The impact of poor sampling of polymorphism on cladistic analysis

Akinobu Watanabe\textsuperscript{a,b,*}

\textsuperscript{a}Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA; \textsuperscript{b}Richard Gilder Graduate School, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

Accepted 15 June 2015

Abstract

Despite its ubiquity in the natural world, polymorphism is commonly disregarded or poorly sampled in phylogenetic analyses due to deliberate sampling strategy, inadequate sampling effort and limited specimen availability. Poor sampling of intraspecific variation engenders differential sampling of morphs within polymorphic species, which could generate conflicting tree topologies by altering the character-based affinity among taxa. To assess the potential magnitude of this impact, Polymorphic Entry Replacement Data Analysis (PERDA) was developed as a new script for the TNT phylogenetic program. This script simulates poor sampling of polymorphic taxa on a matrix of discrete characters by iteratively replacing each polymorphic state (e.g. $[01]$) with a randomly selected single state included in the original polymorphic coding (e.g. 0 or 1). The trees recovered from these subsampled data sets provide a distribution of tree distances, which indicates the level of incongruent trees resulting from different combinations of single states. Performing PERDA on empirical data sets shows alarming frequencies and magnitudes of conflicting tree topologies, demonstrating that poor sampling within polymorphic taxa could yield highly incompatible trees in many data sets. This troubling outcome undermines phylogenetic inferences based on data with poor intraspecific sampling, which is typical for palaeontological studies. With trees obtained from subsampled data sets, PERDA also generates a metaconsensus tree revealing interspecific relationships that become ambiguous due to documented levels of intraspecific variation. These collapsed clades point to taxa for which evidence should be sought to justify their taxonomic classification.

Polymorphism generally denotes the stable existence of two or more distinct, heritable forms within populations of a species, exclusive of intraspecific variation due to sexual dimorphism, ontogeny, geographical location and plasticity (Ford, 1945; Wiens, 1999). In cladistics, the treatment of polymorphism remains a contentious issue without resolution (e.g. Archie et al., 1989; Nixon and Davis, 1991; Campbell and Frost, 1993; Wiens, 1995, 1998, 2000; Nixon, 1996; Kornet and Turner, 1999; Simmons and Geisler, 2002). Previous studies (Kornet and Turner, 1999; Wiens, 1999) summarized several different approaches including one that replaces every polymorphic state with “?” (missing coding) and frequency coding, in which character states are defined by the relative abundance of morphs in each species (e.g. state “a” if the relative abundance of a morph is 0–10%, state “b” if it is 11–20%). Despite higher degrees of homoplasy compared to fixed characters, both empirical and simulation studies show that the inclusion of polymorphic characters recovers greater proportions of established and true clades than removing them entirely from cladistic analyses (Wiens and Servedio, 1997, 1998; Wiens, 1998). These same studies also found that frequency coding generally performs better than other approaches. Nevertheless, polymorphic coding, which uses multistates (e.g. [01]), continues to be the most commonly employed method for characterizing polymorphisms for discrete characters partly because frequency coding requires a robust intraspecific sampling that is often impractical for many taxonomic groups.

By necessity, the detection of polymorphisms requires at least two individuals or a heterozygous individual for genetic polymorphisms. Yet, phylogenetic studies often sample only single individuals for each species due to multiple factors. First, a robust sampling within species...
generally requires prohibitive amounts of time and expense. Second, specimens may be restricted in availability and access, such as those of critically endangered species. Particularly in vertebrate palaeontology, depauperate sampling is a pervasive problem, in which a single specimen or few fragmentary fossils represent an extinct species. Due to these factors, many phylogenetic studies have largely ignored or failed to recognize polymorphisms in their taxonomic sampling.

With polymorphic taxa, this situation is troubling because poor intraspecific sampling could hinder the ability to reliably infer the phylogenetic relationships of organisms. Given polymorphic coding, phylogenetic reconstructions are expected to be sensitive to poor sampling because the character-based affinities among terminals depend on which morph of polymorphic taxa are sampled. For example, a polymorphic species comprising individuals with states “0” and “1” (i.e. [01]) will either share a state with taxa that bear state “0” or “1” when only a single individual is sampled, potentially influencing the resulting tree topology (Fig. 1). Nixon and Davis (1991), as well as Simmons (2001), demonstrated that polymorphic coding could lead to clades that do not occur in the most parsimonious trees (MPTs) constructed from analysing monomorphic units as terminals. In these situations, a poor sampling of polymorphisms could create false synapomorphies, which may spuriously alter the phylogenetic relationships among not only the polymorphic species, but also those of other taxa in the data.

To evaluate the extent of these effects across multiple data sets, I developed the Polymorphic Entry Replacement Data Analysis (PERDA), a script implemented in the TNT phylogenetic program (Goloboff et al., 2008). Given a data set that includes multistates, the analysis simulates poor intraspecific sampling by selecting single morphs of each polymorphic species in a character matrix. Here, I demonstrate the utility of PERDA for investigating (i) the magnitude of conflicting tree topologies that result from differential sampling of morphs, and (ii) additional ambiguities in phylogenetic relationships associated with poor sampling within polymorphic species. For systematists, the latter objective presents a unique strategy for identifying taxa that require further taxonomic evaluation.

Materials and methods

PERDA

The script (Appendix 1, Data S1) performs the following steps using functions in TNT version 1.1 (Goloboff et al., 2008; Fig. 1). First, it reads in a matrix of discrete characters, in which polymorphisms are encoded as multistates (e.g. “[01]”). It then conducts a parsimony analysis on this unmodified matrix and saves the single MPT or the strict consensus of MPTs as the reference tree. In this study, a traditional search strategy was employed with 1000 replications, holding 10 trees per replication, followed by an additional round of tree bisection and regrafting (TBR) branch swapping on trees in memory from the initial search. Branches were collapsed if they lacked support in any of the MPTs, i.e. minimum length = 0 (”rule 1” of Coddington and Scharff, 1994). TNT does not incur additional steps for changes from an unambiguous ancestral state to a polymorphic set if it contains the ancestral state (i.e. [0,1] is treated as 0 or 1). The program does add a step, or possibly multiple steps for ordered characters, if the polymorphic set does not include the ancestral state. Whether various cost assignments for polymorphic coding influence the resulting trees is outside the scope of this study and previous studies have examined this topic (e.g. Campbell and Frost, 1993; Simmons and Geisler, 2002).

