BMC Evolutionary Biology

Research article

Assessing what is needed to resolve a molecular phylogeny: simulations and empirical data from emydid turtles

Phillip Q Spinks*[1,2], Robert C Thomson[1,2], Geoff A Lovely[1,3] and H Bradley Shaffer[1,2]

Address: [1]Department of Evolution and Ecology, Davis, USA, [2]Center for Population Biology, University of California, Davis, USA and [3]Present address Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, USA

Email: Phillip Q Spinks* - pqspinks@ucdavis.edu; Robert C Thomson - rcthomson@ucdavis.edu; Geoff A Lovely - galovely@caltech.edu; H Bradley Shaffer - hbshaffer@ucdavis.edu

* Corresponding author

Published: 12 March 2009

BMC Evolutionary Biology 2009, **9**:56 doi:10.1186/1471-2148-9-56

Received: 24 July 2008
Accepted: 12 March 2009

This article is available from: http://www.biomedcentral.com/1471-2148/9/56

Abstract

Background: Phylogenies often contain both well-supported and poorly supported nodes. Determining how much additional data might be required to eventually recover most or all nodes with high support is an important pragmatic goal, and simulations have been used to examine this question. Most simulations have been based on few empirical loci, and suggest that well supported phylogenies can be determined with a very modest amount of data. Here we report the results of an empirical phylogenetic analysis of all 10 genera and 25 of 48 species of the new world pond turtles (family Emydidae) based on one mitochondrial (1070 base pairs) and seven nuclear loci (5961 base pairs), and a more biologically realistic simulation analysis incorporating variation among gene trees, aimed at determining how much more data might be necessary to recover weakly-supported nodes with strong support.

Results: Our mitochondrial-based phylogeny was well resolved, and congruent with some previous mitochondrial results. For example, all genera, and all species except *Pseudemys concinna*, *P. peninsularis*, and *Terrapene carolina* were monophyletic with strong support from at least one analytical method. The Emydinae was recovered as monophyletic, but the Deirochelyinae was not. Based on nuclear data, all genera were monophyletic with strong support except *Trachemys*, and all species except *Graptemys pseudogeographica*, *P. concinna*, *T. carolina*, and *T. coahuila* were monophyletic, generally with strong support. However, the branches subtending most genera were relatively short, and intergeneric relationships within subfamilies were mostly unsupported.

Our simulations showed that relatively high bootstrap support values (i.e. ≥ 70) for all nodes were reached in all datasets, but an increase in data did not necessarily equate to an increase in support values. However, simulations based on a single empirical locus reached higher overall levels of support with less data than did the simulations that were based on all seven empirical nuclear loci, and symmetric tree distances were much lower for single versus multiple gene simulation analyses.

Conclusion: Our empirical results provide new insights into the phylogenetics of the Emydidae, but the short branches recovered deep in the tree also indicate the need for additional work on this clade to recover all intergeneric relationships with confidence and to delimit species for some problematic groups. Our simulation results suggest that moderate (in the few-to-tens of kb range) amounts of data are necessary to recover most emydid relationships with high support values. They also suggest that previous simulations that do not incorporate among-gene tree topological variance probably underestimate the amount of data needed to recover well supported phylogenies.

Background

In molecular phylogenetic analysis, it is often the case that some relationships are robust, and relatively "easy" to recover while others are difficult to resolve, leading to phylogenetic hypotheses that consist of a patchwork of well and poorly supported nodes. When difficult nodes are encountered, the next logical step is to add taxa and/or data under the reasonable assumption that additional taxa or characters might enable resolution and/or provide support for poorly supported nodes. Whether it is better to add taxa or data is often dependent on the particular situation. When unresolved nodes are related to a long branch and additional unsampled taxa are available, adding taxa might be preferable to adding characters since including additional taxa can help break up long branches [1-3]. On the other hand, if difficult nodes are encountered among closely related taxa, or if taxon sampling is complete or nearly so, then adding additional characters is probably the better strategy [4]. The amount of data required for resolution of difficult phylogenetic problems associated with short internodes, especially those deep in a tree can represent a particularly difficult challenge [5-7] that often requires massive amounts of sequence data to resolve. However, this is not always the case, and robust species trees can sometimes be recovered from moderate amounts of data. For example, Rokas et al. [8] analyzed 106 genes from eight *Saccharomyces* species and found that data from ≥ 20 genes were sufficient to recover a fully resolved and well-supported species tree, with little additional gain in accuracy as more data were added to the analysis. In general, the amount of data required for a given level of resolution, and the gain in phylogenetic accuracy for an increase in data sampling, depends on the true species tree, the rate of evolution for a particular marker, and the fit of the selected model of evolution to the actual substitution pattern of the data.

Interacting with this general question of data quantity is the more elusive problem of data quality. Individual gene trees may or may not accurately reflect overall phylogenetic trees, rate heterogeneity can lead to long branch attraction [9], and anomalous gene trees can lead to positively misleading phylogenetic results [6]. When combined with the low phylogenetic signal in many nuclear gene sequences, even a few such renegade gene trees can lead to great phylogenetic uncertainty and the need to sample many independent nuclear markers to recover well supported phylogenies. These problems have been further exacerbated in metazoan phylogenetics because of a very heavy reliance on mitochondrial DNA (mtDNA) as a single workhorse molecule; mtDNA is appealing because it is a single-locus genome that often yields very high phylogenetic support, but it is also subject to gene tree-species tree conflicts [5,10] that may require massive amounts of nuclear data to overcome.

How much data?

Determining how much, and what kind of molecular data will, on average, yield a satisfactory increase in support values for a given phylogeny has been approached in at least two ways. The first is the brute force approach–keep collecting data and track how some measure of precision or accuracy (bootstrap support and among-gene topological concordance are two such measures) does or does not increase. The appeal of this approach is that it conveys a sense of how real data collected on real organisms advances phylogenetic knowledge. However, it has drawbacks: large volumes of sequence data are expensive to collect, and marker development remains a significant technical challenge for many taxa (but see [11-14]). A related approach is to subsample (jackknife) large empirical data sets to determine the minimum amount of data that would have been necessary to recover well-supported trees. In these analyses, "target" phylogenies are generated from large amounts of sequence data, and then subsamples of the full data set are analyzed to determine the fraction of the full data set required to recover the target phylogeny [8,15,16]. Both of these approaches are obviously limited to clades for which large amounts of sequence data are available or can be easily acquired. As a result, the taxa that have been examined in this manner have often been separated by large evolutionary distances and long phylogenetic branches because large sequence resources tend to be distributed widely across clades in a few model organisms.

The alternative strategy, and the one used in this study, is to use a modest multi-gene dataset combined with phylogenetic simulations to explore the predicted gain in support values as more (simulated) data are added to a study. Several previous analyses have used simulations to estimate the amount of data necessary to resolve difficult phylogenetic problems. In some cases, sequence data were "grown" such that one or a few empirical data partitions were bootstrapped to generate progressively larger data sets. These pseudoreplicate data sets were then subjected to phylogenetic analyses to estimate the amount of data potentially required to resolve a phylogeny or recover particular node(s) at a predetermined level of support [4,17-23]. These kinds of approaches also have their strengths and weaknesses. On the positive side, simulated data are essentially free, allowing one to determine ahead of time whether a major sequencing effort might be worth the cost of data acquisition. However, simulated data are never a substitute for the real thing, and the reliability of simulations depends on the models of evolution that are used and idiosyncratic features of the dataset on which the simulations are based. Critically, simulation studies performed to date have not accounted for the effects of topological variation in gene trees on phylogenetic inference, and these effects can be profound. For example, when

data from a single locus are pseudoreplicated to create additional simulated characters, even weak signal, when multiplied, can eventually become strong signal, leading to well-supported phylogenies (e.g. [21,23]). Conversely, multigene data sets frequently contain conflicting phylogenetic signal, and analyses performed on these types of data sets will often result in polytomies or recover nodes with poor support. Thus, to realistically simulate data from multiple markers, variation in gene tree topology should be incorporated.

Here, our primary goal is to examine the effect of increasing data on multilocus phylogenetic inference when variation in gene tree topologies is incorporated. We examine this issue using a multi-gene empirical dataset for the new-world pond turtles (family Emydidae) coupled with a simulation approach. Emydids are typical of many real-world clades; some relationships have been well supported since molecular and morphological approaches were first applied to the group, some remain contentious, and our molecular knowledge is based almost entirely on mtDNA (Thomson and Shaffer, in review). Particularly vexing is the frequent conflict among data partitions/analytical methods such that analyses based on morphology or mitochondrial DNA (mtDNA), or employing different methodologies (i.e. maximum parsimony (MP) vs maximum likelihood (ML)) are incongruent. These conflicts are best seen in a recent attempt to summarize previous hypotheses of emydid relationships with a supertree approach. A matrix representation with parsimony (MRP) supertree analysis utilizing all available phylogenies resulted in a tree that was nearly completely unresolved [24].

Simulation approach

Our goals in this paper are two-fold. First, we present a multi-gene phylogeny of emydid turtles to bring greater resolution to this important vertebrate clade. Given the impending full-genome sequence of the emydid *Chrysemys picta* http://www.genome.gov/10002154, the importance of emydids as model systems in ecological, evolutionary and developmental studies [25-29], and the endangered status of many contained species [30], we view this as an important goal in its own right. Second, we use extensive simulation analyses to examine the expected gain in phylogenetic accuracy and precision that will result from an order of magnitude increase in sequence data. We explicitly model the effects of variation in gene tree topology in our work, and explore both the overall gains in phylogenetic resolution, and the ability to recover particular problematic nodes with high support values with a substantial increase in the quantity of sequence data. We did not assess the impact of adding additional taxa because our taxon sampling for deeper nodes in the Emydidae is complete; most of the taxa missing from our analysis represent members of closely-related species groups (*Graptemys*, *Pseudemys*, and *Trachemys*) that offer little opportunity to dissect long branches on the emydid tree. However, species boundaries and intraspecific relationships within these deirocheline genera have long been problematic [31-35], thus these groups will undoubtedly require extensive taxon and data sampling from each putative species as well as the use of newer species-delimitation methods [36-38] for complete resolution.

Emydid biology

The Emydidae is a clade of 48 currently recognized species [39] of freshwater aquatic, semi-terrestrial, and terrestrial turtles distributed across much of the northern hemisphere from central South America and the West Indies to southern Canada. In addition, one species (*Emys orbicularis*) is distributed across Europe, parts of North Africa, and the Middle East [40], and *Emys trinacris*, which was recently removed from *E. orbicularis*, is narrowly distributed on Sicily, and adjacent mainland Italy [41]. Emydid species diversity is highest in the southeastern United States, where about half of this species diversity is found. Emydids occupy a wide diversity of aquatic and terrestrial habitats including relatively cool lakes, ponds, and streams in southern Canada and the northeastern US (*Chrysemys picta*, and *Emys* [= *Emydoidea*] *blandingii*), brackish coastal habitats along the eastern US seaboard (*Malaclemys terrapin*), freshwater streams and rivers in Mediterranean climates of California (*Emys* [= *Actinemys*]*marmorata*), and terrestrial desert grasslands of the southwestern US/northern Mexico (*Terrapene ornata*) [40].

