Syst. Biol. 59(1):42–58, 2010

DOI:10.1093/sysbio/syp075
Advance Access publication on November 11, 2009

Sparse Supermatrices for Phylogenetic Inference: Taxonomy, Alignment, Rogue Taxa, and the Phylogeny of Living Turtles

ROBERT C. THOMSON* AND H. BRADLEY SHAFFER

Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, CA 95616, USA;
**Correspondence to be sent to: Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, CA 95616, USA; E-mail: rcthomson@ucdavis.edu.*

Received 7 January 2009; reviews returned 18 April 2009; accepted 28 September 2009
Associate Editor: Thomas Buckley

Abstract.—As phylogenetic data sets grow in size and number, objective methods to summarize this information are becoming increasingly important. Supermatrices can combine existing data directly and in principle provide effective syntheses of phylogenetic information that may reveal new relationships. However, several serious difficulties exist in the construction of large supermatrices that must be overcome before these approaches will enjoy broad utility. We present analyses that examine the performance of sparse supermatrices constructed from large sequence databases for the reconstruction of species-level phylogenies. We develop a largely automated informatics pipeline that allows for the construction of sparse supermatrices from GenBank data. In doing so, we develop strategies for alleviating some of the outstanding impediments to accurate phylogenetic inference using these approaches. These include taxonomic standardization, automated alignment, and the identification of rogue taxa. We use turtles as an exemplar clade and present a well-supported species-level phylogeny for two-thirds of all turtle species based on a ∼50 kb supermatrix consisting of 93% missing data. Finally, we discuss some of the remaining pitfalls and concerns associated with supermatrix analyses, provide comparisons to supertree approaches, and suggest areas for future research. [Alignment; GenBank; phyloinformatics; rogue taxa; supermatrix; taxonomy; Testudines; turtle phylogeny.]

As sequence data for a wide variety of organisms have become easier to acquire, interest in inferring large-scale phylogenies has increased dramatically (Bininda-Emonds 2004; Cracraft and Donoghue 2004; Hodkinson and Parnell 2006). The growth of phylogenomic studies that amass data sets to infer phylogeny based on tens to thousands of markers has dramatically increased character sampling, although typically for a limited number of model taxa (Rokas et al. 2003; Wildman et al. 2007; Dunn et al. 2008). Supertree and supermatrix approaches, on the other hand, have emphasized increased taxon sampling, usually with more modest gains in character sampling (Driskell et al. 2004; McMahon and Sanderson 2006; Bininda-Emonds et al. 2007). All of these approaches have spurred methodological advances for dealing with large amounts of character data (reviewed by de Queiroz and Gatesy 2007) as well as handling large numbers of taxa (Roshan et al. 2004; Stamatakis 2006; Zwickl 2006). Although some controversy concerning the relative merits of alternative approaches has arisen (particularly for supermatrix vs. supertree approaches; Gatesy et al. 2002, 2004; Bininda-Emonds et al. 2003), analyses based on hundreds to thousands of taxa and genes are an important element of phylogenetic analyses that have already resulted in substantial phylogenetic synthesis.

As we move toward a true tree of life utilizing genomic data, a number of methodological and technical challenges remain to be solved. The gold standard for such analyses remains the "complete-matrix" approach to data set construction, with the aim of scoring all characters for all taxa. Such studies provide information on both character and clade evolution, are free from potential biases imposed by missing data, and are well suited to existing phylogenetic methods. However, large data sets also present inherent difficulties; primers cease to work across distantly related taxa, accurate alignment becomes more difficult, and genes or gene regions are gained and lost.

As databases grow both with respect to numbers of sequences (Benson et al. 2007) and taxa (Sanderson 2007), recent efforts have focused on mining the information contained in GenBank and similar large sequence databases (Driskell et al. 2004; McMahon and Sanderson 2006). These "sparse supermatrix" studies have not only met with substantial success but also encountered significant challenges. The problem is that the heterogeneous nature of the data can mask the phylogenetic information contained in these data sets and give the misleading appearance of great uncertainty. The difficulty, then, is that we must separate this apparent uncertainty, caused by heterogeneity in the data, from the true lack of phylogenetic signal.

Two general issues are at the forefront of efforts to combine disparate, often sparse data sets together for supermatrix analyses. The first centres on so-called "rogue taxa" (Sanderson and Shaffer 2002) loosely defined as phylogenetically unstable taxa that can move widely across trees with little or no effect on the tree's score. The diverse and complex patterns of taxonomic and data overlap that lead to rogue taxa make a priori screening difficult, although post-analysis measures of "rogue-iness" provide a possible strategy for identifying (and eliminating) rogues and their effects. Taxonomic instability and errors are one common facet of supermatrix construction that can lead to rogue taxa. When taxonomic names change, the result is often that isolated names represented by few sequences accumulate in GenBank. In a sparse supermatrix, this increases data fragmentation and can lead to one or both of these

"pseudotaxa" behaving as rogues in the resulting tree. Such taxonomic changes can masquerade as phylogenetic uncertainty when none actually exists.

The difficulty associated with accurate alignment forms a second major challenge for supermatrix approaches. The sheer amount of data in GenBank make highly automated approaches necessary; however, the degree of heterogeneity present in the data makes automation difficult for at least 2 reasons. First, sequences in databases vary widely in length for the same gene region. Because most alignment strategies assume homology along the length of the entire sequence, they handle extreme length heterogeneity poorly. Second, the large evolutionary distances involved in many supermatrix-level analyses suggest that large molecular divergences must be managed. Strategies for dealing with this heterogeneity have involved working at the protein level (Driskell et al. 2004), controlling the heterogeneity allowed in clusters by controlling the criteria under which clustering occurs (McMahon and Sanderson 2006), or extracting only complete or nearly complete matrices with little or no missing data (Sanderson et al. 2003; Yan et al. 2005).

In this study, we have 2 primary goals. First, we examine several of the methodological issues involved in going from a large sequence database to a useable phylogeny. In particular, we integrate (semi) automated solutions to rogue taxon identification, nomenclatural changes, and alignment challenges into a single study to examine the phylogenetic signal contained in large sparse supermatrices. Second, we provide an empirical case study by summarizing the available phylogenetic data for turtles. Turtles are a relatively ancient clade and the living crown group captures a long period (~210 million years) of evolutionary history (Gaffney 1990). Thus, the molecular divergence observed in turtles (which leads to many of the difficulties in supermatrix construction) is substantial enough to provide a reasonable test case for our methods. With 321 species, turtles encompass a modest level of species diversity, allowing for a more thorough examination of issues relating to data combinability and practical issues of taxonomic standardization than is possible in much larger clades. Although the resulting data matrix is quite sparse, we report a well-resolved well-supported species-level tree for crown-group turtles containing two-thirds of all described turtle species. More generally, our results indicate that sparse supermatrices constructed from large DNA databases can be a valuable tool for many groups if appropriate strategies are used to deal with the fragmented and heterogeneous nature of these data.

Materials and Methods

Informatics and Matrix assembly

The process of going from raw GenBank sequence data to a polished supermatrix involves several informatic challenges; some of these are easily automated, others require more direct data manipulation. Throughout the data-handling pipeline, we sought to maximize automation and build in checks that would call for human involvement only when necessary. Most of the pipeline was implemented in a series of Perl scripts (available from http://www.eve.ucdavis.edu/rcthomson) and the general strategy is outlined in Figure 1.

Starting data.—We downloaded all clusters for root node "Testudines" that contained at least 4 species-level taxon IDs from the PhyLoTA GenBank browser (http://loco.biosci.arizona.edu/pb/) to use as preliminary clusters in our analysis (Sanderson et al. 2008). PhyLoTA assembles these clusters via single linkage BLAST clustering for all sequences in the GenBank flat files (release 159). Thus, each cluster contains sequences that exhibit a specified degree of BLAST similarity to at least one other sequence within that cluster. These clusters contain sequences from a wide range of mitochondrial genes, nuclear genes that are commonly used in phylogenetic analysis (e.g., RAG-1, R35), as well as SINE flanking region sequences. We added several additional turtle sequences that were released since release 159 of GenBank (Praschag et al. 2007; Thomson et al. 2008; Spinks et al. 2009).

Taxonomy.—The first step in the analysis pipeline was to standardize taxon names across all clusters. This is typically a slow and error prone process but is essential to correctly assemble existing data into supermatrices and supertrees. Typical problems that must be corrected include the numerous errors present in GenBank submissions (largely misspellings and incorrect suffixes on Latin names), updating names where taxonomy has changed between submissions, dealing with "name duplications" where single species are represented by multiple names (usually when unstable taxonomy leads multiple workers to attribute a single species to different genera) and ensuring that terminal taxa (usually species vs. subspecies) used in the analysis are entered at the same taxonomic level. Because these issues are widespread, it is not possible to simply extract the genus and species name or taxon ID from each sequence record and use these for subsequent analyses without introducing a large number of errors.

