Syst. Biol. 58(4):425–438, 2009
Copyright © Society of Systematic Biologists
DOI:10.1093/sysbio/syp043
Advance Access publication on August 18, 2009

Experimental Design in Caecilian Systematics: Phylogenetic Information of Mitochondrial Genomes and Nuclear *rag1*

DIEGO SAN MAURO[1]*, DAVID J. GOWER[1], TIM MASSINGHAM[2], MARK WILKINSON [1], RAFAEL ZARDOYA[3], AND JAMES A. COTTON[1,4]

[1]*Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK;*
[2]*The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK;*
[3]*Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales - CSIC, José Gutiérrez Abascal 2, 28006 Madrid, Spain;*
[4]*School of Biological and Chemical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK;*
**Correspondence to be sent to: Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK;*
E-mail: d.san-mauro@nhm.ac.uk.

Abstract.—In molecular phylogenetic studies, a major aspect of experimental design concerns the choice of markers and taxa. Although previous studies have investigated the phylogenetic performance of different genes and the effectiveness of increasing taxon sampling, their conclusions are partly contradictory, probably because they are highly context specific and dependent on the group of organisms used in each study. Goldman introduced a method for experimental design in phylogenetics based on the expected information to be gained that has barely been used in practice. Here we use this method to explore the phylogenetic utility of mitochondrial (mt) genes, mt genomes, and nuclear *rag1* for studies of the systematics of caecilian amphibians, as well as the effect of taxon addition on the stabilization of a controversial branch of the tree. Overall phylogenetic information estimates per gene, specific estimates per branch of the tree, estimates for combined (mitogenomic) data sets, and estimates as a hypothetical new taxon is added to different parts of the caecilian tree are calculated and compared. In general, the most informative data sets are those for mt transfer and ribosomal RNA genes. Our results also show at which positions in the caecilian tree the addition of taxa have the greatest potential to increase phylogenetic information with respect to the controversial relationships of *Scolecomorphus*, *Boulengerula*, and all other teresomatan caecilians. These positions are, as intuitively expected, mostly (but not all) adjacent to the controversial branch. Generating whole mitogenomic and *rag1* data for additional taxa joining the *Scolecomorphus* branch may be a more efficient strategy than sequencing a similar amount of additional nucleotides spread across the current caecilian taxon sampling. The methodology employed in this study allows an a priori evaluation and testable predictions of the appropriateness of particular experimental designs to solve specific questions at different levels of the caecilian phylogeny. [Experimental design; Gymnophiona; mitochondrial genes; mitochondrial genome; phylogenetic information; rate of evolution; *rag1*; taxon sampling.]

Taxon and character sampling is fundamental in phylogenetics, but this aspect of experimental design is considered complex (e.g., Graybeal 1998; Cummings and Meyer 2005; Rokas and Carroll 2005). Given limited time and resources, it is important to sample taxa and characters efficiently so as to maximize phylogenetic accuracy, precision, and robustness. This issue has most often been dealt with by comparing the benefits of adding more taxa versus more characters, with contrasting conclusions (e.g., Kim 1996, 1998; Graybeal 1998; Hillis 1998; Rannala et al. 1998; Poe and Swofford 1999; Pollock and Bruno 2000; Rosenberg and Kumar 2001; Pollock et al. 2002; Zwickl and Hillis 2002; Rokas and Carroll 2005). Nevertheless, there is a general consensus on the importance of completeness of data sets (Cummings and Meyer 2005), and the need for judicious sampling of taxa and characters (Soltis et al. 2004; Hedtke et al. 2006).

In molecular phylogenetics, the favoring of particular genes or genomic regions has reflected the availability of primers, perceived general utility, and the historical legacy of data and alignments that can be expanded, rather than any special demonstration of their appropriateness for a particular phylogenetic question (Cummings and Meyer 2005). Several empirical studies have investigated the efficacy of some markers in reconstructing phylogeny under various inference frameworks. In particular, studies have compared the performance of different mitochondrial (mt) genes using the mitogenomic (Curole and Kocher 1999) tree as a reference (Cummings et al. 1995; Russo et al. 1996; Zardoya and Meyer 1996; Miya and Nishida 2000; Hardman and Hardman 2006; Mueller 2006) and/or compared, either directly or indirectly, the utility of nuclear and mt genes (Graybeal 1994; Groth and Barrowclough 1999; Springer et al. 2001; San Mauro et al. 2004; Townsend et al. 2008), or used simulations to explore how rates of molecular evolution influence phylogenetic reconstruction (Yang 1998). Results have supported some general conclusions, such as the relatively good performance of mt ribosomal genes and poor performance of *nad4L*, but have not provided universal guidance other than to sample several different genes (e.g., Cummings et al. 1995; Russo et al. 1996; Zardoya and Meyer 1996; Miya and Nishida 2000; Mueller 2006). Unsurprisingly perhaps, previous studies have demonstrated that best practice in character sampling is context specific (Russo et al. 1996) and contingent upon taxon sampling, method of analysis, and measures of performance.

