Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 47 (2008) 129–142
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: An example from squamate reptiles

Ted M. Townsend [a,*], R. Eric Alegre [a,1], Scott T. Kelley [a], John J. Wiens [b], Tod W. Reeder [a]

[a] Department of Biology, 5500 Campanile Drive, San Diego State University, San Diego, CA 92182-4164, USA
[b] Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA

Received 10 May 2007; revised 27 December 2007; accepted 3 January 2008
Available online 24 January 2008

Abstract

Recently, as genome-scale data have become available for more organisms, the development of phylogenetic markers from nuclear protein-coding loci (NPCL) has become more tractable. However, new methods are needed to efficiently sort the large number of genes from genomic databases into more limited sets appropriate for particular phylogenetic questions, while avoiding introns and paralogs. Here we describe a general methodology for identifying candidate single-copy NPCL from genomic databases. Our method uses information from reference genomes to identify genes with relatively large continuous protein-coding regions (i.e., $\geqslant$700 bp). BLAST comparisons are used to help avoid genes with paralogous copies or close relatives (i.e., gene families) that might confound phylogenetic analyses. Exon boundary information is used to identify appropriately spaced potential priming sites. Using this method, we have developed over 25 novel NPCL, which span a variety of desirable evolutionary rates for phylogenetic analyses. Although targeted for higher-level phylogenetics of squamate reptiles, many of these loci appear to be useful across and within other vertebrate clades (e.g., amphibians), and some are relatively rapidly evolving and may be useful for closely-related species (e.g., within genera). This general method can be used whenever large-scale genomic data are available for an appropriate reference species (not necessarily within the focal clade). The method is also well suited for the development of intron regions for lower-level phylogenetic and phylogeographic studies. We provide an online database of alignments and suggested primers for approximately 85 NPCL that should be useful across vertebrates.

Keywords: Genes; Genomics; Phylogenetics; Primers; Reptiles; Vertebrates

1. Introduction

Traditionally, most molecular phylogenetic studies in animals used only mitochondrial genes (e.g., Burns, 1997; Heise et al., 1995; Honeycutt and Adkins, 1993; Ritchie et al., 1997) and numerous phylogenetic studies continue to be published that are based on mitochondrial data alone (e.g., Hyman et al., 2007; Klicka et al., 2007; Lemmon et al., 2007). The ease of amplification and relatively fast evolutionary rate of mitochondrial sequences have made them extremely useful to systematists and population biologists (Avise, 1986; Ballard and Rand, 2005; Brown, 1985; Funk and Omland, 2003; Harrison, 1989; Simon et al., 2006).

However, because the mitochondrial genome is inherited as a unit, the individual genes within it cannot be regarded as independent sources of phylogenetic information (Brown, 1985; Harrison, 1989). The use of mitochondrial data alone is therefore potentially problematic at lower taxonomic levels because of issues such as introgression and incomplete lineage sorting (Funk and Omland, 2003 and references therein). At the same time, many empirical studies suggest that mitochondrial genes may often evolve too rapidly and heterogeneously to be effective for many higher-level phylogenetic analyses. For example, phylogenetic analyses based on mitochondrial DNA that examined deep relationships within salamanders (Weisrock et al.,

* Corresponding author. Fax: +1 619 594 5676.
E-mail address: ttownsend@projects.sdsu.edu (T.M. Townsend).
[1] Present address: Computational Biosciences Program, Arizona State University, Tempe, AZ 85287, USA.

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.01.008

2005), mammals (Arnason et al., 2002), and reptiles (Douglas et al., 2006; Zhou et al., 2006) have all recovered controversial relationships at odds with strongly supported nuclear phylogenies (Murphy et al., 2001b; Townsend et al., 2004; Vidal and Hedges, 2005; Wiens et al., 2005). These problems of high and heterogeneous rates of change in mitochondrial genes may even create problems of long-branch attraction at lower taxonomic levels (e.g., among genera within vertebrate families; Wiens and Hollingsworth, 2000).

The nuclear genome contains protein-coding, RNA-coding, and non-coding regions, and offers a wealth of independent and unlinked markers evolving at a variety of rates. However, development of nuclear genes for phylogenetic analysis has historically been more difficult than for mitochondrial genes. Non-coding regions (e.g., introns) and loop regions of rRNA genes generally evolve more rapidly, thus making them potentially useful among closely-related species (e.g., Dolman and Phillips, 2004; Gaines et al., 2005; Sequeira et al., 2006; Weibel and Moore, 2002; Willows-Munro et al., 2005). Unfortunately, such regions are also prone to marked length variation that makes alignment generally more difficult, especially at higher taxonomic levels (Matthee et al., 2001; Sequeira et al., 2006; Sotoadames et al., 1994). In contrast, nuclear protein-coding loci (NPCL) can be far easier to align because they are less prone to excessive length variation (Boekhorst and Snel, 2007), any length variation present must occur in multiples of three, and nucleotide sequences can be translated to (more conserved) amino acid sequences to help constrain and guide alignment. These advantages make NPCL an attractive alternative to data from mitochondrial genes or nuclear RNA or non-coding regions, especially for analyses of higher-level phylogeny.

There are nonetheless several obstacles to developing NPCL as phylogenetic markers. Perhaps the greatest problem is the widespread presence of introns within these genes. Messenger RNA (mRNA) sequence data have long been available for many nuclear proteins from a diversity of taxa, making possible the design of primers complementary to conserved exon-coding regions. However, without the corresponding genomic sequence (within which the coding regions of a gene are interspersed), determining the exon boundaries of a particular gene can be difficult. Without knowledge of these exon boundaries, primer design is a very hit-or-miss process (i.e., primers designed to amplify a few hundred bases of exon sequence may actually span several thousand bases of non-coding intron sequence).

Another obstacle is the difficulty of detecting paralogous gene copies or members of closely-related gene families. If these paralogs are inadvertently amplified in some taxa, the resulting gene trees may not reflect the true species histories, and there may be strong statistical support for a misleading species phylogeny (Downie and Gullan, 2004; Maddison, 1997; Mitchell and Wen, 2004; Sword et al., 2007).

A final obstacle is the sheer size of the nuclear genome. In recent years, the amount of genomic sequence data for animals has risen dramatically, and many whole genomes are now completed in at least draft form (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj). But given that tens of thousands of potential loci are available, identifying particular loci with desirable properties using non-automated methods is somewhat impractical (or at least daunting).

Perhaps because of these obstacles, most phylogenetic studies of animals incorporating NPCL have been based on a few "stock" genes (e.g., *CMOS, RAG1*), with only a few exceptions (e.g., Bardeleben et al., 2005; Li et al., 2007; Murphy et al., 2001a; Roelants et al., 2007; Vidal and Hedges, 2005). Many of these "stock" loci are single exon genes that, due to their lack of introns, can be developed without genomic information. However, methods are clearly needed that can extract large numbers of useful phylogenetic loci from nuclear genomic databases.

Li et al. (2007) recently described a method of identifying NPCL for phylogenetic analyses using ray-finned fishes (Actinopterygii) as their study system. Their method involved automated BLAST comparisons of whole genome sequences of two fish, *Danio rerio* and *Fugu* (*Takifugu*) *rubripes*. Homologous exon regions were identified and aligned, and consensus primers were designed from these two species. The authors succeeded in developing primers for 10 relatively conserved NPCL that appear to be useful for higher-level fish systematics.

We have developed a similar approach for generating new nuclear loci for phylogenetic analysis using genomic databases. Although we illustrate this approach with a particular group of vertebrates (squamate reptiles = lizards and snakes), our general approach should be applicable to almost any group of organisms for which one or more complete nuclear genomes are available. Furthermore, many of the loci and associated primers that we have developed specifically for squamates seem to be broadly applicable across vertebrates.

The squamate Tree of Life project (Deep Scaly) is a multidisciplinary effort funded by the US National Science Foundation to resolve the phylogenetic relationships among the major groups of squamate reptiles. A major component of this project is the development of 50 NPCL not previously used for phylogenetic analyses in Squamata. At the time this study was initiated, the chicken (*Gallus gallus*) was the closest relative to squamate reptiles for which the nuclear genome had been sequenced and made available (Hillier et al., 2004). We have used information from the *Gallus* genome (along with that from the pufferfish [*Fugu rubripes*] and several mammalian species) in conjunction with search tools on the NCBI website to develop a number of nuclear loci for phylogenetic analysis over the past three years. Here we describe the relatively simple and straightforward method that we used to identify and develop these loci. This general method can be used to develop novel loci for a variety of taxonomic groups and hierarchical levels.

2. Materials and methods

2.1. Overview of method

The general strategy of our method was to first identify NPCL likely to be present across vertebrates, based on their presence in the genomes of both *Homo sapiens* and *Fugu rubripes* (pufferfish). These NPCL were then filtered to retain only those of appropriate size and evolutionary rate for our phylogenetic analyses, and that seemed to be single-copy. Finally, these candidate genes were compared to their homologs in other amniotes to develop primers for loci useful for squamate phylogenetic studies. Importantly, although squamates were the focal group, the primers were used to amplify outgroup taxa from all other major amniote groups. These outgroup taxa included mouse (*Mus musculus*), echidna (*Tachyglossus aculeatus*), snapping turtle (*Chelydra serpentina*), giant Amazon river turtle (*Podocnemis expansa*), crocodile (*Crocodylus* sp.), American alligator (*Alligator mississippiensis*), emu (*Dromaius novaehollandiae*), and tuatara (*Sphenodon punctatus*).