To proceed, we downloaded the NCBI XML taxonomy file for Testudines and extracted all names of rank species or subspecies and their corresponding taxon IDs. This list contains all names that workers have used to submit sequences to GenBank for turtles, including misspellings, names denoting hybrids (e.g., "*Caretta caretta* × *Lepidochelys kempii*"), and names for which the species had not yet been named or unknown (e.g., "*Elseya* sp. Albagula"). We then created a second "updated" list of species binomials only based on the most recent complete checklist produced for turtles (Turtle Taxonomy Working Group 2007). Where species binomials in the updated list exactly matched a name in the NCBI list, we mapped the taxon ID from the NCBI list to the corresponding name in the updated list. For species

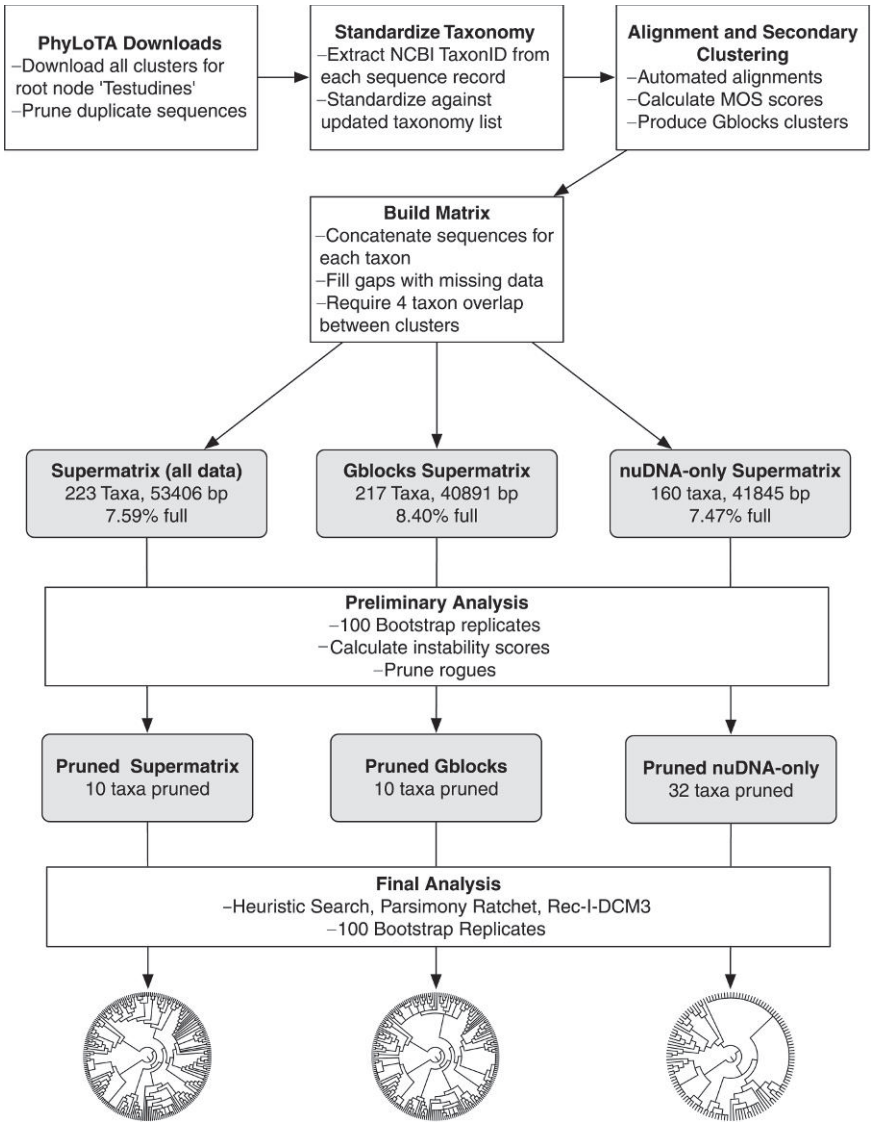

FIGURE 1. A flowchart outlining the informatic pipeline developed for this study.

that appeared in the updated list but not in the NCBI list, we created new unique taxon IDs.

We extracted the taxon name from the definition line of each sequence record and replaced it with the appropriate taxon ID from the NCBI list. For each record, we then asked if that taxon ID appeared in the updated list; if so, no further action was taken on that sequence. If it did not, the name was returned to the user for direction on how to proceed. In most cases, the correction needed was obvious (e.g., change *Pelea* to *Palea*). The user inputs the correction, which causes that taxon ID to be associated with the correct name in the updated list. The next time that particular error is encountered, the taxon ID is recognized in the updated list and the correction takes place automatically. At the end of this process, the researcher has dealt with each "correction" exactly once, all clusters contain the correct names, and the resulting updated list contains all synonymies needed to rectify the NCBI list with the updated list. If each correction tends to be unique, this approach contributes relatively little to the automation of the process. In practice, because the majority of changes are the result of either changing taxonomy or very common misspellings, each change was necessary many times and the automation process was quite efficient.

Alignment and secondary clustering.—Most commonly used alignment approaches are geared toward global alignment problems (those where sequences are similar in length and homologous across the entire sequence) and are poorly suited to dealing with sequences that vary widely in length and overlap (Morgenstern 2004; McMahon and Sanderson 2006). For these more problematic sequences, local alignments may improve overall alignment quality (Pearson and Lipman 1988; Huang

and Miller 1991; Morgenstern 2004). We experimented extensively with several methods and eventually decided on a combined approach using a local alignment algorithm (implemented in DIALIGN2, Morgenstern 1999) to bring together broad regions of homology followed by the refinement algorithm from MUSCLE (Edgar 2004) to clean up the small unaligned regions that DIALIGN2 can leave behind (personal observation; McMahon and Sanderson 2006).

To further identify problematic alignments, we performed independent alignments of all clusters using DIALIGN2 alone, MUSCLE alone (doing full alignments, rather than the refinement-only algorithm), and MUSCLE-refined DIALIGN2 alignments and compared the results using the multiple overlap score (MOS) implemented in MUMSA (Lassmann and Sonnhammer 2005). This score is a measure of consistency between alignments produced under a set of methods. If the score for all 3 alignments is uniform and high, we can infer that all methods agreed and that the alignment for that cluster was not problematic. When MOS scores did not agree, we targeted that cluster for additional effort including realignment using different settings and manual adjustments.

Because the initial PhyLoTA screen used relatively stringent BLAST criteria to determine cluster membership, sequences from a single gene were sometimes parsed into multiple clusters (e.g., there were 4 preliminary clusters for the mitochondrial DNA (mtDNA) control region in our data set). Often, these multiple clusters consisted of different segments of a single gene that overlapped only slightly or not at all; other times they consisted of different taxonomic groups for the same gene that were separated because of great sequence divergence. Combining these clusters into single alignments can, in principal, increase the density of the data set. We therefore conducted a round of secondary clustering aimed at forcing these separate clusters together. We combined clusters based on gene identity information contained in the definition line for each sequence and attempted to align them using our alignment pipeline. These alignments were then evaluated using MOS scores and checked by eye. We performed adjustments where necessary, and in some cases decided that the data did not support secondary clustering because of limited overlap between sequences with deep molecular divergences. In these cases, we left the sequences as separate clusters. Finally, to examine what effects any remaining poorly aligned regions of the data set might have on our analyses, we produced a second set of alignments that had been pruned of poorly aligned and very sparse regions by the Gblocks program (Castresana 2000). By comparing the pruned versions of the alignments to the nonpruned alignments (Fig. 1), we were able to gauge the effect of the most difficult-to-align parts of the data set on the overall results.

Matrix construction.—Clusters were combined into several different supermatrices that allowed us to explore issues of taxonomic overlap, efficacy of Gblocks, and mitochondrial versus nuclear data composition. We first assembled the aligned clusters produced by the DIALIGN2 + MUSCLE into 4 concatenated supermatrices such that each cluster overlapped with one or more other clusters by at least 3, 4, 5, or 6 taxa, respectively. McMahon and Sanderson (2006) found that even with a minimum 4-taxon overlap, many problematic taxa were included in the data set; we attempted to bracket this by increasing and decreasing the stringency of the taxon overlap requirement. Next, we assembled 2 other comparative matrices based on 4-taxon overlap only. The first used the Gblocks clusters to examine the gain or loss in phylogenetic precision when the most difficult parts of an alignment were eliminated. The second consisted of the nuclear DNA (nuDNA) data only to compare with that based on mtDNA + nuDNA. Supermatrices were assembled by concatenating sequences from each cluster for individual species and filling in the rest of the matrix with missing data.

Phylogenetic Analysis

We employed a 2-step analysis procedure aimed at first identifying rogue taxa and then performing a more thorough set of analyses after rogue taxa had been removed from a given data set. Because the best methods for analysing sparse supermatrices are not well understood, we employed several and compared the results.

Rogue taxa.—To identify rogues, we searched 100 parsimony bootstrap replicates of each supermatrix using heuristic searches in PAUP*4.0b10 with a single random sequence addition starting tree and limiting each replicate to 30 min. We used the resulting collection of trees to calculate a taxonomic instability index (I) for each taxon (i) in the set of trees using the following, implemented in Mesquite (Maddison W.P. and Maddison D.R. 2007):

$$I = \sum_{(x,y),j \neq i} \frac{|D_{ijx} - D_{ijy}|}{(D_{ijx} - D_{ijy})^2},$$

where D_{ijx} is the patristic distance between taxa i and j in tree x and D_{ijy} is the same distance for tree y. Their difference is summed across all taxa and tree pairs. Because rogue taxa, by definition, have many divergent placements in the tree, this summation will be larger for rogue taxa than for taxa whose placement in the tree is relatively consistent. Based on instability scores, we experimented with pruning apparent rogues from the collection of initial bootstrap trees and recalculating reduced consensus trees after each taxon was deleted. If removal of a taxon led to a large increase in resolution in the reduced consensus, we pruned that taxon from the data set as a rogue; if its removal caused no, or modest, changes, the taxon was left in the data set. This process was carried out starting with the most unstable taxon and proceeded down the list. Pruning was conservative in that only clearly rogue taxa were removed from the

data set. Our goal was to include the maximum number of taxa in the tree, with the sacrifice that we might have included somewhat "rogue-y" taxa that contributed to low support values. This criterion is clearly arbitrary; an alternative would be to employ a fixed cutoff (e.g., prune the 5% or 10% least stable taxa). There is relatively little risk associated with overpruning because the removal of nonrogue taxa has little effect on the overall tree resolution (though support values should increase slightly in the vicinity of taxa that are unnecessarily pruned).