Results

Empirical mtDNA Phylogeny

Visual inspection of the cyt*b* sequencing chromatograms of four samples revealed the presence of multiple peaks at some nucleotide positions, potentially indicating the presence of nuclear mitochondrial pseudogenes (numts) [42]. Since we were unable to confidently determine the actual cyt*b* sequences for these individuals (despite numerous attempts to sequence them), we excluded these sequences from our analysis (Additional file 1). All of the remaining sequences showed the typical mitochondrial composition bias for guanine nucleotides (A = 30%, C = 31%, G = 12%, T = 27%), and the coding region was conserved. Thus, we consider these sequences to represent authentic mtDNA. Our cyt*b* data was composed of up to 1070 base pairs (bp) for 66 taxa. This matrix was mostly complete with ~7% percent missing data. Of the 1070 characters, 567 were constant while 429 were parsimony informative. ML analysis recovered a single tree with a -ln*L* score of 8318.58321. Fig 1 is the ML reconstruction with ML and MP bootstrap values and Bayesian posterior probabilities (BPP) as indicated.

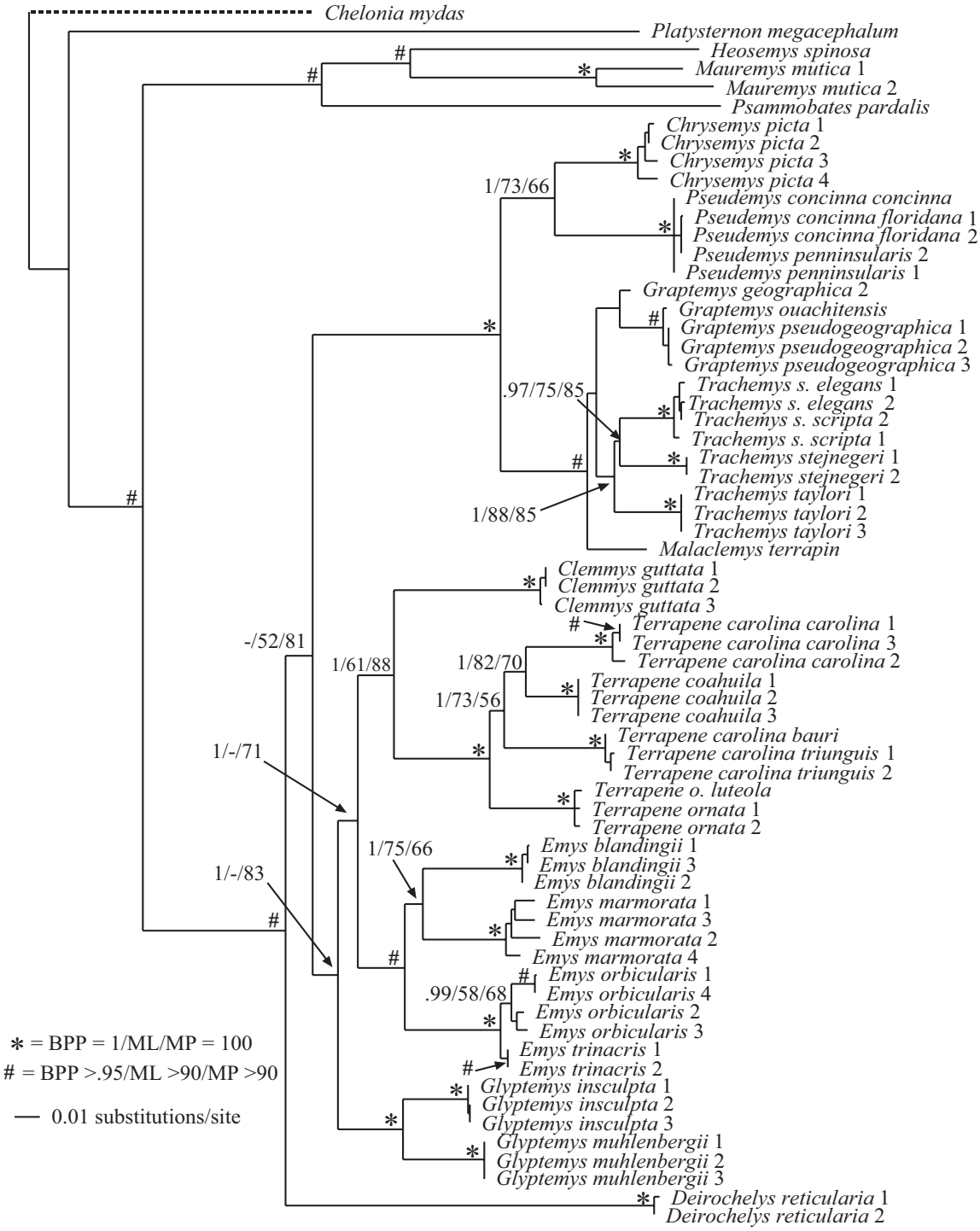


Figure 1

Maximum-likelihood phylogeny of the 66-taxon mitochondrial cytochrome *b* data set (1070 bp). Estimated ML model parameters conform to the GTR+G+I model of sequence evolution. -ln *L* = 8318.58321, rate matrix: A-C = 1.9225, A-G = 16.9646, A-T = 0.9616, C-G = 0.391, C-T = 16.9646, G-T = 1. Base frequencies: A = .30, C = .31, G = .12, T = .27. Proportion of invariant sites (I) = 0.407, and γ-shape parameter = 1.0758. # indicate nodes with Bayesian posterior probabilities (BPP) of 1, and ML and MP bootstrap values of 100. * indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 90. Numerical values indicate BPP/ML/MP support values.

Overall, the mtDNA phylogeny was fairly well supported with most nodes receiving strong support from at least one analytical method (Fig. 1). In addition, all genera and species (for which we had >1 sample) were well supported as monophyletic except *Pseudemys concinna*, *P. peninsularis*, and *Terrapene carolina*. Comparisons of our mitochondrial results to previous analyses were complicated by the fact that phylogenies from previous analyses often vary as a function of analytical method, data partition, and combination of data partitions (see [27]). Thus, depending on the data partition/method(s) used, parts of our mtDNA-based tree were congruent with those from previous analyses while others were novel. For example, the Emydinae was monophyletic, but with support from BPP and MP bootstrap values only (100 and 83, respectively). Relationships among emydine genera recovered here were the same as the ML results, but differed from the MP results, of Feldman and Parham (2002) [43] in their analysis of mitochondrial ND4 and cytochrome *b* gene sequences. Finally, relationships among the Deirochelyinae recovered here were novel, and not congruent with those from previous analyses (e.g. [27,43,44]).

A particularly troubling result from our mtDNA analysis is the placement of *Deirochelys reticularia* as sister to the remaining Emydidae [27], rendering the subfamily Deirochelyinae non-monophyletic. However, *D. reticularia* is on a relatively long branch, thus the phylogenetic position of this taxon might be an artifact due to composition bias of mtDNA sequences. Phylogenetic analyses of R-Y coded data are less susceptible to systematic biases such as composition bias in mtDNA sequences [45] that can lead to spurious phylogenetic results. We used R-Y coding to pool third codon position purines (adenine/guanine: R) and pyrimidines (cytosine/thymine: Y) into two-state categories (R and Y), and performed MP and ML phylogenetic analyses on this data set [45]. Under MP, *Deirochelys* remained sister to the remaining Emydidae. However, under ML *Deirochelys* shifted to a new position within the Emydidae, but the Emydinae was rendered paraphyletic (not shown). Thus, while problematic and in need of final resolution, the relative position of *Deirochelys* based on mtDNA does not appear to be an artifact of composition bias.

Empirical single-locus nuclear phylogenies

PCR or sequencing reactions failed for seven sequences (despite multiple attempts) so these sequences were coded as missing data (Additional file 1). Patterns from the sequencing chromatograms from all nuclear loci except RAG-1 indicated that some individuals were heterozygous for length polymorphisms [46]. However, by sequencing each gene fragment in both directions, we were able to generate sequence data from most of each locus for the length-polymorphic individuals. In addition, the TB73 locus contained a poly A/T region that was difficult to align confidently so we excluded a 13-bp region of this locus from these phylogenetic analyses [TreeBase S2303].

Individual loci ranged in size (590 bp – 1104 bp), and in number of parsimony-informative characters (22 – 72), with the average locus ~850 bp in length, and containing ~50 parsimony-informative characters (see legends Figs 2, 3, 4, 5, 6, 7, 8).

To assess the relative phylogenetic performance of individual loci, we generated phylogenies for each locus independently, and under the assumption that current taxonomy is accurate, counted clades recovered from each locus including Emydidae, Deirochelyinae, and Emydinae as well as all genera and species from which we had > 1 sample (Table 1). Phylogenies generated from all loci except RAG-1 recovered the Emydidae as monophyletic with high MP or ML bootstrap support values (Figs 4, 5, 6, 7, 8, 9, Table 1), and the Deirochelyinae was recovered with strong support from HNF-1α, R35, and TGFB2. Similarly, the Emydinae was recovered from HNF-1α, R35, and RELN, but with strong support from HNF-1α only. (Figs 2, 3, 5, 8). Support levels for other clades varied across genes. For example, Deirochelys and E. blandingii were recovered as monophyletic at all loci, whereas P. concinna, and G. pseudogeographica as well as two species of Terrapene (carolina, coahuila) were never recovered as monophyletic (Figs 2, 3, 4, 5, 6, 7, 8, Table 1).

Empirical concatenated nuclear phylogeny

Our concatenated nuDNA data set was composed of seven loci and up to 5961 bp of which 4912 were invariant (or excluded). Among ingroup taxa, 350 of these were parsimony informative. Again, this matrix was mostly complete with ~7% missing data. Fig 9 shows the ML tree with BPP and ML/MP bootstrap support values as indicated. The concatenated nuclear data recovered the Deirochelyinae (inclusive of *D. reticularia*), and the Emydinae as reciprocally monophyletic with strong support from all analytical methods. All genera were monophyletic with strong support except *Trachemys*, and all species except *Graptemys pseudogeographica*, *P. concinna*, *T. carolina*, and *T. coahuila* were monophyletic, mostly with strong support (Fig. 9). However, the branches subtending most genera were relatively short, and intergeneric relationships within subfamilies were mostly unsupported.