Goldman (1998; Massingham and Goldman 2000) proposed a general method for constructing efficient sampling designs, on a case-by-case basis, by using a likelihood framework. This approach has almost never been applied to real phylogenetic problems (Goldman 1998; Geuten et al. 2007), so that its considerable potential for molecular phylogenetics remains largely unexplored. Goldman's approach is based on the estimation of Fisher, or expected, information for the likelihood function. Other concepts of "phylogenetic information" have been introduced elsewhere (Ronquist 1996; Thorley et al. 1998; Wilkinson et al. 2004; Gauthier and Lapointe 2007; Townsend 2007; Wägele and Mayer 2007; Cotton and Wilkinson 2008), but here we use the term information exclusively to mean the "Fisher information" of Goldman (1998).

Fisher information is easiest to understand in the context of a model with a single parameter, where it is the second derivative (the rate of change of the slope) of the likelihood function with respect to the parameter in question. Evaluated at the maximum-likelihood (ML) value of the parameter, this is known as the observed information, and the negative inverse of the observed information is, asymptotically, the variance of the parameter estimate, and so is used in constructing "support intervals" (approximate confidence intervals) for the ML parameter estimate. The expected value of the information, where the given parameter estimate is assumed to be its true value, is called the expected information. Both the observed information and expected information (evaluated at the ML estimate) are measures of the variance of ML estimates (Efron and Hinkley 1978). Here, we wish to compare the information conveyed by a set of genes about a particular phylogeny with a view to predicting which loci will be most appropriate for solving similar phylogenetic problems, so we follow Goldman (1998) in employing only the expected information. For more complex models that have more than 1 parameter, the Fisher information is a matrix of partial derivatives, containing information about both the variance and covariances of the likelihood function for each model parameter. Goldman (1998) proposes using the determinant of this matrix as a measure of the information an experiment can provide about all the parameters, which we shall refer to as the total phylogenetic information.

As a tool for experimental design in molecular systematics, the expected information measure has strengths and potential weaknesses. Data only influence the information matrix through the estimates of the tree and model parameters, so the method can easily be used to investigate the effects of differences in the substitution process, such as variation in base composition or in rates across sites, as well as being able to quantify the effect of different levels of divergence and of adding additional taxa (Goldman 1998). One potential drawback in phylogenetics is that the tree topology is part of the structure of the model (Yang et al. 1995), rather than a parameter of the model, so the information matrix does not directly estimate the uncertainty in the tree itself, which is the usual aim of molecular systematics. This also means we need to assume a particular tree topology in making information calculations. A final important characteristic of the expected information is that information is calculated per site, so, as long as it is estimated on the same underlying tree, information matrices can be summed across sites and even across partitions, allowing comparisons between different sets of loci even where different loci are evolving under different models.

We recently determined the complete mt genome and partial nuclear *rag1* sequences of several caecilian amphibians (Gymnophiona), and used them to infer phylogenetic relationships of families within the group (San Mauro et al. 2004, 2006). We demonstrated the potential of these molecular markers, leading us to suggest (San Mauro et al. 2004) that "expanded taxon sampling" was the way forward for additional insights. Here we apply Goldman's methods to critically evaluate our specific recommendation and to illustrate how his approach can be used to develop sampling strategies more generally.

