Molecular Phylogenetics and Evolution 49 (2008) 514–525

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Developing markers for multilocus phylogenetics in non-model organisms: A test case with turtles

Robert C. Thomson [a,b,*], Andrew M. Shedlock [c], Scott V. Edwards [c], H. Bradley Shaffer [a,b]

[a] Department of Evolution and Ecology, University of California, 2320 Storer Hall, Davis, CA 95616, USA
[b] Center for Population Biology, University of California, Davis, CA 95616, USA
[c] Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 26 February 2008
Revised 1 August 2008
Accepted 5 August 2008
Available online 12 August 2008

Keywords:
BAC library
Comparative genomics
Marker development
Multilocus phylogenetics
PCR primers
Turtle

ABSTRACT

We present a strategy for phylogenetic marker development in non-model systems. Rather than using the traditional approach of comparing distantly related taxa to develop conserved primers for unknown species, we explore an alternative strategy that builds primers directly from a single, relatively well characterized species and applies those primers to increasingly distantly related taxa. We develop and test our protocol with turtles. Using a single BAC end-sequence library consisting of 3461 sequences totaling 2.43 million base pairs of data, we outline a procedure to flag repeat elements, followed by a BLAST approach to categorize sequences into high, low, and no similarity compartments compared to GenBank sequences. We developed and tested a panel of 96 primer pairs with a set of turtle tissues that forms a series of increasingly distantly related taxa with respect to the BAC reference species. Finally, we sequenced 11 of these newly discovered markers across a diverse set of 18 turtle species that spans the 210 million years of chelonian crown-group history and that includes representatives of most of the major clades of extant turtles. Our results indicate that large numbers of new, phylogenetically informative markers can be developed quickly and inexpensively from a single BAC, EST, or similar genomic resource, and that those markers provide reliable phylogenetic information across both shallow and deep levels of phylogenetic history. Our results also highlight the importance of screening for and managing repetitive elements found in randomly sequenced DNA fragments. We presume that our strategy should work well across any similarly divergent clade, suggesting that many-marker datasets can be developed quickly and efficiently for phylogenetic analysis.

1. Introduction

The field of systematics has seen rapid progress in applying molecular tools to the study of phylogeny and phylogeography. This progress has been spurred, in part, by the introduction of new or modified tools that have overcome difficulties faced broadly by the community. Examples include the introduction of universal-primers (Kocher et al., 1989), the introduction of faster and cheaper automated sequencing (Sanger et al., 1977; Smith et al., 1986) and the introduction of exon-primed intron-crossing primers (Palumbi and Baker, 1994). While these developments have eliminated or lessened many of the issues traditionally faced by systematists, significant roadblocks to further progress remain. Prominent among these are the limited availability of genetic markers for non-model systems, and the importance of using multiple unlinked markers to test single-marker results. Given the paramount goal of assembling the tree of life, much systematic work in the coming years will necessarily focus on little-studied organisms for which few molecular tools are available. Emerging genome-level tools are becoming more widely available, and they may help to solve this problem, although their use in systematics has so far been limited. If these tools are broadly useful across the clades in which they were developed, then they potentially constitute a vast resource for systematic research. However, the utility of these resources for phylogenetics is a largely unexamined question.

Until very recently, genome-level resources were restricted to only a few model organisms, which severely limited their general utility throughout the systematics community. In recent years, this limitation has begun to change. According to the National Center for Biotechnology Information, 27 vertebrate genomes are now complete or in the draft assembly stage. These resources, coupled with the more numerous BAC, EST, and other smaller-scale resources associated with these projects constitute an enormous amount of information that has yet to be widely employed outside the field of comparative genomics. For example, the 2002

* Corresponding author. Address: Department of Evolution and Ecology, University of California, 2320 Storer Hall, Davis, CA 95616, USA. Fax: +1 530 752 1449.
E-mail address: rcthomson@ucdavis.edu (R.C. Thomson).

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.08.006

NSF-funded "First 100 BACs" initiative resulted in the completion of 62 bacterial artificial chromosome libraries for non-model taxa, some of which were subsequently end-sequenced (Couzin, 2002). These end-sequence libraries contain a wealth of data by current systematic biology standards, often amounting to several million base pairs of sequence information. The same is true for cDNA, EST, YAC, and other resources. As of late 2007, GenBank contained 21.2 million sequences from 705 organisms in its Genome Survey Sequence (GSS) database (which contains reads originating from BAC, YAC, Cosmid, and other genomic libraries; dbGSS release 110207) and 46.5 million sequences from 1422 organisms in its Expressed Sequence Tag (EST) database (which contains reads originating from mRNA; dbEST release 102607). These resources have obvious potential to benefit systematic research, yet they are rarely employed for these ends. The resource utilized in this study (a turtle end-sequenced BAC library) has seen use in comparative genomics (Shedlock, 2006; Shedlock et al., 2007), but very little use in phylogenetics per se.

Throughout the history of molecular phylogenetics, single locus datasets, and mitochondrial DNA in particular, have proven extraordinarily successful at elucidating phylogenetic patterns at many levels. However, as the nature of multilocus phylogenetic analysis, and of phylogenies themselves, becomes better appreciated, phylogenetic analysis with any single locus—whether mtDNA or nuclear DNA—has been questioned (Brito and Edwards, 2008). To be sure, mtDNA is able to resolve its gene tree better than most nuclear genes can resolve their own gene trees. But single loci, even mtDNA, are subject to issues of non-concordance between gene and species trees due to introgression or lineage sorting (Funk and Omland, 2003), natural selection (Ballard and Kreitman, 1995), and arbitrary divergence masquerading as real population structure (Irwin, 2002).

In response to these issues, there is a growing trend in molecular systematics toward multiple-nuclear-marker datasets for phylogeny reconstruction. In some cases these datasets are collected in order to explicitly test earlier mtDNA-based estimates of phylogeny (Ballard et al., 2002; Bardeleben et al., 2005). In other cases, workers have attempted to make use of information contained in the variation among gene trees in order to estimate a species tree (Edwards et al., 2007; Maddison and Knowles, 2006). Because nuclear markers are often characterized by lower rates of substitution than mtDNA (Brown et al., 1979), more nuclear data are generally required to resolve a phylogeny with strong support than with mitochondrial data; as a consequence, these many-marker studies are often limited by a paucity of nuclear primers. This is particularly true when the taxon of interest is poorly known or very rare (and therefore of high conservation value) and tissue availability is limited, making it difficult to test and optimize new markers. Because of this, methods for rapid and inexpensive development of nuclear markers have become increasingly important.

New nuclear-sequence markers are generally developed using some version of the universal-primer approach. This strategy was pioneered nearly two decades ago, and exploits existing sequences from divergent taxa that are aligned and used to develop conserved primers that should amplify across taxa (Kocher et al., 1989). The universal-primer approach has been widely used, and it is clearly effective for developing new markers. However, it suffers from the limitation of requiring homologous sequence data for several organisms. If many-markers are required, then large amounts of homologous sequence data are required, which may simply not exist. Some workers have employed a scaled-up universal-primer approach by using comparative genomic data to develop markers in clades that have large amounts of homologous sequence data available (Backstrom et al., 2008; Li et al., 2007; Lyons et al., 1997; Townsend et al., 2008). However, in less-studied clades, genomic resources often exist for only a single-species. Here, the universal-primer strategy may be impossible to use for many potentially informative genomic regions.

In this study, we used turtles as an exemplar clade to examine strategies for developing many-marker datasets from a single genomic resource. A single end-sequenced BAC library constructed from a western painted turtle (*Chrysemys picta bellii*) is currently the most extensive genomic resource available for turtles, along with fully sequenced mitochondrial genomes for several turtle species (Kumazawa and Nishida, 1999; Mindell et al., 1999; Parham et al., 2006; Zardoya and Meyer, 1998). Thus far, the universal-primer approach has produced primers for only 11 nuclear-sequence markers for turtles, while more than 200 primers spanning the entire mtDNA genome have been described (Engstrom et al., 2007). In a previous study, we outlined a strategy for using this end-sequenced BAC resource to discover markers that could be used in species-delimitation studies across turtles (Shaffer and Thomson, 2007). Here, we more fully develop this strategy, sequence a subset of the resulting markers across a phylogenetically broad set of living turtles, and examine the phylogenetic performance of these BAC-derived sequences for both deep and shallow problems in turtle phylogeny. By intentionally developing sets of markers for characterized genes (based on BLAST hits to GenBank) and for apparently anonymous regions of the genome, we also examine the phylogenetic breadth over which these two marker classes produce useful phylogenetic information in our turtle test panel.

Turtles are a reasonable case study because they are subject to increasing phylogenetic interest and currently lack the genetic resources necessary to make substantial systematic progress on the more difficult parts of their radiation (Shaffer et al., 2007). However, the family-level phylogeny of turtles is becoming reasonably well understood (Krenz et al., 2005; Near et al., 2005; Shaffer et al., 1997), allowing us to ensure that our sampling spans the breadth of living turtle diversity. We designed 96 nuclear primer pairs using the BAC library, and used PCR to screen them across a phylogenetically representative panel of 18 turtle species, including at least one representative from most of the families of turtles in the world. We also sequenced exemplar species to examine the phylogenetic utility of the resulting sequences. Our results demonstrate that genomic resources like a single end-sequenced BAC library are valuable sources of phylogenetically informative markers across the turtle tree of life, and suggest that a similar strategy should work for many other taxa.

