer, R., Rouget, M., Davies, T.J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., Proches, S., van der Bank, M., Reeves, G., Hedderson, T.A.J. & Savolainen, V. (2007) Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature*, **445**, 757–760.

Goodman, G. & Walsh, V. (2001) *The story of taxol*. Cambridge University Press, Cambridge.

Gower, J.C. (1974) Maximal predictive classification. *Biometrics*, **30**, 643–654.

Graham, C.H., Parra, J.L., Tinoco, B.A., Stiles, F.G. & McGuire, J.A. (2012) Untangling the influence of ecological and evolutionary factors on trait variation across hummingbird assemblages. *Ecology*, **39**, S99–S111.

Harmon, L.J., Losos, J.B., Davies, T.J., Gillespie, R.G., Gittleman, J.L., Jennings, B.W., Kozak, K.H., McPeek, M.A., Moreno-Roark, F., Near, T.J., Purvis, A., Ricklefs, R.E., Schluter, D., Schulte, J.A. II, Seehausen, O., Sidlauskas, B.L., Torres-Carvajal, O. & Weir, J.T. (2010) Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, **64**, 2385–2396.

Harvey, P.H. & Pagel, M.D. (1991) *The comparative method in evolutionary biology*. Oxford University Press, Oxford.

Hennig, W. (1966) *Phylogenetic Systematics*. University of Ilinois Press, Urbana.

Hillis, D.M. (1987) Molecular versus morphological approaches to systematics. *Annual Review of Ecology and Systematics*, **18**, 23–42.

Hillis, D.M. (1995) Approaches for assessing phylogenetic accuracy. *Systematic Biology*, **44**, 3–16.

Hillis, D.M., Huelsenbeck, J.P. & Cunningham, C.W. (1994) Application and accuracy of molecular phylogenies. *Science*, **264**, 671–677.

Huelsenbeck, J.P. & Crandall, K.A. (1997) Maximum likelihood in phylogenetics. *Annual Review of Ecology and Systematics*, **28**, 437–466.

Huelsenbeck, J.P., Bull, J.J. & Cunningham, C.W. (1996) Combining data in phylogenteic analyses. *Trends in Ecology and Evolution*, **11**, 152–158.

Ingram, T., Harmon, L.J. & Shurin, J.B. (2012) When should we expect early bursts of trait evolution in comparative data? Predictions from an evolutionary food web model. *Journal of Evolutionary Biology*, **25**, 1902–1910.

Isaac, N.J., Turvey, S.T., Collen, B., Waterman, C. & Baillie, J.E. (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE*, **2**, e296.

Jansky, S.H., Simon, R. & Spooner, D.M. (2006) A test of taxonomic predictivity: resistance to white mold in wild relatives of cultivated potato. *Crop Science*, **46**, 2561–2570.

Judd, W.S., Campbell, C.S., Kellog, E.A. & Stevens, P.F. (1999) *Plant Systematics: a phylogenetic approach*. Sinauer, Sunderland, MA.

Kluge, A.G. & Wolfe, A.J. (1993) Cladistics: whats in a word? *Cladistics*, **9**, 183–200.

Kraft, N.J.B., Cornwell, W.K., Webb, C.O. & Ackerly, D.D. (2007) Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *The American Naturalist*, **170**, 271–283.

Krajewski, C. (1994) Phylogenetic measures of biodiversity: a comparison and critique. *Biological Conservation*, **69**, 33–39.

Lankester, E.R. (1870) On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Annals and Magazine of Natural History*, **6**, 34–43.

Lewis, P.O. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**, 913–925.

Losos, J.B. (2008) Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters*, **11**, 995–1003.

Maddison, D.R. & Schulz, K.S. (2007) *The tree of life web project*. Available at: http://tolweb.org (accessed 1 July 2011).

May, R.M. (1990) Taxonomy is destiny. *Nature*, **347**, 129–130.

Minh, B.Q., Klaere, S. & Haeseler, A.v., (2006) Phylogenetic diversity within seconds. *Systematic Biology*, **55**, 769–773.

Mishler, B.D. (2009) Three centuries of paradigm changes in biological classification: is the end in sight? *Taxon*, **58**, 61–67.