Materials and Methods

Taxon Sampling and DNA Sequencing

This study includes 9 species of caecilian amphibians, representing all 6 families recognized by Wilkinson and Nussbaum (2006). San Mauro et al. (2004) indicated that Caeciliidae (the largest, most diverse, and cosmopolitan caecilian family; see Taylor 1968; Nussbaum and Wilkinson 1989; Wilkinson and Nussbaum 2006) was particularly inadequately represented by a single species, particularly given its paraphyly with respect to the Typhlonectidae (Nussbaum 1979; Hedges et al. 1993; Wilkinson 1997; Wilkinson et al. 2003; Frost et al. 2006; Roelants et al. 2007), and perhaps also the Scolecomorphidae (Wilkinson et al. 2003; Frost et al. 2006). Thus, we also include here 3 species that are considered to represent different major caeciliid lineages (Taylor 1968; Wilkinson and Nussbaum 2006): the East African *Boulengerula taitanus*, the West African *Geotrypetes seraphini*, and the South American *Siphonops annulatus*. The nucleotide sequence of the complete mt genome of *S. annulatus* was determined by San Mauro et al. (2006), and those of *B. taitanus* and *G. seraphini* were newly determined for this study. A 1509 base pair (bp) long fragment of the nuclear *rag1* was also determined for each of these 3 species.

In all cases, total DNA was purified from ethanol-preserved liver with standard phenol/chloroform extraction (Sambrook et al. 1989), and nucleotide sequences were determined using the primers, conditions, and methods reported by San Mauro et al. (2004). Details of the species, voucher specimens, and GenBank accession numbers are given in Table 1. Distinct structural features of the mt genomes of *B. taitanus* and *G. seraphini* are presented in the Supplementary material, Appendix 1 (available at http://sysbio.oxfordjournals.org/).

Sequence Alignments, Phylogeny Reconstruction, and Support

Various data partitions were prescribed (Table 2) and alignments were prepared for each. Nucleotide

TABLE 1. Caecilian samples employed in this study

Species	Taxonomic assignment[a]	Voucher number	Collection locality	GENBANK accession nos. (mt genomes, *rag1*)
Rhinatrema bivittatum	Gymnophiona: Rhinatrematidae	BMNH 2002.6	Kaw, French Guiana	AY456252, AY456257
Ichthyophis glutinosus	Gymnophiona: Ichthyophiidae	MW 1733	Peradeniya, Sri Lanka	AY456251, AY456256
Uraeotyphlus cf. *oxyurus*	Gymnophiona: Uraeotyphlidae	MW 212	Payyanur, India	AY456254, AY456259
Scolecomorphus vittatus	Gymnophiona: Scolecomorphidae	BMNH 2002.100	Amani, Tanzania	AY456253, AY456258
Typhlonectes natans	Gymnophiona: Typhlonectidae	BMNH 2000.218[b]	Potrerito, Venezuela[b]	AF154051, AY456260
Gegeneophis ramaswamii	Gymnophiona: Caeciliidae	MW 331	Thenmalai, India	AY456250, AY456255
Boulengerula taitanus	Gymnophiona: Caeciliidae	NMK A/3112	Wundanyi, Kenya	AY954504[c], DQ320062[c]
Geotrypetes seraphini	Gymnophiona: Caeciliidae	BMNH 2005.2	Cameroon (no locality – pet trade)	AY954505[c], DQ320063[c]
Siphonops annulatus	Gymnophiona: Caeciliidae	BMNH 2005.9	Dominguez Martins, Brazil	AY954506, DQ320064[c]

BMNH, The Natural History Museum, London (UK); MW, field series of the Zoology Department, University of Kerala (India) and the Department of National Museums, Colombo (Sri Lanka); NMK, National Museums of Kenya, Nairobi (Kenya).

[a] Taxonomy of Wilkinson and Nussbaum (2006).

[b] Only for the specimen used to sequence *rag1*. Collection data for the voucher used to sequence the mt genome are unknown (pet trade).

[c] Determined for this study.

TABLE 2. Names and included genes of each data partition employed in this study

Name	Genes included
AT6	*atp6* without third-codon positions
AT8	*atp8* without third-codon positions
CO1	*cox1* without third-codon positions
CO2	*cox2* without third-codon positions
CO3	*cox3* without third-codon positions
CYB	*cob* without third-codon positions
ND1	*nad1* without third-codon positions
ND2	*nad2* without third-codon positions
ND3	*nad3* without third-codon positions
ND4	*nad4* without third-codon positions
ND4L	*nad4L* without third-codon positions
ND5	*nad5* without third-codon positions
ND6	*nad6* without third-codon positions
PROTS-NO3	mt protein-coding genes without third-codon positions
PROTS-ALL	mt protein-coding genes - all positions
3rdPOS	third-codon positions of mt protein-coding genes
12S	mt *rrnS*
16S	mt *rrnL*
tRNAs	All mt tRNA genes except *trnF*
mtGENOME-NO3	All single mt data sets combined, excluding third-codon positions
RAG1	nuclear *rag1*