After the consensus tree from an unmodified matrix is recorded, PERDA simulates sampling a single morph of each polymorphic species. It scans the character matrix, and each time a polymorphic state is encountered, the script substitutes the cell with a single state selected equiprobably from the states included in the polymorphic coding. Hence, for a
two-state polymorphism (e.g. “[01]”), it has a 50% probability of choosing each state (e.g. “0” or “1”). PERDA retains all missing data (i.e. “?”) in the data set. Once it replaces all polymorphic states in the matrix with randomly chosen single states, the script conducts a parsimony analysis on the modified matrix with the same search strategy as described above. The strict consensus tree is then constructed based on the resulting MPTs and subsequently recorded in a tree file (“perda_trees.tnttre”). This study did not consider the implications of selecting morphs with probability concomitant with their frequencies in populations due to the lack of such data for most phylogenetic data sets.

To explore different combinations of single morph sampling, the replacement of polymorphic states in the original character matrix to single states is repeated a specified number of times. For this study, I performed 10,000 iterations for each data set, generating 10,000 matrices with variable combinations of monomorphic sampling and 10,000 strict consensus trees from analysing these modified matrices. The number of polymorphic characters determines the minimum number of iterations required to evaluate every combination of single state substitutions. For polymorphisms comprising two states, the number of different combinations is $2^n$, in which $n$ is the number of multistates in the character matrix. In PERDA, the combinations of morphs are determined randomly, not exhaustively, because a data matrix with 20 binary polymorphic characters, for instance, would require over million rounds ($2^{20} = 1,048,576$) of separate parsimony analyses. As such, 10,000 iterations were performed on all data sets for consistency and sufficient exploration of parameter space as assessed by observed stability in the results.

When all iterations are completed, PERDA proceeds to compute the dissimilarity between each consensus tree from the modified matrices and the reference tree constructed from the original character matrix with polymorphic states. To clarify, the analysis does not assume that the reference tree is the correct or the preferred phylogeny for the sampled species. It uses the reference trees to reveal ambiguities in phylogenetic inference associated with differential sampling of morphs beyond the ambiguity from polymorphic coding. Here, tree dissimilarity was measured based on the commonly used Robinson–Foulds (RF) distance, which is equal to the proportion of bipartitions defined by a branch in one tree that is lacking in another (Bourque, 1978; Robinson and Foulds, 1981). RF distances of 0 and 1 signify perfectly congruent and incongruent tree topologies, respectively. However, RF distances count compatible, as well as conflicting, groups between two trees. This calculation would therefore inflate the degree of dissimilarity when the resolution of one tree is substantially lower than the other. Because this study concerns conflicting tree topologies generated from differential sampling of morphs, the calculation of tree distances was modified to account for only incompatible groups.

After comparing all trees from subsampled data with the reference tree, the script outputs a list of tree distances (“perda_treedist.txt”), which was then used to create a histogram in R statistical software (R Development Core Team, 2012). The histogram visualizes the distribution of different trees that results from differential sampling of single morphs. If trees from subsampled data sets are mainly congruent with the reference tree (tree distance ~ 0), then the inferred phylogenetic relationships are robust against poor sampling within polymorphic species. Alternatively, if trees disparate from the reference tree (tree distance $\gg 0$) are abundant, then this study would undermine the reliability of phylogenetic trees reconstructed from poor sampling within polymorphic species, particularly palaeontological data sets.

The histograms provide information beyond previous studies that varied the sample size within polymorphic species and recorded the “accuracy” of phylogenetic inference. These studies define accuracy as the average proportion of nodes in the true tree (for simulation studies) or well-established tree (for empirical studies) that are retained in the trees obtained from subsampled data (Archie et al., 1989; Wiens and Servedio, 1997, 1998). However, this metric does not measure the distribution of deviations from a reference tree. Thus, a moderate value for accuracy could be obtained from similar frequencies of highly variable tree topologies or high frequencies of both perfectly congruent and incongruent topologies. In this regard, histograms provide a more nuanced view of the magnitude and abundance of different trees generated from differential sampling within polymorphic species. Furthermore, as noted by Simmons (2001), the “accuracy” metric used in these previous studies does not differentiate between incongruent groups from differences in tree resolution.

Besides tree distances, PERDA also constructs a metaconsensus tree from the set of consensus trees recovered from single morph sampling (“perda_metatree.tnttre”). A comparison between the metaconsensus tree and the reference tree (i.e. consensus tree of MPTs recovered from the unmodified character matrix) reveals additional ambiguities in phylogenetic relationships that could potentially result from poor sampling within polymorphic taxa. Additional branches that collapse under differential sampling of morphs indicate that a morph with a particular combination of single states is considered most closely related to a different species than analysing with other morphs of the species.
Data

Analyses were conducted on published character matrices. One is a morphological data set of the anguid lizard genus *Abronia* (Campbell and Frost, 1993) with 23 terminals in the ingroup, 30 characters and 73 instances of polymorphism (~10% of the total number of cells in the ingroup). As in the original study, a hypothetical ancestor was used to root the tree. The authors of the study note that the recorded intraspecific variation reflects polymorphisms and excludes variation due to sexual dimorphism, teratology or environmental effects (Campbell and Frost, 1993, pp. 60–61). These polymorphisms, coded as a separate state in each character (e.g. [01] as state “8”), were converted to multistates for this study.

PERDA was also conducted on a largely morphological data set of the hylidae frog genus *Scinax*, which consists of osteological, myological, larval, reproductive and karyotypic characters (Faiivovich, 2002). The matrix includes 36 species of *Scinax* and seven additional species in three genera (*Aplastodiscus, Hyla, Smilisca*) as outgroup although only one (*H. cinerea*) was used to root the trees. Of the 86 characters, 27 include polymorphic states (0.75% of the total number of cells excluding *H. cinerea*). Because all but three behavioural characters are morphological, this data set was characterized here as morphological data.

Both the *Abronia* and the *Scinax* data sets considered polymorphisms in their respective system by sampling several individuals for most species. The studies sampled, on average, 11.3 and 12.6 individuals for the *Abronia* and *Scinax* data sets, respectively, which exceed the numbers suggested for capturing the intraspecific variation of even continuous characters (Wiens and Servedio, 1997; Tixier, 2012). As such, these data sets were considered to adequately account for polymorphisms that exist in nature and thus allow PERDA to simulate scenarios in which single morphs are sampled within each polymorphic species. When PERDA is performed on data sets with a robust intraspecific sampling, it examines if and to what extent incompatible tree topologies originate from single morph sampling, given that the species classification and the documented level of polymorphism are accurate.

In addition, elongation factor 1-alpha 1 (EF1A1) sequence data of 26 *Polistes* wasps (Santos et al., 2014) were analysed. This matrix contains IUPAC ambiguity codes for nucleotide bases (Cornish-Bowden, 1985). These ambiguities originate from mixed signals during sequencing, which commonly originate from the presence of multiple gene copies in the genome, suboptimal amplification reactions or heterozygous single nucleotide polymorphisms. Although only a single individual was sequenced for each species, PERDA was conducted on this molecular data set to examine the extent to which ambiguity associated with intraspecific genetic variation, combined with genuine error from a typical sequencing procedure (i.e. Sanger sequencing on an ABI machine), could alter the topology of resulting phylogenetic trees in a cladistic framework.