Simulations

Results from the ML simulations (Fig. 10) were qualitatively very similar to our MP simulation results (Fig. 11) in that an increase in data generally resulted in an overall increase in support values. Bootstrap support values of ≥ 95 for all nodes were eventually reached in some datasets

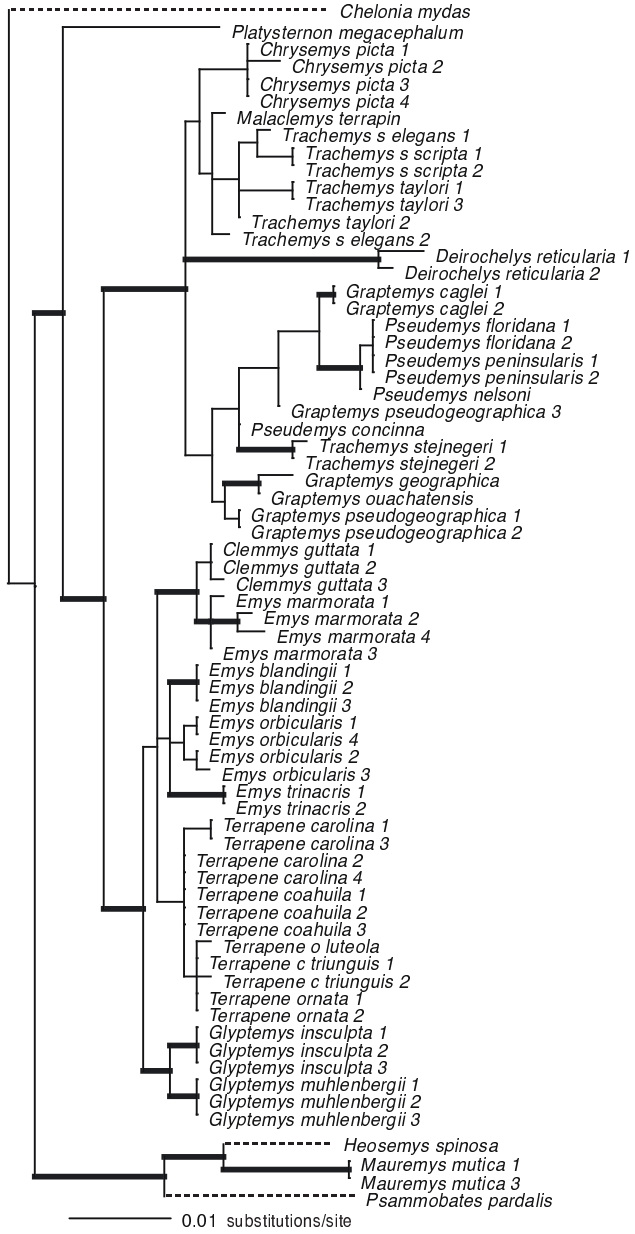
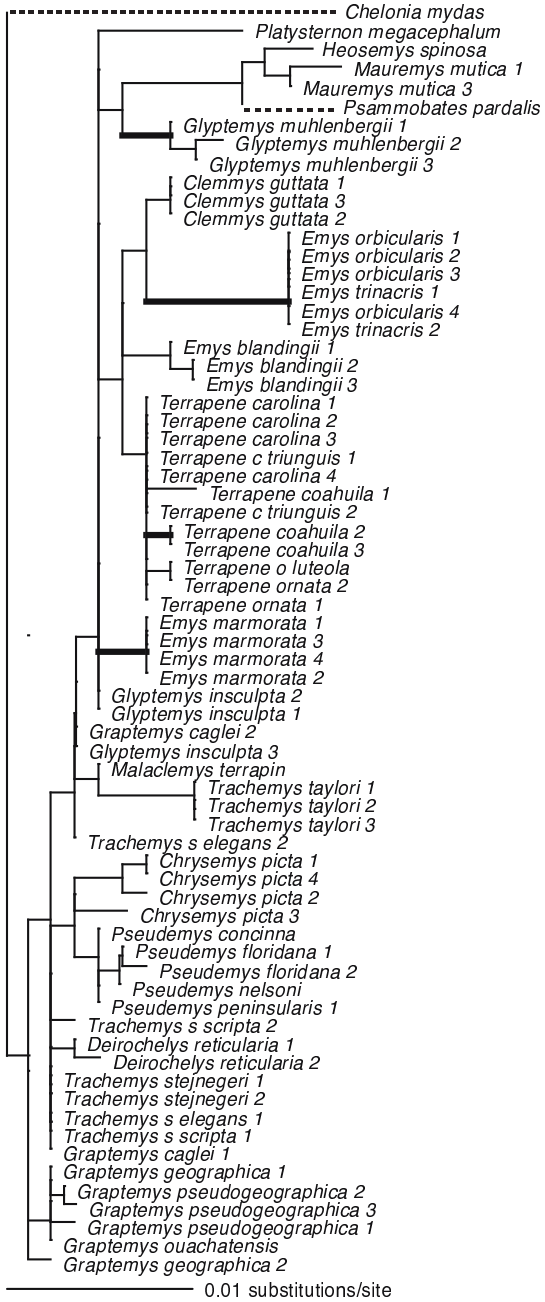


Figure 2

Maximum-likelihood phylogeny of the 69-taxon HNF-1α data set. This data set was composed of up to 768 bp. Among the ingroup, 72 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G model of sequence evolution. -ln *L* = 2533.85397, rate matrix: A-C = 0.70584, A-G = 2.497433, A-T = 0.273807, C-G = 0.59001, C-T = 2.972687, G-T = 1. Base frequencies: A = 0.28, C = 0.23, G = 0.22, and T = 0.27, and γ-shape parameter = 0.670155. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.

Figure 3

Maximum-likelihood phylogeny of the 69-taxon R35 data set (978 bp). This data set was composed of up to 978 bp. Among the ingroup, 60 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G model of sequence evolution. -ln *L* = 2571.763999, rate matrix: A-C = 0.939205, A-G = 2.365954, A-T = 0.608604, C-G = 0.875824, C-T = 3.090148, G-T = 1. Base frequencies: A = 0.28, C = 0.18, G = 0.22, and T = 0.32, and γ-shape parameter = 0.582180. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.



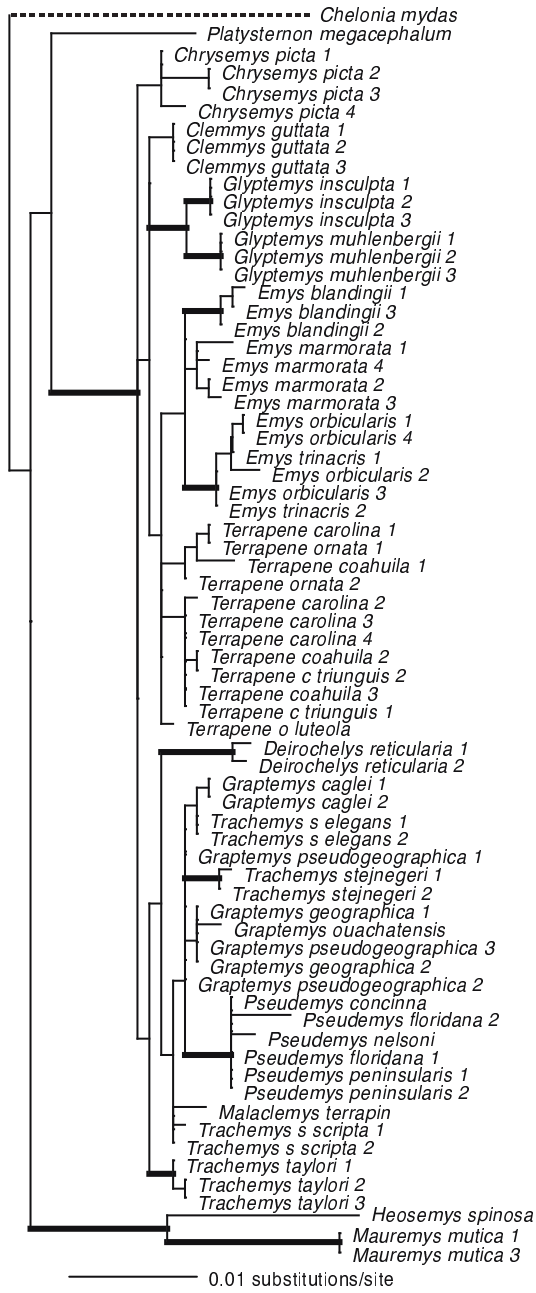


Figure 4

Maximum-likelihood phylogeny of the 69-taxon RAG-1 data set. This data set was composed of up to 788 bp. Among the ingroup, 33 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G+I model of sequence evolution. -ln *L* = 1742.011947, rate matrix: A-C = 0.781722, A-G = 2.299258, A-T = 0.337743, C-G = 0.880761, C-T = 4.054761, G-T = 1. Base frequencies: A = 0.32, C = 0.23, G = 0.22, and T = 0.23, and γ-shape parameter = 0.2.501963. Proportion of invariable sites (I) = 0.462491. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.

Figure 5

Maximum-likelihood phylogeny of the 69-taxon RELN data set. This data set was composed of up to 1104 bp. Among the ingroup, 48 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G+I model of sequence evolution. -ln *L* = 2847.341935, rate matrix: A-C = 1.068197, A-G = 3.028768, A-T = 0.454495, C-G = 0.612036, C-T = 2.836311, G-T = 1. Base frequencies: A = 0.32, C = 0.17, G = 0.18, and T = 0.33, and γ-shape parameter = 1416.809681. Proportion of invariable sites (I) = 0.482101. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.

except for the ML simulations, where the maximum proportion of nodes recovered was 96% and 90% at bootstrap support levels of ≥ 70 and ≥ 95, respectively (Fig. 10). However, the single-model simulation based on RAG-1 only (Fig. 12) reached higher overall levels of support with less data than did the "full" simulations (i.e. those based on all seven nuclear loci). For example, of the 70 data sets from the single-locus simulation, 57 had > 90% of nodes supported at the ≥ 70 bootstrap support level. In contrast, 28 of the 70 data sets from the full simulation had > 90% of nodes supported at the ≥ 70 bootstrap support level. Results were more one-sided for the ≥ 95 bootstrap support level where 32 of 70 data sets from the single-locus simulation, but only 2 of 70 data sets from the full simulation had > 90% of nodes supported at the ≥ 95 bootstrap support level (Figs 11, 12).

To examine these effects more quantitatively, we 1) show symmetric tree distances plotted as a function of total data for 1 kb incremental increases from 1–70 kb of simulated data, and 2) compare MP phylogenies generated from empirical data with those generated from simulated data. The key result from the symmetric tree distances plots is that the symmetric tree distances are much lower for single versus multiple gene simulations. For example, the average symmetric tree distance for the single-locus simulation (1.5) was almost an order of magnitude lower than the average from the full simulation (10.7) (Fig. 13). In addition, trees generated from simulated data were also similar to those from our empirical data. To compare trees from simulated vs empirical data, we performed an MP bootstrap analysis on the empirical data (31-taxon, 5974 bp), and compared these trees to results from 6000 bp of simulated data (trees not shown). Support values from the empirical data were slightly lower than those from simulated data where 61% of nodes were recovered at the ≥ 70 bootstrap support level, compared to 75% and 71% of nodes for the single-locus and full simulations, respectively. However, at the ≥ 95 support level about 50% of nodes were recovered from analyses of empirical and simulated data (Fig. 11).

Figure 6
Maximum-likelihood phylogeny of the 69-taxon TB29 data set. This data set was composed of up to 590 bp. Among the ingroup, 22 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G model of sequence evolution. -ln *L* = 1485.985421, rate matrix: A-C = 0.724474, A-G = 2.026332, A-T = 0.393397, C-G = 0.419566, C-T = 1.269648, G-T = 1. Base frequencies: A = 0.31, C = 0.21, G = 0.19, and T = 0.29, and γ-shape parameter = 0.8245. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.

Discussion

Our results speak both to general issues in phylogenetic data requirements and specific progress in the phylogeny of emydid turtles. Overall, our results argue strongly for the insights that can be gained from a combination of multi-gene empirical results which summarize our current state of phylogenetic knowledge, and biologically realistic simulations to gain a sense of the data required to further clarify these phylogenetic results.