sequences of mt *rrnS* (12S) and *rrnL* (16S) genes were aligned using CLUSTAL X version 1.83 (Thompson et al. 1997) with default penalties for gap opening and gap extension, and changed by eye to correct for obvious misalignments. The CLUSTAL alignments were checked against secondary structure models using the VIENNA Webserver for RNA secondary structure prediction and comparison (Hofacker et al. 1994; Hofacker 2003). Sequences of each mt tRNA gene (except *trnF*, which is absent in *G. ramaswamii*; San Mauro et al., 2004) were aligned manually based on inferred cloverleaf secondary structures and concatenated to form a single partition. Deduced amino acid sequences of all 13 mt protein-coding genes were aligned manually against a previous database (San Mauro et al. 2004), and the alignments imposed upon the corresponding nucleotide sequences (used in all subsequent analyses). *Rag1* nucleotide sequences were aligned manually against San Mauro et al.'s (2004) database. In all cases, gaps and alignment ambiguities were excluded from partitions using GBLOCKS version 0.91b (Castresana 2000) with default parameter settings.

Caecilian phylogeny was estimated from a combined data set excluding third-codon positions of mt protein-coding genes because transitions were saturated as judged by plots (not shown) of pairwise uncorrected (transition and transversion) differences versus corrected sequence divergence (measured as ML distance). Rooted trees assume the Rhinatrematidae to be the sister group of all other caecilians based on previous molecular (Hedges et al. 1993; San Mauro et al. 2004, 2005; Frost et al. 2006; Roelants et al. 2007) and morphological (Nussbaum 1977, 1979; Wilkinson 1992, 1996, 1997; Wilkinson and Nussbaum 1996) data.

Phylogeny was estimated using ML (Felsenstein 1981) and Bayesian Inference (BI; Huelsenbeck et al. 2001). ML analysis was performed with PAUP* version 4.0b10 (Swofford 1998) and RAxML version 7.0.4 (Stamatakis 2006). PAUP* used heuristic searches with 10 random stepwise addition sequences of taxa and tree bisection and reconnection branch swapping. RAxML used the rapid hill-climbing algorithm (Stamatakis et al. 2007) computing 10 distinct ML trees starting from 10 distinct randomized maximum-parsimony starting trees. BI was performed with MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) running 4 simultaneous Markov chains for 10 million generations, sampling every 1000 generations, and discarding all samples during a 1 million generation burn-in period to reduce dependence on the initial starting point. Adequate convergence of the Bayesian Markov chain Monte Carlo runs was judged by plots of ln L scores and low standard deviation of split frequencies (as implemented in MrBayes), as well as using the convergence diagnostics implemented in the online tool AWTY (Nylander et al. 2008). Two independent BI runs were performed as an additional check that the chains mixed well and so converged.

Best fit models of nucleotide substitution were identified using the Akaike information criterion (AIC; Akaike 1973) as implemented in Modeltest version 3.7 (Posada and Crandall 1998). For ML using PAUP*, a single model of nucleotide substitution was selected: general time reversible (GTR; Rodríguez et al. 1990) with gamma-distributed among-site rate heterogeneity approximated with 4 categories (Γ_4; Yang 1994) and a proportion of invariable sites (I; Reeves 1992). For BI and RAxML, 6 alternative partitioning schemes (of 1, 2, 4, 7, 17, and 32 partitions, respectively; see Supplementary material, Appendix 2) were compared using the AIC, the Bayesian information criterion (Schwarz 1978), and standard Bayes factors (Nylander et al. 2004), as employed in recent studies (McGuire et al. 2007; Li et al. 2008). The 7-partition strategy (first codon positions of mt protein-coding genes, second codon positions of mt protein-coding genes, first codon positions of *rag1*, second codon positions of *rag1*, third-codon positions of *rag1*, mt ribosomal genes, and mt tRNA genes) was the preferred for both BI and ML frameworks (see Supplementary material, Appendix 2). For BI, the models employed for each of the 7 partitions were: GTR + Γ_4 + I (first codon positions of mt protein-coding genes), GTR + Γ_4 + I (second codon positions of mt protein-coding genes), GTR + I (first codon positions of *rag1*), GTR + I (second codon positions of *rag1*), GTR + Γ_4 (third-codon positions of *rag1*), GTR + Γ_4 + I (mt ribosomal genes), and GTR + Γ_4 (mt tRNA genes). In the case of RAxML, the GTR + Γ_4 model was employed for each of the 7 partitions. Support for internal branches was evaluated by non-parametric bootstrapping with 2000 replicates (ML) and posterior probabilities (BI). The combined data alignment used to infer the phylogenetic relationships of caecilians has been placed in TreeBASE under accession number S2403.