Furthermore, an analysis was performed on the Theropod Working Group (TWiG) matrix (Turner et al., 2012) to illustrate the extended utility of PERDA for palaeontological data given their deficient record of intraspecific variation and the taxonomic uncertainty associated with the fossil record. The data set consists of 110 species of coelurosaurian dinosaurs and 475 characters, of which 0.29% (154) of 55 809 cells in the ingroup taxa are polymorphic. With palaeontological data, species-level classification of fossil specimens is inherently uncertain due to the confounding and compounding factors of taphonomic processes. As such, polymorphic states in extinct taxa are more likely to be misinterpreted as occurrences of multiple species (“lumping”) and, conversely, variable forms considered as multiple species are more likely to be multiple forms of a single species (“oversplitting”). Furthermore, polymorphic states could represent intraspecific variation due to ontogeny or sexual dimorphism. While the histograms from PERDA on palaeontological data are useful for simulating even poorer intraspecific sampling of fossil taxa than the fossil record affords, the metaconsensus tree incorporates these taxonomic uncertainties, facilitating a more conservative practice for systematic palaeontology.

To seek key factors that contribute to the susceptibility of discrete character data to differential sampling of morphs, PERDA was performed on additional published data sets with variable levels of polymorphism. Selected matrices include morphological data of nemesiid spiders (Goloboff, 1995), *Solenopsis* fire ants (Pitts et al., 2005), crocodylomorphs (Turner and Sertich, 2010) and early dinocephalians (Peecook et al., 2013). Results from these data were used to investigate whether the abundance of polymorphic states, number of terminals and number of characters are concomitant with the impact of single morph sampling on phylogenetic inference.

Results

Tree distribution

Collectively, the histograms of tree distances from PERDA illuminate three general consequences of single morph sampling (Fig. 2; tree distances recorded in Table S1). First, for all four data sets, single morph sampling often produces consensus trees with groupings that are incompatible with respective reference trees (i.e. tree distance > 0). The proportion of per-
fectly compatible trees (i.e. tree distance = 0), however, varies widely among data sets. With the Polistes EF1A1 data, for instance, 13.3% of consensus trees from PERDA fully agree with the topology of the reference tree. The TWiG coelurosaur data yield trees of which 43.2% are completely compatible with the reference tree. Therefore, over half of the trees from single morph sampling contain conflicting groupings, despite a matrix with only 0.29% of cells that are polymorphic. Similarly, 54.0% of consensus trees generated by the Abronia data are fully compatible with the reference tree and thus, nearly half of the trees contain conflicting topologies. Although the Scinax frog data recovered a large proportion of perfectly compatible trees (75.1%), one-quarter of the trees from PERDA contain incompatible clades. These results demonstrate the vulnerability of phylogenetic inference to poor sampling of polymorphic taxa.

Second, the degree of incompatible topologies is variable. Although the grand mean of all 10000 tree distances is near zero for most data sets, the mean distance of incongruent trees (i.e. tree distance > 0) exceeds 0.1 for several data sets. The mean distance for the Abronia data, for example, is above 0.15, indicating that on average over 15% of the groupings in trees with any conflicting clades (46.0% of the trees) are incompatible with the reference tree. With 3.4% (979) of codings that are polymorphic, the Polistes wasp data yield mean distances of 0.144 and 0.166 for all trees and only incompatible trees, respectively. Meanwhile, both the Scinax and the TWiG data generate mean tree distances near zero, implying that differential morph sampling led to proportionately fewer conflicting topologies for these two data sets compared to the Abronia and Polistes data.

Third, the histograms show large spreads in tree distances, in which values deviate greatly from zero. Among the four data sets analysed, the Abronia lizard data produced the largest range, with maximum tree distance of 0.56, indicating that up to half of the

![Fig. 2. Histograms of tree distances between the reference tree and consensus trees recovered from PERDA on: (a) Abronia lizard morphological data (Campbell and Frost, 1993); (b) Scinax frog data (Faivovich, 2002); (c) Polistes EF1A1 data (Santos et al., 2014); and (d) Theropod Working Group (TWiG) morphological data (Turner et al., 2012). Tree distance is defined as the proportion of incompatible groupings between two phylogenetic trees. $Mean_{total}$ and $Mean_{dist>0}$ are calculated from all strict consensus trees and a subset comprising all incompatible trees (tree distance > 0), respectively.](image-url)
Properties of data sets

groupings conflict with those of the reference tree with
certain combinations of single morphs. Despite the
greatest proportion of perfectly compatible trees,
PERDA on the Scinax frog data set generated trees
with up to 23.0% incompatible clades. The maximum
tree distance for the Polistes and TWiG data is 0.41
and 0.242, respectively. While the distribution of non-
zero tree distances from the Abronia and TWiG data
resembles an exponential probability distribution, the
Scinax and Polistes data generate more irregular pat-
terns in their distributions.

Tree resolution

The strict metaconsensus trees from PERDA exhibit
lower resolutions in comparison with their respective
reference trees. With the Abronia lizard data, only the
sister group relationship between A. bogerti and
A. chiszar remains after rounds of single morph sam-
ping (Fig. 3a). Although several relationships are
retained, the metaconsensus tree of the Scinax frog
data shows collapsed branches across the tree at vari-
ous hierarchical levels when compared to the reference
tree (Fig. 3b). The number of internal nodes in the ref-
cence tree is 29, whereas the number decreases almost
by half to 16 nodes in the metaconsensus tree. The
metaconsensus tree of the Polistes wasp data retains
the phylogenetic resolution in the strict consensus tree
with the exception of (i) the sister group relationship
of P. geminatus to the clade of P. major, P. marginalis
and P. cinerascens; (ii) clade of P. bicolor, P. sagitar-
ius, P. nimpha and P. dominula; and (iii) another clade
comprising P. metricus, P. carolina, P. dorsalis and
P. perplexus (Fig. 4).

Analogous to the Polistes data, the metaconsensus
tree from the TWiG data set shows a mosaic of col-
lapsed and persistent clades (Fig. 5). Major coeluro-
saurian clades collapse, including Compsognathidae,
Maniraptoriformes and Paraves. The disintegration of
these clades forms a notable 40-tomy that defines
Maniraporta, but with the exclusion of Ornitholestes,
considered a member of Maniraporta in the reference
tree. Meanwhile, several key groups maintain much of
the phylogenetic structure, such as Tyrannosauroida,
Alvarezsauroidea (with the exclusion of Haplocheirus),
Therizinosauria, Jinfengopterygidae and Avialae. The
clades Ornithomimosauria, Oviraptorosaura and
Pygostylia are also retained, but the phylogenetic rela-
tionships within these groups collapse into major poly-
tomies, thus losing much of the resolution observed in
the reference tree.