Mouquet, N., Devictor, V., Meynard, C.N. *et al.* (2012) Ecophylogenetics: advances and perspectives. *Biological Reviews*, **87**, 769–785.

Nixon, K.C. & Carpenter, J.M. (1996) On simultaneous analysis. *Cladistics*, **12**, 221–241.

Olmstead, R.G., dePamphilis, C.W., Wolfe, A.D., Young, N.D., Elisons, W.J. & Reeves, P.A. (2001) Disintegration of the Scrophulariaceae. *American Journal of Botany*, **88**, 348–361.

Pagel, M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences*, **255**, 37–45.

Pagel, M. & Meade, A. (2006) Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*, **167**, 808–825.

Patterson, C., Williams, D.M. & Humpries, C.J. (1993) Congruence between molecular and morphological phylogenies. *Annual Review of Ecology and Systematics*, **24**, 153–188.

Piel, W.H., Auman, J., Chan, L., Dominus, M.J., Grapeyev, V., Gujral, M., Guo, Y., Lapp, H., Ruan, J., Shyket, H., Vos, R.A. & Tannen, V.. (2010) *A database of phylogenetic knowledge*. Available at: http://treebase.org (accessed 1 July 2011).

Posada, D. & Crandall, K.A. (1998) MODELTEST: testing the model of dna substitution. *Bioinformatics*, **14**, 817–818.

Rodrigues, A.S.L., Brooks, T.M. & Gaston, K.J. (2005) Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference? *Phylogeny and conservation* (ed. by A. Purvis, J.L. Gittleman and T.M. Brooks), pp. 101–119. Cambridge University Press, Cambridge.

Rodrigues, A.S.L., Grenyer, R., Baillie, J.E.M., Bininda-Emonds, O.R.P., Gittlemann, J.L., Hoffman, N.M., Safi, K., Schipper, J., Stuart, S.N. & Brooks, T. (2011) Complete, accurate, mammalian phylogenies aid conservation planning, but not much. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2652–2660.

Ronquist, F. & Huelsenbeck, J.P. (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–4.

Ronsted, N., Symonds, M.R., Birkholm, T., Christensen, S., Meerow, A., Molander, M., Molgaard, P., Petersen, G., Rasmussen, N., van Staden, J., Stafford, G. & Jager, A. (2012) Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of amaryllidaceae. *BMC Evolutionary Biology*, **12**, 182.

Sanderson, M.J. & Donoghue, M.J. (1996) The relationship between homoplasy and confidence in a phylogenetic tree. *Homoplasy: the recurrence of similarity in evolution* (ed. by M.J. Sanderson and L. Hufford), pp. 67–89. Academic Press, San Diego.

Saslis-Lagoudakis, C.H., Klitgaard, B.B., Forest, F., Francis, L., Savolainen, V., Williamson, E.M. & Hawkins, J.A. (2011) The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: an example from *Pterocarpus* (Leguminosae). *PLoS ONE*, **6**, e22275.

Saslis-Lagoudakis, C.H., Savolainen, V., Williamson, E.M., Forest, F., Wagstaff, S.J., Baral, S.R., Watson, M.F., Pendry, C.A. & Hawkins, J.A. (2012) Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences USA*, **109**, 15835–15840.

Scotland, R.W., Olmstead, R.G. & Bennett, J.R. (2003) Phylogeny reconstruction: the role of morphology. *Systematic Biology*, **52**(4), 539–548.

Simpson, G.G. (1953) *The major features of evolution*. Columbia University Press, New York.

Simpson, M.G. (2006) *Plant systematics*. Elsevier, Burlington, MA.

Smith, W.L. & Wheeler, W.C. (2006) Venom evolution widespread in fishes: a phylogenetic road map for the bioprospecting of piscinevenoms. *Journal of Heredity*, **97**, 206–217.

Sneath, P.H.A. (1989) Predictivity in taxonomy and the probability of a tree. *Plant Systematics and Evolution*, **167**, 43–57.

Sneath, P.H.A. & Hansell, R.I.C. (1985) Naturalness and predictivity of classifications. *Biological Journal of the Linnean Society*, **24**, 217–231.