Evaluation of Alternative Tree Topologies

Five alternative tree topologies (see Results) were evaluated using parametric bootstrapping (PB; Efron 1985; Goldman 1993; Huelsenbeck et al. 1996) and the nonparametric approximately unbiased (AU; Shimodaira 2002) test. Each PB was conducted using Paml version 4.2 (Yang 2007) with 2000 simulated data sets under 7 independent GTR+Γ_5 models, assigned to the same partitions defined for the BI and RAxML analyses. A Holm–Bonferroni multiple-test correction (Holm 1979) was applied to maintain the experimentwise type I error rate at the nominal level of 5%. AU tests were carried out using CONSEL version 0.1i (Shimodaira and Hasegawa 2001) with sitewise log likelihoods calculated by PAML with independent GTR + Γ_5 models assigned to the same partitions used for BI and RA × ML, and 1 million multiscale bootstrap replicates.

Estimation of Phylogenetic Information

Best fit models of nucleotide substitution for 20 mt and 1 nuclear *rag1* data partitions (Table 2) were selected using the AIC in Modeltest. Details on partition length, best fit models, and associated parameters are shown in Supplementary material, Appendix 3. EDIBLE (Massingham and Goldman 2000) was employed to calculate the expected phylogenetic information (derived from the Fisher information matrix; Edwards 1972; Atkinson and Donev 1992) given the model parameters of each data set, and the ML tree inferred from the combined data. Phylogenetic information is quantified per site. To obtain the information for each partition, the per-site information matrices were multiplied by the partition length (or alternatively, the total phylogenetic information, being the determinant of the information matrix, is multiplied by the partition length to the power of the number of branches). Total phylogenetic information is not additive between partitions, although the sitewise information matrices are when the branch lengths are common across partitions. We can also compare information between partitions that vary only in their rate of evolution (so that all branch lengths in the tree are multiplied by a constant factor s for each partition). If the information matrix for a partition with rate s is I, then an information matrix comparable between partitions can be found by multiplying by the rate (i.e., sI). Again, this is equivalent to multiplying the phylogenetic information by the rate to the power of the number of branches. To compare information between loci, the relative branch lengths for the tree were fixed at those for the full data set employed in the phylogenetic reconstruction analyses (i.e., mt ribosomal, tRNA, and protein-coding genes, and nuclear *rag1* combined), as it encompasses the variation of all source genes. Phylogenetic information scores were also estimated per branch of the caecilian unrooted tree for each partition.

Geuten et al. (2007) extended Goldman's (1998) method to allow calculation of changes in phylogenetic information upon addition of new hypothetical taxa to a nonclock-like tree such as the caecilian phylogeny studied here. The diagonal elements of the Fisher information matrix describe the information gained about the corresponding branch assuming the lengths of all other branches are known. In that case, the information about a branch can be found by inverting the Fisher information matrix, extracting the appropriate element and then taking its reciprocal (see the "generalized D criteria" for experimental design; Atkinson and Donev 1992).

We compared changes in information to identify branches where addition of a new taxon produces the greatest increase in phylogenetic information for the least well-supported branch of our caecilian phylogeny (see Results). We added a hypothetical new taxon separately to all 9 terminal and 5 internal branches at 12 (evenly distributed) different positions along each branch of the rooted ML tree for the combined data. Each node of the phylogeny was assigned a height equal to the mean distance to all its descendents, or 0 if it is a tip of a terminal branch. The length of an additional branch added between 2 nodes was determined by linear interpolation of their heights, hence such branches are longer the closer to the root of the tree they are added. To check the effect of this experimental regime, we also estimated phylogenetic information for the 3 most informative sister-taxon additions in terminal branches (see Results) when the newly added branch was half or twice the length of its sister.