Properties of data sets

With outputs from four additional data sets, the
results point to properties of character matrices that
are concomitant with their vulnerability to differential
sampling of morphs (Table 1; Fig. 6). Based on the
point distribution in bivariate plots (Fig. 6), the num-
ber and proportion of polymorphic states in data sets
appear to be associated logarithmically or asymptoti-
cally with the maximum tree distance from PERDA,
along with the mean tree distances. Similarly, the
number of polymorphic states seems to covary negatively
with the percentage of fully compatible trees. Least
squares regression analyses indicate three notable rela-
tionships: (i) a positive correlation between maximum
tree distance and both raw and log-transformed per-
centage of polymorphic states ($P = 0.030$ and 0.003,
respectively), (ii) a positive correlation between grand
mean of tree distance and log-transformed number of
polymorphic states ($P = 0.097$), and (iii) a negative
correlation between the percentage of perfectly com-
patible trees and both raw and log-transformed num-
ber of polymorphic states ($P = 0.097$ and 0.055,
respectively).

Although these correlations may be genuine, addi-
tional regression analyses reveal that their significance
depends on the inclusion of the results from the Polis-

tes sequence data set, which produces large tree dis-
tances and a very small percentage of fully compatible
trees relative to morphological data. The only relation-
ship that remains significant ($P < 0.05$) with the exclud-
ion of Polistes data is the positive correlation between
log-transformed percentage of polymorphic states and
maximum tree distance ($P = 0.008$). With only mor-
phological data, the raw and log-transformed number of
terminals correlates negatively with the percentage of
perfectly compatible trees ($P = 0.058$ and 0.066,
respectively). Moreover, regression analysis weakly
points to a negative correlation between the number of
characters and mean distance of incompatible trees
($P = 0.055$).

Discussion

Impact of poor sampling of polymorphism

Results from performing PERDA on select data sets
clearly show that differential sampling of morphs jeop-
dardizes the integrity of inferred phylogenetic relation-
ships. The reference (i.e. strict consensus) trees from
unmodified data sets all exhibit unresolved topologies,
which are expected to reduce both the tree distance
values of trees from single morph sampling and the
number of incompatible trees. Nevertheless, the analys-
eses produced variable proportions of incongruent tree
topologies, in which most data sets yielded discourag-
ing proportions of trees ($\leq 75\%$) that are fully compa-
rible with the reference trees (Table 1; Fig. 2). The
potential of data sets, such as the Solenopsis fire ant
data, with a seemingly negligible number of polymorphic states (18; 0.03% of cells) to produce conflicting tree topologies is alarming, although this could be attributed to the relatively small number of characters. Large tree distance values may be due to topologically unstable (“wildcard”) terminals, which inflate RF distances. The prunnelsen function in TNT (Goloboff et al., 2008), however, indicates that prun-
ing up to three terminals from consensus trees does not improve the resolution of the metaconsensus tree for the data sets analysed in this study, with the exception of the TWiG matrix, which produced one additional node with the removal of three terminals. This outcome therefore discredits the possibility that these high tree distance values are primarily due to few wildcard taxa.

The tree distance values reported here are concordant with previous studies that demonstrated a substantial decrease in accuracy (i.e. average proportion of nodes in a reference tree retained in the trees obtained from subsampled data) when sample size was reduced to one specimen per terminal taxon (Archie et al., 1989; Wiens and Servedio, 1997, 1998; Wiens, 1998). Wiens and Servedio (1998), for instance, observed a sharp decline in accuracy from 91 to 24% when the sample size of each species was reduced from eight to one based on a simulation study. Taken together with the present study, the implication is clear —poor intraspecific sampling of polymorphic taxa could easily lead to incongruent, and potentially inaccurate, inference of their phylogenetic history.

Besides averages and percentages, the histograms (Fig. 2) provide additional insights into the impact of single morph sampling. The tree distances of incompatible trees from the Abronia and TWiG matrices produce distributions that appear to follow an exponential probability function (Fig. 2a, d). In contrast, the histograms from the Scinax and Polistes data exhibit irregular contours (Fig. 2b, c). Histograms of additional data sets (Fig. S1) also show exponential (i.e. crocodylomorph data set) as well as irregular distribution patterns (Nemesiid spider and Solenopsis fireant data). This variation in distribution patterns probably reflects the idiosyncratic character-based interactions that underlie each phylogenetic data set.

Second, variables, such as mean tree distances (e.g. “accuracy” in previous studies), are not commensurate with other metrics that describe the overall impact of single morph sampling. For instance, Abronia and Polistes data sets generate similar mean distances for incompatible trees (0.152 and 0.166, respectively), yet these data produce starkly different proportions of fully compatible trees and distribution of incompatible tree topologies (Fig. 2a, c). In addition, the Scinax and TWiG data sets share similar ranges in tree distance (0.230 and 0.242, respectively), but show a drastic difference in the proportion of fully compatible trees (75.05 and 43.20%, respectively). Therefore, investigations on factors that produce different tree topologies should examine histograms in tandem with basic descriptive parameters to formulate more informed descriptions of their impact on tree construction.

The variability observed in tree distances also extends to tree resolution. Comparisons between the reference trees and respective metaconsensus trees indicate that the extent to which clades collapse varies among data sets. For instance, the Abronia lizard data produce a metaconsensus tree that is almost entirely unresolved (Fig. 3a). In contrast, the metaconsensus trees from Polistes and TWiG matrices feature a more mosaic pattern in clade dissolution, in which some
Fig. 5. Reference and metaconsensus trees from PERDA on the Theropod Working Group (TWiG) morphological data (Turner et al., 2012). The reference tree is the strict consensus tree recovered from unmodified matrix. Numbers above and below internal nodes denote bootstrap (1000 pseudoreplications) and Bremer support values, respectively.
transitions in Characters 5, 8