2. Materials and methods

2.1. Bac library

The BAC library employed in this study was constructed from an individual *C. p. bellii* from Grant County, Washington, USA (MVZ #238119) and was produced as part of the NSF-funded "First 100 BAC" initiative (Couzin, 2002). Subsequent end sequencing of the library produced 3461 reads averaging ~700 base pairs each, for a total of 2.43 mb of sequence data. Each end-sequence was assigned a unique identifier and placed into one large FASTA file containing all 3461 sequences (Genbank numbers 84109150–84112610, Shedlock et al., 2007).

2.2. Identification of known repeat sequences

Because a large proportion of the turtle genome (and most genomes) is composed of interspersed repetitive elements and low-complexity repeats (microsatellites) that are unsuitable for sequence-based analysis, we first attempted to identify these known repetitive elements in the turtle BAC end-sequence library. CR1-like long interspersed nuclear elements (LINEs) are inappropriate

for phylogenetic analysis but exist in very high copy numbers in the turtle genome (Shedlock, 2006; Shedlock et al., 2007), and so we initially sought to remove these elements by comparing all end-sequences to a chicken repeat database using *RepeatMasker* (Smit et al., 1996–2004). We also removed microsatellites and other regions of low-complexity DNA in this repeat-removal step. Regions identified as repetitive were masked with 'N's and subsequently removed with a Perl script. After this, we attempted to flag remaining repetitive elements by comparing all end-sequences to both a vertebrate repeat database and a database of transposable-element encoded proteins using *RepeatMasker*. For all comparisons, we used *RepeatMasker*'s default settings and processing time set to 'slow' to maximize the number of repetitive elements identified.

2.3. Primer design and testing

Primer design follows the strategy outlined in Shaffer and Thomson (2007). From the total pool of reads remaining after CR1 and low-complexity DNA removal, we chose candidates for primer design by selecting a read at random, comparing it to existing GenBank sequences using *tBLASTX* and binning that read into one of three categories based on similarity to existing GenBank sequences: high similarity, defined as e-value $< 10^{-5}$ over at least a 350 bp region; low similarity, defined as e-value $> 10^{-5}$ over at least 350 bp; and no similarity, where no *BLAST* hits were returned. We also kept track of all cases in which the *BLAST* results suggested that a read might be repetitive based on close similarity to a repetitive element (or a gene that signified one, e.g. reverse transcriptase). We continued choosing reads in this way until we had assembled a panel of 96 candidates for primer design consisting of 48 high similarity, 24 low similarity, and 24 no similarity reads, each of which had been screened for repetitive regions using three methods (vertebrate repeat database, transposable-element protein database, and *BLAST* results).

For each selected read, we designed a single primer pair using the program *Primer3* (Rozen and Skaletsky, 2000) setting the optimal annealing temperature to 60 °C and the optimal primer size to 20 bp, and favoring the largest possible primed region with the requirement that the entire primed region (including both primers) had to lie within the region of similarity identified by our *BLAST* screening.

Each primer set was tested in a single 25 μl PCR reaction without optimization for each of 18 phylogenetically diverse turtles including the individual *C. picta* from which the BAC library was constructed. PCRs were performed with an initial denaturation step of 6 min at 95 °C, followed by 40 cycles of a 30 s denaturation at 95 °C, a 60 s annealing at 60 °C, and a 2 min elongation at 72 °C, followed by a final 10 min at 72 °C. Each of the 1728 PCR experiments was visualized on 1% agarose gels that were stained with ethidium bromide and photographed. Each reaction was scored as amplifying a single band of the expected size, amplifying multiple bands or a 'smear', or amplifying nothing.

2.4. Taxa

We chose a set of 18 individual turtles representing 16 species that provide broad phylogenetic coverage across the living crown-group of chelonians. About half of these are representatives of the set of 23 species that we examined previously (Fujita et al., 2004; Near et al., 2005; Shaffer et al., 1997). Based on our current state of phylogenetic knowledge, this sampling provides representation of Cryptodira and Pleurodira (the two major clades that define the root node of living turtles), as well as most of the major lineages (families and/or superfamilies) of both clades. We also sampled additional taxa from the family Emydidae more intensively to examine the fall-off in primer performance as one moves out phylogenetically from the original BAC resource within a family. Within the subfamily Deirochelyinae, we sampled another individual of the BAC subspecies, *C. p. bellii*, another *Chrysemys* subspecies/species (*C. p. dorsalis*), and three additional species drawn from the diverse genera *Graptemys* and *Pseudemys*. We also sampled two species from the subfamily Emydinae, the putative sister group to Deirochelyinae (Shaffer et al., 1997) representing a divergence time of over 30 million years (Hutchison, 1996).

2.5. Sequencing

A subset of markers was sequenced to investigate variation at each locus as well as to verify homology among similar-sized fragments. For each of the three primer classes (high, low, and no similarity) we chose two markers that amplified single products in the largest number of taxa and two markers that amplified single products (as much as possible) throughout emydids, but ceased to do so in more distant relatives of *C. picta*. We included two markers that were flagged as repetitive in this set to allow comparisons with putatively non-repetitive markers. We attempted to sequence each of these 12 total markers in both directions for each individual in the test panel. For primer pairs that did not work on individual taxa in the original PCR experiments, we performed further optimization of PCR conditions and used gel extractions when appropriate. Whenever this further optimization succeeded in producing a single clean band, we attempted to sequence it, regardless of the result of the earlier PCR experiments. Any reactions that failed were repeated at least once to verify that the failure was not the result of a lab error. Base-calling was performed using *phred*, and sequences were aligned to the expected end-sequence from the BAC library using *ClustalX* and aligned to each other using *phrap* and *ClustalX*. All alignments were checked by eye, adjusted where necessary, and any ambiguous positions were excluded from further analysis.

We calculated uncorrected pairwise genetic distance between *C. picta* and *Terrapene carolina* for each non-repetitive marker that we sequenced and compared this to the same calculation for five other markers that have been previously employed for turtle systematics (HNFL, REELIN, RAG-1, R35, and TGFB; Fujita et al., 2004; Krenz et al., 2005; Spinks and Shaffer, 2007) to provide a relative measure of variation in the BAC-derived markers.

2.6. Phylogenetic analysis

Each marker was analyzed separately in both maximum-likelihood (ML) and Bayesian frameworks. Maximum-likelihood trees were obtained by selecting the best-fitting model of molecular evolution from a set of 56 models with *ModelTest3.6* using the Akaike information criterion (AIC) (Posada and Crandall, 1998). Parameters used for the selected model were fixed at the values found using *ModelTest3.6*. Searches were implemented with TBR branch swapping and 100 random-addition sequence replicates in *PAUP*4.0b10*. Support was assessed using 100 bootstrap replicates with the same settings, except that the number of random-addition sequence replicates was reduced to 10 to save computation time. Bayesian analysis was performed using *MrBayes3.1* with 2 simultaneous searches, each with three heated chains and one cold chain run for 10^7 generations and sampled every 10^3 generations, discarding the first 10^6 generations as burn-in (Ronquist and Huelsenbeck, 2003). The model implemented was the best-fitting model selected from a set of 24 models of molecular evolution with *MrModelTest2.2* (Nylander, 2004). We judged that the chain had reached stationarity at the end of 10^7 generations by checking that the average standard deviation of split frequencies for the two runs approached zero. We also verified that the

potential scale reduction factor (PSRF) approached one for each parameter that we were estimating and that the likelihoods had reached a stationary value.

We combined our multimarker dataset using two different strategies. First, likelihood and Bayesian analyses were repeated for a single concatenated dataset containing all genes, where missing data due to individual-marker failure in specific taxa was coded with question marks and treated as missing data. *ModelTest3.6* was again employed to choose a new best-fitting model and to estimate parameters for the ML analysis. We then partitioned the dataset by marker and performed a mixed-model Bayesian analysis using the same search strategy and implementing the models and parameters from *MrModelTest2.2* for each individual gene.

Second, we analyzed our data using the new BEST—Bayesian Estimation of Species Trees—algorithm (Liu and Pearl, 2007). This method been shown to perform well in computer simulations (Edwards et al., 2007) and in the few empirical datasets to which it has been applied (Belfiore et al., 2008; Brumfield et al., in press). We applied the same gene-specific substitution models as in the Bayesian analysis, used default settings on all the priors, and ran a single chain (10 million generations) for the gene tree search, and a single chain (20,000 generations) for the species tree search. In both parts of the program we assumed priors specifying equivalent mutation rates for each gene; genes not sequenced for particular taxa were designated with question marks.

3. Results

3.1. Identification of known repeat sequences

Our initial repeat-removal step removed 86,080 bp identified as CR1-like elements, 9250 bp identified as simple repeats, and 9930 bp identified as regions of low-complexity DNA, leaving 2.31 mb of sequence data in 2788 reads for subsequent analysis. Our repeat flagging step found 464 reads containing repetitive elements, which were composed mostly of short interspersed nuclear elements (SINEs), leaving 2324 putatively non-repetitive reads.

3.2. Primer design and testing

Primer design produced 96 primer pairs that primed regions ranging from 320 to 896 base pairs long, with an average length of 639 bp. The markers are numbered from TB01 (Turtle BAC 01) to TB96 and are listed in Table 1. We included roughly equal numbers of markers derived from sequences that had been flagged as repetitive and those that had not (47 were flagged as repetitive by at least one method, 49 were not flagged by any method). The three methods we used agreed in many cases that markers were repetitive, though in other cases markers were flagged by only one or two of the methods. These markers are tagged in Table 1, along with which method(s) flagged them as repetitive.