Phylogenetic resolution: empirical and simulation results

At least for problems of the size and complexity of Emydidae, ~6 kb of nuclear sequence data were clearly insufficient to recover well-supported relationships among many genera or species. Roughly half of the nodes were recovered with ≥ 95 MP bootstrap support (Fig. 9), and those nodes were spread across both shallow and deep nodes of the tree. About 1 kb of mitochondrial DNA yielded similar support levels (Fig. 1). However, except for

Figure 7
Maximum-likelihood phylogeny of the 70-taxon TB73 data set. This data set was composed of up to 668 bp. Among the ingroup, 60 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G model of sequence evolution. -ln *L* = 2197.656119, rate matrix: A-C = 0.96077, A-G = 2.179468, A-T = 1.20066, C-G = 0.881978, C-T = 2.497604, G-T = 1. Base frequencies: A = 0.29, C = 0.18, G = 0.20, and T = 0.33, and γ-shape parameter = 0.741199. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.

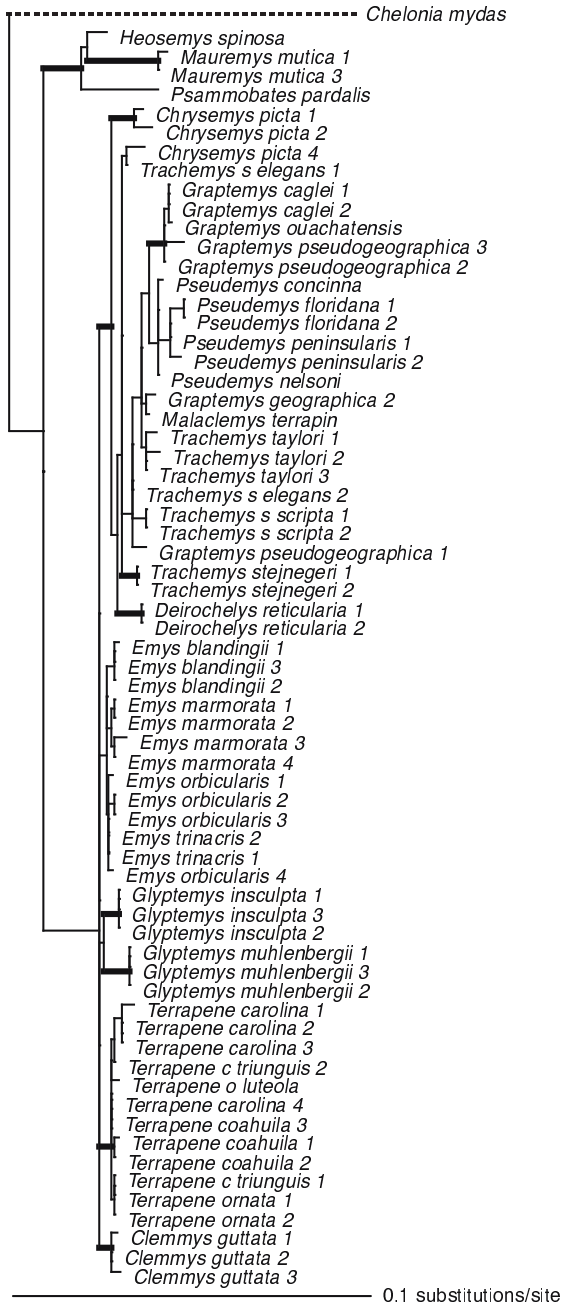


Figure 8
Maximum-likelihood phylogeny of the 67-taxon TGFB2 data set. This data set was composed of up to 1078 bp. Among the ingroup, 58 characters were parsimony-informative. Estimated ML model parameters conform to the GTR+G model of sequence evolution. -ln *L* = 3186.405935, rate matrix: A-C = 0.883433, A-G = 2.425602, A-T = 0.769195, C-G = 0.592286, C-T = 1.898448, G-T = 1. Base frequencies: A = 0.29, C = 0.21, G = 0.20, and T = 0.30, and γ-shape parameter = 0.862394. Thick branches indicate nodes with ≥ .95 BPP and ML and MP bootstrap values ≥ 70.

species and genus monophyly, there was only a single relationship (*Glyptemys* as sister to the remaining emydines) shared by both analyses. Although it remains unclear whether this incongruence has a biological basis (i.e. introgression/hybridization), is due to incomplete lineage sorting, or results from some combination of fac-

tors, it certainly confirms the growing suspicion that phylogenetic conclusions and taxonomic changes based purely on mitochondrial gene trees may often be premature, and that multiple markers are required for this level of analysis. However, determining how much data might be required remains challenging.

Overall, our simulations suggest that, although the percentage of nodes supported is always a decreasing function of additional data analyzed, relatively high overall support values can probably be obtained with moderate amounts of data. However, previous simulations based on single (or a few) loci–because they ignore variation among gene trees–probably underestimate the amount of data necessary for recovering robust phylogenies. For example, with 20 kb sequence data, 82% and 67% of nodes were recovered with bootstrap support values of ≥ 70 and ≥ 95, respectively, but increasing the data sampling by 50% had little effect on support values. With 30 kb, 87% and 72% of nodes were recovered with bootstrap support values of ≥ 70 and ≥ 95, respectively; with 40 kb, each increased by another percent or two. Thus, all else being equal, there was relatively little gain in overall number of supported nodes after the first ~24 loci (assuming a standard locus is about 850 bp). Rokas et al. (2003) found empirically that data from ≥ 20 genes was sufficient to recover the phylogeny of *Saccharomyces* yeast species with strong support, based on subsamples of their 106-locus dataset. Although ~24 independent nuclear markers currently stretches the limits of most non-model organism datasets, this need not be the case in the near future. As additional genomic resources become available, assembling 24 or more markers should become feasible for many metazoan, plant, and fungal taxa, using either traditional universal-primer [14,47,11,13] or more clade-restrictive strategies [12] for primer development.

Among our empirical loci, HNF-1α, was one of the shortest, but had the most parsimony-informative characters of all loci (see Fig. legends). Therefore we wanted to determine if HNF-1α (or any other locus) might have had a disproportionate impact on our multi-locus simulations. For example, our a priori expectation is that if HNF-1α were driving the simulations, then simulated trees should be similar or identical to the empirical tree generated from HNF-1α. We tested this prediction using the SH test. To carry this out, we compared the 70 kb MP 50% majority rule consensus tree from the simulations with empirical 50% majority rule consensus trees generated from each locus. The 70 kb MP tree had the lowest -ln *L* score and trees generated from all seven empirical loci were significantly longer than the 70 kb MP tree ($P \leq 0.048$). Thus, the simulations did not appear to be overly influenced by any one marker.

As in previous simulation studies, our single-locus simulation results (based on RAG-1) recovered relatively high support values and low symmetric distances compared to the full simulations (Figs 12, 13). The RAG-1 data were simulated using a single input tree and model of molecular evolution (see below). In contrast, each dataset from the full simulation was compiled from markers drawn independently from the pool of simulated data. Consequently, our nuDNA data was simulated from a minimum of one input tree/model of molecular evolution up to a maximum of 70 input trees, and seven models of molecular evolution. Thus, the high variance in support values and symmetric tree distances among data sets from the full simulation suggest that the phylogenetic performance of data drawn from individual loci may be conveying a somewhat false sense of encouragement compared to more thorough multi-gene simulations. In other words, a well-supported tree generated from a single locus, either in a simulation or empirical framework, does not necessarily mean that the organismal phylogeny is known with confidence. Essentially, the largely stochastic variance among genes and their associated gene trees is never challenged by independent data in single gene analyses, which can leave one with greater support for idiosyncratic gene tree results than one might obtain based on a more comprehensive sampling of gene tree histories, and this was born out by the SH tests. The message is clear–empirical studies of long reads from single genes, and simulations based on single genes, will often yield overly optimistic views of the certainty of organismal phylogenies.

Emydid phylogeny

Generally, our empirical phylogenetic results were in line with our *a priori* expectations in that the mtDNA tree was well-resolved, well supported, and consistent with previous mtDNA results while our nuDNA tree was not as well supported, and contained relatively short branches among most emydine and deirochelyine genera. Direct comparisons of our mtDNA and nuDNA-based trees was complicated by differences in number of terminals, but at the mitochondrial level (excluding outgroups) 86% of nodes (19/22) were well supported from at least one analytical method, but only about half (11/23) of interspecific nodes were supported based on nuDNA (Figs 1, 9). Both datasets agree on the monophyly of most genera (some of which contain only a single species), but not on relationships among those genera. For example, the nuclear dataset strongly supported the monophyly of the traditionally-recognized subfamily Deirochelyinae; within it, a *Graptemys-Pseudemys-Malaclemys-Trachemys* clade was the only strongly supported node. Both of these groups were rejected by the mtDNA dataset (SH test, $P < 0.01$). Similarly, the nuclear dataset supported the monophyly of the traditional Emydinae, the sister-group relationship of *Glyptemys* to all remaining emydines, *Clemmys*

Table 1: Table of clades recovered from phylogenetic analyses of individual and concatenated nuclear loci. + indicates clades that were recovered while – indicates clades that were not recovered.

Clade	Locus							Concatenated nuDNA
	HNF-1α	RAG	R35	RELN	TB29	TB73	TGFB	
Emydidae	+	-	+	+	+	+	+	+
Deirochelyinae	+	-	+	-	-	-	+	+
Chrysemys	+	+	+	+	+	-	-	+
Deirochelys	+	+	+	+	+	+	+	+
Graptemys	-	-	+	-	-	-	-	+
caglei	+	-	-	+	-	-	-	+
geographica	*	-	+	-	+	-	-	+
pseudogeographica	-	-	-	-	-	-	-	-
Pseudemys	-	+	-	+	+	-	+	+
concinna	-	-	-	-	-	-	-	-
peninsularis	-	-	-	-	-	-	-	+
Trachemys	-	-	-	-	-	-	-	-
scripta	-	-	-	-	-	+	-	+
stejnegeri	+	-	+	+	-	+	+	+
taylori	-	+	+	+	-	-	+	+
Emydinae	+	-	+	+	-	-	-	+
Clemmys	+	+	+	+	-	+	+	+
Emys	-	-	-	+	-	-	-	+
blandingii	+	+	+	+	+	+	+	+
marmorata	+	+	-	+	-	+	+	+
orbicularis	+	-	-	-	-	-	-	+
trinacris	+	-	-	-	-	-	-	+
Glyptemys	+	-	+	+	-	-	-	+
insculpta	+	-	+	+	+	+	+	+
muhlenbergii	+	+	+	+	+	-	+	+
Terrapene	-	-	-	+	+	-	+	+
carolina	-	-	-	-	-	-	-	-
coahuila	-	-	-	-	-	-	-	-
ornata	-	-	+	-	-	-	-	+
Totals	15	8	15	16	9	8	12	25

* The monophyly of *G. geographica* could not be assessed for HNF-1α because we were not able to sequence both specimens for this locus.

as sister group to the *Emys-Terrapene* clade, and *E. marmorata* as sister to the remaining species of *Emys*; of these, only the position of *Glyptemys* was also supported by the mtDNA analysis. Further, a more detailed examination of the *Emys* species relationships revealed at least in that case, that the mtDNA phylogeny was most likely the result of an ancient hybridization and mitochondrial gene capture event [48].