The procedure can be divided into three general phases: Phase 1 was the identification of candidate vertebrate protein-coding genes by BLASTing the pufferfish genome against the human genome. Note that in this paper all genes are referred to by the official abbreviations of their respective human homologs (as approved by the HUGO Gene Nomenclature Committee, http://www.gene.ucl.ac.uk/nomenclature/). Phase 2 was the identification of the homologs of these Phase 1 candidate genes in the chicken genome, examination of exon boundaries, and identification of potential primer sites flanking variable areas within individual exons. Phase 3 was the alignment of all available amniote sequences to allow primer design. Fig. 1 gives a schematic overview of the entire NPCL discovery procedure.

2.2. Identification of potential vertebrate loci

The vertebrate genes (represented by their human homologs) found in Phase 1 had to meet several criteria. To maximize efficiency, an effort was made to develop the longest gene fragments possible that could be sequenced completely in both directions with only two sequencing reactions (i.e., approximately 500–800 bp). Because each amplified fragment had to be contained within a single exon, candidate genes were limited to those containing at least one exon $\geqslant$ 250 amino acids (aa) long (Fig. 1, Phase 1.2). The genes also needed sufficient variability to be potentially useful for phylogenetic analyses in squamates. Because it was not clear initially what level of *Homo–Fugu* divergence would correspond to the desired level of variability in squamates, results were sorted into multiple bins based on three different levels of aa divergence (Fig. 1, Phase 1). Finally, to lessen the chances of developing genes with paralogs or other close relatives, a gene was excluded if the *Fugu* protein sequence significantly matched more than one distinct *Homo* protein (Fig. 1, Phase 1.3).

Phase 1 included three steps that were largely automated using Python scripts written by REA with help from STK (see Fig. 1). All of these programs are publicly available on our website (http://www.fieldmuseum.org/deepscaly/data.html). Step 1 in Phase 1 involved identifying appropriately sized human proteins and sorting them based on evolutionary rate (inferred from levels of aa divergence between *Fugu* and *Homo*). To accomplish this, *Fugu* and *Homo* genome protein databases (db) were downloaded. The *Fugu* protein db (ftp://ftp.jgi-psf.org/pub/JGI_data/Fugu/fugu_v3_prot.fasta.Z) was generated from expressed sequence tags (EST). This db contained both complete and incomplete protein sequences, and different portions of the same protein were sometimes present under several sequence identification numbers. The *Homo* genome protein db (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.34.3/protein/protein.fa.gz) contained complete sequence for all known human proteins. This *Homo* db was reformatted (instructions at http://www.ncbi.nlm.nih.gov/Class/BLAST/blast_course.short.html#STAND), and the *Fugu* EST db was BLASTed against it (BLAST files downloaded from ftp://ftp.ncbi.nih.gov/blast/). The *Fugu* db was then filtered to retain only those protein sequence fragments that were significant matches to human proteins $\geqslant$ 250 aa long, and the fragments were parsed to three files based on the degree of *Fugu–Homo* aa similarity (40–60%, 61–80%, and 81–90%; Fig. 1, Phase 1.1). These levels of aa similarity were chosen somewhat arbitrarily. Previous experience with BLAST searches suggested that accepting similarity scores <40% often leads to non-homologous pairings. Furthermore, we considered levels of aa similarity of >90% between the distantly-related *Homo* and *Fugu* to be unlikely to yield a large number of informative characters within squamates, especially for a sequence fragment only a few hundred base pairs long. In summary, this step substantially reduced the size of the *Fugu* db file for the subsequent steps.

The second step in Phase 1 was to further filter the list of matching *Homo–Fugu* aa sequences by incorporating exon boundary information for *Homo*. To accomplish this, we downloaded a db of human proteins that contained information on the number and size of all exons (Human ExInt file, http://sege.ntu.edu.sg/wester/exint138/). The three *Fugu* dbs from the previous step were each BLASTed against this human db, and only those *Homo–Fugu* matches containing a human protein with at least one exon $\geqslant$ 250 aa long were retained (Fig. 1, Phase 1.2).

The third step in Phase 1 was to filter out genes with potential paralogous copies or very close gene family relatives. This was accomplished by simply discarding any *Homo–Fugu* pairs containing a *Fugu* sequence that significantly matched more than one *Homo* accession number in a BLAST search (Fig. 1, Phase 1.3). This step undoubtedly eliminated some potentially useful genes, because the

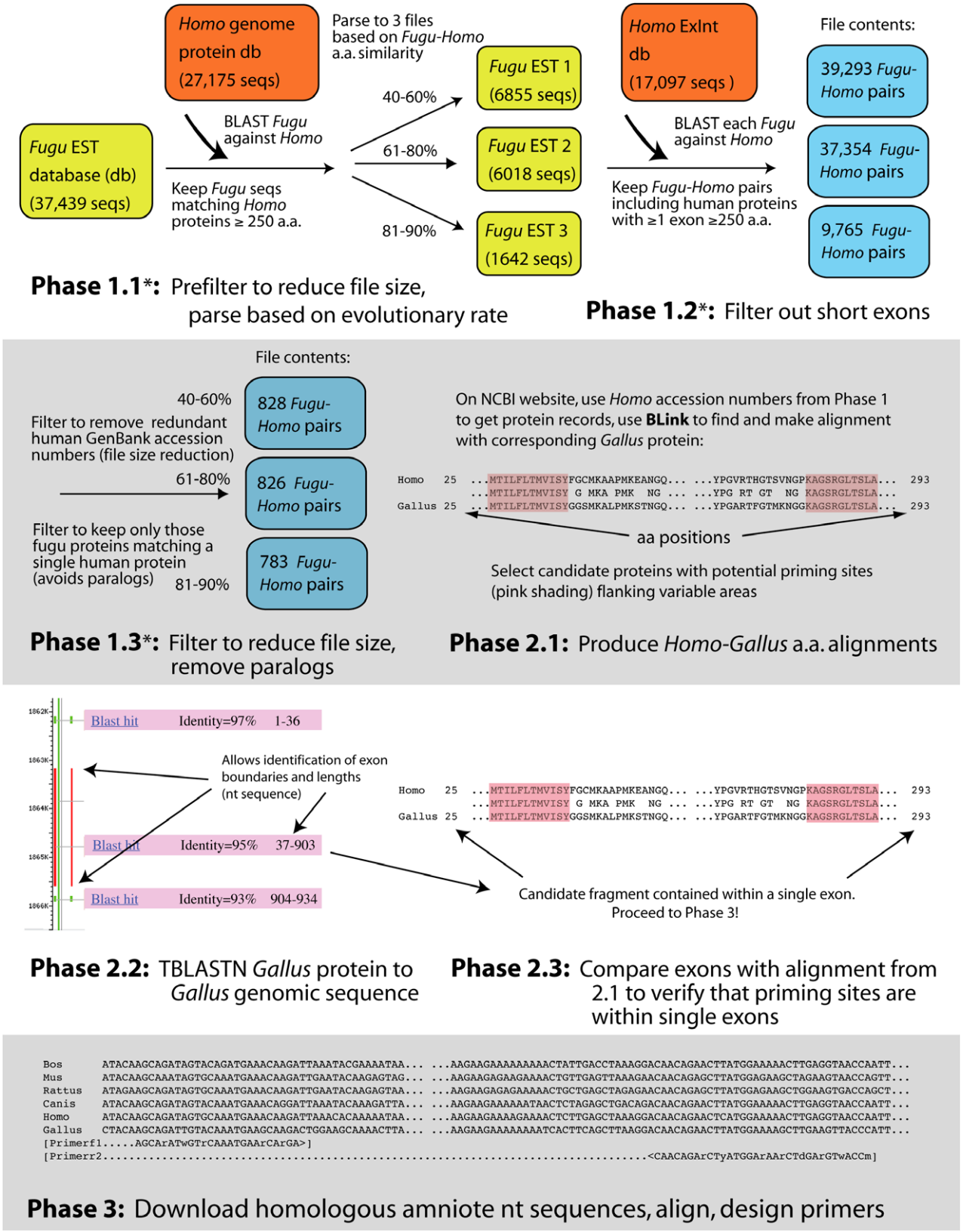

Fig. 1. Schematic overview of the method described here for development of novel nuclear loci for phylogenetics studies in squamates and other vertebrates. Asterisks indicate steps that were automated via Python scripts.

same human protein is often represented in GenBank by more than one separate submission (and therefore accession number). However, we felt it was important to be conservative in this step to reduce potential paralogy problems

in future phylogenetic analyses. Ideally, one would eliminate matches to multiple protein names or symbols instead of multiple accession numbers. Unfortunately, this is impractical because neither protein names nor symbols are standardized across taxa. Each entry in the final files thus consisted of the *Fugu* sequence fragment number, the GenBank accession number of the *Homo* gene, and the name of the *Homo* gene as given to GenBank by the submitting researcher.