The results of our PCR screening step are presented in Fig. 1. Seventy three primer sets (76%) amplified a single band for the BAC *C. p. bellii* specimen, while the phylogenetically most-distant clade from the BAC resource, the pleurodires, amplified single bands in 10–34 markers (10–35%), depending on the particular species. Overall, the high similarity markers amplified single bands across a broader phylogenetic span than the low and no similarity marker groups, respectively. On average, the high similarity markers amplified single bands in 10.6 out of 18 turtles, the low similarity in 7.7 of the 18, and the no similarity in 8.2 of the 18. Repetitive and non-repetitive markers amplified single products in similar numbers of species (10.9 and 10.2 average for repetitive and non-repetitive high similarity markers respectively; 8.7 and 7.3 for low similarity, and 8.6 and 8.1 for no similarity).

3.3. Sequencing

From our panel of 12 markers selected for sequencing, 11 yielded useable sequence (Table 2; GenBank Accession Nos. EU907492–EU907599). One marker (TB64) failed to yield any useable sequence even though it produced strong, single PCR bands. The remaining 11 markers yielded useable sequence in most of the taxa for which they produced single bands at the PCR stage. TB29 produced clean sequence in the greatest number of taxa (16 out of 17 expected, 94%), while TB17 did so in the fewest (11 out of 16 expected, 69%). In addition, we attempted to sequence several markers that had yielded multiple bands at the PCR stage by using gel extraction to isolate the appropriate sized band. In several cases this yielded good sequence (gray cells marked 'yes' in Table 2). We also attempted further optimization of PCR conditions for reactions that we could not successfully gel extract or did not produce bands in the initial PCR experiments, and we were mostly unsuccessful in getting additional markers to amplify (empty gray cells in Table 2).

TB17 was flagged as a repetitive marker by two repeat-detection methods (Table 1). However, we sequenced it because it also amplified a single band in a large number of taxa. The sequences for TB17 contained seven positions that we scored as heterozygous within many or most taxa sequenced, based on even-sized, double peaks at a position. We interpret these as sites that vary among the different gene copies that were being sequenced. No other markers that we sequenced showed this pattern. Based on these results, we excluded TB17 from all other analyses. TB95 was also flagged as repetitive, although only by the vertebrate repeat database. Sequencing this marker yielded normal-appearing sequence data that were easily alignable and produced a phylogenetic signal that was congruent with our understanding of turtle phylogeny.

Uncorrected genetic divergence between *C. p. bellii* and *T. carolina* provide a rough indication of the amount of sequence divergence that accumulates over 30–40 million years of divergent evolutionary history (Hutchison, 1996), and values measured for each marker are presented in Table 3. Pairwise divergence ranged from 1% in TB29 to 2.9% in TB59 with a mean of 1.9% (standard deviation 0.7%). The mean value for five other commonly used nuclear markers in turtles was 1.2% (standard deviation 0.5%). High similarity markers were characterized by a lower pairwise divergence than the low or no similarity markers (Table 3).

3.4. Phylogenetic analyses

Maximum-likelihood and Bayesian analysis recovered similar topologies for all individual-marker analyses. Some trees differed in the resolution of nodes, but the two methods never strongly supported conflicting topologies. Overall, individual-marker analyses were less well resolved than the tree based on the concatenated data (compare Figs. 2 and 3 with Fig. 4). No genes supported highly unexpected groupings based on well supported clades from the literature, with the exception of TB81, which recovered a sister grouping of *Chelus fimbriatus* with *Podocnemis expansa*, rendering the Chelidae non-monophyletic (Figs. 2 and 4). At the shallowest phylogenetic levels, we sampled two genera for multiple individuals (three individual *C. picta* and two *Pseudemys*, one each of two species). In many cases, individual genes supported the monophyly of these genera (particularly *Chrysemys*), sometimes with strong support, and in no case did they support non-monophyly of either genus (Figs. 2 and 3). Of the two genes that resolved a relationship between the three *Chrysemys* in our dataset, neither resolved a