Tree searches.—Optimal phylogenetic analysis strategies for sparse supermatrices are not well understood. Previous studies have focused on maximum parsimony (MP) for computational speed and simplicity (McMahon and Sanderson 2006) and we follow this here. For the 4-taxon overlap matrix, we also carried out maximum likelihood (ML) analyses and attempted a Bayesian analysis to investigate the feasibility of model-based approaches for sparse matrices. We also employed matrix representation using parsimony (MRP; Baum 1992; Ragan 1992) and gene tree parsimony (GTP; Slowinski and Page 1999) supertree approaches to allow variation in gene trees among clusters and to provide a comparison between supermatrix and the supertree approaches.

Parsimony searches were implemented using standard heuristic searches, parsimony ratchet searches (Nixon 1999), and Rec-I-DCM3 boosted parsimony searches (Roshan et al. 2004). The heuristic searches were carried out in PAUP*4.0b10 using tree bisection and reconnection (TBR) branch swapping with stepwise addition starting trees and 10 random sequence addition replicates. Searches were allowed to swap to completion. Parsimony ratchets were implemented in PAUP* with 10 replicates of 200 iterations each, upweighting alternatively 10% or 25% of the matrix by weight 1 in each iteration followed by a final heuristic search using the best parsimony ratchet trees as starting trees. We then filtered the best trees found from all phases of the ratchet search. Finally, we employed Rec-I-DCM3 boosted parsimony searches in PAUP* using the CIPRES portal. These searches used a 50% maximum taxon subset and 50 iterations of the search, performing TBR branch swapping for both small and large tree inferences. Allowing large numbers of trees to be saved at each step of the search had little effect on the results, greatly increased the duration of searches, and sometimes crashed the run and/or the machine it was being carried out on due to excessive memory demands. This was a much larger problem when attempting to analyse data sets that we knew contained rogue taxa than in the pruned data set. We therefore limited most searches between 10 and 10,000 trees. We took the shortest trees found with each search strategy and used them to construct both strict and majority rule consensus trees. We assessed phylogenetic support via nonparametric bootstrapping in PAUP*. We performed 100 bootstrap replicates using TBR branch swapping with stepwise addition starting trees and 10 random sequence addition replicates holding 10 trees at each replicate.

Although recent work has suggested that model-based phylogenetic methods may be quite sensitive to missing data (Lemmon et al. 2009), we felt that it was important to examine our data in ML and Bayesian frameworks. We ran analyses under a GTR + I + Γ model of molecular evolution, which was selected as the best-fitting model using DTModSel (Minin et al. 2003). We carried out maximum likelihood analyses and bootstrapping in RAxML version 7.0.4 (Stamatakis 2006) using the combined bootstrapping and ML search algorithm. Bayesian analyses were carried out under MrBayes version 3.1.2 (Ronquist and Huelsenbeck 2003) with 4 independent runs, each with 4 chains. We sampled from the chains for every 1000 generations and assessed convergence using the average standard deviation of split frequencies.

We carried out MRP supertree analysis by first inferring a parsimony tree for each cluster by itself (after rogue taxa had been pruned) and using those trees to assemble a single MRP matrix in Mesquite. We analysed the matrix in PAUP* using 10 random addition sequence replicates, allowing up to 1000 trees to be stored for each replicate, and constructed a strict consensus of the resulting trees. GTP analyses are more difficult to carry out on a tree of this size. The only existing algorithm, to our knowledge, that can carry out GTP on a tree this large is implemented in the program DupTree (Wehe et al. 2008). The method requires fully bifurcating trees, which are not attainable with confidence for most of our clusters (or for most phylogenetic data sets). We followed the methods used by the developers of the algorithm, employing neighbor joining trees as approximations for each cluster's fully resolved tree (Bansal et al. 2007). We carried out a batch of analyses alternatively supplying no starting tree (using the leaf adding heuristic to generate a starting tree) and supplying the majority rule consensus tree from the MP search as a starting tree. We allowed the algorithm to attempt all gene tree rerootings and used the full queue-based heuristic. We replicated the no starting tree analyses 10 times and kept the overall best scoring tree.

RESULTS

Informatics and Matrix Assembly

Starting data.—We downloaded the 79 clusters from the PhyLoTA GenBank browser release 1.01 (which uses data from GenBank release 159). To these, we added 14 additional clusters of newly released data (Praschag et al. 2007; Thomson et al. 2008; Spinks et al. 2009). We opted not to include the 19 existing turtle mitochondrial genome sequences because they largely duplicate existing GenBank data, adding little new information. Combining all the mitochondrial clusters into one larger cluster (that included the mtDNA genomes) would alleviate this data duplication problem at the expense

of greatly increased data heterogeneity, a decrease in alignment quality, and several technical problems associated with handling extremely long sequences. In addition, the length of the sequences was problematic for much of the automation and the local alignment algorithm, primarily by causing excessive memory demands. For similar reasons, we also removed 1 cluster from the analysis that was composed of a large piece of the mitochondrial genome including cytochrome *b*, the mtDNA control region, and several transfer RNAs. The 93 clusters contained 6331 sequences. We filtered sequences so that each cluster contained only 1 sequence per species, keeping the longest sequence for each species when duplicates existed. After filtering, the clusters contained 1096 sequences. Most of the duplication was due to the presence of large population-level sampling for a few species.

Taxon and character sampling was heterogeneous among higher level turtle clades (Table 1). Overall, the tortoises (Testudinidae) and Old World pond turtles (Geoemydidae) had the most complete taxon sampling, whereas the New World pond turtles (Emydidae) had the most extensive character sampling. The side-necked turtles (Pleurodira) and the mud turtles (Kinosternidae) were characterized by both poor taxon and character sampling.

Taxonomy.—Overall, 50 unique corrections were required to standardize the NCBI names with the updated names, and most corrections were applied several times. If we think of this as the number of unique changes required to bring the NCBI taxonomy into line with the current taxonomy, then 15.6% of species labels (50 changes/321 species) are incorrect in GenBank. Of these, 37 were the result of taxonomic changes at the species or genus level, 7 were the result of spelling errors when sequences were submitted to GenBank, and 6 were attributable to the submission of sequences for "species" that are currently viewed as invalid hybrids (Parham et al. 2001; Stuart and Parham 2007). In total, 187 of the 1096 sequences (17.1%) required changes to standardize them with the current taxonomy.

Alignment and secondary clustering.—MOS scores comparing the DIALIGN-only, MUSCLE-only, and the MUSCLE-refined DIALIGN alignments gave the same score for all 3 alignments in 48 of the 93 clusters, indicating no alignment problems. These scores ranged (on a 0–1 scale) from 0.67 (in a single cluster) to 1.00 (in 44 clusters) and were higher than 0.80 (the recommended cutoff for reasonable alignment quality) for all but one cluster. When clusters received different scores, they were usually very similar (within 0.01 or 0.02), only 6 clusters showed differences between alignment methods >0.02. In all cases, the MUSCLE-refined DIALIGN alignments were the best, or one of the best, alignments, so this strategy was used for the remainder of the study. Gblocking the primary clusters resulted in the removal of between 0% and 84% of the alignment depending on the cluster. Eight clusters had no data removed, whereas 9 had >50% of the alignment removed (mean = 24%; median = 13%). In a few cases, Gblocks removed what appeared to be well-aligned informative data, though this tended to occur in smaller clusters or where sequences only partially overlapped; such clusters tend to contribute little phylogenetic information.

Secondary clustering yielded 5 secondary clusters resulting from the combination of 12 total primary clusters (1 secondary cluster composed of 4 primaries and 4 composed of 2 primaries). We attempted to combine a few more clusters based on information contained in the definition line but rejected these alignments as unreasonable. The most problematic case was the mitochondrial control region clusters, where high molecular

TABLE 1. Taxon and character sampling for each family of turtles in the supermatrix

Family (total no. of species)	Taxon sampling[a]			Character sampling[b]		
	Nuclear	Mitochondrial	Total	Nuclear	Mitochondrial	Total
Cryptodira (235)						
Carettochelyidae (1)	1.00	1.00	1.00	6009	2260	8269
Cheloniidae (6)	0.33	1.00	1.00	3673	2624	6297
Chelydridae (4)	0.50	0.50	0.50	3647	2398	6045
Dermochelyidae (1)	1.00	1.00	1.00	3720	1973	5693
Dermatemydidae (1)	1.00	1.00	1.00	3733	1521	5254
Emydidae (48)	0.48	0.56	0.67	**6275**	1627	**7902**
Geoemydidae (64)	0.67	0.91	0.92	3297	2080	5377
Kinosternidae (25)	0.20	0.40	0.36	1669	1164	2833
Platysternidae (1)	1.00	1.00	1.00	7202	1279	8481
Testudinidae (54)	0.63	**0.93**	**0.93**	2444	**2222**	4666
Trionychidae (30)	**0.77**	0.80	0.80	1720	1965	3685
Pleurodira (86)						
Chelidae (59)	0.27	0.41	0.44	1739	1348	3087
Pelomedusidae (19)	0.21	0.16	0.26	2498	1777	4275
Podocnemidae (8)	0.50	0.50	0.63	2263	1287	3550
Total (321)	0.50	0.66	0.69			

Note: The highest nontrivial value is highlighted in bold in each column.

[a]Proportion of species sampled for each family. Includes rogue taxa that were pruned from the final tree.

[b]Total base pairs per family divided by number of taxa sampled per family. This removes the effect of varying taxon sampling between families.

divergence made alignments problematic. We only combined clusters when the alignments appeared reasonable across the length of the sequence included. For protein-coding regions of the genome, it may be possible to increase the accuracy of alignments further by employing a translation/amino acid alignment step.

As a final quality control, we examined all primary and secondary alignments by eye to ensure that our automated strategy was working as expected. We made minor adjustments to 9 of the 84 total clusters, and more substantial adjustment to a single length heterogeneous cluster, where the misaligned region was associated with the end of a single short sequence.