2.3. Evaluation of variation and exon boundaries in Gallus

Phase 2 was the identification of candidate proteins from Phase 1 that in *Gallus* (our proxy for a reference squamate genome): (1) contained at least one exon $\geqslant$ 250 aa long, (2) contained potential primer sites within these exons to amplify-coding fragments $\geqslant$ 500 bp across amniotes, and (3) had no evidence of paralogous copies or close relatives elsewhere in the chicken genome. The first step in this process was to bring up the nucleotide-level record associated with a *Homo* accession number from one of the three resultant files from Phase 1. Next, a link from this record was followed to the corresponding *Homo* protein sequence, which was then BLASTed against other proteins in GenBank using the BLink function accessible from within each protein record. The resulting list of protein matches (ranked by level of similarity) was examined for certain favorable patterns. Specifically, the list would ideally begin with several mammalian matches, then a chicken match, a match to one or more other amniotes (e.g., a turtle, crocodilian, squamate, or other bird sequence, although these were rarely encountered), and then to one or more non-amniote vertebrates (e.g., frog, salamander, and/or fish). Also, we required that the above matches all referenced the same named protein. If there were intermingled matches to two obviously distinct proteins, for example, this could signal undesirable paralogs or closely-related genes. This last point often required a little further exploration, because as mentioned above, the gene names associated with GenBank records are not standardized, and not all researchers use the same name for a given gene or protein.

If the above criteria were satisfied, a link was followed to the *Homo–Gallus* alignment for that protein (Fig. 1, Phase 2.1). This alignment was examined for level of aa divergence and distribution of aa variation along the length of the protein in relation to potential primer regions (i.e., conserved sequence blocks of $\geqslant$10 identical aa). Because the goal was to sequence fragments ~500–800 bp long, if suitable regions $\geqslant$ 250 aa long were not found, the protein was discarded.

Note that the BLink protein alignments contain no information about exon numbers and boundary positions. Even an "ideal" candidate fragment identified at this step could still have one or more introns, each potentially thousands of bp long, dividing the fragment into multiple exons (and making the gene impractical for our purposes). To determine exon boundaries in the chicken genes, under the assumption that these boundaries are similar in squamates, an online BLAST search was conducted using the *Gallus* protein sequence against the *Gallus* genomic sequence using the TBLASTN program (http://www.ncbi.nlm.nih.gov/genome/seq/GgaBlast.html). As in Phase 2.1, the result was a ranked list of matches. However, because the BLAST search was against genomic sequence, this time any continuous run of protein sequence that scored a significant match had to represent an individual exon or portion thereof (Fig. 1, Phase 2.2, illustrating NCBI's MapViewer function).

In addition to providing exon boundaries, this step also served as a final check for paralogous genes. The presence of paralogs was inferred if the *Gallus* protein sequence scored a significant match to more than one *Gallus* chromosome or genomic region. If no paralogs were detected by this step, the start and stop positions of any suitably sized exon(s) were compared to the start and stop positions of the candidate fragment identified in the human–chicken BLAST search (Fig. 1, Phase 2.3). If the candidate fragment was fully contained within a single exon, it was selected for primer development.

It should be noted that any gene duplications occurring within squamates after their split from archosaurs (i.e., the clade composed of the bird and crocodilian lineages) would not be detected by our protocol. However, the recent completion of the genome of *Anolis carolinensis* (an iguanian lizard) should partially mitigate this concern for future work by allowing BLAST searches for paralogs against that genome.

2.4. Primer development

Phase 3 (primer development) began with an alignment of all available amniote homologs for each gene retained up to this point, which were identified from the BLink output described above and downloaded as full GenBank nucleotide (nt) sequence files. These files were loaded into the VectorNTI program (Invitrogen) and aligned using the Clustal W algorithm (Thompson et al., 1994) with gap-opening and gap-extension parameters set at their respective defaults for both pairwise (10.0 and 0.1) and multiple (10.0 and 0.2) alignments. Vector NTI was further used to design preliminary primer sets using its PCR Analysis Protocol. Because the alignments from which these primers were designed usually consisted of several mammals and the chicken, the resulting primers tended to be biased toward mammals. These preliminary primer sets were therefore used mainly as a means of easily locating the most conserved regions, and the final hand-tuned primers were (in many cases) deliberately biased toward the *Gallus* sequence. Whenever possible, multiple sets of nested primers were designed to maximize the chances of successful amplification.

Next, our primers were used to amplify a set of 10 squamate "test taxa" chosen to represent several well-established clades and encompass a range of divergence levels within squamates, thus allowing an evaluation of each gene's potential for resolving higher-level phylogeny. Specifically, we chose two geckos (*Coleonyx variegatus, Gekko gecko*) representing a putative basal clade of squamates (Townsend et al., 2004; Vidal and Hedges, 2005), two acrodont, agamid iguanians (*Agama agama, Physignathus cocincinus*) and two pleurodont, phrynosomatid iguanians (*Phrynosoma platyrhinos, Uta stansburiana*) representing highly nested and well-established squamate clades (Iguania, Acrodonta, and Pleurodonta: Estes et al., 1988; Townsend et al., 2004; Vidal and Hedges, 2005; Agamidae: Frost and Etheridge, 1989; Macey et al., 2000; Phrynosomatidae: Frost and Etheridge, 1989; Schulte et al., 2003), two snakes (*Boa constrictor, Lampropeltis getula*) (Estes et al., 1988; Vidal and Hedges, 2004), and two varanid anguimorphs (*Varanus acanthurus, Varanus exanthematicus*) (Ast, 2001; Estes et al., 1988). Loci that could be readily amplified and sequenced and showed variation across these taxa were considered good candidates for our study. Those loci that did not were either discarded or, in some cases, refinements were made to the primer sequences and they were tried again. Detailed amplification protocols for the loci from this study are available on our website (http://www.fieldmuseum.org/deepscaly/data.html).

As a final check for phylogenetic usefulness and potential paralogy problems, we also conducted preliminary maximum-likelihood (ML) analyses of each gene to identify whether representatives of six well-established groups were placed together (i.e., geckos, snakes, varanids, iguanians, phrynosomatid iguanians, agamid iguanians). ML analyses were performed using GARLI v0.95 (Zwickl, 2006), which conducts heuristic searches using a genetic algorithm approach. Each gene was analyzed under the General Time Reversible (GTR, Tavaré, 1986) model or one of its submodels, as determined under the AIC criterion using ModelTest (Posada and Crandall, 1998), with starting trees built by stepwise random addition. Bootstrap analyses were performed under these same conditions in GARLI using 100 pseudoreplicates. Maximum-likelihood topologies and bootstrap results for each gene were compared to results from similar analyses of the entire, 26-gene concatenated data set. We assumed that if there were problems of paralogy or inappropriate evolutionary rates with particular genes, then analyses of these genes would fail to support many of these well-established groups, or might contradict other strongly supported results from the combined analyses.

2.5. Potential predictors of variation within squamates

The ability to easily screen genomic databases for genes that might be best suited for phylogenetic studies at particular hierarchical levels would be very helpful to researchers seeking to develop loci for specific projects (i.e., "slow" genes for higher-level studies and "fast" genes for analyses of closely-related species). Therefore, several parameters were evaluated for their usefulness in predicting general levels of divergence within squamate reptiles, focusing on those parameters that could be estimated before any laboratory work was performed. In other words, given that we obtained sequence data from 10 test taxa, we evaluated what parameters accurately predicted the levels of divergence among them (but only considering parameters estimated without having squamate sequences). As a standard for intra-squamate variability, we used average genetic distances between our 10 test taxa. The evolutionary model and parameter values for each full data set were determined by the AIC criterion using the program Modeltest, version 3.7 (Posada and Crandall, 1998). Average intra-squamate distances were calculated using PAUP* (Swofford, 2002), using the % maximum-likelihood (ML) distance. These average distances were then compared to various aa- and nt-level divergences between non-squamate taxa in our original alignments of GenBank amniote sequences to test the predictive value of these measures.

3. Results and discussion

3.1. New loci

Approximately 2500 *Homo–Fugu* homology matches resulted from the BLAST and filter procedures of Phases 1.1–1.3 (Fig. 1). From this list, over 270 *Homo* protein records were retrieved and BLASTed against GenBank records (Fig. 1, Phase 2.1; Table 1). About 190 of these BLAST searches either returned no close *Gallus* matches (suggesting the gene might be absent in squamates), returned close matches to multiple distinct proteins (suggesting the gene was not single-copy), or yielded *Gallus* proteins lacking conserved potential priming sites, These genes were discarded (Table 1). Approximately 85 NPCL fit our selection criteria (Fig. 1, Phases 2.1–2.3; Table 1). For these loci, additional vertebrate sequences were downloaded and primers were designed (Fig. 1, Phase 3; Table 1). As of October 2007, 42 loci were amplified and sequenced in at least some squamate taxa (26 loci for the test taxa) and 21 of these were amplified and sequenced for most of the project's 143 ingroup taxa. Tables 1 and 2 summarize these results.