Table 1
Ninety-six primer pairs designed from the *C. p. bellii* BAC library

Primer Set	BAC End-sequence ID	Similarity class	F-primer sequence (5′–3′)	R-primer sequence (5′–3′)	Fragment size	Repeat type
TB01	LQCX369TJ	High	CCGGGCTGACATATACGAAA	GACTGCTGTTGGGTTGTTGA	719	0
TB02	LQCX508TV	High	GCAGCCAATTAAGGGGGTAT	CCGGCTTATTGAGGGCTTC	602	0
TB03	LQCX308TJ	High	CGGTCCTTCCTGCATTAAAA	CGCACAGATGAGCTATTGGA	738	1,2,3
TB04	LQCX784TV	High	TGGGTTCTCCCATGCTTAAT	ATGCCTGGACTGTGGAAAAA	609	0
TB05	LQCX112TV	High	CTTGTGCACTTGGAGGCTCT	GCCTGCTACGCCATCAAT	403	1,3
TB06	LQCXC82TJ	High	TCACAAACCAGTGCAAGTCA	TTTGTCAGTAATGTGCAAGAATTG	621	0
TB07	LQCX341TV	High	GGGGATTTTTCCGAGACTTC	TAGGTTTGCAAGGTCGCTCT	766	0
TB08	LQCX159TV	High	CAGTCCCCTGGGAACATCT	ACAGCCACCGTGAGGTCTAC	605	0
TB09	LQCXJ75TJ	High	ACACAGATGGGCTGAGTCCT	GTGATGCTCCCTTTGACGTT	405	0
TB10	LQCXB44TV	High	CCAAAGACTGCAGCAACATT	CGGAGTGGTAGGTGGTGAGA	411	0
TB11	LQCXJ87TV	High	AGACCAAAGGGAGGACCCTA	GCGGGCAGTGATACTCCTTA	535	1,2,3
TB12	LQCX129TJ	High	TAAATGCCTTGTCGGCTTCT	GGAGGACAAAAGGAGGGTTA	518	1,3
TB13	LQCX675TV	High	CCCAGGTACAGGCTGAATGT	ACACACCCCATTTCCCACT	607	1,2,3
TB14	LQCX061TV	High	GGAAACACCCACAGACTGGT	CTGCACTTGCTCAGCCTGTA	753	2,3
TB15	LQCXF05TV	High	ACGAAGGCTGACCTTTCCTT	CCATGTTAATTCTCTTAAGCCCTTT	606	1,2,3
TB16	LQCX048TJ	High	CCTCGACCAGGAGTGTCAGT	GAACAGCAGACAGAGAGCACA	604	1,3
TB17	LQCXE85TJ	High	GGCGATAGATGGAACCCATA	AATCGCCAGCGACCTTTTA	417	2,3
TB18	LQCX377TV	High	CTACTGTGGAGGCTGCTGTG	GCCATGGCATTAATGGGTAG	432	2,3
TB19	LQCXI17TJ	High	GATTCTGGGCATGTGAAACA	CATGAGCCATGGCATTAGC	701	2,3
TB20	LQCX723TV	High	TCTGGGCCCGAGAAACAA	CCTTGCTGCAGCCGTTTA	601	2,3
TB21	LQCX420TV	High	GAGCTGAGGAAGCGTTGTTC	GTCCCATAACCTCAGCCGTA	750	1,2,3
TB22	LQCX371TV	High	CCTCCGAAGAACATGTTGAA	CGGGCTACTACGACCTCTTG	355	0
TB23	LQCXD83TV	High	GAGCTGAGGAAGCGTTGTTC	CATCCACAGCCTCAGAAACA	717	1,2,3
TB24	LQCX062TV	High	TTGCAATGCCAGACAGCTAC	GGGAAGTAAGTCCCTTGCTG	406	2,3
TB25	LQCX510TV	High	GGGCAATCTCTGGGACCT	ATGCATTCTGGTGAACTGGA	401	1,3
TB26	LQCXC15TJ	High	TGTTACTGGGGCTGGTTCAT	AAAAGTACCCCTTGCGGTTT	504	3
TB27	LQCXF13TJ	High	TGTTACTGGGGCTGGTTCAT	AAAAGTACCCCTTGCGGTTT	504	3
TB28	LQCXE10TV	High	CCCAGTTTCTCCAATAGCTCA	CCATGTTAATTCTCTTAAGCCCTTT	516	1,2,3
TB29	LQCX027TV	High	GGTACCAAGCATACCCATTTG	GGTTCAATAAGAATGGGGAAGA	609	0
TB30	LQCX169TV	High	GGAAAAATGAGTGCCCTGAA	GGGGGTAAGGTGAGGGTTAT	409	1,3
TB31	LQCXI12TV	High	GTGAACATCATGGTGGCACT	ATGATGACCTGCACGTTTCC	404	3
TB32	LQCXB23TV	High	GAAAAGATGAGTGTGTTGTCCA	CTCATTACTGTCTGCAGTCATCC	602	1,2,3
TB33	LQCX538TV	High	GCGTTGTTTTAAGTCAGCCATA	CCTGACAAAGCACAGGAACA	509	1,2,3
TB34	LQCX406TV	High	ATAGCAGTGCCGCTGACTTG	CCCCGACCAGGGTATTCTAT	406	1,2,3
TB35	LQCX876TJ	High	AAGTCTGAGAGGGGCACAGA	CTCAACGATTGCATTGGCTA	405	3
TB36	LQCX810TV	High	GGCTTCAGGCATTAAGCAAC	GGGCTGGCCAGATTCACTAT	504	1,3
TB37	LQCX127TV	High	CCCAAAAGAGAGCTGTGGAG	GTTTTGGTCTTCCCATCTAGG	601	3
TB38	LQCXE49TV	High	ATCCTGACCACATCGGTAGC	GGCGAGAAGATTTCAGCCTA	513	3
TB39	LQCXI88TV	High	CCTGGATGCATCAGTTCTAGC	TGGGAATTGAGTCCATGTTG	501	1,3
TB40	LQCXC94TJ	High	CAGGCGTACCAGCACCTC	AGGCGAGATCACCCCTGT	529	1,3
TB41	LQCXH41TJ	High	CAGCATGGACTGATCGTGAA	AGATTCTGTGTCCGGGGTAA	506	0
TB42	LQCXI71TJ	High	CAGCATGGACTGATCGTGAA	ATTCTGTGTCCGGGGTAATG	504	0
TB43	LQCXD03TJ	High	CAGCATGGACTGATCGTGAA	AGATTCTGTGTCCGGGGTAA	506	0
TB44	LQCX121TV	High	TGGATGCCCCATCATCTC	GGAGGACCCGACCCTAAGTA	508	1,3
TB45	LQCX554TV	High	TGCCACCGACTTGAAGGTAT	CCATCATCTCAGGCATTGGT	324	1,3
TB46	LQCXF20TJ	High	TTTAAGGTGACACTGGAAGACC	TGAACCACTGGCTACTTGCTT	401	3
TB47	LQCXC69TJ	High	CTGTCCGCGACTGTGATAAA	CTGGTGGCAATTACCTCTGC	320	3
TB48	LQCX048TV	High	GCCATCAACCTCAGCCTAAA	TACATAGCAGAGGGGCATTG	719	3
TB49	LQCXC69TV	Low	CCCACACATCTCCCTTAAGAA	TTCCAAGGTCACGTCTTTAAAAT	701	0
TB50	LQCXF25TV	Low	CTACAGCTCCGTGACAGCAA	TGTTCCCTGTCTGCACTCTG	745	0
TB51	LQCXD80TV	Low	CACAGCTTACTGCAGCACAGAG	ACTATTGAAAAATTGTCAGTGGCTTG	851	1
TB52	LQCX653TV	Low	GCCAATCTATATAAGATGCTTGATG	GGGGGACAAAATAAGGTACTTAAA	859	1
TB53	LQCX652TJ	Low	TGGGTCTCTGTGCTTGTATCA	CTCAGACAGCCCAAAAGGAG	714	0
TB54	LQCXH43TV	Low	CAAGCTGTTGTCTATGGAGTACTTTC	CCTGCTATTGAATGACATATACTGC	714	0
TB55	LQCX305TJ	Low	AAGGGGGTTGAATGAATGTTT	CCCGCAAAACAAAAACAAAA	888	0
TB56	LQCXD50TV	Low	TCACAGGCAGGAGTCTGATG	GGATGAAAATTGGGATTTCG	854	0
TB57	LQCXE56TJ	Low	TTAAGCGAAACGCCTCAACT	GGTTTCCTGGGGCTGTAAGA	859	0
TB58	LQCXG74TV	Low	TGTGATCTCTTCCAAAGTTGATG	ACACCTCCCCACCATAATGA	503	0
TB59	LQCXA20TV	Low	AATGAGATGGGGGAACTGC	AGTCCGGTCAAAGCCCTAAT	896	0
TB60	LQCX660TV	Low	GTGCTCAAAGGTGCAGAACA	CCTTCGCTTTGATCACGTCT	770	2
TB61	LQCX813TJ	Low	TTTCTTGTGACCCCAAAAGC	TAGTTGCCCCTCCTATGTGC	718	0
TB62	LQCXJ20TJ	Low	AACTTGCTGACTGACGTAAGAAAA	GGAGTTATCTTGTTTGAAGTTAAAGG	703	1
TB63	LQCX675TJ	Low	GGCAGTTTTGACCATTTTGG	AAGTAGCAGCTTCAGAATGTGG	707	2
TB64	LQCX725TJ	Low	TGCACAAATCTCTTATGGACACA	AGCTCCCATTTATGGACACC	601	0
TB65	LQCX474TJ	Low	TGCACCCAAGTCTGACATTT	CCAGACATTTTGCACACTCTG	620	0
TB66	LQCXB85TV	Low	GAGGGAGAGTCAGTGGTTCCT	TGCTGCCTTTTTGAGACCTT	719	2
TB67	LQCXB42TV	Low	GCTGGAGTGGGTCCTGTAAT	AGCTCAATCAGAGCTAGTGTGG	860	0
TB68	LQCX333TV	Low	GGAATCCCATTTTGAGTGTGA	TGAATAGAGACTGGGAATGCTG	875	0
TB69	LQCXC64TJ	Low	GCCAAGACTATGGGGTTTTG	CCTCCAAAAACACCAATCAC	723	2
TB70	LQCXF06TV	Low	ATTTTGGGTCATCCAGGACA	AGGCGTATCTGCCTCTCTCA	881	0
TB71	LQCX238TV	Low	GAATGCAACTCTGGGCTTTC	GGTCTGGTTTGCATTCTGGT	707	0
TB72	LQCXB18TV	Low	GCTGACTCTTCCCCAACAAA	AGGACAGGAGCCAACTCTGA	867	0
TB73	LQCX606TJ	No	TCCAGCAAGATTTAAGGTTTCA	TGCAAAATCTACTCCAGCTTAGG	704	0
TB74	LQCX437TJ	No	GAAGTCTGCAGGAAAAGAATTGA	GCCTCCACACTGAAATGGTT	708	0
TB75	LQCXI73TJ	No	AAATGTTGCGTTGACAATTCAG	GTGGCTGGCTTTTTGTAGGT	608	0

Table 1 (*continued*)

Primer Set	BAC End-sequence ID	Similarity class	F-primer sequence (5′–3′)	R-primer sequence (5′–3′)	Fragment size	Repeat type
TB76	LQCX562TJ	No	CCAGGAGTGCTGAATTATGCT	ATGGTAGGAGACGCACTGCT	713	0
TB77	LQCXI61TJ	No	ACAGGGAAGCATGGGATACA	CGTCTTAGGAACTAATACACCATTCC	501	0
TB78	LQCX930TV	No	TGTGACATGACCTCGAAATAGC	TGGTGATTTCAGCCAAGATG	739	0
TB79	LQCXE44TV	No	CTGAGAGCCTCCCGACAG	CAGATGCATTTTAAGGTTCGTG	875	3
TB80	LQCXD90TJ	No	ACTACATGGGGCCCCTAATC	ACTGGTGAAGCTCAGCACAA	891	0
TB81	LQCXA17TJ	No	AGGCTCTCTTCTGCCATTCA	GAGCCAAAATTTTTCCTTTGC	745	0
TB82	LQCXF48TV	No	GTTTGGGGTTTGCCCTATCT	ACCAAAACACAATGGGCTTC	718	0
TB83	LQCX190TV	No	TTTGCACCATGCATATAGGG	ACGTGTATTTCATGCCACCA	859	2
TB84	LQCX178TV	No	TGTAGTTTATGTCCTGGATCTATGGT	TCCTCTGCATTAACCAGTGC	852	0
TB85	LQCX520TV	No	TATGCCGGAGTTCACAGATG	CGTCTCTCCACAGGCTTTGT	710	0
TB86	LQCX971TV	No	TGGCAATGGGCTAATAGCTT	CACCAACAAACAGAGCTTGG	709	0
TB87	LQCXC32TV	No	TACAACGACCCTGTGGGTTT	CCTGCTGCCAAGCTCTTACA	779	2
TB88	LQCXA09TV	No	CCCAAAGCATGAATGGAGAA	ACGGCTACCCCCGATACTTA	619	0
TB89	LQCX524TV	No	CATCTGCCTCACATCCTTGA	GGATGAAACCGTCTCAGGAA	789	0
TB90	LQCXD82TV	No	ACTGCAGGGAATTGAGCTGT	CCTGGGATGGTATCAAGCTG	858	0
TB91	LQCX974TV	No	CAGCCCTTACTTGAAATCTGG	TCTACCCATGTGGGCTTTTC	508	0
TB92	LQCX430TV	No	GCACCTTGAAGTCTTCCCTTT	TCAGGTGTAGGCAACCTATGG	757	0
TB93	LQCX610TV	No	GGTTGCACAGACACAAATGC	GGGGAGGGCTCTGATTTACT	710	2
TB94	LQCX652TV	No	GGTACGTATACACACACACACACC	TTCCAGACTCCTGACCCAAG	612	0
TB95	LQCX243TJ	No	GATTGATTCGGGGAAGTCCT	GCTTGATGCAAGAACAACCA	727	2
TB96	LQCX774TV	No	CCTCAGCTTTCAACCACCTC	CAACCCTTGAAGGAGGAAATC	735	0

Fragment sizes are in base pairs and include the length of both primers. Repeat types: 0, no detection; 1, BLAST detection; 2, vertebrate repeat database detection; 3, transposable-element database detection. TB stands for 'Turtle BAC'.