Conclusion

Overall, short branches among emydine and deirochelyine genera, particularly for nuclear genes, suggests that recovering relationships among genera and species of this clade of turtles will continue to be a difficult problem. We are making important progress towards our understanding of emydid phylogeny and taxonomy, both in terms of species boundaries and interspecific phylogeny, but that progress has been slow and clearly requires both new data and new approaches. In particular, the exceedingly problematic genus *Pseudemys*, which has been a source of taxonomic uncertainty for over 150 years, while the box turtles (*Terrapene*) and map turtles (*Graptemys*) all remain problematic both in terms of species delimitation and within-genus interspecific phylogenetics [31-35]. All of these groups may well require a more population genetic approach to fully resolve. On the other hand, each of the three *Trachemys* species examined here were monophyletic with strong support based on both mtDNA and nuDNA sequences, although relationships among them remain obscure.

Incongruence between mtDNA vs nuDNA phylogenies is not uncommon, and generating additional data is a logical step towards understanding this incongruence. Although simulations do not inform us as to the causes of among-tree disagreements, they can be useful for determining the utility of generating additional empirical data to solve remaining, difficult parts of phylogenies. In order

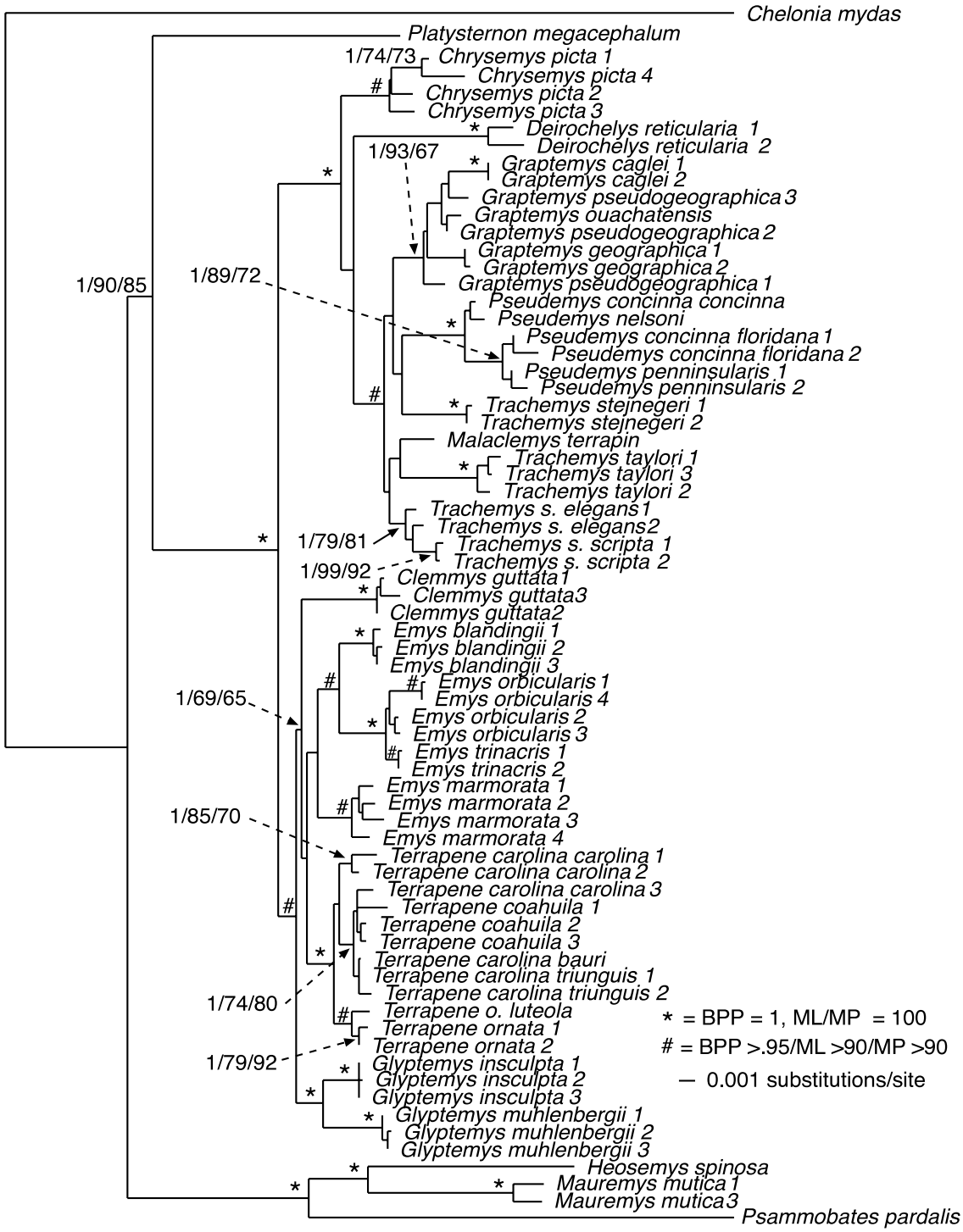


Figure 9

Maximum-likelihood phylogeny based on the 70-taxon seven-locus nuclear DNA data set. This data set was composed of up to 5961 bp. Among the ingroup, 350 characters were parsimony-informative. Nuclear loci included HNF-1α, RAG, RELN, R35, TB29, TB73, and TGFB2. Estimated model parameters conform to the GTR+G+I model of sequence evolution. -ln *L* = 17485.22059, rate matrix: A-C = 1, A-G = 2.7416, A-T = 0.6815, C-G = 0.6815, C-T = 2.7416, G-T = 1. Base frequencies: A = 0.30, C = 0.20, G = 0.20, and T = 0.30. Proportion of invariant sites (I) = 0.3263, and γ-shape parameter = 0.9023. Node symbols as in Fig. 1.

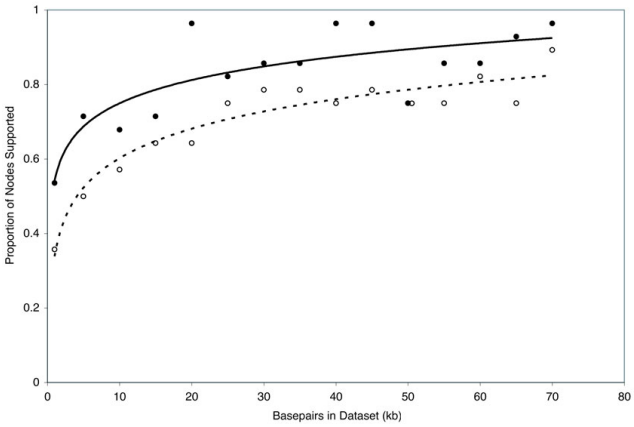


Figure 10
Maximum likelihood simulations showing proportion of nodes with bootstrap support values of ≥ 70 (filled circles) and ≥ 95 (open circles). Due to computational constraints, we analyzed every 5th data set only (i.e. in 5 kb, 10 kb 15 kb etc.) under maximum likelihood.

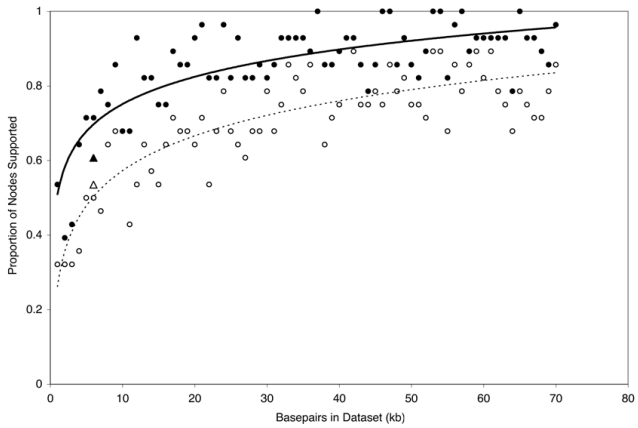


Figure 11
Maximum parsimony simulations showing proportion of nodes with MP bootstrap support values of ≥ 70 (filled circles) and ≥ 95 (open circles). Also shown are support values recovered from analyses of a 31-taxon empirical nuDNA data set (filled triangle = ≥ 70, open = ≥ 95).

for these simulations to provide accurate expectations for future studies, they should embrace realistic levels of among-gene-tree variation by simulating many genes rather than being based on one or a few empirical markers.

Methods

Taxon Sampling

Our analysis incorporated a total of 70 individuals including 64 ingroup, and six outgroup taxa representing all genera and 25 of 48 emydid species (Additional file 1). Our

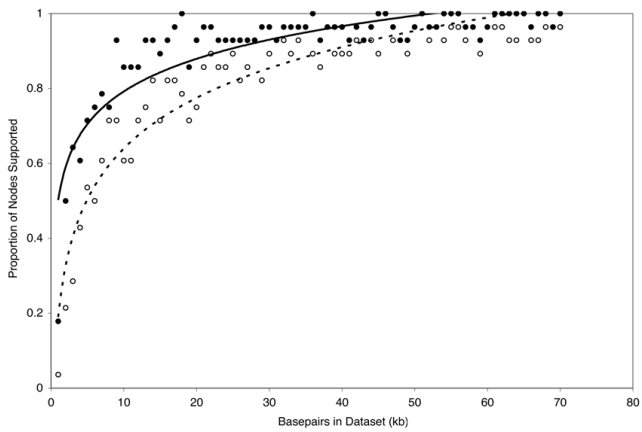


Figure 12
Maximum parsimony simulations of the RAG-1 data set showing proportion of nodes with MP bootstrap support values of ≥ 70 (filled circles) and ≥ 95 (open circles).

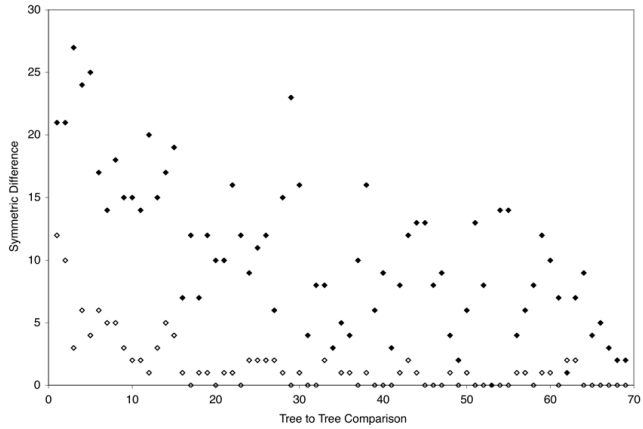


Figure 13
Symmetric tree distances generated from the single-gene simulation (open diamonds), and the full MP simulation (filled diamonds).

analyses included > two individuals of each species except *Malaclemys terrapin*. Of the two recognized subfamilies of emydids, we were missing one of the eleven species of Emydinae (the Mexican box turtle *Terrapene nelsoni*) and several species each from the diverse deirochelyine genera *Graptemys*, *Pseudemys* and *Trachemys*.