Matrices.—Changing the degree of taxonomic overlap between clusters had little effect on the resulting supermatrices. The minimum 3-taxon overlap allowed all clusters to be included in the matrix, including an additional 3 taxa that were excluded with greater overlap levels. However, these 3 extra taxa behaved as some of the worst rogues in the resulting trees and were eventually pruned from the matrix. Requiring 4-, 5-, or 6-taxon overlap included the same set of 223 taxa. Requiring greater taxon overlap excluded an increasing number of the smaller clusters, reducing both the total amount of data and the overall proportion of missing data (Table 2). Because the 4-taxon overlap matrix included the most data and performed well, we used it for the rest of this study. All matrices were very sparse (between 7.5% and 10% full; Fig. 2 and Table 2). The matrices constructed from Gblocks clusters and the nuDNA-only clusters contained fewer taxa and characters than the matrix based on all the data (Table 2). The matrices were deposited in TreeBASE (accession #S2504).

Phylogenetic Analysis

Rogue taxa.—Our first parsimony ratchet and bootstrap runs produced unresolved consensus trees (Table 3). Using the taxon instability index, we identified the least stable (rogue) taxa and found that these were generally taxa with very little data (Fig. 3). We pruned the top scoring rogues from each of the collections of trees (all clusters/4-taxon overlap, Gblocks/4-taxon overlap, and nuDNA-only/4-taxon overlap) and recalculated the reduced consensus trees after each taxon was removed to examine the improvement in tree topology. In the 4-taxon overlap supermatrix containing all data, pruning the 9 least stable taxa resulted in substantial increases in resolution outside of their local placement in the tree. For the 10th and 11th scoring taxa, although their placement in the tree seemed erroneous, removing them did not result in improvements to the topology and so they were left in the trees. We continued this procedure, excluding taxa and checking the effect on the reduced consensus tree and found one more taxon after the 10th and 11th that behaved as a rogue (*Cylindraspis triserrata*, whose removal brings several genera of tortoises into monophyly). In all, we pruned 10 taxa from the supermatrix containing all data and the Gblocks supermatrix (9 of 10 of which were the same taxa in both matrices). Pruning the 5% least stable taxa in the data set would have resulted in the removal of 12 taxa, including the 10 taxa that we removed and the 10th and 11th scoring taxa (*Pseudemys peninsularis* and *Kinosternon subrubrum*, respectively) that we opted not to remove.

Identifying rogue taxa in the nuDNA-only supermatrix was more problematic because there was no clear demarcation between rogues and nonrogues. The nuDNA-only data set had a broader distribution and higher average instability scores than the other 2 matrices (Fig. 3). Most of the taxa that were pruned in the first 2 data sets were not present in this data set (because they only had mtDNA data) but those that were remained rogues. The rest of the taxa were less obvious. Without a clear tail in the distribution of instability scores, we sequentially pruned species, eventually pruning 20% of the total taxa.

Tree searches.—After identifying rogue taxa with the reduced consensus trees, we removed them from the data sets and ran a more thorough set of tree searches. Parsimony ratchet and heuristic searches all produced identically scoring trees (parsimony score for all clusters, Gblocks clusters, and nuDNA clusters: 30117, 21533, and 6955, respectively), suggesting that the parsimony searches were effective in finding optimal trees.

These searches produced strikingly better resolved consensus trees than the analyses including rogue taxa (Table 3 and Fig. 4). The nuDNA-only matrix performed relatively poorly, suggesting that either rogue taxa still exist in the data set that we failed to remove (specifically, in the clade of tortoises and geoemydids, Fig. 4c) or too little nuclear data exist for this approach to be effective. Overall, the matrix using all data performed best, as measured by its ability to unambiguously support a reasonable topology and its higher consensus fork index (Colless 1980) compared with the Gblocks or nuDNA-only data set.

TABLE 2. Summary of each of the 3 main supermatrix data sets

	3-Taxon overlap	4-Taxon overlap			5-Taxon overlap	6-Taxon overlap
		All clusters	Gblocks clusters	nuDNA clusters		
Taxa	226	223	217	160	223	223
Base pairs	53,898	53,406	40,891	41,845	51,212	37,686
Total clusters	91	90	90	76	85	66
Secondary clusters	84	83	84	72	78	59
% missing data	92.56	92.41	91.6	92.53	92.14	90.06

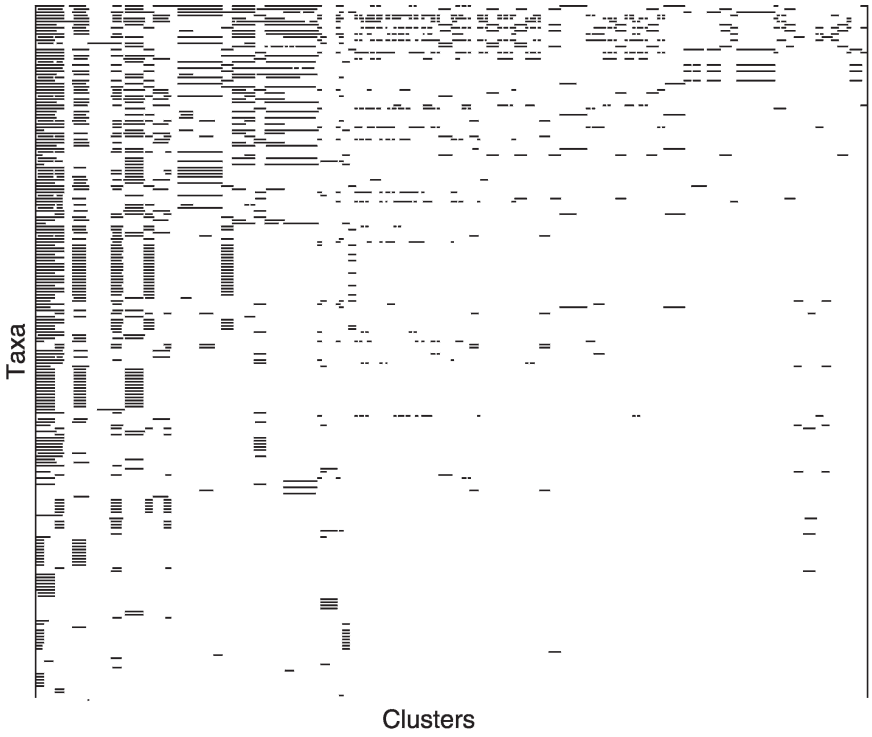

FIGURE 2. A schematic overview of data density in the all data/4-taxon overlap supermatrix. Taxa are ordered vertically (1 taxon per row) from highest data density to lowest and clusters ordered left to right from highest taxon sampling to lowest. Each pixel in the chart represents a 100 bp bin for a single taxon. Filled pixels have data for that bin. White areas are missing data.

The ML analysis recovered a tree that was broadly congruent with our MP tree (Fig. S1, available from http://www.sysbio.oxfordjournals.org/) but differed in important respects. First, the bootstrap values for the ML tree were far higher than for the MP tree, with 116 nodes receiving >95% support for ML analysis versus only 66 nodes for the MP. Both strategies recovered all families as monophyletic and essentially agreed on interfamilial relationships (the placement of Chelydridae was the sole exception). The topological differences that did exist tended to be relatively deep within families, particularly within the Testudinidae and the Geoemydidae.

After 30 million generations, the Bayesian runs had not converged using default settings for the Markov chain Monte Carlo (MCMC). This has been observed in other data sets, particularly those containing large amounts of missing data (Smith and Donaghue 2008). Though this lack of convergence could likely be surmounted by tuning the parameters for the MCMC, we were concerned that bias due to missing data was particularly problematic for Bayesian analyses (more so than for ML, see Discussion section) and—given these concerns—we did not pursue these analyses further.

We also compared our MP and ML trees with MRP and GTP supertrees. Our MRP tree search recovered

TABLE 3. Summary of trees reconstructed from each of 3 main supermatrix data sets

	All clusters		Gblocks clusters		nuDNA clusters	
	With rogues	After pruning	With rogues	After pruning	With rogues	After pruning[c]
Trees[a]						
CFI[b] strict	0.00	0.819	0.093	0.765	0.089	0.408
CFI 95% MJR	0.695	0.848	0.738	0.819	0.713	0.720
Bootstrap support						
CFI 95% MJR	0.241	0.319	0.187	0.230	0.057	0.056
CFI 70% MJR	0.491	0.629	0.411	0.490	0.325	0.408
CFI 50% MJR	0.714	0.805	0.603	0.686	0.497	0.664

Notes: CFI = consensus fork index; MJR = majority rule consensus.
[a]Trees used were a set of 10,000 maximum parsimony trees found during parsimony ratchet runs followed by TBR branch swapping.
[b]Calculated as the number of resolved nodes divided by the total number of nodes possible.
[c]The 32 most unstable taxa were pruned from this set of trees, and 10 taxa were pruned in the other 2 cases.