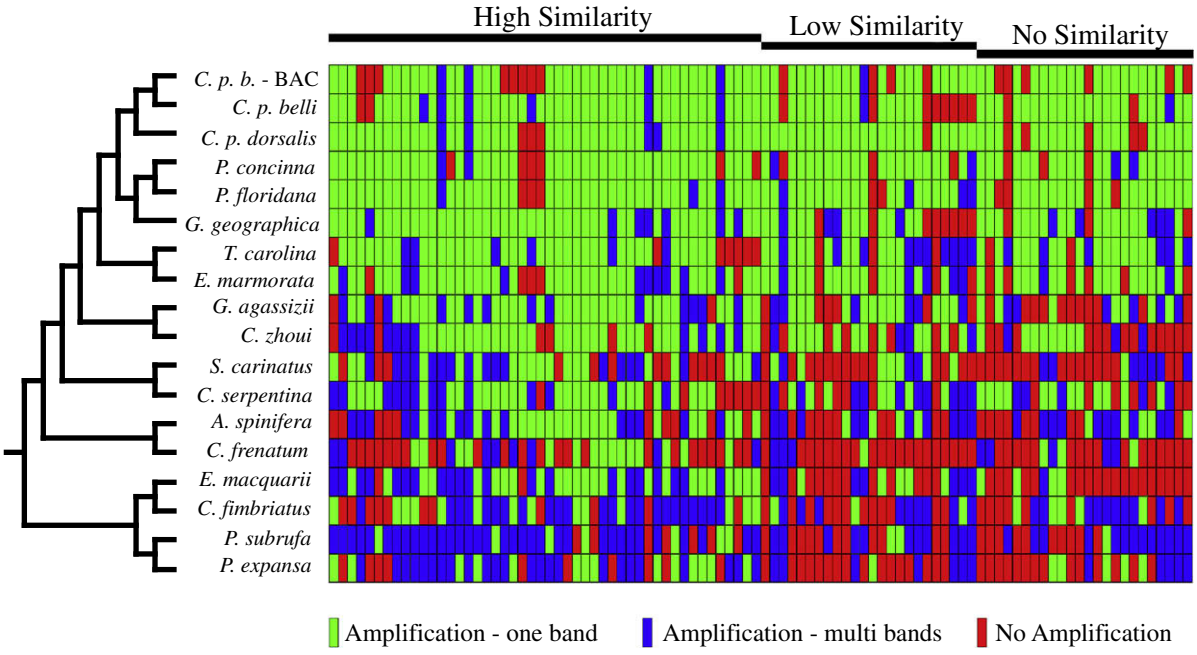

Fig. 1. Results of 1728 PCR reactions testing the utility of each marker on a panel of 18 turtles. Green denotes a single band of the expected size, blue denotes multiple bands or a 'smear', and red indicates no amplification. Markers are numbered from left to right from TB01-TB96 and broken down into 'high similarity', 'low similarity', and 'no similarity' classes.

monophyletic *C. p. bellii*, but they supported alternate groupings of the three turtles. TB59 supported the BAC turtle + *C. p. dorsalis*, and TB86 supported our second *C. p. bellii* + *C. p. dorsalis*, with low bootstrap proportions but high Bayesian posterior probabilities (BPP) in both cases.

Bayesian and ML analyses of the concatenated dataset recovered similar topologies (SH test $p > 0.05$) and so we report a single tree (from the ML analysis) with support values from both analyses (Fig. 4). This tree also agrees with the current understanding of the turtle family-level tree in most respects (Krenz et al., 2005; Shaffer et al., 1997), even though there is no overlap in gene coverage between the current and previous analyses. The only disagreement is our recovery of a non-monophyletic Chelidae. This result is due to TB81. This was the only marker that we sequenced that amplified single bands in the PCR stage in two members of the Chelidae, and so our data on this point are limited to this single-marker.

Our concatenated tree was reasonably well resolved, with nine of 17 nodes receiving bootstrap proportions higher than 90 and BPPs of 100 (Fig. 4), and three others receiving strong support from either ML bootstraps or BPP but not both. The five relatively poorly supported parts of the tree included the relationships among the major cryptodiran clades (Testudinoidea, Chelydridae, Kinosternidae, and Trionychidae), as well as the monophyly of the

Table 2
Sequencing results for 12 markers sequenced across the test panel

Taxon	Sample Number	TB17	TB29	TB53	TB64	TB81	TB82	TB02	TB07	TB49	TB59	TB86	TB95
C. p. bellii—BAC	MVZ238119	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
C. p. bellii	HBS27688	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
C. p. dorsalis	HBS27637	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P. concinna	HBS15761	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P. floridana	HBS108683	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
G. geographica	HBS23396	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes	No	Yes	No
T. carolina	HBS27360	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
E. marmorata	HBS2341	Yes	Yes	Yes	No	Yes	Yes		No	Yes	Yes	No	Yes
G. aggasizzi	HBS109173	No	No	Yes	No	No		Yes	Yes				Yes
C. zhoui	HBS41855	No	Yes	No		No	Yes	Yes			Yes		
S. carinatus	HBS107736	No	Yes	Yes	No								
C. serpentina	HBS23551	Yes	Yes			Yes	Yes		Yes				
A. spinifera	HBS107750	Yes	Yes		No								
C. frenatum	TNE722		Yes				No						
E. macquarii	HBS100714	No	No	No	No	Yes	No		No				
C. fimbriatus	HBS109170	No	Yes			Yes	No				No		
P. subrufa	HBS16420		Yes		No								
P. expansa	HBS109191	Yes	Yes			Yes	No						

Grayed cells indicate markers that were not expected to amplify single bands based on PCR experiments. Successful sequencing is indicated by 'yes', unsuccessful as 'no'. Grayed cells marked 'yes' refer to successful sequencing after additional PCR optimization and/or gel extractions.

Geoemydidae + Testudinidae. All of these have been identified as problematic parts of the turtle family-level tree previously (Krenz et al., 2005; Near et al., 2005; Shaffer et al., 1997), and may reflect actual phylogenetic uncertainty and short branch lengths in this region of the tree, as well as our limited data (sometimes a single gene) for these nodes.

The analysis using BEST produced a result similar to the concatenated analysis, although the posterior probabilities of most nodes were very low. Because the program requires a single sequence as an outgroup, we analyzed only the Cryptodira as an ingroup and use the loci from *P. expansa* as outgroups. In contrast to the concatenated dataset, the consensus of the posterior distribution of 90,020 species trees generated by BEST returned high posterior probabilities only for the monophyly of Emydidae (PP = 1.00) and Trionychidae (PP = 0.88); all other values were at or below 0.50.

Table 3
Uncorrected pairwise divergence values between *C. p. bellii* and *T. carolina* for TB markers compared to other markers used for turtles (see text for citations)

	Uncorrected distance
TB markers	
TB02 (H)	0.018
TB07 (H)	0.012
TB29 (H)	0.010
TB49 (L)	0.028
TB53 (L)	0.017
TB59 (L)	0.029
TB81 (N)	0.026
TB82 (N)	0.022
TB86 (N)	0.014
Average	0.019
High similarity avg.	0.013
Low similarity avg.	0.025
No similarity avg.	0.018
Other nuclear markers	
HNFL (L)	0.018
R35 (L)	0.017
RAG-1 (H)	0.008
REELIN (H)	0.007
TGFB (L)	0.011
Average	0.012
Mitochondrial marker	
Cytochrome—B (H)	0.133

(H), high similarity markers; (L), low similarity markers; and (N), no similarity markers.

4. Discussion

4.1. Marker development in non-model organisms

As might be expected, the number of primers that amplified single bands and produced useable sequence decreased as phylogenetic distance from the BAC turtle increased, and as degree of similarity to other taxa as revealed in *BLAST* results decreased. How this result applies to other clades is an open question. The crown-group of turtles is relatively old (~210 mya, Gaffney, 1990) and the fall-off in performance that we found (i.e. compare the density of green cells in the upper left of Fig. 1 to the density of red cells in the lower right) may be less of a concern if our approach is applied to younger clades, and more of an issue for older clades. For example, within the family Emydidae (the first eight taxa in Fig. 1, with an estimated most recent common ancestor at 34 million years ago, Hutchison, 1996) the fall-off in performance between high, low, and no similarity taxa is relatively slight (80% of high similarity markers yielded a single band, compared to 69% and 75% for low/no similarity markers).

To visualize this fall-off, we plotted the proportion of markers that yielded a single band as a function of BLAST similarity class and estimated divergence time from the BAC turtle (Fig. 5). All three similarity classes show a qualitatively similar intercept and fall-off in marker performance across the diversity of living turtles, although the somewhat lower slope of the high similarity marker class suggests that they may be the markers of choice for distantly related taxa. Most studies that have examined this fall-off in amplifiability across taxa have focused on microsatellite markers and have found that the number of primer pairs showing amplification declines to around 50% of the total over the first few 10s of millions of years of divergence between the taxon for which the markers were developed and the taxon in which they were tested (shown in fishes by Carreras-Carbonell et al. (2007); and in birds and mammals by Primmer et al. (1996)). Here, we find a much slower fall-off in the number of primers that amplify. For these markers, a decline to 50% of primers amplifying took between 70 and 130 million years (depending on similarity class). This difference could be due to a faster rate of substitution of microsatellite flanking regions in the genome