Data Sampling

We downloaded 103 sequences from GenBank, most of which were generated previously by us (Spinks and Shaffer in press). New sequences generated for this study were generally from the same individual specimen represented in GenBank (Additional file 1). Genomic DNA was extracted from blood or other soft tissue samples using a

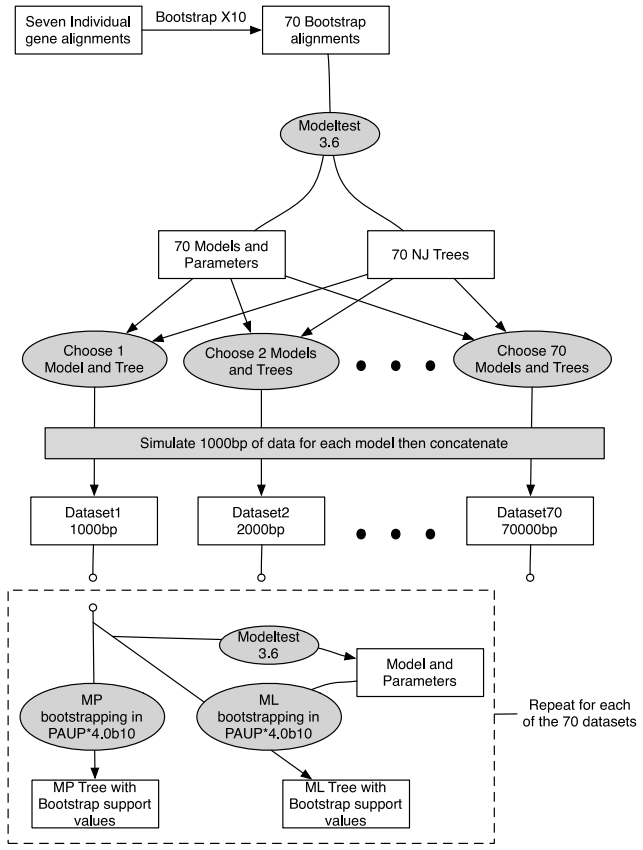


Figure 14
Flow chart detailing methodology for generating simulated data sets. See text for full description.

salt extraction protocol [49], and sequences from multiple markers were almost always generated from the same individual (Additional file 1). We sequenced fragments of one mitochondrial gene and up to seven nuclear loci. PCR products were amplified using 15–20 μL volume Taq-mediated reactions with an initial heating of 95° for 60 seconds, followed by 35 cycles of denaturation at 94°C; 30 seconds, annealing at 58°C-65°C; 30 seconds, and extension at 72°C; 45–90 seconds followed by a final extension of 72° for 10 minutes. Annealing temperatures, extension times, and primer sequences can be found in references following each locus. For mtDNA, we used cytochrome *b* (cyt*b*, [50]) while our nuclear DNA (nuDNA) included intron 2 of the hepatocyte nuclear factor 1α (HNF-1α, [51]) the nuclear recombination activase gene 1 (RAG-1, [52]), intron 61 of the Reelin gene (RELN, [53]), intron 1 of the fingerprint protein 35 (R35, [54]), intron 5 of the transforming growth factor beta-2 (TGFB2, [51]), and two anonymous nuclear loci: TB29 and TB73 [12]. All PCR products were sequenced in both directions on ABI 3730 automated sequencers at the UC Davis Division of Biological Sciences sequencing facility http://dnaseq.ucdavis.edu/.

Phylogenetic Analyses

The cyt*b* sequences were translated using MacClade 4.06 [55] to check for potential sequencing errors and orthology problems, including nuclear mitochondrial pseudogenes (numts). Sequences were initially aligned using MacClade 4.06 with final editing of the alignments by hand in PAUP* 4.0b10 [56], and the alignment was deposited in TreeBase [http://www.treebase.org, accession # S2303]. We performed phylogenetic analyses on mitochondrial and nuclear loci separately. Phylogenetic analyses of the mitochondrial and empirical concatenated nuclear sequence data sets were performed under ML, MP and Bayesian Inference. ML and MP analyses were performed using PAUP* 4.0b10 [56] with ten random stepwise heuristic searches and tree bisection-reconnection (TBR) branch swapping (for MP), or subtree pruning-regrafting (SPR) branch swapping (for ML). Model parameters for ML and Bayesian analyses were estimated in PAUP* 4.0b10, and selected under the Akiake Information Criterion (AIC). Modeltest 3.06 [57] was used to report model parameters for use in ML analyses. We bootstrapped each data set with 100 pseudoreplicates [58], limiting each ML bootstrap replicate to one hour of computation time. For individual nuclear loci, we used RaxML [59] and MrBayes through the CIPRES web portal http://www.phylo.org to carry out ML bootstrap and Bayesian analyses.

For the mitochondrial and concatenated nuDNA analyses, we used MrBayes V3.1.1 [60,61] to perform partitioned model Bayesian analyses. The mitochondrial sequences were partitioned by codon position while the nuDNA sequences were partitioned by locus. All Bayesian analyses were performed with two replicates and four chains for 10^5 generations. Chains were sampled every 10^3 generations, and stationarity was determined when the -log likelihood (-ln *L*) scores plotted against generation time visually reached a stationary value, and when the potential scale reduction factor (PSRF) equaled 1. Trees sampled prior to stationarity were discarded as burn-in.

Simulations

Our simulation approach was of the data-growing type. Importantly, we generalized our simulation procedure in such a way as to provide a more biologically realistic estimate. In particular, we increased the among-gene variance of the simulated data by incorporating variation in models of molecular evolution, model parameter values, and gene tree topologies derived from our empirical nuclear sequence data set (see below).

Our simulation procedure is shown as a flow chart in Fig 14. For the simulations, we did not include the mitochondrial sequences since they are not representative of loci sampled from the nuclear genome. We assembled a 31-taxon data set that included all seven nuclear loci and one

exemplar of each emydid species plus all outgroup taxa. *Graptemys geographica* had a great deal of missing data, and so was excluded from the simulation analysis. For each data partition we produced 10 nonparametric bootstrap replicated datasets using the Seqboot module of the Phylip 3.66 package [62], yielding 70 unique datasets. For each of these datasets, we selected the best fitting model of molecular evolution using Modeltest 3.6 [57] and selected appropriate models via the AIC. In addition, as part of the model selection procedure the Modeltest 3.6 program uses a Modelblock batch file (executed in PAUP* 4.0b10). This batch file generated neighbor-joining (NJ) trees that were used in the model determination procedure. We modified the Modelblock batch file to save these NJ starting trees from each of the 70 simulated data sets. This yielded a pool of 70 models of nucleotide sequence evolution plus their corresponding parameter estimates and NJ trees. The nonparametric bootstrapping step was employed in order to generate a large number of models and trees that were similar to the seven models and trees inferred from the empirical data, and yet incorporated a degree of variation as might be observed if one were to sample additional similar loci.

Next, we constructed simulated datasets by selecting a model, and its corresponding NJ tree at random (with replacement) from the pool of 70, and used Seq-Gen1.3.2 [63] to simulate 1000 nucleotide characters for each model/tree combination. This process was repeated and each simulated block of nucleotide characters was concatenated to produce simulated datasets ranging in size from 1000 bp (one model/tree only) to 70000 bp (70 models/trees). In addition, each dataset was simulated independently with a new model drawn from the pool for each subsequent data set (i.e. the 17th dataset did **not** consist of the 16th dataset with 1000 more base pairs added).

Phylogenies and bootstrap support values (100 pseudoreplicates) were generated using PAUP* 4.0b10 for each simulated dataset under MP. However, due to computational constraints, we analyzed every fifth data set only under ML (i.e. 5000 bp, 10000 bp, 15000 bp, etc.). MP bootstrap replicates were carried out using random sequence addition and TBR branch swapping, with the multiple trees option in effect. For ML bootstrap analyses, we chose a new model of molecular evolution for each simulated dataset using PAUP* 4.0b10, and Modeltest 3.6 with AIC-selected parameters, and employed the same heuristic search settings as in the MP bootstrap analyses. Support values for each simulated phylogeny were determined by counting the total number of nodes in each tree that were supported at the ≥ 70 and ≥ 95 level. To quantitatively compare trees among simulations, we used the symmetric tree distance [64,65] (symmetric difference test implemented in PAUP* 4.0b10). Trees were compared sequentially such that the tree generated from the 1000 bp data set (tree 1) was compared to the tree generated from 2000 bp (tree 2) then tree 2 was compared to tree 3 and so forth.

Our empirical loci varied in length and number of parsimony-informative characters, thus our multi-locus simulation might have been overly influenced by one or a few markers. For example, the average locus was ~850 bp, and contained 50 parsimony informative characters, but at 768 bp, HNF-1α had the most parsimony informative characters (72) of any locus. Therefore, HNF-1α might have had a disproportionate influence on our simulations. We used SH tests (conducted in PAUP* 4.0b10) to assess the impact of this among-locus variation on our simulations. Based on the empirical 31-taxon, seven-locus data set, we compared the 50% majority rule consensus MP tree generated from the 70 kb simulated data to the 50% majority rule consensus MP trees generated from each empirical locus, but with the trees pruned of all but a set of 29 taxa common to all loci. Our reasoning was that if a single locus were driving our simulations then that gene tree would not be significantly different from the tree based on 70 kb of simulated data.

Finally, we compared our simulation procedure to previous methods where data were simulated from a single locus. To carry this out, we repeated our simulation strategy, but used the single empirical model/NJ tree from the RAG-1 locus as input parameters in Seq-Gen to generate 70 datasets ranging from 1000 bp to 70000 bp. We chose RAG-1 since it is one of the most commonly employed phylogenetic markers for vertebrate taxa.

The simulations and tallies of support values were largely automated using a system of PERL, BASH, and R scripts, as well as several PAUP batch files (available from RCT's website, http://www.eve.ucdavis.edu/rcthomson).

Authors' contributions

PQS, RCT and HBS developed the project and designed the simulations. PQS and GAL collected the data. RCT performed the simulations. All authors contributed to writing the ms. All authors read and approved the final manuscript.

Additional material

Additional File 1

Sample identification, and GenBank accession numbers for all sequences used in this study. *All of the GenBank Accession numbers used in this analysis are listed in this table.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2148-9-56-S1.doc]

Acknowledgements

For tissue samples we thank Dan Holland, Eskandar Pouyani (University of Heidelberg, Germany), Cesar Ayres, (Universitario A Xunqueira), Carla Cicero (Museum of Vertebrate Zoology, Berkeley), Matt Aresco, Paul Moler, Raymond Farrell and Robert Zappalorti (Herpetological Consultants Inc.), Travis LaDuc (University of Texas, Austin), Charles Innis (New England Aquarium), Tibor Kovács, Suzanne McGaugh (Iowa State University), Steve Mockford (Dalhousie University), Tamás Molnár (Kapsovár University), Matt Osentoski, and Chris Tabaka (Binder Park Zoo). Dave Starkey, and three anonymous reviewers provided valuable comments. Early discussions with Brian O'Meara contributed to the development of the simulation procedure. This work was supported by grants from the NSF (DEB 0213155, DEB 0516475, DEB 0507916. DEB 0817042), an NSF Doctoral Dissertation Improvement grant (to RCT; DEB-0710380), and funding from the UC Davis Center for Population Biology and the UC Davis Agricultural Experiment Station.