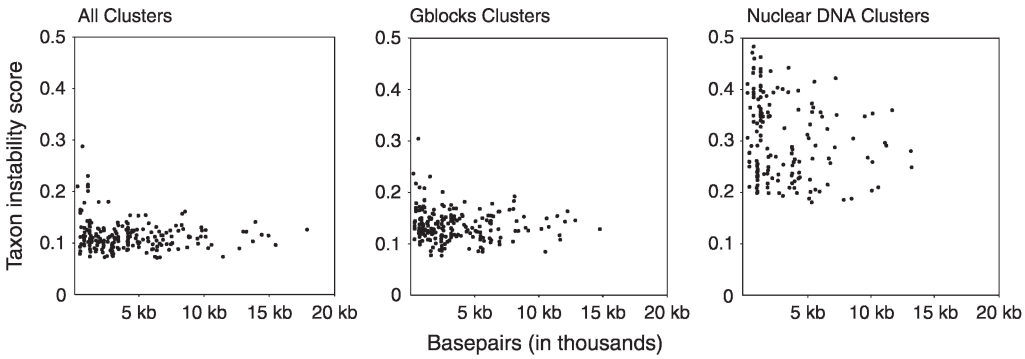

FIGURE 3. Amount of data plotted against instability score for each of the 3 supermatrices. Each point represents a single species, with its instability score for a collection of 100 bootstrap trees on the *y*-axis and the number of base pairs (in kilo base pairs) on the *x*-axis. Instability scores have been divided by the number of taxon comparisons in order to standardize them across data sets with varying taxon sampling.

several thousand equally parsimonious trees (score = 1345), and their strict consensus was poorly resolved (Fig. S2). Several families were not recovered as monophyletic, and some taxa had highly suspect placements. The GTP analyses also performed poorly for this data set. When no starting tree was supplied, the program recovered phylogenies that placed several taxa in clearly erroneous regions of the tree, far from the expected placement (Fig. S3) (score = 336). The trees improved somewhat when our MP consensus tree was supplied as a starting tree. However, the trees still contained many unexpected results (e.g., *Emys* was nonmonophyletic, *Rhinoclemmys* nested deep inside of the Geoemydidae) that are at odds with more traditional molecular analyses (Spinks et al. 2004, 2009), and we suspect that there were too few data to converge to a meaningful solution.

DISCUSSION

Phylogenetic trees based on sparse supermatrices can produce important summaries of existing data and, at least occasionally, novel phylogenetic insights. Building on important initial attempts to automate and streamline supermatrix-based analyses (McMahon and Sanderson 2006), we focused our methodological attention on developing an informatics pipeline that deals with inconsistent/changing taxonomy, difficulties associated with alignment of heterogeneous sequences, and rogue taxa to streamline the construction of well-supported phylogenies from sparse supermatrices. By confronting these issues simultaneously, our pipeline successfully extracted a very well-resolved sparse-matrix phylogeny for the world's turtle fauna. Methodologically, our study demonstrates that failure to correct for rogue taxa and inconsistent taxonomy can mask the latent phylogenetic signal in a sparse data set.

Informatics

Our informatics pipeline automates most stages of the data curation and assembly in going from sequence data to a "finished" matrix. Exceptions include the initial clustering steps (available for most clades from the PhyLoTA browser; Sanderson et al. 2008), secondary clustering, and the taxonomic correction step (although each unique correction must be manually handled only once). Our results suggest that we can achieve a large degree of automation without trading off phylogenetic accuracy.

Clustering.—The primary clusters available from the PhyLoTA browser (that we updated with more recent sequences) negated the need to reextract all sequences of interest from GenBank and perform primary clustering. Secondary clustering is an important step in the pipeline that we handled manually. McMahon and Sanderson (2006) present an automated strategy relying on relaxed BLAST clustering parameters to try to force clusters together. In principle, the McMahon and Sanderson method could be seamlessly integrated into our pipeline; in practice, the extent to which an automated approach can effectively cluster divergent homologous sequences remains an area for future work.

Taxonomy.—To date, studies that have attempted to mine sequence data from large genetic databases have simply used the taxonomy employed in that database, even though it is clear that the errors present in these taxonomies negatively impact the resulting trees. This problem has been recognized and discussed (Page 2004), but to our knowledge has yet to be satisfyingly resolved. Our strategy of updating the entire taxonomy to current names worked well and is easily extended to other clades. The relatively small number of turtle species, combined with the availability of a recent taxonomic list (Turtle Taxonomy Working Group 2007) simplified this task to some degree, although the strategy that we developed is scalable to arbitrarily sized clades. The value of our approach hinges on the frequency distribution of taxonomic errors in the database; if most are single-occurrence errors, then our approach reduces to hand curation. However, given the very high frequency

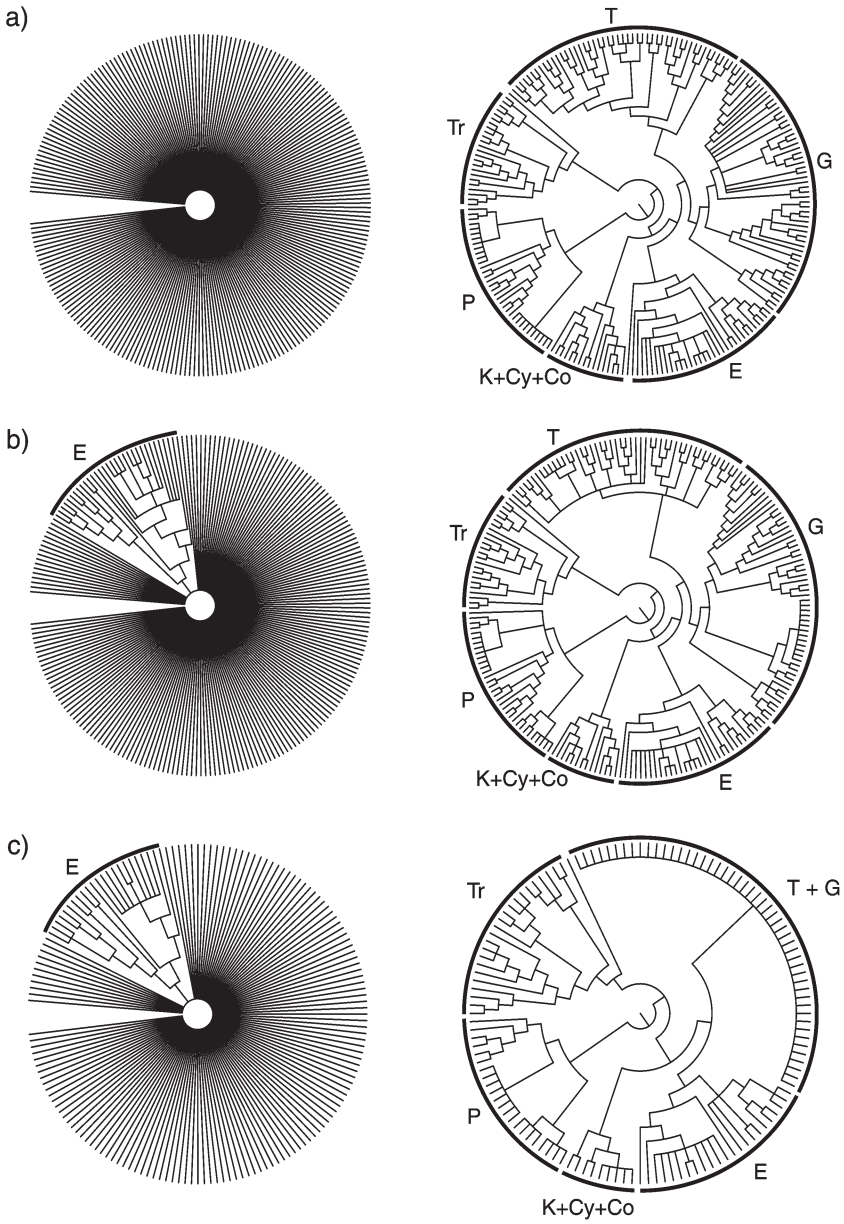

FIGURE 4. Strict consensus trees of all maximally parsimonious topologies from parsimony ratchet runs comparing pre-rogue-pruning data sets (left column) to post-rogue-pruning data sets (right column). a) supermatrix constructed from all data, b) supermatrix constructed from Gblocks clusters, c) supermatrix constructed from nuDNA only. Major clades of turtle denoted with bars if they appear in the trees; Co—Cheloniidae, Cy—Chelydridae, E—Emydidae, G—Geoemydidae, K—Kinosternidae, P— Pleurodira, T—Testudinidae, and Tr—Trionychidae. K + Cy + Co clade also contains the monotypic Dermatemydidae and Dermochelyidae.

(15–20%) of taxonomic errors that we discovered, even for a well-known vertebrate clade like turtles, our approach is a reasonable compromise between automation and hand curation.

As efforts to build linkages among the many large biological databases continue, we expect nomenclatural inconsistency to become an increasingly large problem faced by the biodiversity informatics community. This is particularly true for clades in which taxonomy itself is changing rapidly in the face of new phylogenetic data and conflicting views on taxonomic philosophies. An invaluable tool would be a mature database of taxonomic synonymies that could automatically update and synonymize such taxonomic inconsistencies (e.g., http://www.ubio.org).

Alignment.—Accurate alignment of length heterogeneous sequences is clearly a significant challenge for the supermatrix approach. We outlined a novel strategy that attempts to account for both local and global alignment problems, eliminating some of the manual adjustments

that have previously been required (McMahon and Sanderson 2006).

In practice, our strategy appeared to work well. We performed manual adjustments in only 10 of the 83 total clusters, and in most cases, these were minor adjustments involving only a handful of misaligned base pairs (the largest involved ∼30 bp). In a 50,000+ bp data matrix, this seems unlikely to cause significant problems, and comparisons of output trees based on matrices without manual corrections to the postediting trees show no apparent differences. Similarly, the Gblocks alignments produced a reasonable tree that closely matched the topology found using the entire data set, particularly after rogue taxa were removed (cf. Fig. 4a,b).

Our alignment results argue that a useful starting point for this type of study is an automated initial alignment, incorporating local and global algorithms, and using MOS scores for consistency checks. These automated alignments will likely contain some small misaligned regions that could be corrected by hand, though the potential for improvement to the resulting topology from these corrections is likely to be small.

Cluster combinability.—Our results suggest that the degree of taxonomic overlap between clusters is not a critical factor in supermatrix construction. We excluded over a third of the least overlapping clusters in our most stringent, 6-taxon overlap matrices, and the resulting trees were similar in resolution to those based on 4 and 5-taxon overlap. Although this result may not be a general feature of all data sets, it appears that the 4-taxon overlap suggested by McMahon and Sanderson (2006) is a reasonable starting point for cluster combinability.