Table 1
Information and results from PERDA on phylogenetic data sets ordered by percentage of polymorphic states in the matrix (% PS)

| Matrix                      | No. (%) of PS | No. of terminals | No. of characters | Max. TD | Mean TD (total) | Mean TD (TD > 0) | % TD = 0 |
|-----------------------------|---------------|------------------|-------------------|---------|-----------------|-----------------|----------|
| Solenopsis ants             | 18 (0.03)     | 17               | 36                | 0.133   | 0.009           | 0.125           | 93.14    |
| Dinosaursauriform           | 6 (0.06)      | 33               | 291               | 0       | 0               | 0               | 100.00   |
| TWiG                        | 154 (0.29)    | 110              | 475               | 0.242   | 0.028           | 0.050           | 43.20    |
| Nemesiid spiders            | 49 (0.53)     | 83               | 112               | 0.400   | 0.081           | 0.109           | 25.02    |
| Scinax frogs                | 27 (0.75)     | 42               | 86                | 0.230   | 0.020           | 0.079           | 75.05    |
| Crocodylomorph              | 4 (0.84)      | 84               | 301               | 0.297   | 0.037           | 0.073           | 49.49    |
| Polistes EF1A1              | 979 (3.42)    | 26               | 1101              | 0.405   | 0.144           | 0.166           | 13.27    |
| Abronia lizards             | 73 (10.58)    | 23               | 30                | 0.556   | 0.070           | 0.152           | 54.02    |

Number of terminals indicates number of ingroup terminals. Percentage of polymorphic states (% PS) computed as the number of polymorphic states over the total number of cells in the ingroup (i.e. product of the number of terminals and characters). Percentage of perfectly compatible trees (% TD = 0) is based on the number of trees with zero tree distance (TD) to the reference tree over the number of iterations (i.e. 10 000).

major clades are resistant to single morph sampling and others collapse completely (Figs 4 and 5). As expected, persistent clades are strongly associated with high bootstrap (Felsenstein, 1985) and Bremer support (Bremer, 1988, 1994) values (Figs 3–5). In the Polistes data, for example, nodes with bootstrap values below 100 collapsed, with the exception of the sister-group relationship between P. stigma and P. japonicas (Fig. 4). Likewise, clades in the reference tree of Scinax frogs generally collapsed for those with bootstrap values at or below 50 (Fig. 3).

Although the pattern is more nuanced, PERDA collapses major clades in the TWiG data with relatively low bootstrap or Bremer support values, including the relationships among paravian groups, which has been reported in previous studies (Xu et al., 2011; Turner et al., 2012; Spencer and Wilberg, 2013). These mixed outcomes among data sets further corroborate the notion that the effect of single morph sampling is not unilateral and depends on the underlying character-based structure of the data being analysed. In contrast to the consensus trees from MERDA in which missing entries (i.e. “?”) are replaced with randomly chosen states (Norell and Wheeler, 2003), no new clades were established by PERDA for the select data sets. Nevertheless, PERDA frequently constructs unique clades that do not occur in the reference tree at each replicated sampling of single morphs and may, for some data sets, produce a metaconsensus tree that includes a new clade.

Besides new clades, differential sampling of morphs also produces different sets of synapomorphies that define clades that exist in the reference tree. Certain combinations of morphs could lead to a removal of a synapomorphy. In the reference tree of the Abronia data set (Fig. 3a), for example, 11 synapomorphies define the group comprising A. bogerti and A. chizsari (transitions in Characters 5, 8–12, 17, 20, 23, 24, 26), but a transition in Character 26 is not considered a synapomorphy in some of the consensus trees from PERDA. In addition, unambiguous synapomorphies can be replaced by a new set of synapomorphies that arise from differential morph sampling. The clade that includes A. anzuetoi, A. aurita and A. lythrochila is supported by transitions in Characters 18, 27 and 30, whereas the same clade is defined by transitions in Characters 2, 16, 19 and 30 in one of the consensus trees from PERDA. Therefore, differential sampling of morphs could both add and remove synapomorphies identified through phylogenetic analysis of data with polymorphic states.

Why different trees?

Polymorphic states in phylogenetic data tend to introduce additional ambiguities when reconstructing phylogenetic trees because they often allow for more solutions that are equally parsimonious. Missing states (“?”), for instance, allow terminals to be placed anywhere on a cladogram without additional cost for that character. Likewise, a polymorphism with states “0” and “1” (i.e. [01]) has the same cost associated with transitions to and from states “0” or “1” unless specified otherwise by the user. In fact, a polymorphic state that includes every state in an unordered character is equivalent to a missing state with respect to methodological treatment. When a consensus summarizes these trees, it will generally exhibit a lower resolution than equivalent data sets of taxa without polymorphism unless every polymorphic state contains all possible states.

Previous simulation studies (Wiens and Servedio, 1997, 1998) have shown that the accuracy of phylogenetic trees reconstructed from various treatments of polymorphic characters is influenced by different parameters (e.g. number of individuals or sequences sampled, number of characters, level of polymorphism). Here, the results from multiple empirical data
sets variably support the conclusion that these properties are associated with the extent of conflicting trees recovered from single morph sampling (Table 1; Fig. 6). With the inclusion of Polistes sequence data, the abundance of polymorphic states and number of characters are potential predictors of the vulnerability of phylogenetic data to differential sampling of morphs, as indicated by solid regressions in Fig. 6.

However, besides the positive correlation between log-transformed percentages of polymorphic states and maximum tree distance, these regressions become insignificant ($P > 0.1$) when only the morphological data are considered. This outcome suggests that the extreme values from the Polistes data drive these correlations, which include relatively large mean and maximum tree distances, as well as a very small percentage of perfectly compatible trees. These values imply that sequence data are more vulnerable to single morph sampling than morphological data, probably due to their far greater number of polymorphic states. In fact,
these extreme values prohibit the identification of some of the correlations that exist with morphological data sets (e.g. negative correlation between the mean distance of incompatible trees and number of characters).

When the Polistes sequence data set is removed, regression analyses point to limited predictive power among properties of morphological data for the resulting phylogenetic ambiguity under single morph sampling: (i) a positive logarithmic relationship between percentage of polymorphic states in the data and maximum tree distance, (ii) a negative correlation between raw and log-transformed number of terminals and percentage of fully compatible trees, and (iii) a negative correlation between the number of characters and mean tree distance of incompatible trees, denoted by the dashed regression in Fig. 6. While the number of polymorphic states does not correlate significantly with any of the tree metrics recorded in this study, other properties of character data predict different aspects of the results from PERDA. For instance, the extent of disparate tree topologies from single morph sampling is dependent on the relative abundance of polymorphic states in the data set, but does not appear to influence the mean tree distance or the proportion of fully compatible trees. In contrast, increasing the number of terminals decreases merely the percentage of compatible trees due to the greater number of possible tree topologies that reduce the relative proportion of fully compatible trees.

Although the lack of significant negative correlation between the number of characters and overall mean distance of incompatible trees contrasts with the general pattern found in previous studies (Wiens and Servedio, 1997, 1998; Wiens, 1998), the results here show that greater numbers of characters generally reduce the mean distance of incompatible trees under single morph sampling. Collectively, PERDA indicates that each property of morphological data is a limited and specific predictor of the sensitivity of morphological data sets to differential morph sampling. Compared to results from simulation studies, the restricted number of clear relationships between properties of empirical data sets and tree metrics is probably due to the complex interplay of characters, which determines their robustness to differential sampling of morphs.