References

1. Graybeal A: **Is it better to add taxa or characters to a difficult phylogenetic problems?** *Systematic Biology* 1998, **47(1):**9-17.
2. Kim J: **General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa.** *Systematic Biology* 1996, **45(3):**363-374.
3. Hillis DM: **Infering complex phylogenies.** *Nature* 1996, **383(12):**130-131.
4. Mitchell A, Mitter C, Regier JC: **More taxa or more characters revisited: Combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera).** *Systematic Biology* 2000, **49(2):**202-224.
5. Maddison WP: **Gene trees in species trees.** *Systematic Biology* 1997, **46:**523-536.
6. Degnan JH, Rosenberg NA: **Discordance of species trees with their most likely gene trees.** *Plos Genetics* 2006, **2(5):**762-768.
7. Whitfield JB, Lockhart PJ: **Deciphering ancient rapid radiations.** *Trends in Ecology & Evolution* 2007, **22(5):**258-265.
8. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies.** *Nature* 2003, **425(6960):**798-804.
9. Felsenstein J: **Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.** *Systematic Zoology* 1978, **27(4):**401-410.
10. Pamilo P, Nei M: **Relationships between Gene Trees and Species Trees.** *Molecular Biology and Evolution* 1988, **5(5):**568-583.
11. Backstroem N, Fagerberg S, Ellegren H: **Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome.** *Molecular Ecology* 2008, **17(4):**964-980.
12. Thomson RC, Shedlock AM, Edwards SV, Shaffer HB: **Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles.** *Molecular Phylogenetics and Evolution* 2008, **49(2):**514-525.
13. Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW: **Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: An example from squamate reptiles.** *Molecular Phylogenetics and Evolution* 2008, **47(1):**129-142.
14. Li C, Orti G, Zhang G, Lu GQ: **A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study.** *BMC Evol Biol* 2007, **7(44):**.
15. Bonett RM, Macey JR, Boore JL, Chippindale PT: **Resolving the tips of the tree of life: how much mitochondrial data do we need?** *Lawrence Berkeley National Laboratory* 2005.
16. Rokas A, Carroll SB: **More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy.** *Molecular Biology and Evolution* 2005, **22(5):**1337-1344.
17. Lecointre G, Philippe H, Van Le HL, Le Guyader H: **How many nucleotides are required to resolve a phylogenetic problem?: The use of a new statistical method applicable to available sequences.** *Molecular Phylogenetics and Evolution* 1994, **3(4):**292-309.
18. Berbee ML, Carmean DA, Winka K: **Ribosomal DNA and resolution of branching order among the ascomycota: How many nucleotides are enough?** *Molecular Phylogenetics and Evolution* 2000, **17(3):**337-344.
19. Chojnowski JL, Kimball RT, Braun EL: **Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes.** *Gene* 2008, **410(1):**89-96.
20. Bremer B, Jansen RK, Oxelman B, Backlund M, Lantz H, Kim KJ: **More characters or more taxa for a robust phylogeny – Case study from the coffee family (Rubiaceae).** *Systematic Biology* 1999, **48(3):**413-435.
21. DeFilippis VR, Moore WS: **Resolution of phylogenetic relationships among recently evolved species as a function of amount of DNA sequence: An empirical study based on woodpeckers (Aves: Picidae).** *Molecular Phylogenetics and Evolution* 2000, **16(1):**143-160.
22. de Queiroz A, Lawson R, Lemos-Espinal JA: **Phylogenetic relationships of North American garter snakes (Thamnophis) based on four mitochondrial genes: How much DNA sequence is enough?** *Molecular Phylogenetics and Evolution* 2002, **22(2):**315-329.
23. Wortley AH, Rudall PJ, Harris DJ, Scotland RW: **How much data are needed to resolve a difficult phylogeny? Case study in Lamiales.** *Systematic Biology* 2005, **54(5):**697-709.
24. Iverson JB, Brown RM, Akre TM, Near TJ, Le M, Thomson RC, Starkey DE: **In search of the tree of life for turtles.** *Chelonian Research Monographs* 2007, **4:**85-106.
25. Bull JJ, Vogt RC: **Temperature-Dependent Sex Determination in Turtles.** *Science* 1979, **206(4423):**1186-1188.
26. Valenzuela N, Lance V: **Temperature-dependent sex determination in vertebrates.** Washington, D.C.: Smithsonian Books; 2004.
27. Stephens PR, Wiens JJ: **Ecological diversification and phylogeny of emydid turtles.** *Biological Journal of the Linnean Society* 2003, **79(4):**577-610.
28. Stephens PR, Wiens JJ: **Explaining species richness from continents to communities: The time-for-speciation effect in emydid turtles.** *American Naturalist* 2003, **161(1):**112-128.
29. Congdon JD, Nagle RD, Kinney OM, Sels RCV: **Hypotheses of aging in a long-lived vertebrate, Blanding's turtle (Emydoidea blandingii).** *Experimental Gerontology* 2001, **36:**4-6.
30. **2008 IUCN Red List of Threatened Species** [http://www.iucnredlist.org]
31. LeConte J: **Description of the species of North American tortoises.** *Annual Lyceum Natural History, New York* 1830, **3:**91-131.
32. Carr AF: **Handbook of turtles; the turtles of the United States, Canada, and Baja California.** Ithaca, N.Y.,: comstock Pub. Associates; 1952.
33. Seidel ME: **Morphometric analysis and taxonomy of Cooter and Red-Bellied Turtles in the North American genus Pseudemys (Emydidae).** *Chelonian Conservation and Biology* 1994, **1(2):**117-130.
34. Jackson DR: **Systematics of the *Pseudemys concinna-floridana* complex (Testudines: Emydidae): an alternative interpretation.** *Chelonian Conservation and Biology* 1995, **1(4):**329-333.
35. Lamb T, Lydeard C, Walker RB, Gibbons JW: **Molecular systematics of map turtles (Graptemys): A comparison of mitochondrial restriction site versus sequence data.** *Systematic Biology* 1994, **43(4):**543-559.
36. Shaffer HB, Thomson RC: **Delimiting species in recent radiations.** *Systematic Biology* 2007, **56(6):**896-906.
37. Knowles LL, Carstens BC: **Delimiting species without monophyletic gene trees.** *Systematic Biology* 2007, **56(6):**887-895.
38. Wiens JJ: **Species delimitations: new approaches for discovering diversity.** *Systematic Biology* 2007, **56:**875-878.
39. Turtle Taxonomy Working Group: **An annotated list of modern turtle terminal taxa, with comments on areas of taxonomic instability and recent change.** In *Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises Volume 4*. Edited by: Shaffer HB, Georges A, Fitzsimmons NN, Rhodin AGJ. Chelonian Research Monographs; 2007:173-199.
40. Ernst CH, Barbour RW: **Turtles of the world.** Washington: Smithsonian Institution Press; 1989.
41. Fritz U, Fattizzo T, Guicking D, Tripepi S, Pennisi MG, Lenk P, Joger U, Wink M: **A new cryptic species of pond turtle from southern Italy, the hottest spot in the range of the genus Emys (Reptilia, Testudines, Emydidae).** *Zoologica Scripta* 2005, **34(4):**351-371.

42. Thalmann O, Hebler J, Poinar HN, Paabo S, Vigilant L: **Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes.** *Molecular Ecology* 2004, **13(2):**321-335.
43. Feldman CR, Parham JF: **Molecular phylogenetics of emydine turtles: Taxonomic revision and the evolution of shell kinesis.** *Molecular Phylogenetics and Evolution* 2002, **22(3):**388-398.
44. Bickham JW, Lamb T, Minx P, Patton JC: **Molecular systematics of the genus Clemmys and the intergeneric relationships of emydid turtles.** *Herpetologica* 1996, **52(1):**89-97.
45. Phillips MJ, Lin Y-H, Harrison GL, Penny D: **Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials.** *Proceedings of the Royal Society Biological Sciences Series B* 2001, **268(1475):**1533-1538.
46. Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA: **Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes.** *Human Molecular Genetics* 2005, **14(1):**59-69.
47. Lyons LA, Laughlin TF, Copeland NG, Jenkins NA, Womack JE, OBrien SJ: **Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes.** *Nature Genetics* 1997, **15(1):**47-56.
48. Spinks PQ, Shaffer HB: **Conflicting mitochondrial and nuclear phylogenies for the widely disjunct *Emys* (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization.** *Systematic Biology* in press.
49. Sambrook J, Russell DW: **Molecular cloning: a laboratory manual.** 3rd edition. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2001.
50. Spinks PQ, Shaffer HB, Iverson JB, McCord WP: **Phylogenetic hypotheses for the turtle family Geoemydidae.** *Molecular Phylogenetics and Evolution* 2004, **32(1):**164-182.
51. Primmer CR, Borge T, Lindell J, Saetre GP: **Single-nucleotide polymorphism characterization in species with limited available sequence information: High nucleotide diversity revealed in the avian genome.** *Molecular Ecology* 2002, **11(3):**603-612.
52. Krenz JG, Naylor GJP, Shaffer HB, Janzen FJ: **Molecular phylogenetics and evolution of turtles.** *Molecular Phylogenetics and Evolution* 2005, **37(1):**178-191.
53. Spinks PQ, Shaffer HB: **Conservation phylogenetics of the Asian box turtles (Geoemydidae, Cuora): mitochondrial introgression, numts, and inferences from multiple nuclear loci.** *Conservation Genetics* 2007, **8(3):**641-657.
54. Fujita MK, Engstrom TN, Starkey DE, Shaffer HB: **Turtle phylogeny: insights from a novel nuclear intron.** *Molecular Phylogenetics and Evolution* 2004, **31(3):**1031-1040.
55. Maddison DR, Maddison WP: **MacClade 4 analysis of phylogeny and character evolution.** 4.03th edition. Sunderland, MA: Sinauer Associates Inc; 2001.
56. Swofford DL: **PAUP*: phylogenetic analysis using parsimony (*and other methods).** Sunderland, MA: Sinauer Associates; 1998.
57. Posada D, Crandall KA: **MODELTEST: Testing the model of DNA substitution.** *Bioinformatics (Oxford)* 1998, **14(9):**817-818.
58. Felsenstein J: **Confidence-Limits on Phylogenies – an Approach Using the Bootstrap.** *Evolution* 1985, **39(4):**783-791.
59. Stamatakis A, Hoover P, Rougemont J: **A Rapid Bootstrap Algorithm for the RAxML Web Servers.** *Systematic Biology* 2008, **57(5):**758-771.
60. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics (Oxford)* 2001, **17(8):**754-755.
61. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics (Oxford)* 2003, **19(12):**1572-1574.
62. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** Seattle, WA: Distributed by the author. Department of Genome Sciences. University of Washington; 2005.
63. Rambaut A, Grassly NC: **Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Cabios* 1997, **13(3):**235-238.
64. Penny D, Hendy MD: **The Use of Tree Comparison Metrics.** *Systematic Zoology* 1985, **34(1):**75-82.
65. Robinson DF, Foulds LR: **Comparison of Phylogenetic Trees.** *Mathematical Biosciences* 1981, **53(1–2):**131-148.

Additional file 1. Sample identification, and GenBank accession numbers for all sequences used in this study.