Rogue taxa.—Rogue taxa probably represent the most insidious problem for supermatrix phylogenetics. Even a few rogue taxa can lead to the complete collapse of a tree, obscuring the otherwise considerable phylogenetic content of a data set (compare the left and right columns of Fig. 4). Ideally, we would like to be able to identify and eliminate the rogue problem before, rather than after, phylogenetic analysis. However, doing so requires identifying unique nonphylogenetic features of rogues, and we have thus far been unable to do so. Rogues appear to be unstable because of the absolute amount of data and their specific pattern of data overlap within and between clusters. For example, 2 taxa that are distantly related and overlap by a certain number of base pairs in a certain set of clusters may be well behaved because they have sufficient overlap with closer relatives that control their appropriate placement in the tree. On the other hand, 2 taxa with the exact same pattern of overlap that are closely related may be much more problematic because fewer closely related taxa are likely to exist that can "pull" them into the correct placement on the tree. Without reference to overlap between "closely related" individuals (which requires some prior knowledge of phylogeny), it is difficult to decide which taxa will behave as rogues, leading us to the less efficient, but operational approach that we developed for rogue identification. Visually checking the effect of each potential rogue's removal on the tree was a time-consuming step. In hindsight, inspecting the distribution of instability scores (Fig. 3) and drawing a cutoff for rogue removal would have resulted in essentially the same tree, allowing for essentially complete automation of this important step. A histogram of instability scores could be compiled, allowing the user to choose a cutoff based on a gap in the distribution, as occurred for our full set of clusters at about a score of 0.175 (Fig. 3).

Tree searches.—Because analysis strategies for sparse supermatrices are still poorly studied, we employed several approaches to infer a phylogeny from our supermatrices. Concerns over missing data have been discussed extensively in the literature, with some authors advocating the inclusion of all available data (Wiens 2006) and others finding that the negative consequences of extensive missing data should be avoided (Dunn et al. 2003). Recent work clarifies and highlights several issues that can arise as a result of missing data, particularly as they relate to ML and Bayesian phylogenetics (Lemmon et al. 2009). For phylogenetic analyses of complete matrices, the attractive statistical properties of ML and Bayesian approaches makes them the method of choice. However, these properties depend on a reasonably accurate model of the data. When missing data are introduced into the data set, obtaining accurate parameter estimates for the model becomes increasingly problematic (Lemmon et al. 2009). Because the implicit model employed by parsimony is fixed (not parameterized based on the data), it does not suffer from this inability to correctly parameterize the model.

Here, we have focused primarily on parsimony. Although we acknowledge that the parsimony model often does not fit molecular data particularly well, there are no other models that fit these data particular well either. In most ways, the topologies produced by MP and ML were similar, differing primarily in regions of the tree that were poorly supported by both (cf. Figs. 5 and S1). However, MP returned generally more conservative bootstrap values (66 vs. 116 nodes supported at 95% or greater). It is possible that introducing complex partitioning schemes could lessen the effect of a missing data bias, although this would need to be balanced against the paucity of data available for each partition. In this case, we favour the more conservative parsimony estimates, at least until more work is done on the impact of extreme levels of missing data on phylogeny estimation.

Our supertree-based approaches met with limited success. The MRP analysis produced a largely unresolved strict consensus of many thousands of equally parsimonious supertrees. These results are in line with a previous supertree analysis of turtles (Iverson et al. 2007), although that study was missing some heavily

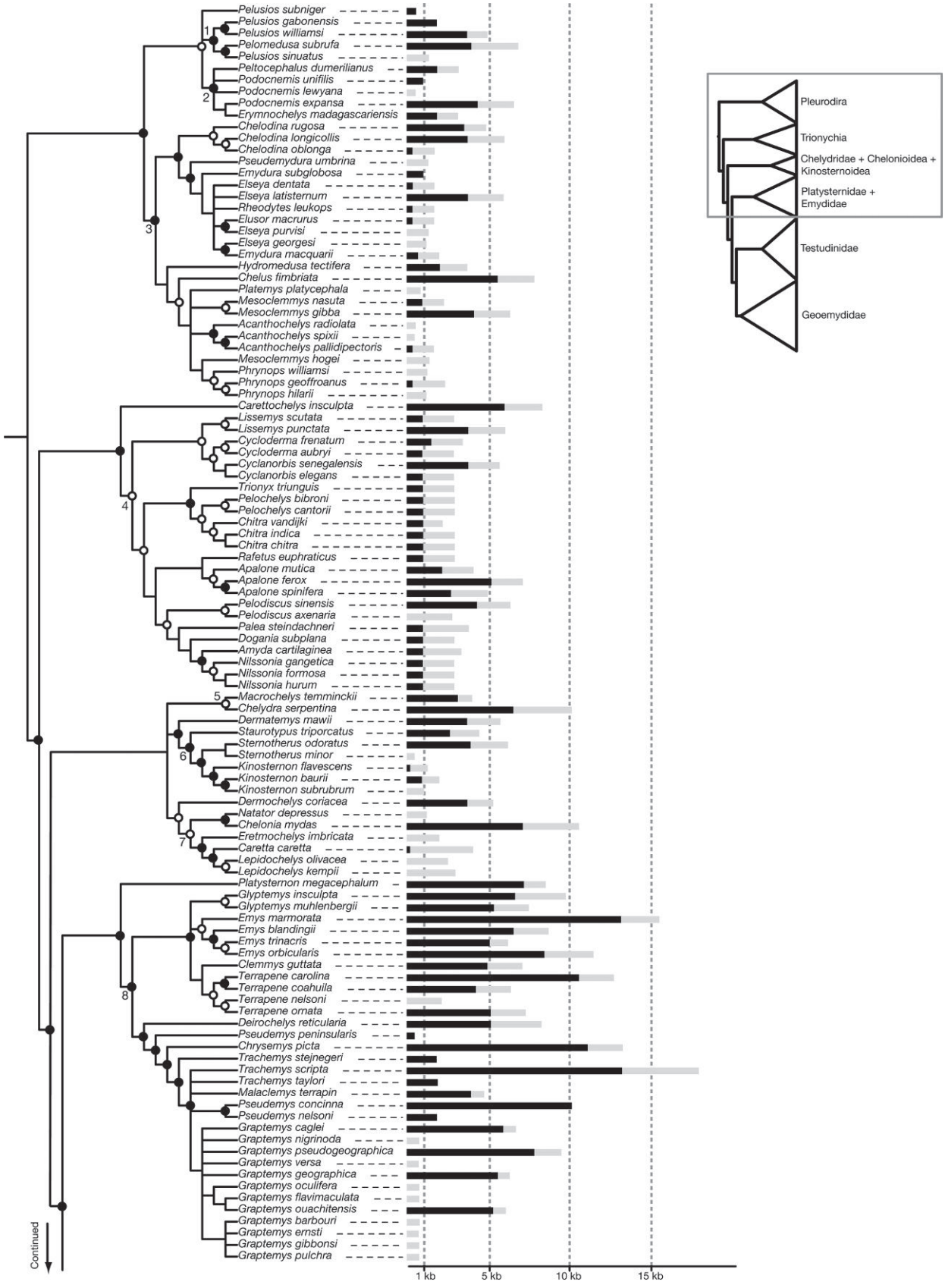

FIGURE 5. (Continued)

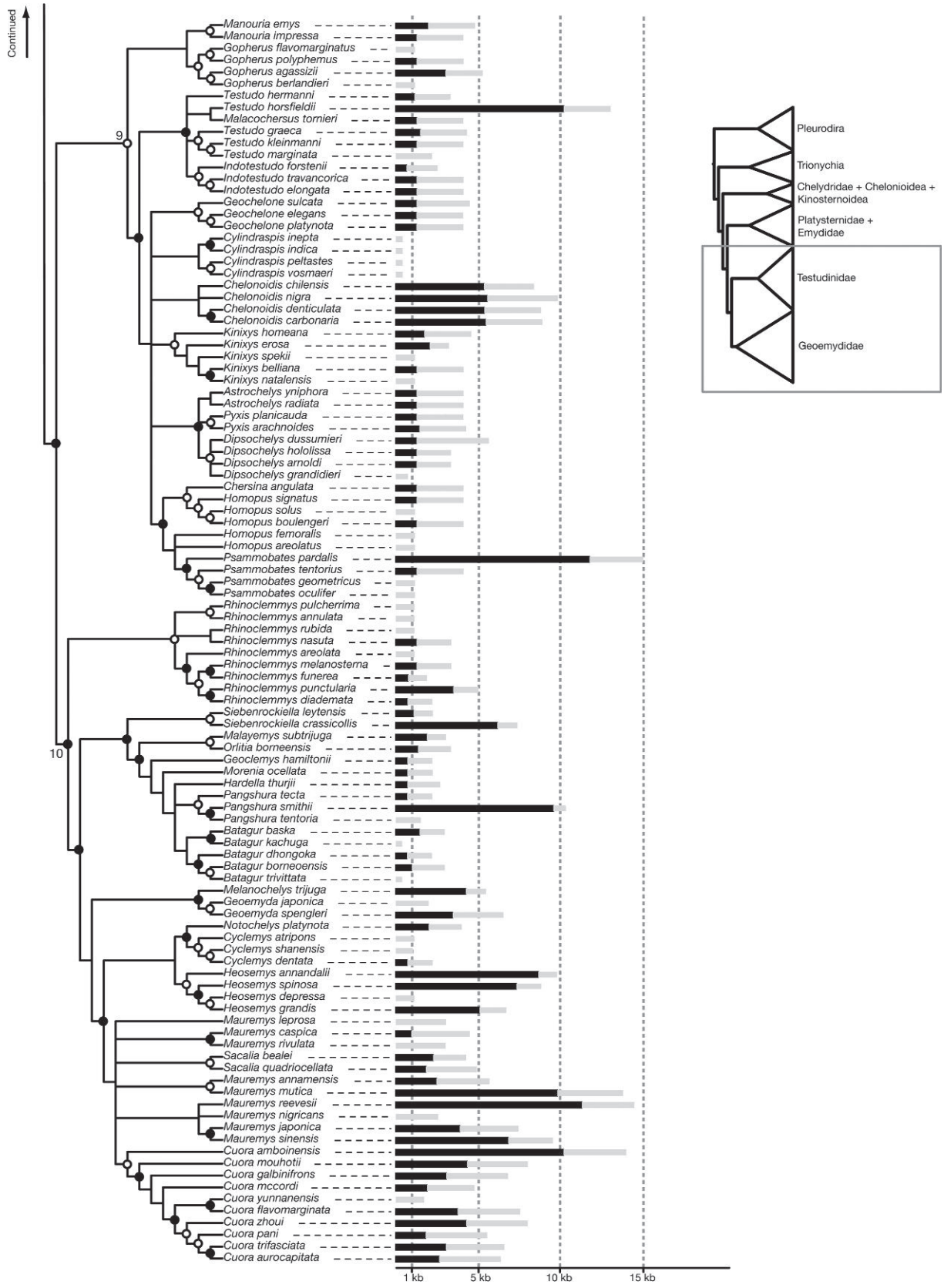

FIGURE 5. (Continued)