**Comparison with missing coding**

An approach described and criticized by Nixon and Davis (1991) replaces all polymorphic states with missing states ("?") to potentially reveal phylogenetic ambiguities associated with polymorphism. Although this missing coding approach would probably increase the number of MPTs (unless all polymorphic characters in the data contain every state), this inflated tree count and any additional ambiguity in inferred phylogenetic relationships would be arbitrary and decoupled from conditions actually observed in nature. In addition, this practice generally reveals considerably less phylogenetic uncertainty compared to PERDA, constructing a greater number of perfectly compatible trees and lower mean tree distances (Table 2). When missing coding is applied to the selected data sets, three of the seven data sets (i.e. Abronia, nemesiid and Solenopsis data sets) generate MPTs that are all perfectly compatible with the reference tree, whereas PERDA produced incompatible tree topologies.

For the Polistes and TWiG data, the resulting MPTs from missing coding show greater concordance with the reference trees than the strict consensus trees recovered from single morph sampling (Table 2; Figs 7–9). The Scinax, Polistes and crocodylomorph data sets yield smaller mean tree distances under missing coding than from PERDA (Table 2). Despite greater mean tree distances between MPTs and reference trees from missing coding for the TWiG data set, its strict consensus trees from PERDA include tree topologies that are more disparate (0.242 and 0.151, 0.151).

### Table 2

Information and results from applying “missing” coding on phylogenetic data sets

| Matrix              | No. (%) of PS | % MS  | No. of MPTs | Max. TD | Mean TD (total) | Mean TD (TD > 0) | % TD = 0 |
|---------------------|---------------|-------|-------------|---------|-----------------|------------------|----------|
| Solenopsis ants     | 18 (0.03)     | 6.54  | 6           | 0       | 0               | 0                | 100.00   |
| Dinosauriform       | 6 (0.06)      | 36.7  | 9           | 0       | 0               | 0                | 100.00   |
| TWiG                | 154 (0.29)    | 57.7  | 10 000*     | 0.151   | 0.076           | 0.073            | 3.25     |
| Nemesiid spiders    | 49 (0.53)     | 11.1  | 72          | 0       | 0               | 0                | 100.00   |
| Scinax frogs        | 27 (0.75)     | 9.3   | 36          | 0.029   | 0.005           | 0.029            | 83.33    |
| Crocodylomorph      | 4 (0.83)      | 39.4  | 912         | 0.044   | 0.003           | 0.021            | 84.85    |
| Polistes EF1A1      | 979 (3.42)    | 0     | 20          | 0.048   | 0.010           | 0.048            | 80.00    |
| Abronia lizards     | 73 (10.58)    | 0     | 21          | 0       | 0               | 0                | 100.00   |

Percentage of polymorphic states (% PS) and missing states (% MS) are computed as the number of polymorphic states and missing states over the total number of cells in the ingroup, respectively. Percentage of perfectly compatible trees (% TD = 0) is based on the number of trees with zero tree distance (TD) to the reference tree over the number of iterations (i.e. 10 000). MPTs, most parsimonious trees.

*Overflow of trees saved in memory.
respectively). Moreover, the strict consensus trees from missing coding show considerably greater resolution than the metaconsensus trees from PERDA (Figs 3, 4, 8 and 9). Notably, the strict consensus tree of *Polistes* wasp data from missing coding is more resolved than the reference tree due to the clade that places *P. sagittarius* as sister group to *P. dominula* and *P. nimpha* (Figs 3 and 8). Clearly, missing coding exposes less ambiguity in phylogenetic relationships that originate from intraspecific character-based interactions. Furthermore, this approach comes with several problems outlined by Nixon and Davis (1991), as well as Platnick et al. (1991), including the possibility that none of the MPTs from missing coding is congruent with MPTs supported by any combination of single morph sampling.

**Taxonomic implications**

With the construction of metaconsensus trees, the script becomes a tool for taxonomic classification by revealing ambiguities in phylogenetic relationships at the intersection of intra- and interspecific variation. Clades that collapse under PERDA signify groups in which variation within purported “species” could overcome interspecific relationships. When this occurs, the species-level classification within this group must be re-examined and corroborated based on additional lines of evidence because the morphs of polymorphic “species” characterized by polymorphic states, in reality, could be multiple species or, alternatively, other terminals in the clad could be additional morphs of the same species. For extant taxa, population-level studies could allow more accurate assessments of their species classification. In palaeontological studies, however, species-level delimitation cannot be confirmed, but may be supported by stratigraphic and biogeographical evidence. If the species classification within collapsed clades cannot be justified, then the practice of establishing taxonomic names for and within these groups should be discouraged. Admittedly, PERDA is a harsh sampling strategy. However, it exposes charac-
ter-based taxonomic ambiguities that originate from intraspecific variation, allowing a more conservative classification of not only species, but also at higher taxonomic levels.

Using the TWiG coelurosaur data set as an example, its metaconsensus tree collapses several key clades that are present in the strict consensus tree of the original matrix (Fig. 5). For instance, the clade Deinonychosauria, which includes dromaeosaurids and troodontids, and is a sister group to birds (Avialae), collapses to form a notable polytomy with 40 branches. Hence, variation considered to be intraspecific is able to overcome clades established by the consensus tree recovered from the unmodified data. In this case, the metaconsensus tree undermines the establishment of “Paraves”, in addition to all other collapsed clades, when the a priori distinction between intra- and interspecific variation is relaxed. When compared to the reference tree, the metaconsensus tree points to clades for which taxonomic classification of their members should be corroborated based on independent evidence, such as stratigraphic occurrence or geographical distribution.

**Strategies for the treatment of polymorphism**

Given the ubiquity of polymorphism in nature and phylogenetic data, strategies for maintaining precision in our phylogenetic inference warrant discussion. One may argue that phylogenetic data should exclude polymorphic characters in the first place because they tend to be homoplastic (Nixon and Wheeler, 1990). Both empirical and simulation studies, however, demonstrate that the removal of polymorphic characters consistently leads to underperformance compared to data that incorporate polymorphic characters (Wiens, 1995, 1998; Wiens and Servedio, 1997). This mirrors the result that highly homoplastic characters collectively contain high resolving power (Källersjö et al., 1999). As such, the recognition and inclusion of polymorphic characters are critical to the practice of constructing phylogenetic trees.