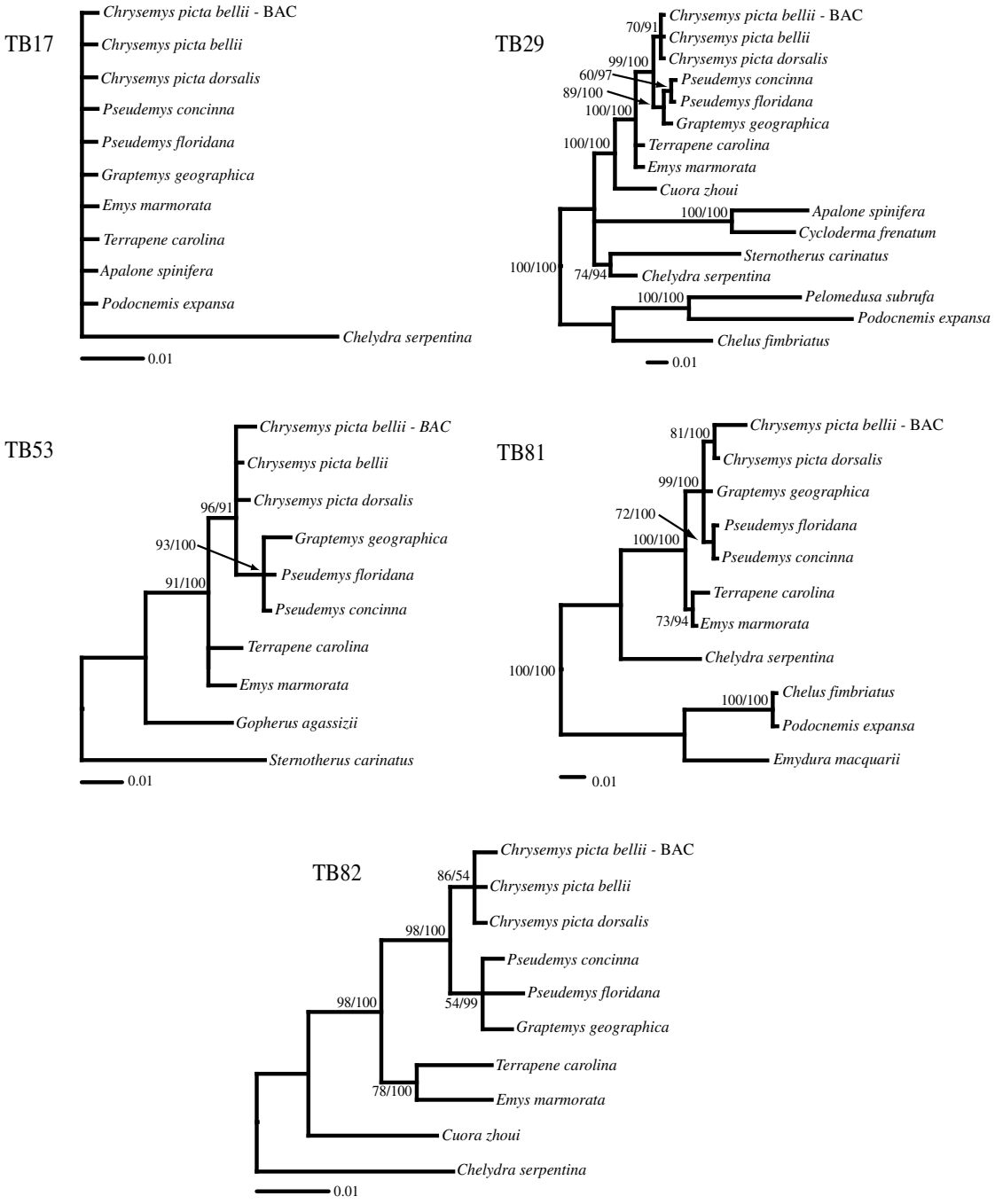

Fig. 2. Individual gene trees for each sequenced marker that produced useable sequence across the most turtles. Numbers on the nodes refer to ML bootstraps and posterior probabilities, respectively.

or due to a slower overall rate of substitution in turtles (Avise et al., 1992). The one comparable study of which we are aware in turtles seems to argue for the latter explanation. FitzSimmons et al. (1995) found that all six sea-turtle microsatellite primer pairs that they examined successfully amplified in *Trachemys scripta*, a species that last shared a common ancestor with sea-turtles approximately ~95 million years ago (Near et al., 2005), which agrees closely with our findings.

No marker amplified single bands across all taxa, though many-markers did for each taxon, implying that the panel contains phylogenetically useful markers for any clade of interest. To develop primers for the deepest nodes of the turtle tree, the taxonomic span over which primers amplify could almost certainly be increased through additional optimization, either by redesigning primers, further altering PCR conditions or cloning. However, for turtles and many other taxa, much of the remaining phylogenetic progress to be made will occur below the family-level, and our BAC library approach should provide a useful marker set for these studies. Our ongoing unpublished work at the within-family and within-genus levels confirms that TB markers that yield reliable sequence for one representative of a clade tend to be useful across that clade, and we are currently employing these markers within several turtle lineages. In addition, these markers appear to be a rich resource for the discovery of within-species

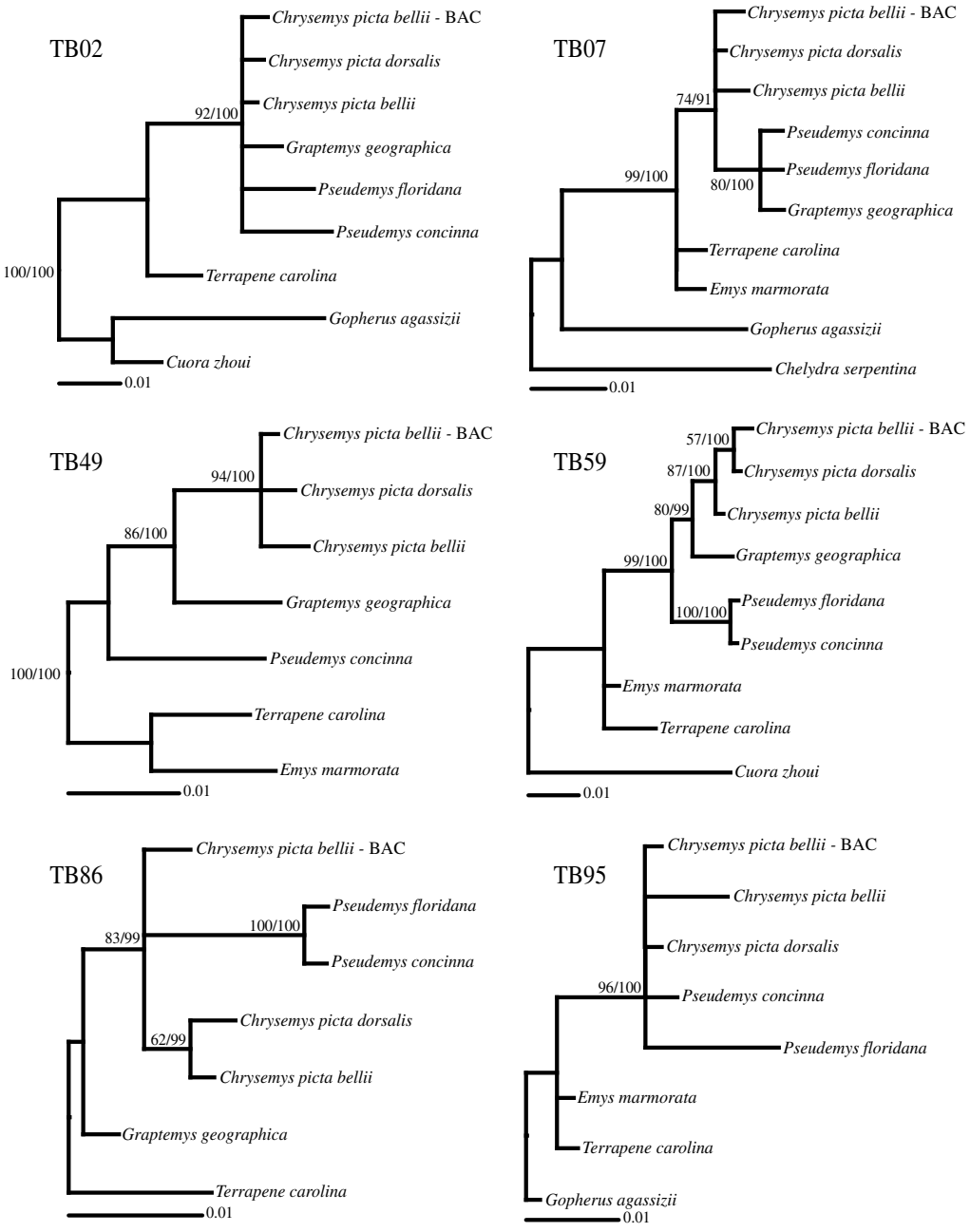

Fig. 3. Individual gene trees for each sequenced marker that produced useable sequence across emydids. Numbers on the nodes refer to ML bootstraps and posterior probabilities, respectively.

single nucleotide polymorphisms, given that the BAC-identified sequences yield a SNP about every 250 base pairs across loci and species (Shaffer and Thomson, 2007).

Our limited sequencing results suggest that the low and no similarity marker sets may be more variable than the high similarity set (Table 3). If this is the case, it could reflect the high similarity markers sitting in more conserved regions of the genome than the other two marker classes. However, even if this difference is real, it appears to be a relatively small effect, and all markers were phylogenetically informative (Figs. 2 and 3). The single exception was marker TB17, a repetitive marker that we sequenced, which provided no phylogenetic information (Fig. 2). Given that the high similarity markers have at least some annotation information, it may be desirable to favor these for phylogenetics even if they evolve at a somewhat slower rate.