sampled phylogenies that have recently become available. Bininda-Emonds and Sanderson (2001) showed that when large sets of highly nonoverlapping trees are combined, the performance of MRP degrades relative to the supermatrix approach. Our analysis, and that of Iverson et al. (2007), both use more trees that share fewer taxa than the cases simulated by Bininda-Emonds and Sanderson (2001), suggesting that the poor supertree performance we observed may stem from this problem. Alternative MRP approaches, such as weighted MRP, may improve resolution by more heavily weighting the portions of the source trees that are strongly supported (Bininda-Emonds and Sanderson 2001), although the Iverson et al. (2007) study attempted this and still recovered an unresolved supertree. Our GTP analysis also failed to produce a well-resolved supertree (Fig. S3), suggesting that GTP does not perform well in the face of large amounts of missing data. An attractive property of GTP is that it attempts to deal with incongruity among gene trees. For data sets where this is a concern, an alternative strategy would be to implement a paralogy test in the early data filtering stages of the analysis as was done in McMahon and Sanderson (2006). Although GTP always produces a fully resolved tree, we found the topology of the best species tree to be riddled with obviously erroneous placements of some taxa compared with the supermatrix results and with the published literature on turtle phylogenetics (Fig. S3). As in the MRP case, we expect that this effect is largely due to nonoverlapping taxa among the input trees. Edwards (2009) argues that species tree approaches are likely to be affected by missing data more strongly than supermatrix approaches because in the former, missing data implies that the genealogy for certain taxa is entirely unknown for certain "genes" (in our case, clusters). However, in supermatrix analyses, the signal present in other genes provides at least some estimate of the genealogy and can help make up for the missing data. Taken together, our results suggest that supermatrix approaches are potentially more powerful for extracting phylogenetic information from very sparse data sets.

Turtle Phylogeny

Because this analysis essentially synthesizes data from existing studies, we should reasonably expect that the tree recovered here (Fig. 5) closely matches the trees in those previous studies. This appears to be the case. At the deepest levels, both Pleurodira and Cryptodira are well supported, and with the one exception of the poorly sampled Pelomedusidae, all currently recognized families are monophyletic. These results have been found in several previous studies on different families, though few studies include enough outgroups to rigorously test monophyly (Dutton et al. 1996; Spinks et al. 2004; Le et al. 2006). In addition, the relationships between families are consistent with current understanding (Shaffer et al. 1997; Fujita et al. 2004; Krenz et al. 2005; Parham et al. 2006; Thomson et al. 2008). Within families, many genera are monophyletic. In turtles, generic-level taxonomy is undergoing many changes as nonmonophyletic groups continue to be discovered and revised (Turtle Taxonomy Working Group 2007). Thus, some of the nonmonophyly is likely real and reflects inappropriate taxonomy (the tortoise genus *Testudo* is a potential case), whereas much of it is likely due to insufficient data and/or a history of hybridization (e.g., the emydid genus *Pseudemys*).

The most problematic parts of our tree are also those relationships that have never been convincingly resolved in the literature. For instance, the relationships between the snapping turtles (Chelydridae), sea turtles (Chelonoidea), and mud turtles (Kinosternoidea) are unresolved in our MP tree and in other recent analyses (Krenz et al. 2005). Other poorly resolved parts of our tree include the relationships among the 3 most diverse genera of the Emydidae (*Pseudemys*, *Trachemys*, *Graptemys*) and those among the Australian "short-necked" chelids (particularly *Emydura* and *Elseya*). In these cases, the existing data (Lamb et al. 1994; Stephens and Wiens 2003) suggest that these may be rapid recent radiations that will be difficult to resolve using standard phylogenetic methods.

One of the key controversies in turtle systematics is the phylogenetic position of the big-headed turtle, *Platysternon megacephalum*. Initial morphological and mtDNA assessments tentatively placed it sister to the snapping turtles (family Chelydridae) and found that its phylogenetic placement depended on the specific combination of data sets employed (Shaffer et al. 2008). Later studies using nuDNA and larger pieces of mtDNA unambiguously place *Platysternon* in, or sister to, the Testudinoidea (Emydidae + Geoemydidae + Testudinidae), though the specific position varied depending on the study. Cervelli et al. (2003) and Krenz et al. (2005) placed *Platysternon* in a polytomy with the other major clades of Testudinoidea using U17 snoRNA and RAG-1, respectively. Parham (2006) employed whole mtDNA genomes to examine the relationship of *Platysternon* and found a sister relationship between Emydidae and *Platysternon* with strong support. Our study recovered this same relationship with weak support (Figs. 5 and S1), but only for the full supermatrix after pruning rogues (*Platysternon* was pruned in both the Gblocks

FIGURE 5. Majority rule consensus of 100 bootstrap replicates for the full supermatrix after pruning rogues. Open circles denote bootstrap proportion >95 and closed circles represent bootstrap proportions >70. Histograms denote the total amount of data in kilo base pairs for each taxon, with each bar coloured proportional to the amount of each data type; black—nuDNA, gray—mitochondrial DNA. Numerals on nodes refer to nonmonotypic families: 1—Pelomedusidae, 2—Podocnemidae, 3—Chelidae, 4—Trionychidae, 5—Chelydridae, 6—Kinosternidae, 7—Cheloniidae, 8—Emydidae, 9—Testudinidae, and 10—Geoemydidae.

and the nuclear-only trees due to instability and is not included in Fig. 4b,c). Because we avoided including the mtDNA genomes in our study, this resolution does not simply reflect an overwhelming influence of a single large mitochondrial data set and so can be viewed as an independent confirmation of this result. This conclusion is bolstered by the strong taxon and character sampling present across the Testudinoidea (Table 1). We do, however, note that the internal branches resolving the major lineages of the Testudinoidea are short, and thus, incongruity among genes and anomalous gene trees may remain a concern (Degnan and Rosenberg 2006).

The fact that the nuDNA-only supermatrix produced very little resolution despite our aggressive efforts to remove rogue taxa suggest that there are not yet enough nuclear data to recover a well-supported tree for turtles. This implies that much of what we know about the turtle tree of life depends on mitochondrial data. The problems associated with inferring phylogeny from single genes are well known (Funk and Omland 2003; Ballard and Rand 2005) and argues that we need more extensive nuclear gene sampling for turtles. It also emphasizes that even "multilocus" datasets are often dominated by mitochondrial (or chloroplast) effects, as a comparison of Figure 4a,c shows. Given the extremely dire conservation status of the world's turtle and tortoise fauna (Shaffer et al. 2007; IUCN 2008) and the potential for sound phylogenies to contribute to conservation prioritization strategies, there is a strong need for developing a comprehensive multilocus phylogeny for Testudines.

Conclusions

Whether viewed as comprehensive summaries of available information or as a distinct phylogenetic strategy, sparse supermatrices clearly have a role to play in helping guide large multimarker projects on the tree of life. Certain informatic challenges, including taxonomic errors and identification of rogue taxa, will plague any attempts to use large multi-user databases in the construction of phylogenies; and automated solutions to these challenges are necessary to streamline the use of these data.

Eventually, we hope that sufficient data will be available that phylogenies based on nearly complete matrices for large sets of taxa will be possible. In the meantime, it does seem that even extremely sparse supermatrices perform well. These approaches have the potential to allow for the identification of key sequencing to be performed to reduce fragmentation and produce increasingly dense data matrices. Likewise, supermatrices can help make clear where taxon sampling is most deficient. Finally, certain parts of a tree may have strong character and taxon sampling and yet remain largely unresolved, indicating that traditional phylogenetic efforts are failing. By examining characteristics of both the data sets and the trees, supermatrix approaches will likely serve as useful guides for maximizing the benefit of future research efforts.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

This work was supported by a National Science Foundation Doctoral Dissertation Improvement Grant (DEB-0710380), a Society of Systematic Biologists Graduate Student Research Award and funding from the UC Davis Center for Population Biology (to R.C.T.), grants from the National Science Foundation (DEB-0507916, DEB-0213155, and DEB-0817042 to H.B.S.), and the UC Davis Agricultural Experiment Station.

Acknowledgments

We thank Phil Spinks, Mike Sanderson, and anonymous reviewers for comments on an earlier version of this manuscript. The study was substantially improved through discussions with Levi Gray, Ian Wang, the members of a Monte Carlo discussion group at UC Davis, and the Shaffer Lab discussion group.