Yet, only single individuals are sampled for many species in phylogenetic studies, impeding the recognition of multiple morphs within species. Even in the taxonomic literature, singletons (species described from one specimen) and uniques (species described from one locality) are common (Dayrat, 2005; Lim et al., 2012). An obvious strategy for alleviating this issue is to increase the intraspecific sample size. For continuous characters, Tixier (2012) recommends sampling at least ten individuals for each species to determine the breadth of intraspecific variation. When using polymorphic coding for discrete characters, however, Wiens and Servedio (1997, 1998) have shown that intraspecific sample size beyond two generally has a marginal effect on the accuracy of trees. Therefore, the detection of multiple morphs does not necessarily require a robust intraspecific sampling. In fact, for sequence data, oversampling within species would probably increase the number of multistates associated with amplification and sequencing error. Error rates for the particular
Fig. 9. Strict consensus tree from missing coding of polymorphic states and metaconsensus tree from PERDA on theropod working group (TWiG) morphological data (Turner et al., 2012). Bootstrap (1000 pseudoreplications) and Bremer support values are indicated above and below the internal nodes, respectively, in cases where the Bremer support value differs from the reference tree (Fig. 5).
sequencing protocol should be consulted to determine the optimal intraspecific sample size.

A better strategy may be to increase the number of characters (e.g., number of loci or morphological characters), which generated greater accuracy for every treatment of polymorphic characters under virtually any condition (Wiens and Servedio, 1997, 1998). Despite these outcomes from simulation studies, results from PERDA indicate that, for empirical data, increasing the number of characters generally does not reduce the phylogenetic ambiguity associated with polymorphism under single morph sampling, although it may lower the overall disparity in tree topology for morphological data (Fig. 6). This suggests that the single morph sampling must first be resolved before the increase in number of characters leads to appreciable benefit to phylogenetic inference in empirical data sets.

One potential approach for overcoming these issues associated with polymorphic coding is to treat each sampled morph as terminals in a phylogenetic analysis. This practice avoids non-monophyletic groups and generalization of character states over multiple individuals (Vrana and Wheeler, 1992), while retaining polymorphic characters in the data. In addition, it removes ambiguities in phylogenetic relationships associated with the treatment of polymorphisms. With respect to taxonomic polymorphism, Nixon (1996) recommends analysing at the level of morphs to prevent the reduction of character congruence among taxa. As with the metaconsensus tree, any paraphyly of terminals purported to be a single species compels further investigation of their taxonomic assignment. Consequently, this strategy has implications for diversity assessments and conservation. Nonetheless, it has several benefits over the commonly employed polymorphic coding. In particular, this strategy could be employed subsequent to PERDA for polymorphic taxa that are members of clades that become collapsed in the metaconsensus tree. If all morphs of a purported species form a monophyletic group defined by a set of character states, then this would justify their assignment as a valid species (Nixon and Wheeler, 1990). Conversely, if the morphs do not form a monophyletic group, then their status as a single species is undermined.

Conclusions

PERDA is a new TNT script that illuminates the potential impact of poor intraspecific sampling of polymorphic taxa on phylogenetic inference. The results presented here demonstrate alarming numbers and frequencies of disparate trees with clades that conflict with the strict consensus trees constructed from original data. In addition, PERDA results in the dissolution of major clades that occur in the reference trees. Even data sets with seemingly negligible amounts of polymorphic states lead to conflicting tree topologies. This outcome therefore undermines the typical practice of collecting single or very few specimens for each taxon in systematic studies, which is true for most palaeontological studies.

Compared to the strict consensus trees from unmodified data, the metaconsensus tree from PERDA reveals taxonomic groups that are vulnerable to differential sampling of morphs. For both extant and extinct taxa, this method exposes collapsed clades, in which species-level taxonomy should be re-evaluated and justified based on other lines of evidence. In fact, PERDA is superior to missing coding of polymorphisms for exposing ambiguities in phylogenetic positions of taxa because it explores a greater set of character-based interactions by subsampling actual observations. Furthermore, PERDA provides an automated approach for exploring the ambiguities associated with the intersection of intra- and interspecific variation on any data with discrete characters. The objective of the analysis, however, is not to give a verdict on the treatment of polymorphic characters, but instead to investigate the behaviour of phylogenetic reconstructions in the practical context of intraspecific sampling.

Acknowledgements

I thank Pablo Goloboff (CONICET) and Mark Siddall (AMNH) for assistance with TNT scripting; the Willi Hennig Society for free access to TNT; and James Carpenter (AMNH), John Denton (AMNH), Pablo Goloboff (CONICET), Ansel Payne (AMNH), Scott Steppan (Florida State University), Dennis Stevenson (NY Botanical Garden), Ward Wheeler (AMNH), Joanna Wolfe (AMNH) and two anonymous reviewers for constructive comments on the study and earlier versions of the manuscript. This study was supported by the Richard Gilder Graduate School Comparative Biology Program (AMNH), ExxonMobil GeoSciences Grant, NSF Graduate Research Fellowship and the NSF Dissertation Improvement Grant (DEB 1406849).

References

Archie, J.W., Simon, C., Martin, A., 1989. Small sample size does decrease the stability of dendrograms calculated from allozyme-frequency data. Evolution 43, 678–683.

Bourque, M. 1978. Arbres de Steiner et réseaux dont certains sommets sont a localisation variable. Ph.D. Dissertation, Université de Montreal, Montreal, Quebec.

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

Bremer, K., 1994. Branch support and tree stability. Cladistics 10, 295–304.
Campbell, J.A., Frost, D.R., 1993. Anguid lizards of the genus Abronia: revisionary notes, descriptions of four new species, a phylogenetic analysis, and key. Bull. Am. Mus. Nat. Hist. 216, 1–121.

Coddington, P., Scharff, N., 1994. Problems with zero-length branches. Cladistics 10, 415–423.

Cornish-Bowden, A., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 13, 3021–3030.

Dayrat, B., 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85, 407–415.

Fairovich, J., 2002. A cladistic analysis of Scinax (Anura: Hylidae). Cladistics 18, 367–393.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 38, 16–24.

Ford, E.B., 1945. Polymorphism. Biol. Rev. 20, 73–88.

Felsenstein, J., 1985. Phylogenies and the comparative method. Am. Nat. 125, 1–15.

Faivovich, J., 2002. A cladistic analysis of

Santos, B.F., Payne, A., Pickett, K.M., Carpenter, J.M., 2014. Phylogenetic and historical biogeography of the paper wasp Polistes (Hymenoptera: Vespidae): implications for the overwintering hypothesis of social evolution. Cladistics, DOI: 10.1111/cla.12103.

Wiens, J.J., Norell, M.A., Wheeler, W.C., 2003. Missing entry replacement data in cladistic analysis. Cladistics 7, 233–241.

Nixon, K.C., Wheeler, Q.D., 1990. An amplification of the specimen delimitation methods. Syst. Biol. 61, 165–169.

Nixon, K.C., 1996. Paleobotany in cladistics and cladistics in paleobotany: enlightenment and uncertainty. Rev. Palaeobot. Polynol. 90, 361–373.