Our analysis of the identification and phylogenetic utility of repetitive sequences produced mixed results. Our repeat screening identified about a third of the sequences in our BAC end library as potential repeats. However, these repetitive markers showed no detectable difference in their patterns of PCR success compared to non-repetitive markers. At the sequence level, one sequenced repetitive marker (TB17) showed obvious differences from the putative single copy marker sequences and was phylogenetically uninformative. This was the only marker in our dataset to show this pattern and we find it encouraging that none of the markers that we expected to be single copy showed a similar pattern. However, the other sequenced repetitive marker did not show such a pattern. If many other markers derived from flagged repetitive elements show the pattern of inflated "heterozygosity", and if potential markers are not in short supply, then the most efficient

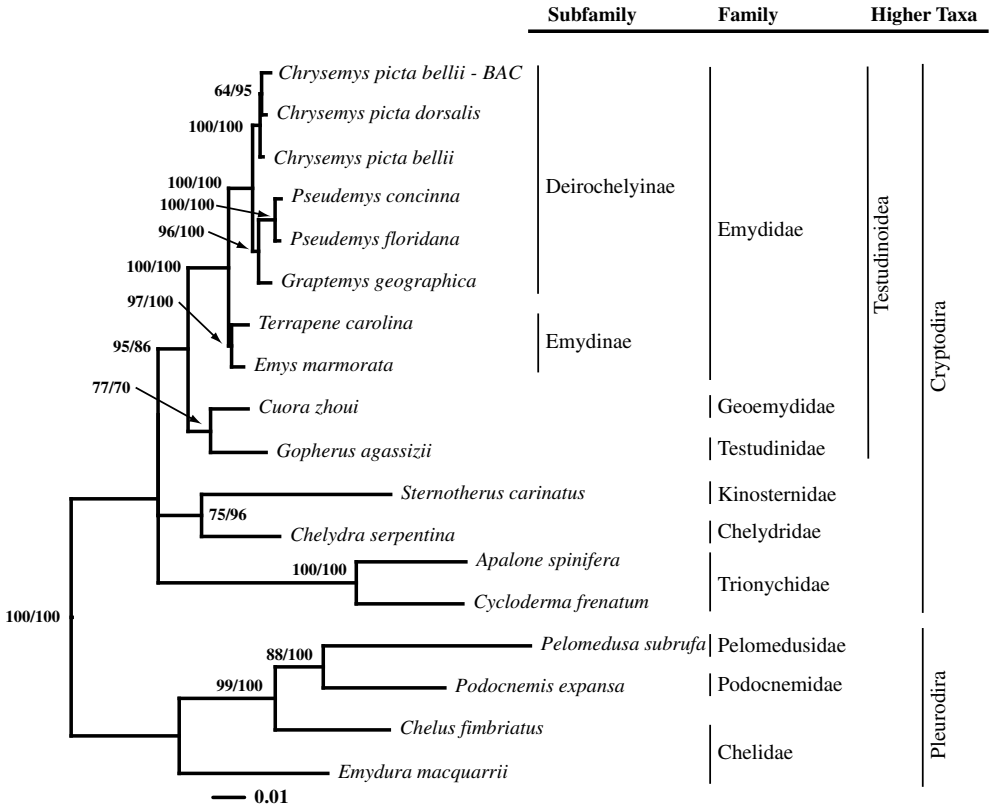

Fig. 4. Single concatenated tree based on all sequence data. Numbers on the nodes refer to ML bootstraps and posterior probabilities, respectively. Higher taxon names follow those used in Turtle Taxonomy Working Group (2007).

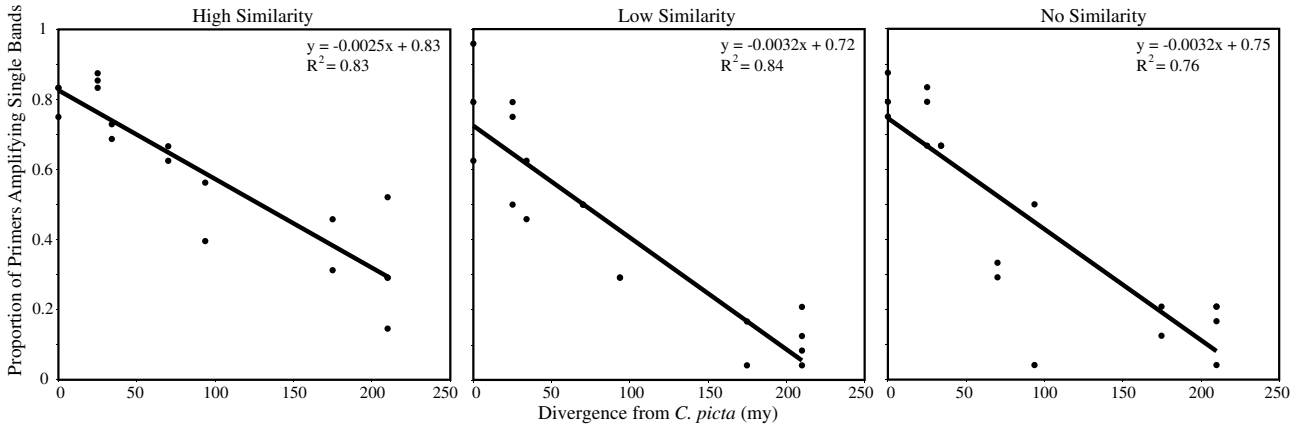

Fig. 5. Regression of the proportion of primers for each similarity class that amplified single bands (green cells in Fig. 1) against the divergence time from *Chrysemys* in millions of years. Divergence time estimates are from (Near et al., 2005) and Spinks et al. (unpublished).

strategy for determining phylogenetic markers is probably to exclude all potential repetitive elements, and to develop and test primers only for non-repetitive sequences.

4.2. Implications for turtle systematics

The results of our phylogenetic analyses are encouraging in that they largely agree with the existing understanding of the family-level turtle tree, thus suggesting that these BAC-derived markers are phylogenetically informative. These results are also encouraging in that they appear to be informative even at low levels of divergence, particularly based on the concatenated gene tree results (Fig. 4). For example, within the Emydidae the widely recognized subfamilies Emydinae and Deirochelyinae were recovered as reciprocally monophyletic (Gaffney and Meylan, 1988). At a closer level, the sister group relationship of *Graptemys* and *Pseudemys* with respect to *Chrysemys* has been suggested previously (Iverson et al., 2007; Stephens and Wiens, 2003) and here received strong support from some individual genes (Fig. 2) and in the concatenated analysis (Fig. 4).

The BEST method returned much lower posterior probabilities than the concatenation analysis, a result that has been observed in empirical datasets before (Belfiore et al., 2008; Brumfield et al., in press; Edwards et al., 2007). It remains to be seen whether this difference in support between concatenated and unconcatenated datasets is a universal phenomenon, and if so, which

approach gives the more accurate estimates of confidence. It also may be that missing data more drastically affect phylogenetic analysis using the BEST method than using concatenation.

At the genus level, we find strong support for the monophyly of *Chrysemys*, although these data do not clearly resolve the within-genus relationships. Starkey et al. (2003) recommended elevation of *C. p. dorsalis* to full species based on analysis of mitochondrial DNA. While our results do not unambiguously support this elevation, they also do not challenge it. Most genes did not strongly conflict with the monophyly of *C. p. bellii* with respect to *C. p. dorsalis*, and the two genes that resolved a relationship between the three *Chrysemys* individuals (TB 59 and TB 86) did not agree on a single grouping, suggesting that any divergence between these taxa is too recent for strong phylogenetic signal to have evolved at many nuclear loci (Jennings and Edwards, 2005). The smaller population size of the mitochondrion would lead one to expect that, under the neutral coalescent, it would sort to monophyly faster than the average nuclear gene (Hudson and Coyne, 2002) and so we do not view these data as conflicting with *C. p. dorsalis* as a full species. Rather, these results suggest that more data are required to resolve this issue. It is possible that *C. p. dorsalis* is in its early stages of speciation and is thus not easily recoverable as a monophyletic group at the nuclear DNA level even though it is already a distinct metapopulation lineage (sensu de Queiroz, 1998). If this is the case, corroboration of the mtDNA results for this group may require approaches that more explicitly incorporate allele frequency variation (Pritchard et al., 2000; Reich et al., 2008) and rely less on demonstrating gene tree monophyly (Knowles and Carstens, 2007; Shaffer and Thomson, 2007).

4.3. Further applications

Because genomic resources are becoming more common, the general strategy outlined here should be applicable for marker development in other taxa. While a large proportion of the sequences contained in the GenBank GSS and EST databases are derived from a relative few model organisms, the taxonomic coverage is broad enough to be useful across a wide-range of taxa. Drawing a few examples from vertebrates, the database currently contains EST and/or GSS sequences from all of the major lineages of Amphibia except for caecilians, all the major lineages of Reptilia, most of the major lineages of Mammalia and many of the major lineages of the Actinopterygii.

We emphasize that although the *Chrysemys* BAC library greatly facilitated the development of our markers it was not specifically required for our study. A basic requirement for the generation of multiple novel nuclear markers is any sort of clone library, including small-insert plasmid libraries such as are routinely used for characterizing microsatellites. Had we been interested in chromosomally linked markers, especially those separated by tens to hundreds of kilobases, then the BAC library would have been essential. But, in general we isolated single loci from each BAC clone, and in principle our dataset could have been generated just as easily from a small-insert plasmid library.