References

Ballard J.W.O., Rand D.M. 2005. The population biology of mitochondrial DNA and its phylogenetic implications. Annu. Rev. Ecol. Evol. Syst. 36:621–642.

Bansal M.S., Burleigh J.G., Eulenstein O., Wehe A. 2007. Heuristics for the gene-duplication problem: a $\Theta(n)$ speed-up for the local search? Lect. Notes Comput. Sci. 4453:238–252.

Baum B.R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon. 41:3–10.

Benson D., Karsch-Mizrachi I., Lipman D., Ostell J., Wheeler D. 2007. GenBank. Nucleic Acids Res. 35:D21.

Bininda-Emonds O.R.P., editor. 2004. Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic Publishing.

Bininda-Emonds O.R.P., Cardillo M., Jones K.E., MacPhee R.D.E., Beck R.M.D., Grenyer R., Price S.A., Vos R.A., Gittleman J.L., Purvis A. 2007. The delayed rise of present-day mammals. Nature. 446: 507–512.

Bininda-Emonds O.R.P., Jones K.E., Price S.A., Grenyer R., Cardillo M., Habib M., Purvis A., Gittleman J.L. 2003. Supertrees are a necessary not-so-evil: a comment on Gatesy et al. Syst. Biol. 52:724–729.

Bininda-Emonds O.R.P., Sanderson M. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst. Biol. 50:565–579.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540–552.

Cervelli M., Oliverio M., Bellini A., Bologna M., Cecconi F., Mariottini P. 2003. Structural and sequence evolution of U17 small nucleolar RNA (snoRNA) and its phylogenetic congruence in chelonians. J. Mol. Evol. 57:73–84.

Colless D.H. 1980. Congruence between morphometric and allozyme data for Menidia species: a reappraisal. Syst. Zool. 29:288–299.

Cracraft J., Donoghue M.J., editors. 2004. Assembling the tree of life. Oxford: Oxford University Press.

de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22:34–41.
Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2. Available from: http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020068
Driskell A.C., Ane C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science. 306:1172–1174.
Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sorensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.
Dunn K., McEachran J., Honeycutt R. 2003. Molecular phylogenetics of myliobatiform fishes (Chondrichthyes: Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood. Mol. Phylogenet. Evol. 27:259–270.
Dutton P.H., Davis S.K., Guerra T., Owens D. 1996. Molecular phylogeny for marine turtles based on sequences of the ND4-Leucine tRNA and control regions of mitochondrial DNA. Mol. Phylogenet. Evol. 5:511–521.
Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.
Fujita M.K., Engstrom T.N., Starkey D.E., Shaffer H.B. 2004. Turtle phylogeny: insights from a novel nuclear intron. Mol. Phylogenet. Evol. 31:1031–1040.
Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.
Gaffney E.S. 1990. The comparative osteology of the triassic turtle *Proganochelys*. Bull. Am. Mus. Nat. Hist. 194:1–263.
Gatesy J., Baker R.H., Hayashi C. 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. Syst. Biol. 53:342–355.
Gatesy J., Matthee C., DeSalle R., Hayashi C. 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51:652–664.
Hodkinson T.R., Parnell J.A.N., editors. 2006. Reconstructing the tree of life: taxonomy and systematics of species rich taxa. Boca Raton (FL): CRC Press.
Huang X., Miller W. 1991. A time-efficient linear-space local similarity algorithm. Adv. Appl. Math. 12:337–357.
IUCN 2008. IUCN Red List of Threatened Species. Version 2008.1. Available from: http://www.iucnredlist.org.
Iverson J.B., Brown R.M., Akre T.M., Near T.J., Le M., Thomson R.C., Starkey D.E. 2007. In search of the tree of life of turtles. In: Shaffer H.B., FitzSimmons N.N., Georges A., Rhodin A.G.J., editors. Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises. Cambridge (MA). Lunenburg (MA): Chelonian Research Foundation. p. 85–106 (Chelonian research monographs).
Krenz J.G., Naylor G.J.P., Shaffer H.B., Janzen F.J. 2005. Molecular phylogenetics and evolution of turtles. Mol. Phylogenet. Evol. 37: 178–191.
Lamb T., Lydeard C., Walker R.B., Gibbons J.W. 1994. Molecular systematics of map turtles: a comparison of mitochondrial restriction site versus sequence data. Syst. Biol. 43:543–559.
Lassmann T., Sonnhammer E.L.L. 2005. Automatic assessment of alignment quality. Nucleic Acids Res. 33:7120–7128.
Le M., Raxworthy C.J., McCord W.P., Mertz L. 2006. A molecular phylogeny of tortoises (Testudines: Testudinidae) based on mitochondrial and nuclear genes. Mol. Phylogenet. Evol. 40: 517–531.
Lemmon A.R., Brown J.M., Stanger-Hall K., Moriarty Lemmon E. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58: 130–145.
Maddison W.P., Maddison D.R. 2007. Mesquite: a modular system for evolutionary analysis [Internet]. Version 2.01. Available from: http://mesquiteproject.org.
McMahon M.M., Sanderson M.J. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55:818–836.
Minin V., Abdo Z., Joyce P., Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.
Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 15:211–218.
Morgenstern B. 2004. DIALIGN: multiple DNA and protein sequence alignment. Nucleic Acids Res. 32:W33–W36.
Nixon K. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics. 15:407–414.
Page R. 2004. Taxonomy, supertrees, and the tree of life. In: Bininda-Emonds O.R.P., editor. Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic Publishing. p. 247–266.
Parham J.F., Feldman C.R., Boore J.L. 2006. The complete mitochondrial genome of the enigmatic bigheaded turtle (Platysternon): description of unusual genomic features and the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA. BMC Evol. Biol. 6:11.
Parham J.F., Simison W.B., Kozak K.H., Feldman C.R., Shi H. 2001. New Chinese turtles: endangered or invalid? A reassessment of two species using mitochondrial DNA, allozyme electrophoresis and known-locality specimens. Anim. Conserv. 4:357–367.
Pearson W.R., Lipman D.J. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA. 85:2444–2448.
Praschag P., Hundsdörfer A.K., Fritz U. 2007. Phylogeny and taxonomy of endangered South and South-east Asian freshwater turtles elucidated by mtDNA sequence variation (Testudines: Geoemydidae: Batagur, Callagur, Hardella, Kachuga, Pangshura). Zool. Scr. 36:429–442.
Ragan M. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53–58.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.
Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.
Roshan U.W, Moret B.M.E., Williams T.L., Warnow T. 2004. Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees. Proceedings 3rd IEEE Computational Systems Bioinformatics Conference (CSB 2004). 98–109.
Sanderson M.J. 2007. Construction and annotation of large phylogenetic trees. Aust. Syst. Bot. 20:287–301.
Sanderson M.J., Boss D., Chen D., Cranston K.A., Wehe A. 2008. The PhyLoTA browser: processing GenBank for molecular phylogenetics research. Syst. Biol. 57:335–346.
Sanderson M.J., Driskell A.C., Ree R.H., Eulenstein O., Langley S. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20:1036–1042.
Sanderson M.J., Shaffer H.B. 2002. Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. 33:49–72.
Shaffer H.B., FitzSimmons N.N., Georges A., Rhodin A.G.J., editors. 2007. Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises; 2005 August 8–12; Cambridge (MA). Lunenburg (MA): Chelonian Research Foundation.
Shaffer H.B., Meylan P., McKnight M.L. 1997. Tests of turtle phylogeny: molecular, morphological, and paleontological approaches. Syst. Biol. 46:235–268.
Shaffer H.B., Starkey D.E., Fujita M.K. 2008. Molecular insights into the systematics of the snapping turtles. In: Steyermark A.C., Finkler M.S., Brooks R.J., editors. Biology of the snapping turtle. Baltimore (MD): Johns-Hopkins University Press. p. 44–49.
Slowinski J.B., Page R.D.M. 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48:814–825.
Smith S.A., Donaghue M.J. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science. 322:86–89.
Spinks P.Q., Shaffer H.B., Iverson J.B., McCord W.P. 2004. Phylogenetic hypotheses for the turtle family Geoemydidae. Mol. Phylogenet. Evol. 32:164–182.

Spinks P.Q., Thomson R.C., Lovely G.A., Shaffer H.B. 2009. Assessing what is needed to resolve a molecular phylogeny: simulations and empirical data from Emydid turtles. BMC Evol. Biol. 9:56.

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

Stephens P.R., Wiens J.J. 2003. Ecological diversification and phylogeny of emydid turtles. Biol. J. Linn. Soc. 79:577–610.

Stuart B.L., Parham J.F. 2007. Recent hybrid origin of three rare Chinese turtles. Conserv. Genet. 8:169–175.

Thomson R.C., Shedlock A.M., Edwards S.V., Shaffer H.B. 2008. Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. Mol. Phylogenet. Evol. 49:514–525.

Turtle Taxonomy Working Group. 2007. An annotated list of modern turtle terminal taxa with comments on areas of taxonomic instability and recent change. In: Shaffer H.B., Georges A., FitzSimmons N.N., Rhodin A.G.J., editors. Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises. Lunenburg (MA): Chelonian Research Foundation. p. 173–199 (Chelonian research monographs).

Wehe A., Bansal M.S., Burleigh J.G., Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24:1540–1541.

Wiens J. 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39:34–42.

Wildman D.E., Uddin M., Opazo J.C., Liu G., Lefort V., Guindon S., Gascuel O., Grossman L.I., Romero R., Goodman M. 2007. Genomics, biogeography, and the diversification of placental mammals. Proc. Natl. Acad. Sci. USA. 104:14395.

Yan C., Burleigh J.G., Eulenstein O. 2005. Identifying optimal incomplete phylogenetic data sets from sequence databases. Mol. Phylogenet. Evol. 35:528–535.

Zwickl D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD thesis]. Austin (TX): University of Texas at Austin.