Table 1
Summary of NPCL development results to date

	40–60% file	61–80% file	81–90% file	Totals
Genes examined[a]	210	60	9	279
Primers designed[b]	65	16	4	85
Successes[c]	26	8	2	42

[a] *Homo–Gallus* alignment made, *Gallus* exon boundaries determined.
[b] Potential priming sites found on *Homo–Gallus* alignment, other amniote sequences downloaded, primers designed.
[c] Sequence obtained for at least some squamate taxa.

Table 2
PCR primer sequences for 26 NPCL developed for this study and performance of individual genes relative to the combined data[a]

Gene[c]	Primers	Sequences	Gallus fragment length (bp)	Percentage of seven squamate clades recovered (supported)[b]
ADNP	ADNP_f5	5′ ATTGAAGACCATGARCGYATAGG 3′	811	71.4 (71.4)
	ADNP_r2	5′ GCCATCTTYTCHACRTCATTGA 3′		
AHR	AHR_f4	5′ CARGATGAGTCTRTKTATCTCT 3′	571	100 (85.7)
	AHR_r3	5′ GYRAACATSCCATTRACTTGCAT 3′		
AKAP9	AKAP9_f6	5′ AGCARATWGTRCAAATGAARCARGA 3′	1481	100 (100)
	AKAP9_r2	5′ TCHAGYTTYTCCATRAGTTCTGTTG 3′		
BACH1	BACH1_f1	5′ GATTTGAHCCYTTRCTTCAGTTTGC 3′	1330	100 (100)
	BACH1_r2	5′ ACCTCACATTCYTGTTCYCTRGC 3′		
BACH2	BACH2_f1	5′ GGKCCRYTGYTACAGTTYGCCTA 3′	562	100 (87.5)
	BACH2_r9	5′ TCTCCDGACAGGCARAGCGTGAT 3′		
BDNF	BDNF_f	5′ GACCATCCTTTTCCTKACTATGGTTATTTCATACTT 3′	670	100 (85.7)
	BDNF_r	5′ CTATCTTCCCCTTTTAATGGTCAGTGTACAAAC 3′		
BMP2	BMP2_f6	5′ CAKCACCGWATTAATATTTATGAAA 3′	590	100 (100)
	BMP2_r2	5′ CGRCACCCRCARCCCTCCACAACCA 3′		
DNAH3	DNAH3_f1	5′ GGTAAAATGATAGAAGAYTACTG 3′	721	100 (100)
	DNAH3_r6	5′ CTKGAGTTRGAHACAATKATGCCAT 3′		
ECEL1	ECEL1_f1	5′ TGACVGCVCACTAYGAYGAGTTCCARGA 3′	677	57.1 (42.8)
	ECEL1_r8	5′ CGGATGACRTAGCGSGAGGWGTTCCTGT 3′		
FSHR	FSHR_f1	5′ CCDGATGCCTTCAACCCVTGTGA 3′	753	87.5 (71.4)
	FSHR_r2	5′ RCCRAAYTTRCTYAGYARRATGA 3′		
FSTL5	FSTL5_f1	5′ TTGGRTTTATTCTTCAYAAAGA 3′	622	100 (85.7)
	FSTL5_r2	5′ YTCTSAACYTCAGTGATYTCACA 3′		
GPR37	GPR37_f7	5′ GCCACCAACGTGCAGATGTACTA 3′	706	85.7 (71.4)
	GPR37_r2	5′ CAATGAGTCCCVACAGARGCAAA 3′		
MKL1	MKL1_f1	5′ GTGGCAGAGCTGAAGCARGARCTGAA 3′	978	85.7 (85.7)
	MKL1_r2	5′ GCRCTCTKRTTGGTCACRGTGAGG 3′		
NGFB	NGFB_f2	5′ GATTATAGCGTTTCTGATYGGC 3′	573	100 (85.7)
	NGFB_r2	5′ CAAAGGTGTGTGTWGTGGTGC 3′		
NT3	NTF3_f1	5′ ATGTCCATCTTGTTTTATGTGATATTT 3′	576	85.7 (71.4)
	NTF3_r1	5′ ACRAGTTTRTTGTTYTCTGAAGTC 3′		
PNN	PNN_f1	5′ TTTGCAGARCARATAAAYAAAATGGA 3′	945	100 (100)
	PNN_r1	5′ AACGCCTTTTGTCTTTCCTGTCTGATT 3′		
PRLR	PRLR_f1	5′ GACARYGARGACCAGCAACTRATGCC 3′	532	100 (100)
	PRLR_r3	5′ GACYTTGTGRACTTCYACRTAATCCAT 3′		
PTGER4	PTGER4_f1	5′ GACCATCCCGGCCGTMATGTTCATCTT 3′	471	85.7 (71.4)
	PTGER4_r5	5′ AGGAAGGARCTGAAGCCCGCATACA 3′		
PTPN12	PTPN12_f1	5′ AGTTGCCTTGTWGAAGGRGATGC 3′	758	100 (100)
	PTPN12_r6	5′ CTRGCAATKGACATYGGYAATAC 3′		
REV3L	REV3L_f1	5′ AATGCTGAARCYGAAGAYTGTGA 3′	1554	85.7 (71.4)
	REV3L_r3	5′ AGARTAMAARCTRCAAAATCCMG 3′		
SLC30A1	SLC30A1_f1	5′ AAYATGCGWGGAGTKTTTCTGC 3′	543	100 (71.4)
	SLC30A1_r2	5′ AAAGATGATTCRGRYTGYAYGTTT 3′		
SNCAIP	SNCAIP_f10	5′ CGCCAGYTGYTGGGRAARGAWAT 3′	481	71.4 (71.4)
	SNCAIP_r13	5′ GGWGAYTTGAGDGCACTCTTRGGRCT 3′		
TRAF6	TRAF6_f1	5′ ATGCAGAGGAATGARYTGGCACG 3′	639	100 (85.7)
	TRAF6_r2	5′ AGGTGGCTGTCRTAYTCYCCTTGC 3′		

(*continued on next page*)

Table 2 (*continued*)

Gene[c]	Primers	Sequences	Gallus fragment length (bp)	Percentage of seven squamate clades recovered (supported)[b]
UBN1	UBN1_f1	5′ CCYCTMAATTTYCTGGCWGARCAGGC 3′	707	100 (85.7)
	UBN1_r2	5′ GGTCAGYAAYTTKGCCACHCCYT 3′		
ZEB2	ZFHX1B_f1	5′ TAYGARTGYCCAAACTGCAAGAAACG 3′	882	100 (85.7)
	ZFHX1B_r2	5′ AGTACAGACATGTGGTCCTTGTATGGGT 3′		
ZFP36L1	ZFP36L1_f1	5′ GCTGTGCCGYCCCTTYGARGARAACG 3′	605	71.4 (71.4)
	ZFP36L1_r2	5′ TCKGAGATGGARAGTCTGCTGAA 3′		

[a] Primers shown were used to amplify at least some squamate taxa. In some cases, these same primers amplified all ingroup and outgroup taxa, and in other cases internal primers were designed to amplify the remaining taxa. See downloadable amniote alignments at http://www.fieldmuseum.org/deepscaly/data.html for exact locations and sequences of all primers designed for this study.

[b] Test taxa alignments for individual genes and combined data were analyzed in an ML framework under the GTR model or its variants. Seven nodes within squamates received 100% bootstrap support using the combined data (see Fig. 4 and text). The percentage of these nodes recovered and those receiving ⩾70% bootstrap support (parentheses) with each individual gene is given.

[c] All genes identified by the official symbol of their respective human homologs in NCBI's Entrez Gene. ADNP, activity-dependent neuroprotector; AHR, aryl hydrocarbon receptor; AKAP9, A kinase (PRKA) anchor protein (yotiao) 9; BACH1, basic leucine zipper transcription factor 1; BACH2, basic leucine zipper transcription factor 2; BDNF, brain-derived neurotrophic factor; BMP2, bone morphogenetic protein 2; DNAH3, dynein, axonemal, heavy chain 3; ECEL1, endothelin converting enzyme-like 1; FSHR, follicle stimulating hormone receptor; FSTL5, follistatin-like 5; GPR37, G protein-coupled receptor 37 (endothelin receptor type B-like); MKL1, megakaryoblastic leukemia (translocation) 1; NGFB, nerve growth factor, beta polypeptide; NT3, 3′-nucleotidase; PNN, pinin, desmosome associated protein; PRLR, prolactin receptor; PTGER4, prostaglandin E receptor 4 (subtype EP4); PTPN12, protein tyrosine phosphatase, non-receptor type 12; REV3L, REV3-like, catalytic subunit of DNA polymerase zeta (yeast); SLC30A1, solute carrier family 30 (zinc transporter), member 1; SNCAIP, synuclein, alpha interacting protein (synphilin); TRAF6, TNF receptor-associated factor 6; UBN1, ubinuclein 1; ZEB2, zinc finger E-box binding homeobox 2; ZFP36L1, zinc finger protein 36, C3H type-like 1.