Nixon, K.C., Davis, J.J., 1991. Polymorphic taxa, missing values and cladistic analysis. Cladistics 7, 233–241.

Norell, M.A., Wheeler, W.C., 2003. Missing entry replacement data analysis: a replacement approach to dealing with missing data in paleontological and total evidence data sets. J. Vertebr. Paleontol. 23, 275–283.

Peecook, B.R., Sidor, C.A., Nesbitt, S.J., Smith, R.M.H., Steyer, J.S., Angielczyk, K.D., 2013. A new silesaurid from the upper Ntawere Formation of Zambia (Middle Triassic) demonstrates the rapid diversification of Silesauridae (Avemetatarsalia, Dinosauriformes). J. Vertebr. Paleontol. 33, 1127–1137.

Pitts, J.P., McHugh, J.V., Ross, K.G., 2005. Cladistic analysis of the fire ants of the Solenopsis saevissima species-group (Hymenoptera: Formicidae). Zool. Scr. 34, 493–505.

Platnick, N.I., Griswold, C.E., Coddington, J.A., 1991. On missing entries in cladistic analysis. Cladistics 7, 337–343.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Santos, B.F., Payne, A., Pickett, K.M., Carpenter, J.M., 2014. Phylogenetic and historical biogeography of the paper wasp genus Polistes (Hymenoptera: Vespidae): implications for the overwintering hypothesis of social evolution. Cladistics, DOI: 10.1111/clad.12103.

Simmons, N.B., 2001. Misleading results from the use of ambiguity coding to score polymorphisms in higher-level taxa. Syst. Biol. 50, 613–620.

Simmons, N.B., Geisler, J.H., 2002. Sensitivity analysis of different methods of coding taxonomic polymorphisms: an example from higher-level bat phylogeny. Cladistics 18, 571–584.

Spencer, M.R., Wilberg, E.W., 2013. Efficacy or convenience? Model-based approaches to phylogeny estimation using morphological data. Cladistics 29, 663–671.

Tixier, M.S., 2012. Statistical approaches to assess intraspecific variations of morphological continuous characters: the case study of the family Phytoseiidae (Acari: Mesostigmata). Cladistics 28, 489–502.

Turner, A.H., Sertich, J.J.W., 2010. Phylogenetic history of Simosuchus clarki (Crocodyliformes: Notosuchia) from the Late Cretaceous of Madagascar. J. Vertebr. Paleontol. 6, 177–236.

Turner, A., Makovicky, P.J., Norell, M.A., 2012. A review of dromaeosaurid systematics and paravian phylogeny. Bull. Am. Mus. Nat. Hist. 371, 1–206.

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

Wiens, J.J., 1995. Polymorphic characters in phylogenetic systematics. Syst. Biol. 44, 482–500.

Wiens, J.J., 1998. Testing phylogenetic methods with tree congruence: phylogenetic analysis of polymorphic morphological characters in phrynosomatid lizards. Syst. Biol. 47, 427–444.

Wiens, J.J., 1999. Polyplasy in systematics and comparative biology. Annu. Rev. Ecol. Syst. 30, 327–362.

Wiens, J.J., Servedio, M.R., 1997. Accuracy of phylogenetic analysis including and excluding polymorphic characters. Syst. Biol. 46, 332–345.

Wiens, J.J., Servedio, M.R., 1998. Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. Syst. Biol. 42, 229–253.

Xu, X., You, H., Du, K., Han, F., 2011. An Archaeopteryx-like theropod from China and the origin of Avialae. Nature 475, 460–465.

Supporting Information

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

Data S1. PERDA script in plain text format.

Figure S1. Histograms of tree distances between reference trees and consensus trees recovered from PERDA for data sets not figured in the main text.

Table S1. Tree distances from PERDA on data sets presented in this study.

Appendix 1

Script for performing PERDA in TNT v1.1 (Goloboff et al., 2008). To run, specify the number of iterations and data matrix in the command line (e.g. "perda matrix.tnt 10000;"). It outputs three files: (i) a tree file ("perda_trees.tnttre") containing the reference strict consensus tree as the first tree (Tree 0) and all trees from single state sampling; (ii) another tree file ("perda_metatree.tnttre") comprising a strict consensus tree from trees recovered from subsampling; and (iii) a text file ("perda_metatree.txt") listing tree distances between the strict consensus tree from the original data matrix and trees obtained from subsampling.

```r
macr
macro "* 10000000;
macro entry1("10\ntax\nchar10000030(1000:2500);"
macro "-
if (argnumber < 2) ermsg ERROR!
Too few arguments! Please rerun with matrix filename, number of iterations.

e.g. >perda matrix.tnt 10000;
```
end;

if (argnumber > 2) error "ERROR! Too many arguments! Please rerun with matrix filename number of iterations.
  e.g. >perda matrix.tnt 10000;
end;

sil =console;
set nseed 0;
proc %1;
hold 10000;
taxname =;
collapse rule 1;
multi = replicate 1000 hold 10; /** modify appropriately ***/
bbreak;
nelsen *;
tsave* perda_trees.tnttre;
save /;
keep 0;
/** VARIABLES ***/
var: state num pick tmp[22] chosen rule dist;
set nnums %2+1;
var: rf['nums'];
/** ANALYSIS ***/
loop 1 %2
  proc %1;
  hold 10000;
  loop 0 ntax
    loop 0 nchar
      set state states[#3 #2];
      if ('state' != missing);
        set nnums numbits('state');
        if ('nums' != 1)
          set pick getrandom [ 1 'nums' ];
          set tmp $bitset; 'pick' 'state';
          set chosen ( Stmp=1 );
xread =3 +2 'chosen';
    end
  end
  tsave /;
  keep 0;
  sil =console;
  proc perda_trees.tnttre;
  nelsen *1.21;
  tsave *perda_metatree.tnttre;
  save /;
  tsave /;
  keep 0;
  proc perda_trees.tnttre;
  set rule collapse;
collapse none;
  loop 1 %2
    tcomp * ];0 ];1; /** based on script by P. Goloboff ***/
    tcomp * ];1 ];0;
    set dist ( [tnodes[ntrees]+tnodes[ntrees-1]])/([tnodes[0]+tnodes[1]]);
    set nnums ntrees - 1;
    keep 'nums';
    set rf[#1] '/.3/dist';
  stop
  collapse 'rule';
sil =all;
  log perda_treedist.tnttre;
loop 1 %2
  tquote 'rf[#1]';
  tquote
-------------------------------------------------------------------------------------------------
* ANALYSIS DONE *

Output files:
1. perda_trees.tnttre (Tree 0: consensus from original matrix)
2. perda_metatree.tnttre (consensus of trees from replicates)
3. perda_treedist.tnttre (list of tree distances)

;