The approach we have employed here could also easily be adapted to find other marker types, depending on the needs of a particular study. We generally attempted to avoid sequencing repetitive elements because we do not expect them to be useful for sequence-based phylogenetic analysis. However the sequences we flagged as repetitive consist of many potentially valuable markers such as microsatellites and interspersed nuclear elements. It would be straightforward to filter out non-repetitive sequences and design primers only for, say, microsatellites instead. SINE and LINE loci could also be easily targeted by extracting retroelement sequence from the databases and designing primers in flanking regions for multilocus insertion analysis (Shedlock et al., 2004), which has been employed successfully to resolve geomydid turtle relationships (Sasaki et al., 2006). Our repeat-removal and filtering steps identified ~206,000 base pairs of such repetitive sequence, potentially representing hundreds of useful markers.

5. Conclusion

The work reported here suggests that BAC end-sequences and other low-coverage genomic resources constitute a rich source of markers for phylogenetic and phylogeographic studies. The techniques we employ require only basic bioinformatics and relatively little effort to go from a single-species resource to a clade-wide phylogenetic database. The development of 96 primer pairs and phylogenetic screening of a dozen genes cost a few thousand dollars, and the bulk of the marker development was accomplished in about a month. It is our hope that these markers can be put to direct use in turtle systematics and that these techniques can be employed in other clades to help alleviate the paucity of nuclear markers available for many organisms.

Acknowledgments

We thank J. Robert Macey and the Joint Genome Institute for supplying DNA from the BAC turtle. Matt Aresco, Tag Engstrom, Elmar Meier, C. Richard Tracy, and Nicole Valenzuela supplied tissues that were used in this study. Phil Spinks, Allan Larson, and two anonymous reviewers provided useful comments on an earlier version of this manuscript. Initial BAC library construction was supported by the NSF (IBN-0207870) and the primer development work was supported in part by Grants from the NSF (DEB-0507916, DEB-0213155, DEB-0817042), an NSF Doctoral Dissertation Improvement Grant to RCT (DEB-0710380), the UC Davis Centers for Population Biology and Center for Biosystematics, and the UC Davis Agricultural Experiment Station.

References

Avise, J.C., Bowen, B.W., Lamb, T., Meylan, A.B., Bermingham, E., 1992. Mitochondrial DNA evolution at a turtle's pace. Evidence for low genetic variability and reduced microevolutionary rate in the testudines. Mol. Biol. Evol. 9, 457–473.

Backstrom, N., Fagerberg, S., Ellegren, H., 2008. Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. Mol. Ecol. 17, 964–980.

Ballard, J.W.O., Chernoff, B., James, A.C., 2002. Divergence of mitochondrial DNA is not corroborated by nuclear DNA, morphology, or behavior in *Drosophila simulans*. Evolution 56, 527–545.

Ballard, J.W.O., Kreitman, M., 1995. Is mitochondrial DNA a strictly neutral marker? Trends Ecol. Evol. 10, 485–488.

Bardeleben, C., Moore, R.L., Wayne, R.K., 2005. A molecular phylogeny of the Canidae based on six nuclear loci. Mol. Phylogenet. Evol. 37, 815–831.

Belfiore, N.M., Liu, L., Moritz, C., 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys*. Syst. Biol. 57, 294–310.

Brito, P.H., Edwards, S.V., 2008. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica. in press, doi:10.1007/s10709-008-9293-3.

Brown, W.M., George Jr., M., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA 76, 1967.

Brumfield, R.T., Liu, L., Lum, D.E., Edwards, S.V., in press. Comparison of species tree methods for reconstructing the phylogeny of Bearded Manakins from multilocus sequence data.

Carreras-Carbonell, J., Macpherson, E., Pascual, M., 2007. Utility of pairwise mtDNA genetic distances for predicting cross-species microsatellite amplification and polymorphism success in fishes. Conserv. Genet. 9, 1572–9737.

Couzin, J., 2002. NSF's ark draws alligators, algae, and wasps. Science 297, 1638–1639.

de Queiroz, K., 1998. The general lineage concept of species, species criteria, and the process of speciation. In: Howard, D.J., Berlocher, S.H. (Eds.), Endless Forms: Species and Speciation. Oxford University Press, pp. 57–75.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936.

Engstrom, T.N., Edwards, T., Osentoski, M.F., Myers, E.M., 2007. A compendium of PCR primers for mtDNA, microsatellite, and other nuclear loci for freshwater turtles and tortoises. In: Shaffer, H.B., FitzSimmons, N.N., Georges, A., Rhodin,

A.G.J. (Eds.), Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Tortoises and Freshwater Turtles. Chelonian Research Monographs, Lunenburg, MA, pp. 116–133.
FitzSimmons, N., Moritz, C., Moore, S.S., 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Mol. Biol. Evol. 12, 432–440.
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.
Gaffney, E.S., Meylan, P.A., 1988. A phylogeny of turtles. In: Benton, M.J. (Ed.), The Phylogeny and Classification of the Tetrapods. Oxford University Press, New York, pp. 157–219.
Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. Evolution 56, 1557–1565.
Hutchison, J., 1996. Testudines. In: Prothero, D.R., Emry, R.J. (Eds.), The Terrestrial Eocene–Oligocene Transition in North America. Cambridge University Press, Cambridge, pp. 337–354.
Irwin, D.E., 2002. Phylogeographic breaks without geographic barriers to gene flow. Evolution 56, 2383–2394.
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. (Eds.), Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises. Chelonian Research Monographs, Cambridge, MA, pp. 85–106.
Jennings, W.B., Edwards, S.V., 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution 59, 2033–2047.
Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895.
Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S.V., Paabo, S., Villablanca, F.X., Wilson, A.C., 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA 86, 6196–6200.
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.
Kumazawa, Y., Nishida, M., 1999. Complete mitochondrial DNA sequences of the green turtle and blue-tailed mole skink: statistical evidence for archosaurian affinity of turtles. Mol. Biol. Evol. 16, 784–792.
Li, C., Ortí, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7, 44.
Liu, L., Pearl, D.K., 2007. Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56, 504–514.
Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E., O'Brien, S.J., 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nat. Genet. 15, 47–56.
Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Mindell, D.P., Sorenson, M.D., Dimcheff, D.E., Hasegawa, M., Ast, J.C., Yuri, T., 1999. Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. Syst. Biol. 48, 138–152.
Near, T.J., Meylan, P.A., Shaffer, H.B., 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. Am. Nat. 165, 137–146.
Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.
Palumbi, S.R., Baker, C.S., 1994. Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. Mol. Biol. Evol. 11, 426–435.
Parham, J.F., Macey, J.R., Papenfuss, T.J., Feldman, C.R., Turkozan, O., Polymeni, R., Boore, J., 2006. The phylogeny of Mediterranean tortoises and their close relatives based on complete mitochondrial genome sequences from museum specimens. Mol. Phylogenet. Evol. 38, 50–64.
Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.
Primmer, C.R., Moller, A.P., Ellegren, H., 1996. A wide-range survey of cross-species microsatellite amplification in birds. Mol. Ecol. 5, 365–378.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.
Reich, D., Price, A.L., Patterson, N., 2008. Principal component analysis of genetic data. Nat. Genet. 40, 491–492.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Rozen, S., Skaletsky, H., 2000. Primer3 on the www for general users and for biologist programmers. In: Krawetz, S., Misener, S. (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 365–386.
Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463–5467.
Sasaki, T., Yasukawa, K., Takahashi, K., Miura, S., Shedlock, A.M., Okada, N., 2006. Extensive morphological convergence and rapid radiation in the evolutionary history of the family Geoemydidae (Old World Pond Turtles) revealed by SINE insertion analysis. Syst. Biol. 55, 912–927.
Shaffer, H.B., FitzSimmons, N.N., Georges, A., Rhodin, A.G.J. (Eds.), 2007. Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises. 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., Thomson, R.C., 2007. Delimiting species in recent radiations. Syst. Biol. 56, 896–906.
Shedlock, A.M., 2006. Phylogenomic investigation of CR1 LINE diversity in reptiles. Syst. Biol. 55, 902–911.
Shedlock, A.M., Botka, C.W., Zhao, S., Shetty, J., Zhang, T., Liu, J.S., Deschavanne, P.J., Edwards, S.V., 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. Proc. Natl. Acad. Sci. USA 104, 2767–2772.
Shedlock, A.M., Takahashi, K., Okada, N., 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol. Evol. 19, 545–553.
Smit, A., Hubley, R., Green, P., 1996–2004 RepeatMasker Open-3.0.
Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., Hood, L.E., 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321, 674–679.
Spinks, P.Q., Shaffer, H.B., 2007. Conservation phylogenetics of the Asian box turtles (Geoemydidae, *Cuora*): mitochondrial introgression, numts, and inferences from multiple nuclear loci. Conserv. Genet. 8, 641–657.
Starkey, D.E., Shaffer, H.B., Burke, R.L., Forstner, M.R.J., Iverson, J.B., Janzen, F.J., Rhodin, A.G.J., Ultsch, G.R., 2003. Molecular systematics, phylogeography, and the effects of Pleistocene glaciation in the painted turtle (*Chrysemys picta*) complex. Evolution 57, 119–128.
Stephens, P.R., Wiens, J.J., 2003. Ecological diversification and phylogeny of emydid turtles. Biol. J. Linn. Soc. 79, 577–610.
Townsend, T.M., Alegre, R.E., Kelley, S.T., Wiens, J.J., Reeder, T.W., 2008. Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. Mol. Phylogenet. Evol. 47, 129–142.
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. (Eds.), Defining Turtle Diversity: Proceedings of a Workshop on Genetics, Ethics, and Taxonomy of Freshwater Turtles and Tortoises. Chelonian Research Monographs, pp. 173–199.
Zardoya, R., Meyer, A., 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. Proc. Natl. Acad. Sci. USA 95, 14226–14231.