3.2. Phylogenetic informativeness and range of evolutionary rates

As Fig. 2 illustrates, NPCL with a broad range of evolutionary rates were developed using our protocol. Average ML-corrected divergences among the squamate test taxa ranged from 9.6% to 61.6%. Two of these loci (*PRLR* and *UBN1*) are considerably more variable than the others, but even without these loci, the range of distances is nearly 4-fold.

Seven nodes received 100% ML bootstrap support in the analysis of the concatenated (26-gene) data (Fig. 4). These nodes represent all six well-established clades discussed above (nodes 2–7 in Fig. 4), as well as a seventh clade represented by all ingroup test taxa except geckos (node 1 in Fig. 4). This node is not supported by morphological data (Estes et al., 1988), but is consistent with a node strongly supported by recent molecular studies that sampled all major squamate lineages (Townsend et al., 2004; Vidal and Hedges, 2005).

All 27 of the genes from Fig. 2 appear to contain considerable phylogenetic information for higher-level squamate relationships (Fig. 4, Table 2, and unpublished data). Because an effort was made to target particularly variable regions, many of the loci developed appear to have evolutionary rates substantially higher than other loci commonly used for squamate phylogenetics. This is important, because one of the main potential drawbacks of NPCL is their greatly reduced variability relative to mitochondrial genes (Hillis et al., 1996). The slower rate of NPCL evolution is certainly advantageous at moderate to deeper levels where saturation of mitochondrial genes is problematic (e.g., Birks and Edwards, 2002; Blouin et al., 1998; Roelants and Bossuyt, 2005; Townsend et al., 2004). However, it can sometimes limit the usefulness of NPCL for resolving species- or intraspecific-level relationships (e.g., Jesus et al., 2002; Leache and McGuire, 2006).

Recombination-activating gene 1 (*RAG1*) is a long (~3 kb), single-copy NPCL that has been successfully used in all major vertebrate groups (e.g., Brinkmann et al., 2004; Groth and Barrowclough, 1999; Hugall et al., 2007; San Mauro et al., 2005; Townsend et al., 2004; Waddell and Shelley, 2003), and its evolutionary rate is comparable to most other NPCL used in published studies of squamate relationships (Table 3). Among the 26 new genes from this study compared in Fig. 2, only eight show divergence levels lower than those of *RAG1*, and the average intra-squamate divergence of the "fastest" locus is almost three times that of *RAG1* (Table 3).

Admittedly, none of our loci approach the evolutionary rate of the mitochondrial protein-coding genes (for comparison, using data downloaded from GenBank, average intra-squamate uncorrected divergence for the mitochondrial *ND2* gene was about 1.5 times the uncorrected divergence of the "fastest" gene in this study). Nevertheless, we have developed several gene regions with relatively rapid rates, and these should prove more useful for resolving

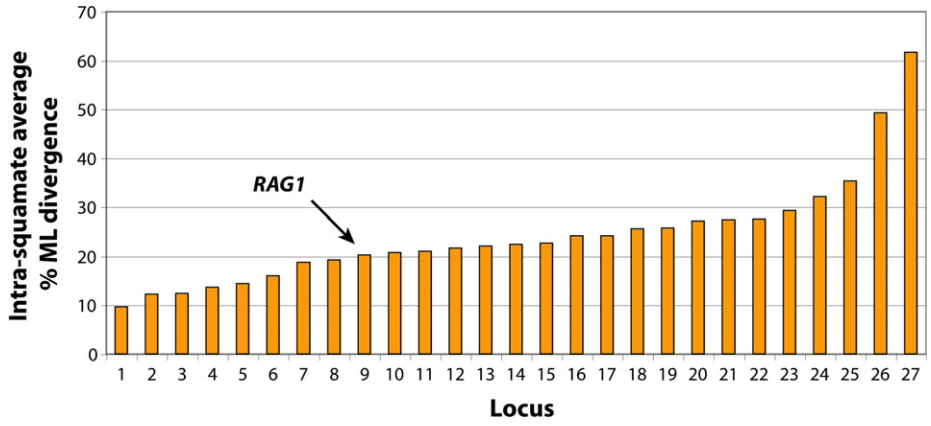

Fig. 2. Variability of 27 NPCL in squamate reptiles. Twenty-six loci from Table 2 in order of increasing variability, plus the commonly used locus *RAG1* for comparison. 1 = *ZEB2*, 2 = *BDNF*, 3 = *FSTL5*, 4 = *ZFP36L1*, 5 = *ADNP*, 6 = *BACH2*, 7 = *PNN*, 8 = *NGFB*, 9 = *RAG1*, 10 = *FSHR*, 11 = *SLC301A*, 12 = *SNCAIP*, 13 = *TRAF6*, 14 = *BMP2*, 15 = *GPR37*, 16 = *ECEL1*, 17 = *PTGER4*, 18 = *AHR*, 19 = *MKL1*, 20 = *DNAH3*, 21 = *AKAP9*, 22 = *REV3L*, 23 = *NT3*, 24 = *BACH1*, 25 = *PTPN12*, 26 = *UBN1*, 27 = *PRLR*.

recent divergences than many currently used NPCL. Furthermore, the limited length variation in these NPCL, coupled with codon constraints, should make these genes easier to consistently amplify and align across an array of taxa than nuclear introns.

3.3. Predictors of variation within squamates

Parsing *Homo–Fugu* BLAST matches by aa divergence levels (Fig. 1, Phase 1) is one way of sorting genes into groupings potentially predictive of their level of variation in squamates. However, as Table 1 shows, most of the NPCL examined were from the file containing genes with *Homo–Fugu* aa-similarities of 40–60%. There are two reasons for this. First, nuclear genes with relatively rapid rates of evolution were specifically targeted for this project, and therefore genes in this file were the first to be examined. Second, and more importantly, it became apparent that the original divisions based on *Homo–Fugu* aa similarity were not especially useful; a great number of the loci found in the 40–60% file were also present in the 61–80% and 81–90% files. This apparently occurred because the *Fugu* db consists of fragmentary protein sequences, often multiple fragments per gene, and each of these fragments was BLASTed against the *Homo* db of complete proteins (Fig. 1, Phase 1). Rate heterogeneity along the length of these proteins led to multiple *Homo–Fugu* matches at different similarity levels.

Several other parameters were also examined as potential predictors of levels of variation within squamates. Whole-gene *Homo–Gallus*% aa divergence is perhaps the most easily acquired of these parameters (its converse, aa similarity, is given with each alignment of these two taxa, Fig. 1, Phase 2.1). However, because only a portion of each gene was sequenced, and rate heterogeneity along the length of the genes was obvious from the alignments, global divergence seemed likely to be a poor predictor. *Homo–Gallus*% aa and nt divergences for only the targeted fragment required downloading, aligning, and truncating these sequences, but was also relatively easily accom-

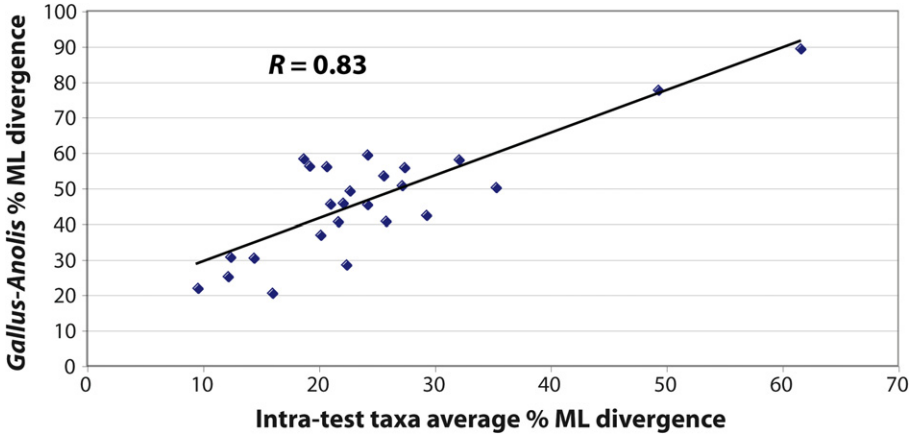

Fig. 3. Evaluation of *Gallus–Anolis* nt-level genetic distances as a predictor of variation within squamates. Intra-squamate nt-level distance values were calculated using the 10 test taxa described in the text.