Sober, E. & Steel, M. (2002) Testing the hypothesis of common ancestry. *Journal of Theoretical Biology*, **218**, 395–408.

Soutullo, A., Dodsworth, S., Heard, S.B. & Mooers, A.O. (2005) Distribution and correlates of carnivore phylogenetic diversity across the Americas. *Animal Conservation*, **8**, 249–258.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Systematics Agenda 2000. (1994) Charting the biosphere. New York: Society of Systematic Biologists, American Society of Plant Taxonomists, Willi Hennig Society, Association of Systematics Collections.

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R.L., McDonald, P., Bennett, A.F., Long, A.D. & Gaut, B.S. (2012) *The molecular diversity of adaptive convergence Science*, **335**, 457–461.

Vander Velde, D.G., Georg, G.I., Gollapudi, S.R., Jampani, H.B., Liang, X.Z., Mitscher, L.A. & Ye, Q.M. (1994) Wallifoliol, a taxol congener with a novel carbon skeleton, from Himalayan *Taxus wallichiana*. *Journal of Natural Products*, **57**, 862–7.

Vane-Wright, R.I., Humphries, C.J. & Williams, P.H. (1991) What to protect?- Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.

Wagner, W.H. (1961) Problems in the classification of ferns. *Recent Advances in Botany*, **1**, 841–844.

Wake, D.B., Wake, M.H. & Specht, C.D. (2011) Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science*, **331**, 1032–1035.

Webb, C.O. (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, **156**, 145–155.

Wiens, J.J. (2004) The role of morphological data in phylogeny reconstruction. *Systematic Biology*, **53**, 653–661.

Winter, M., Devictor, V. & Schweiger, O. (2013) Phylogenetic diversity and nature conservation: where are we? *Trends in Ecology & Evolution*, **28**, 199–204.

Wortley, A.H., Rudall, P.J., Harris, D.J. & Scotland, R.W. (2005) How much data are needed to resolve a difficult phylogeny? *Systematic Biology*, **54**, 697–709.

Yoder, J.B., Clancey, E., Des Roches, S., Eastman, J.M., Gentry, L., Godsoe, W., Hagey, T.J., Jochimsen, D., Oswald, B.P., Robertson, J., Sarver, B.A.J., Schenks, J.J., Spear, S.F. & Harmon, L.J. (2010) Ecological opportunity and the origin of adaptive radiations. *Journal of Evolutionary Biology*, **23**, 1581–1596.

SUPPORTING INFORMATION

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

Appendix S1 Matrix identifiers and references for all binary matrices are available in (Appendix S1).

Appendix S2 Matrix identifiers and references for all multistate matrices are available in Appendix S2.

Table S1 Correlation and p values for Spearman's Rho versus matrix characteristics.

Text S1 Direct quotes from literature exploring feature diversity and phylogenetic distance.

Figures S1 The observed SSRCC for all 113 multistate matrices.

Figure S2 The relationship between publication date and SSRCC between tree distance and similarity for all matrices used in this study.

BIOSKETCHES

Steve Kelly is a Leverhulme Trust Early Career Fellow in the Department of Plant Sciences at the University of Oxford, and his research interests span a broad spectrum of scientific disciplines from the early evolutionary events of life on earth to developing novel mathematical and computational methods for elucidating complex systems of gene regulation. The unifying feature of much of his work is the interplay between mathematical tools and wet-bench science, and how each can be used to inform and direct the other.

Richard Grenyer is a biologist and conservationist who is interested in the fundamental position that space and geographical processes occupy in biodiversity science and modern conservation strategy. He was appointed to a University Lectureship in Biodiversity and Biogeography at the School of Geography and the Environment at the University of Oxford in August 2010. He is also Fellow and Tutor in Physical Geography at Jesus College.

Robert Scotland is a systematic botanist in the Department of Plant Sciences at the University of Oxford with interests in systematic theory, taxonomy and classification, species discovery and homology. He is currently President of the Systematics Association.

Author contributions: RWS had the initial insight, SK did the analyses and RWS, SK and RG designed the approach taken and contributed equally to writing the manuscript.

Editor: Jeremy Austin