		GenBank accession #'s							
Tissue #	**Species**	**Cyt*b***	**HNFL**	**RELN**	**R35**	**RAG-1**	**TB29**	**TB73**	**TGFB2**
HBS 23173	*Chrysemys picta* 1	FJ770586	FJ770625	FJ770758	FJ770669	FJ770714	FJ770804	FJ770850	FJ770896
HBS 26210	*Chrysemys picta* 2	FJ770587	FJ770626	FJ770759	FJ770670	FJ770715	FJ770805	FJ770851	FJ770897
HBS 27134	*Chrysemys picta* 3	FJ770588	FJ770627	FJ770760	FJ770671	FJ770716	FJ770806	FJ770852	NA
HBS 27448	*Chrysemys picta* 4	FJ770589	FJ770628	FJ770761	FJ770672	FJ770717	FJ770807	FJ770853	FJ770898
MVZ 175961	*Clemmys guttata* 2	FJ770590	FJ770629	FJ770762	FJ770673	FJ770718	FJ770808	FJ770854	FJ770899
FMNH 265273	*Clemmys guttata* 3	FJ770591	FJ770630	FJ770763	FJ770674	FJ770719	FJ770809	FJ770855	FJ770900
MVZ 137744	*Dierochelys r. reticularia* 1	FJ770592	FJ770631	FJ770764	FJ770675	FJ770720	FJ770810	FJ770856	FJ770901
HBS 108680	*Dierochelys r. chrysea* 2	FJ770593	FJ770632	FJ770765	FJ770676	FJ770721	FJ770811	FJ770857	FJ770902
IPMB 4706	*Emys trinacris* 1	AJ131415	FJ770633	FJ770766	FJ770677	FJ770722	FJ770812	FJ770858	FJ770903
IPMB 4707	*Emys trinacris* 2	AJ131415	FJ770634	FJ770767	FJ770678	FJ770723	FJ770813	FJ770859	FJ770904
HBS 108712	*Glyptemys insclpta* 2	FJ770594	FJ770635	FJ770768	FJ770679	FJ770724	FJ770814	FJ770860	FJ770905
HBS 108711	*Glyptemys insclpta* 3	FJ770595	FJ770636	FJ770769	FJ770680	FJ770725	FJ770815	FJ770861	FJ770906
HBS 108716	*Glyptemys muhlenbergii* 2	FJ770596	FJ770637	FJ770770	FJ770681	FJ770726	FJ770816	FJ770862	FJ770907
HBS 108717	*Glyptemys muhlenbergii* 3	FJ770597	FJ770638	FJ770771	FJ770682	FJ770727	FJ770817	FJ770863	FJ770908
HBS 108718	*Graptemys caglei* 1	NA	FJ770639	FJ770772	FJ770683	FJ770728	FJ770818	FJ770864	FJ770909
HBS 108719	*Graptemys caglei* 2	NA	FJ770640	FJ770773	FJ770684	FJ770729	FJ770819	FJ770865	FJ770910
HBS 23396	*Graptemys geographica* 1	NA	FJ770641	FJ770774	FJ770685	FJ770730	FJ770820	FJ770866	NA
HBS 23397	*Graptemys geographica* 2	FJ770598	NA	FJ770775	FJ770686	FJ770731	FJ770821	FJ770867	FJ770911
HBS 23347	*Graptemys ouachatensis*	FJ770599	FJ770642	FJ770776	FJ770687	FJ770732	FJ770822	FJ770868	FJ770912
HBS 11150	*Graptemys p. koni* 1	FJ770600	FJ770643	FJ770777	FJ770688	FJ770733	FJ770823	FJ770869	FJ770913
HBS 23217	*Graptemys pseudogeographica* 3	FJ770601	FJ770644	FJ770778	FJ770689	FJ770734	FJ770824	FJ770870	FJ770914
MVZ 137745	*Malaclemys terrapin*	FJ770602	FJ770645	FJ770779	FJ770690	FJ770735	FJ770825	FJ770871	FJ770915
HBS 23325	*Pseudemys c. concinna*	FJ770603	FJ770646	FJ770780	FJ770691	FJ770736	FJ770826	FJ770872	FJ770916
HBS 108683	*Pseudemys c. floridana* 1	FJ770604	FJ770647	FJ770781	FJ770692	FJ770737	FJ770827	FJ770873	FJ770917
HBS 108682	*Pseudemys c. floridana* 2	FJ770605	FJ770648	FJ770782	FJ770693	FJ770738	FJ770828	FJ770874	FJ770918
HBS 108599	*Pseudemys nelsoni*	NA	FJ770649	FJ770783	FJ770694	FJ770739	FJ770829	FJ770875	FJ770919
HBS 108722	*Pseudemys penninsularis* 1	FJ770606	FJ770650	FJ770784	FJ770695	FJ770740	FJ770830	FJ770876	FJ770920
HBS 108723	*Pseudemys penninsularis* 2	FJ770607	FJ770651	FJ770785	NA	NA	FJ770831	FJ770877	FJ770921

HBS 27355	*Terrapene c. carolina* 2	FJ770608	FJ770652	FJ770786	FJ770696	FJ770741	FJ770832	FJ770878	FJ770922
HBS 27366	*Terrapene c. carolina* 3	FJ770609	FJ770653	FJ770787	FJ770697	FJ770742	FJ770833	FJ770879	FJ770923
HBS 108689	*Terrapene c. bauri* 4	FJ770610	FJ770654	FJ770788	FJ770698	FJ770743	FJ770834	FJ770880	FJ770924
HBS 108677	*Terrapene coahuila* 2	FJ770611	FJ770655	FJ770789	FJ770699	FJ770744	FJ770835	FJ770881	FJ770925
HBS 108678	*Terrapene coahuila* 3	FJ770612	FJ770656	FJ770790	FJ770700	FJ770745	FJ770836	FJ770882	FJ770926
HBS 108701	*Terrapene o. luteola*	FJ770614	FJ770657	FJ770791	FJ770701	FJ770746	FJ770837	FJ770883	FJ770927
HBS 27365	*Terrapene c. triunguis* 1	FJ770615	FJ770658	FJ770792	FJ770702	FJ770747	FJ770838	FJ770884	FJ770928
HBS 27385	*Terrapene c. triunguis* 2	FJ770616	FJ770659	FJ770793	FJ770703	FJ770748	FJ770839	FJ770885	FJ770929
HBS 35662	*Terrapene ornata* 2	FJ770613	FJ770660	FJ770794	FJ770704	FJ770749	FJ770840	FJ770886	FJ770930
HBS 27243	*Trachemys s. elegans* 2	FJ770617	FJ770661	FJ770795	FJ770705	FJ770750	FJ770841	FJ770887	FJ770931
HBS 108688	*Trachemys s. scripta* 1	FJ770618	FJ770662	FJ770796	FJ770706	FJ770751	FJ770842	FJ770888	FJ770932
HBS 108687	*Trachemys s. scripta* 2	FJ770619	FJ770663	FJ770797	FJ770707	FJ770752	FJ770843	FJ770889	FJ770933
HBS 108728	*Trachemys stejnegeri* 1	FJ770620	FJ770664	FJ770798	FJ770708	FJ770753	FJ770844	FJ770890	FJ770934
HBS 108729	*Trachemys stejnegeri* 2	FJ770621	FJ770665	FJ770799	FJ770709	FJ770754	FJ770845	FJ770891	FJ770935
HBS 108673	*Trachemys taylori* 1	FJ770622	FJ770666	FJ770800	FJ770710	FJ770755	FJ770846	FJ770892	FJ770936
HBS 108679	*Trachemys taylori* 2	FJ770623	FJ770667	FJ770801	FJ770711	FJ770756	FJ770847	FJ770893	FJ770937
HBS 108674	*Trachemys taylori* 3	FJ770624	FJ770668	FJ770802	FJ770712	FJ770757	FJ770848	FJ770894	FJ770938
HBS 16391	*Clemmys guttata* 1	EU787026	EU787083	EU787297	EU787165	EU787247	EU787273	EU787380	EU787225
HBS 23408	*Emys blandingii* 1	EU787042	EU787096	EU787310	AY905211	EU787253	EU787279	EU787386	EU787231
HBS 108703	*Emys blandingii* 2	EU787037	EU787094	EU787308	EU787176	EU787254	EU787280	EU787387	EU787232
HBS 108702	*Emys blandingii* 3	EU787040	EU787100	EU787314	EU787181	EU787255	EU787281	EU787388	EU787233
HBS 39753	*Emys marmorata* 2	EU787053	EU787110	EU787324	AY905237	EU787257	EU787283	EU787390	EU787235
HBS 39806	*Emys marmorata* 1	EU787044	EU787101	EU787315	AY905217	EU787256	EU787282	EU787389	EU787234
HBS 39814	*Emys marmorata* 3	EU787061	EU787118	EU787332	AY905253	EU787258	EU787284	EU787391	EU787236
HBS 39843	*Emys marmorata* 4	EU787051	EU787108	EU787322	AY905235	EU787259	EU787285	EU787392	EU787237
IPMB 4529	*Emys orbicularis* 1	EOR131412	EU787124	EU787338	EU787184	EU787260	EU787286	EU787393	EU787238
IPMB 4597	*Emys orbicularis* 2	EU787065	EU787125	EU787339	EU787185	EU787261	EU787287	EU787394	EU787239
HBS 108694	*Emys orbicularis* 3	EU787071	EU787152	EU787366	EU787212	EU787262	EU787288	EU787395	EU787240
HBS 108690	*Emys orbicularis* 4	EU787073	EU787154	EU787368	EU787214	EU787263	EU787289	EU787396	EU787241
HBS 108714	*Glyptemys insclpta* 1	EU787027	EU787084	EU787298	EU787166	EU787248	EU787274	EU787381	EU787226
HBS 108715	*Glyptemys muhlenbergii* 1	EU787028	EU787085	EU787299	EU787167	EU787249	EU787275	EU787382	EU787227
HBS 23204	*Graptemys pseudogeographica* 2	EU787025	EU787082	EU787296	EU787164	EU787246	EU787272	EU787379	EU787224
HBS 27240	*Terrapene c. carolina* 1	EU787029	EU787086	EU787300	EU787168	EU787250	EU787276	EU787383	EU787228

HBS 108676	*Terrapene coahuila* 1	EU787030	EU787087	EU787301	EU787169	EU787251	EU787277	EU787384	EU787229
HBS 27363	*Terrapene ornata* 1	EU787031	EU787088	EU787302	EU787170	EU787252	EU787278	EU787385	EU787230
HBS 23001	*Trachemys s. elegans* 1	EU787024	EU787081	EU787295	EU787163	EU787245	EU787271	EU787378	EU787223
HBS 109887	*Chelonia mydas*	EU787021	EU787076	EU787292	EU787159	AY687907	EU787266	EU787373	EU787219
HBS 16255	*Platysternon megacephalum*	NC_007970 MVZ230486 NC_007694	EU787077	EU787293	EU787160	AY687905	EU787267	EU787374	NA
HBS 109888	*Psammobates pardalis*	MVZ241333	EU787078	NA	EU787162	AY687912	EU787269	EU787376	EU787221
HBS 109889	*Heosemys spinosa*	EU787022	EU787079	EU787294	EU787161	AY687913	EU787268	EU787375	EU787220
MVZ 230476	*Mauremys mutica* 1	FJ770585	EF011276	EF011232	EF011426	FJ770713	FJ770803	FJ770849	FJ770895
MVZ 230487	*Mauremys mutica* 3	EU787023	EU787080	EF011233	EF011427	EU787244	EU787270	EU787377	EU787222

FMNH = Field Museum of Natural History, HBS = tissue collection of H. Bradley Shaffer, IPMB = Institute for Pharmacy and Molecular Biotechnology, Heidelberg, Germany, MVZ = Museum of Vertebrate Zoology, Berkeley, California. NA = missing data.