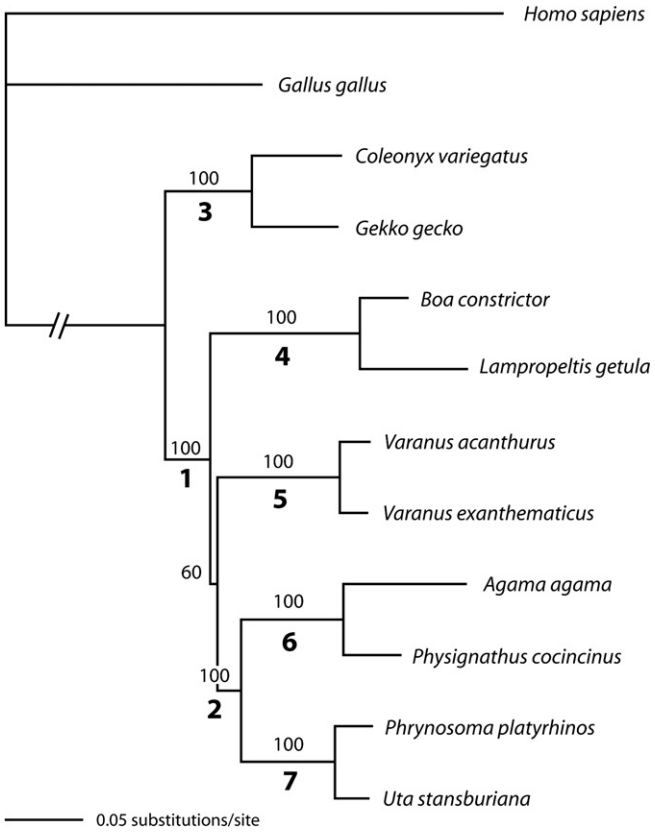

Fig. 4. Maximum-likelihood phylogram from GARLI analysis (GTR+I+G) of the concatenated 26-gene data set (20,474 bp total, 6456 bp parsimony-informative). Bootstrap values are given above each branch. Highly supported clades referenced in Table 2 are numbered: 1 = snakes, varanids, and iguanians, 2 = iguanians, 3 = geckos, 4 = snakes, 5 = varanids, 6 = acrodont, agamid iguanians, 7 = pleurodont, phrynosomatid iguanians.

plished. All three of these measures (whole-gene *Homo–Gallus* aa divergence, targeted fragment aa-level divergence, and targeted fragment nt-level divergence) were compared to average genetic distances among the test taxa using Excel (Microsoft, Inc.). Each measure was positively correlated with average intra-squamate divergences across genes, but the correlation coefficients were not high ($R = 0.61$, 0.67, and 0.53, respectively).

One possible reason for this weak to moderate correlation is the stochasticity inherent in comparing a single mammal (*Homo*) to a single bird (*Gallus*). For any single species, the rate of molecular evolution for a given gene might be anomalously high or low compared to the average rate of a wider sampling of related species (however, *Gallus* is the only bird for which genomic data were available). Another possible reason is that, for some genes, there could be divergent selection at the level of mammals versus reptiles (including birds), but stabilizing selection within each of these respective clades. Therefore, nt-level *Mus–Rattus*% ML distances were also compared to intra-squamate distances. Once again, the correlation was positive but not high ($R = 0.54$). Thus, no strong predictor was found from the resources available when our study was begun.

The recent completion of a first draft of the *Anolis carolinensis* allowed us to do a final analysis in which *Gallus–Anolis* and intra-squamate ML distances were compared. In this case, the correlation coefficient R rose substantially relative to previous analyses ($R = 0.83$; Fig. 3). This measure thus appears to be a potentially useful predictor for researchers interested in developing NPCL for various levels within squamate reptiles. Fossil evidence suggests the bird and squamate reptile lineages diverged approximately 260–300 million years ago, and the bird and mammal lineages diverged approximately 312–330 million years ago (Benton and Donoghue, 2007). The ranges of these estimates are not far from overlapping, and differential selection pressures, as well as generalized lineage-specific differences in evolutionary rate, likely contribute to the poorer correlation of *Homo–Gallus* divergences to intra-squamate divergences. However, the observation that genetic distance between clades separated over 250 million years ago correlate reasonably well with rates within one of the clades may be useful to researchers working on other groups.

Table 3
Levels of variation for several genes previously used in phylogenetic studies of squamate reptiles[a]

Gene	% ML distance[b]	# base pairs
HOX	14.5	444
MAFB	16.3	324
JUN	20.9	330
RAG2	21.6	723
RAG1	**21.7**	**2862**
CMOS	23.8	360
α-Enolase	27.3	81
R35	29.1	732
AMEL	35.4	336

[a] Data for these calculations from Vidal and Hedges, 2005.
[b] Average ML-corrected distances among 18 ingroup taxa representing major squamate lineages, in order of increasing variability. *RAG1* (bold) is the only locus also used in our study.

3.4. Comparison to other recent work

Our approach is similar in many ways to that of Li et al. (2007) (compare their Fig. 2 with our Fig. 1). Both methods begin with the identification of putative homologs between two reference species, and then proceed to the identification of continuous open reading frames within these genes. Both methods employ steps to exclude genes with paralogs that could confound phylogenetic analyses. Finally, both methods can be modified to search for loci evolving at different evolutionary rates.

However, there are also some important differences. Li et al. simply aligned two reference species, designed nested primer sets from conserved regions, and proceeded with amplifications. They reported a 67% (10/15) success rate (single bands of appropriate size) on randomly chosen loci using this method. For our method, we compare the interspecific amino acid variation within the specific fragment(s)

targeted for amplification (Fig. 1, Phase 2), not just of the entire gene or even individual open reading frames (which might be quite large). Our approach also involves alignments of a diverse array of species (all available amniotes in our case) (Fig. 1, Phase 3). The first of these steps allowed us to specifically target highly variable regions, and the second step was a great help in designing "universal" (often highly degenerate) primers for more variable gene regions.

The method of Li et al. (2007) certainly allows flexibility in the search criteria, but similarity comparisons are made between whole exons, which may not be indicative of evolutionary rates in the parts of the exon that will actually be sequenced and used in phylogenetic analyses (this is why we compare only the sequences for the targeted regions). Furthermore, designing primers based on only two taxa may work well for slowly evolving genes, but may be problematic for more rapidly evolving loci (this is why we design primers using a phylogenetically diverse group of organisms).

In general, the loci developed by Li et al. (2007) appear to be more slowly evolving than those developed using our method. For comparison, of the 10 genes (out of 15 attempted) reported as successes by Li et al. (their Table 1), we found homologs in humans and chickens for seven, and the average *Homo–Gallus* aa similarity was 93%. Our own success rate was approximately 49% (42/85, see Table 1) across all loci for which we designed primers, and for the 26 loci presented in this paper (Table 2), the average *Homo–Gallus* aa similarity was only 72%. It seems likely that our lower success rate was a function of the higher variability in the genes we chose to develop. However, our final criterion for success was based on how consistently the genes could be amplified and sequenced across a variety of squamate taxa. It may be that some of the loci considered successes by Li et al. (2007), based only on amplifying a single band of the correct size, would not meet our more strict final criteria.

Li et al. (2007) did not report any obvious paralogy problems for the genes they sequenced, and we likewise sequenced no obvious gene copies. We did occasionally get multiple bands (which were not sequenced) for some genes we tested. These multiple bands may in fact represent paralogs, but they also may have resulted simply from the relative non-specificity of our primers. Each of our primers had on average twice as many degenerate bases as those of Li et al. (see Table 2 of each paper), which once again was a function of the more variable regions we targeted.

Another potential strength of our method is the incorporation of the NCBI Map Viewer function (Fig. 1, Phase 2.2). For researchers wishing to develop intron regions, this step allows easy visualization of the length and relative position of all introns. Li et al. (2007) did not intend their method to be used for intron development, and it was not the main objective of our study either. However, we recognize that our method holds great potential for this purpose, and we have recently begun to develop intron regions using it.

Finally, we note that the Li et al. (2007) method for the original sorting of genes based on relative variability is more elegant than our own, and appears to avoid the problem of redundancy among files that we experienced sorting by *Fugu* (EST) versus *Homo* (whole protein) db comparisons (see above and Fig. 1). Perhaps a combination of the two methods (i.e., theirs for initial sorting paired with ours for more detailed comparisons and primer development) would be ideal.

3.5. Amniote alignments

In the course of this study, amniote alignments have been produced for 85 NPCL containing exons of suitable length and variability for phylogenetic studies at various levels. We have made these alignments publicly available on our website (http://www.fieldmuseum.org/deepscaly/data.html). Most of these alignments include sequences from human (*Homo*), rat (*Rattus*), mouse (*Mus*), cow (*Bos*), dog (*Canis*), and chicken (*Gallus*), and some include a marsupial, crocodilian, turtle, or squamate reptile. Each alignment is annotated with the inferred start and stop positions of its exon(s) of interest, and the positions and sequences of all primers designed for this project are also indicated. Many of these exact primers should be useful to researchers studying phylogenetic relationships within or among various amniote clades. At the least, researchers will have a convenient collection of pre-identified primer locations flanking variable-coding regions for a large number of variable NPCL. Sequences from other available taxa can easily be added to the alignments and primers can be modified to best match the clade of interest.

3.6. Applications to other vertebrate clades

Our results also show the potentially broad utility of the loci developed, in terms of applicability across major clades and to different phylogenetic scales. Our primers have amplified taxa from all major amniote lineages, with only a few exceptions. Furthermore, these genes are also proving useful within other vertebrate clades, including non-amniotes. For example, researchers in Wiens' lab used PTPN, PTGER4, and TNS3 to help resolve phylogenetic relationships among closely-related species of hylid frogs (Smith et al., 2007; note that TNS3 is included in our online alignments, but is not currently being used for squamates). Thus, despite the fact that these genes were developed for resolving higher-level squamate phylogeny, we find that they are also informative among species within genera in a distantly-related clade.

3.7. Ongoing and future work

Thus far, we have only used our method to develop loci encompassing protein-coding regions. However, it is also

ideally suited to the development of intron regions for species-level or intraspecific (e.g., phylogeographic) studies. This is a particularly exciting prospect because there has historically been a paucity of nuclear markers available with sufficient variation for intraspecific studies (but see Dolman and Phillips, 2004; Lyons et al., 1997). As mentioned above in the Methods, once individual exons have been identified by BLASTing a protein sequence against the *Gallus* genomic sequence, NCBI's MapViewer function makes intron identification and size estimation relatively straightforward. Positions of inferred exons relative to genomic sequence are shown graphically, and the user can simply look for adjacent exons (each with conserved sequence for primer design) separated by an appropriately sized intron.

The *Anolis* genome project (http://www.broad.mit.edu/models/anole/) will soon provide a complete, searchable squamate genome, and this will be very valuable for future locus development for squamates and other non-avian reptile groups (i.e., crocodilians, turtles). We have recently downloaded the first assembly released from this project and have been able to incorporate the *Anolis* data into our gene discovery procedure. The addition of the *Anolis* sequence to existing alignments has already helped us successfully redesign primers for multiple genes that did not amplify in squamates using our original primers. However, it should be noted that all of the genes from this paper were developed before the *Anolis* genome became available, demonstrating that it is not necessary to have a completed genome available within the ingroup for this approach to be successful.

3.8. Conclusions

The general method described here is one that can easily be extended to other taxa, including other animals, plants, and fungi. As one example, the first coleopteran genomic draft assembly was recently completed for the red flour beetle *Tribolium castaneum* (NCBI Entrez Genome Project ID 12539). Genomic sequencing is also complete or nearly complete for species from several-related insect orders (Diptera, Lepidoptera, Hymenoptera, as well as other more distantly-related orders). However, molecular phylogenetic studies of beetles have relied almost exclusively on mitochondrial DNA or nuclear ribosomal DNA; only a few very recent coleopteran studies have included one or two NPCL (Sasakawa and Kubota, 2007; Sota and Ishikawa, 2004; Sota et al., 2005). An objective, automated comparison of published insect genomes similar to the one described here would likely identify many other NPCL suitable for phylogenetic analyses within this very large and economically important clade.

This is an exciting time for molecular systematic studies. Practical phylogenomic approaches have become a reality, and the number of accessible independent data sources is set to rise dramatically across all taxonomic groups in the near future. This influx of new data, combined with theoretical and algorithmic advances, should bring us substantially closer toward the goal of a fully resolved Tree of Life.

Acknowledgments

Sarah Smith, Saad Arif, Caitlin Kuzcynski, Carolina Ulloa, Brice Noonan, Dan Mulcahy, Dean Leavitt, Andrew Schlossman, and Alelí Camacho tested many of the primers in the lab, and our estimates of the success of these genes are based largely on their work. This work was funded through a Tree of Life grant from the National Science Foundation (EF 0334923 to JJW; EF 0334967 to TWR). Our initial ideas on the design of this gene search strategy were inspired by correspondence with Cliff Cunningham.

References

Arnason, U., Adegoke, J.A., Bodin, K., Born, E.W., Esa, Y.B., Gullberg, A., Nilsson, M., Short, R.V., Xu, X.F., Janke, A., 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. Proc. Natl. Acad. Sci. USA 99, 8151–8156.

Ast, J.C., 2001. Mitochondrial DNA evidence and evolution in Varanoidea (Squamata). Cladistics 17, 211–226.

Avise, J.C., 1986. Mitochondrial-DNA and the evolutionary genetics of higher animals. Phil. Trans. R. Soc. Lond. B 312, 325–342.

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

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.

Benton, M.J., Donoghue, P.C.J., 2007. Paleontological evidence to date the tree of life. Mol. Biol. Evol. 24, 26–53.

Birks, S.M., Edwards, S.V., 2002. A phylogeny of the megapodes (Aves: Megapodiidae) based on nuclear and mitochondrial DNA sequences. Mol. Phylogenet. Evol. 23, 408–421.

Blouin, M.S., Yowell, C.A., Courtney, C.H., Dame, J.B., 1998. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol. Biol. Evol. 15, 1719–1727.

Boekhorst, J., Snel, B., 2007. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. BMC Bioinform. 8, 7.

Brinkmann, H., Venkatesh, B., Brenner, S., Meyer, A., 2004. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. Proc. Natl. Acad. Sci. USA 101, 4900–4905.

Brown, W.M., 1985. The mitochondrial genome of animals. In: MacIntyre, R.J. (Ed.), Molecular Evolutionary Genetics. Plenum, NY, pp. 95–130.

Burns, K.J., 1997. Molecular systematics of tanagers (Thraupinae): evolution and biogeography of a diverse radiation of neotropical birds. Mol. Phylogenet. Evol. 8, 334–348.

Dolman, G., Phillips, B., 2004. Single copy nuclear DNA markers characterized for comparative phylogeography in Australian wet tropics rainforest skinks. Mol. Ecol. Notes 4, 185–187.

Douglas, D.A., Janke, A., Arnason, U., 2006. A mitogenomic study on the phylogenetic position of snakes. Zool. Scr. 35, 545–558.

Downie, D.A., Gullan, P.J., 2004. Phylogenetic analysis of mealybugs (Hemiptera: Coccoidea: Pseudococcidae) based on DNA sequences from three nuclear genes, and a review of the higher classification. Syst. Entomol. 29, 238–259.

Estes, R., de Queiroz, K., Gauthier, J.A., 1988. Phylogenetic relationships within Squamata. In: Estes, R., Pregill, G. (Eds.), Phylogenetic

Relationships of the Lizard Families. Stanford University Press, Stanford, pp. 119–281.

Frost, D.R., Etheridge, R., 1989. A phylogenetic analysis and taxonomy of iguanian lizards (Reptilia: Squamata). The University of Kansas Museum of Natural History, Miscellaneous Publications, pp. 1–65.

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.

Gaines, C.A., Hare, M.P., Beck, S.E., Rosenbaum, H.C., 2005. Nuclear markers confirm taxonomic status and relationships among highly endangered and closely related right whale species. Proc. R. Soc. Lond. B 272, 533–542.

Groth, J.G., Barrowclough, G.F., 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Mol. Phylogenet. Evol. 12, 115–123.

Harrison, R.G., 1989. Animal mitochondrial-DNA as a genetic marker in population and evolutionary biology. Trends Ecol. Evol. 4, 6–11.

Heise, P.J., Maxson, L.R., Dowling, H.G., Hedges, S.B., 1995. Higher-level snake phylogeny inferred from mitochondrial-DNA sequences of 12S ribosomal-RNA and 16S ribosomal-RNA genes. Mol. Biol. Evol. 12, 259–265.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A.M., Delany, M.E., Dodgson, J.B., Chinwalla, A.T., Cliften, P.F., Clifton, S.W., Delehaunty, K.D., Fronick, C., Fulton, R.S., Graves, T.A., Kremitzki, C., Layman, D., Magrini, V., McPherson, J.D., Miner, T.L., Minx, P., Nash, W.E., Nhan, M.N., Nelson, J.O., Oddy, L.G., Pohl, C.S., Randall-Maher, J., Smith, S.M., Wallis, J.W., Yang, S.P., Romanov, M.N., Rondelli, C.M., Paton, B., Smith, J., Morrice, D., Daniels, L., Tempest, H.G., Robertson, L., Masabanda, J.S., Griffin, D.K., Vignal, A., Fillon, V., Jacobbson, L., Kerje, S., Andersson, L., Crooijmans, R.P.M., Aerts, J., van der Poel, J.J., Ellegren, H., Caldwell, R.B., Hubbard, S.J., Grafham, D.V., Kierzek, A.M., McLaren, S.R., Overton, I.M., Arakawa, H., Beattie, K.J., Bezzubov, Y., Boardman, P.E., Bonfield, J.K., Croning, M.D.R., Davies, R.M., Francis, M.D., Humphray, S.J., Scott, C.E., Taylor, R.G., Tickle, C., Brown, W.R.A., Rogers, J., Buerstedde, J.M., Wilson, S.A., Stubbs, L., Ovcharenko, I., Gordon, L., Lucas, S., Miller, M.M., Inoko, H., Shiina, T., Kaufman, J., Salomonsen, J., Skjoedt, K., Wong, G.K.S., Wang, J., Liu, B., Wang, J., Yu, J., Yang, H.M., Nefedov, M., Koriabine, M., deJong, P.J., Goodstadt, L., Webber, C., Dickens, N.J., Letunic, I., Suyama, M., Torrents, D., von Mering, C., Zdobnov, E.M., Makova, K., Nekrutenko, A., Elnitski, L., Eswara, P., King, D.C., Yang, S., Tyekucheva, S., Radakrishnan, A., Harris, R.S., Chiaromonte, F., Taylor, J., He, J.B., Rijnkels, M., Griffiths-Jones, S., Ureta-Vidal, A., Hoffman, M.M., Severin, J., Searle, S.M.J., Law, A.S., Speed, D., Waddington, D., Cheng, Z., Tuzun, E., Eichler, E., Bao, Z.R., Flicek, P., Shteynberg, D.D., Brent, M.R., Bye, J.M., Huckle, E.J., Chatterji, S., Dewey, C., Pachter, L., Kouranov, A., Mourelatos, Z., Hatzigeorgiou, A.G., Paterson, A.H., Ivarie, R., Brandstrom, M., Axelsson, E., Backstrom, N., Berlin, S., Webster, M.T., Pourquie, O., Reymond, A., Ucla, C., Antonarakis, S.E., Long, M.Y., Emerson, J.J., Betran, E., Dupanloup, I., Kaessmann, H., Hinrichs, A.S., Bejerano, G., Furey, T.S., Harte, R.A., Raney, B., Siepel, A., Kent, W.J., Haussler, D., Eyras, E., Castelo, R., Abril, J.F., Castellano, S., Camara, F., Parra, G., Guigo, R., Bourque, G., Tesler, G., Pevzner, P.A., Smit, A., Fulton, L.A., Mardis, E.R., Wilson, R.K., 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695–716.

Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), 1996. Molecular Systematics. Sinauer Associates, Sunderland, MA.

Honeycutt, R.L., Adkins, R.M., 1993. Higher-level systematics of eutherian mammals—an assessment of molecular characters and phylogenetic hypotheses. Annu. Rev. Ecol. Syst. 24, 279–305.

Hugall, A.F., Foster, R., Lee, M.S.Y., 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Syst. Biol. 56, 543–563.

Hyman, I.T., Ho, S.Y.W., Jermiin, L.S., 2007. Molecular phylogeny of Australian Helicarionidae, Euconulidae and related groups (Gastropoda: Pulmonata: Stylommatophora) based on mitochondrial DNA. Mol. Phylogenet. Evol. 45, 792–812.

Jesus, J., Brehm, A., Harris, D.J., 2002. Relationships of *Tarentola* (Reptilia: Gekkonidae) from the Cape Verde Islands estimated from DNA sequence data. Amphibia-Reptilia 23, 47–54.

Klicka, J., Burns, K., Spellman, G.M., 2007. Defining a monophyletic Cardinalini: a molecular perspective. Mol. Phylogenet. Evol. 45, 1014–1032.

Leache, A.D., McGuire, J.A., 2006. Phylogenetic relationships of horned lizards (*Phrynosoma*) based on nuclear and mitochondrial data: evidence for a misleading mitochondrial gene tree. Mol. Phylogenet. Evol. 39, 628–644.

Lemmon, E.M., Lemmon, A.R., Collins, J.T., Lee-Yaw, J.A., Cannatella, D.C., 2007. Phylogeny-based delimitation of species boundaries and contact zones in the trilling chorus frogs (*Pseudacris*). Mol. Phylogenet. Evol. 44, 1068–1082.

Li, C., Ortí, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinipterygii) as a case study. BMC Evol. Biol. 7.

Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E., Obrien, S.J., 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nat. Genet. 15, 47–56.

Macey, J.R., Schulte, J.A., Larson, A., Ananjeva, N.B., Wang, Y.Z., Pethiyagoda, R., Rastegar-Pouyani, N., Papenfuss, T.J., 2000. Evaluating trans-tethys migration: an example using acrodont lizard phylogenetics. Syst. Biol. 49, 233–256.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.

Matthee, C.A., Burzlaff, J.D., Taylor, J.F., Davis, S.K., 2001. Mining the mammalian genome for artiodactyl systematics. Syst. Biol. 50, 367–390.

Mitchell, A., Wen, J., 2004. Phylogenetic utility and evidence for multiple copies of Granule-Bound Starch Synthase I (GBSSI) in Araliaceae. Taxon 53, 29–41.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryderk, O.A., O'Brien, S.J., 2001a. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614–618.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., Springer, M.S., 2001b. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294, 2348–2351.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Ritchie, P.A., Lavoue, S., Lecointre, G., 1997. Molecular phylogenetics and the evolution of Antarctic notothenioid fishes. Comp. Biochem. Physiol., A: Mol. Integr. Physiol. 118, 1009–1025.

Roelants, K., Bossuyt, F., 2005. Archaeobatrachian paraphyly and Pangaean diversification of crown-group frogs. Syst. Biol. 54, 111–126.

Roelants, K., Gower, D.J., Wilkinson, M., Loader, S.P., Biju, S.D., Guillaume, K., Moriau, L., Bossuyt, F., 2007. Global patterns of diversification in the history of modern amphibians. Proc. Natl. Acad. Sci. USA 104, 887–892.

San Mauro, D., Vences, M., Alcobendas, M., Zardoya, R., Meyer, A., 2005. Initial diversification of living amphibians predated the breakup of Pangaea. Am. Nat. 165, 590–599.

Sasakawa, K., Kubota, K., 2007. Phylogeny and genital evolution of carabid beetles in the genus *Pterostichus* and its allied genera (Coleoptera: Carabidae) inferred from two nuclear gene sequences. Ann. Entomol. Soc. Am. 100, 100–109.

Schulte, J.A., Valladares, J.P., Larson, A., 2003. Phylogenetic relationships within Iguanidae inferred using molecular and morphological data and a phylogenetic taxonomy of Iguanian lizards. Herpetologica 59, 399–419.

Sequeira, F., Ferrand, N., Harris, D.J., 2006. Assessing the phylogenetic signal of the nuclear beta-fibrinogen intron 7 in salamandrids (Amphibia: Salamandridae). Amphibia-Reptilia 27, 409–418.

Simon, C., Buckley, T.R., Frati, F., Stewart, J.B., Beckenbach, A.T., 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 37, 545–579.

Smith, S.A., Arif, S., de Oca, A.N.M., Wiens, J.J., 2007. A phylogenetic hot spot for evolutionary novelty in middle American treefrogs. Evolution 61, 2075–2085.

Sota, T., Ishikawa, R., 2004. Phylogeny and life-history evolution in *Carabus* (subtribe Carabina: Coleoptera, Carabidae) based on sequences of two nuclear genes. Biol. J. Linn. Soc. 81, 135–149.

Sota, T., Takami, Y., Monteith, G.B., Moore, B.P., 2005. Phylogeny and character evolution of endemic Australian carabid beetles of the genus *Pamborus* based on mitochondrial and nuclear gene sequences. Mol. Phylogenet. Evol. 36, 391–404.

Sotoadames, F.N., Robertson, H.M., Berlocher, S.H., 1994. Phylogenetic utility of partial DNA sequences of G6PDH at different taxonomic levels in Hexapoda with emphasis on Diptera. Ann. Entomol. Soc. Am. 87, 723–736.

Swofford, D.L., 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, MA.

Sword, G.A., Senior, L.B., Gaskin, J.F., Joern, A., 2007. Double trouble for grasshopper molecular systematics: intra-individual heterogeneity of both mitochondrial 12S-valine-16S and nuclear internal transcribed spacer ribosomal DNA sequences in *Hesperotettix viridis* (Orthoptera: Acrididae). Syst. Entomol. 32, 420–428.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura, R.M. (Ed.), Some Mathematical Questions in Biology—DNA Sequence Analysis. American Mathematical Society, Providence, pp. 57–86.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Townsend, T.M., Larson, A., Louis, E., Macey, J.R., 2004. Molecular phylogenetics of Squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. Syst. Biol. 53, 735–757.

Vidal, N., Hedges, S.B., 2004. Molecular evidence for a terrestrial origin of snakes. Proc. R. Soc. Lond. B 271, S226–S229.

Vidal, N., Hedges, S.B., 2005. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. C. R. Biol. 328, 1000–1008.

Waddell, P.J., Shelley, S., 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol. Phylogenet. Evol. 28, 197–224.

Weibel, A.C., Moore, W.S., 2002. A test of a mitochondrial gene-based phylogeny of woodpeckers (genus *Picoides*) using an independent nuclear gene, beta-fibrinogen intron 7. Mol. Phylogenet. Evol. 22, 247–257.

Weisrock, D.W., Harmon, L.J., Larson, A., 2005. Resolving deep phylogenetic relationships in salamanders: analyses of mitochondrial and nuclear genomic data. Syst. Biol. 54, 758–777.

Wiens, J.J., Bonett, R.M., Chippindale, P.T., 2005. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander relationships. Syst. Biol. 54, 91–110.

Wiens, J.J., Hollingsworth, B.D., 2000. War of the iguanas: conflicting molecular and morphological phylogenies and long-branch attraction in iguanid lizards. Syst. Biol. 49, 143–159.

Willows-Munro, S., Robinson, T.J., Matthee, C.A., 2005. Utility of nuclear DNA intron markers at lower taxonomic levels: phylogenetic resolution among nine *Tragelaphus* spp.. Mol. Phylogenet. Evol. 35, 624–636.

Zhou, K.Y., Li, H.D., Han, D.M., Bauer, A.M., Feng, J.Y., 2006. The complete mitochondrial genome of *Gekko gecko* (Reptilia: Gekkonidae) and support for the monophyly of Sauria including Amphisbaenia. Mol. Phylogenet. Evol. 40, 887–892.

Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin. Available from: (<http://www.zo.utexas.edu/faculty/antisense/Garli.html>).