Information calculations were performed using the EDIBLE software (Massingham and Goldman 2000) modified to incorporate the GTR model of substitution and sitewise rate variation. Statistical tests such as analysis of variance, analysis of covariance, and linear regression were conducted using STATISTICA version 6.0 (StatSoft Inc. 2001).

Results and Discussion

Caecilian Phylogeny

After exclusion of gaps, alignment ambiguities, and third-codon positions of mt protein-coding genes, the final combined alignment is 11 867 bp, of which 7221 are invariant and 2683 are parsimony informative. ML (both PAUP* [$-\ln L = 57,002.416$] and RAxML [$-\ln L = 55\,619.939$]) and BI ($-\ln L = 55\,825.160$ for run 1; $-\ln L = 55\,827.310$ for run 2) yielded the same inferred relationships among caecilian taxa with differences only in branch lengths and levels of support (Fig. 1). All posterior probabilities are close to 1 (BI) and ML bootstrap support is substantial (>75–100%) for all internal branches except 1 (Fig. 1).

The recovered tree agrees with the most recent molecular (Wilkinson et al. 2002, 2003; San Mauro et al. 2004, 2005; Frost et al. 2006; Roelants et al. 2007) and morphological (Wilkinson and Nussbaum 1996, 1999; Wilkinson 1997) studies in supporting the sister group relationship of Ichthyophiidae and Uraeotyphlidae, and the monophyly of Teresomata (Caeciliidae + Scolecomorphidae + Typhlonectidae = Caeciliidae of Frost et al., 2006). Within Teresomata, there is more uncertainty about inter- and intrafamilial phylogenetic relationships (Wilkinson 1997; Wilkinson et al. 2003; San Mauro et al. 2004; Frost et al. 2006; Wilkinson and Nussbaum 2006; Roelants et al. 2007). Our results agree with those of Roelants et al. (2007) and with more traditional classifications in recovering Scolecomorphidae as the sister group of all other teresomatan caecilians, and Caeciliidae paraphyletic with respect only to Typhlonectidae (Nussbaum 1979; Duellman and Trueb 1986; Nussbaum and Wilkinson 1989; Hedges et al. 1993; Wilkinson and Nussbaum 1996; Wilkinson 1997; Wilkinson et al. 2003). We consider the complete congruence between our and Roelants et al.'s (2007) results to be impressive given the marked differences between the data sets: Roelants et al.'s (2007) being more nuclear-based (1 mt ribosomal gene fragment [10%] + 4 nuclear protein-coding gene fragments [90%]) and including representatives of all amphibian lineages (171 amphibian taxa, of which 24 are caecilians) and some amniote outgroups, and ours being more mt-based (complete mt genome [87%] + 1 nuclear protein-coding gene fragment [13%]) and using exclusively (9) caecilian lineages.

In contrast to our results, previous analyses of different data (Wilkinson et al. 2003; mt ribosomal genes; Frost et al. 2006; mt ribosomal and nuclear protein-coding and ribosomal genes) found Caeciliidae to be paraphyletic with respect to Scolecomorphidae as well as Typhlonectidae, with a *Boulengerula* + *Herpele* clade (part of Caeciliidae) recovered as the sister group of all other teresomatan caecilians. Interestingly, the only internal relationship in our ML tree that is not strongly supported is the basal split within Teresomata (Fig. 1). We used PB and the AU test to evaluate the 3 alternative resolutions of the Scolecomorphidae, the caeciliid *Boulengerula*, and all other teresomatans (Table 3). We also evaluated the subtrees of the phylogenies of Wilkinson et al. (2003) and Frost et al. (2006) that are induced by our more limited taxon sampling (Table 3). PB rejects all constrained topologies, whereas the AU test allows rejection only of the topologies of Wilkinson et al. (2003) and Frost et al. (2006) (topologies 4 and 5 in Table 3).

Discrepancies between results from parametric and nonparametric likelihood-based tests are far from completely understood but may be related to different forms of null hypotheses to model misspecification and/or to uncertainty as to the appropriate selection of alternative hypotheses (Goldman et al. 2000; Strimmer and Rambaut 2001; Buckley 2002). In light of the AU tests, we cannot rule out some uncertainty in our caecilian tree (particularly regarding the resolution of *Scolecomorphus*, *Boulengerula*, and other teresomatans). However, our resolution of these relationships receives considerable additional support from the recent molecular study of Roelants et al. (2007) and from morphological phylogenies (Wilkinson and Nussbaum 1996, 1997; Wilkinson 1997).

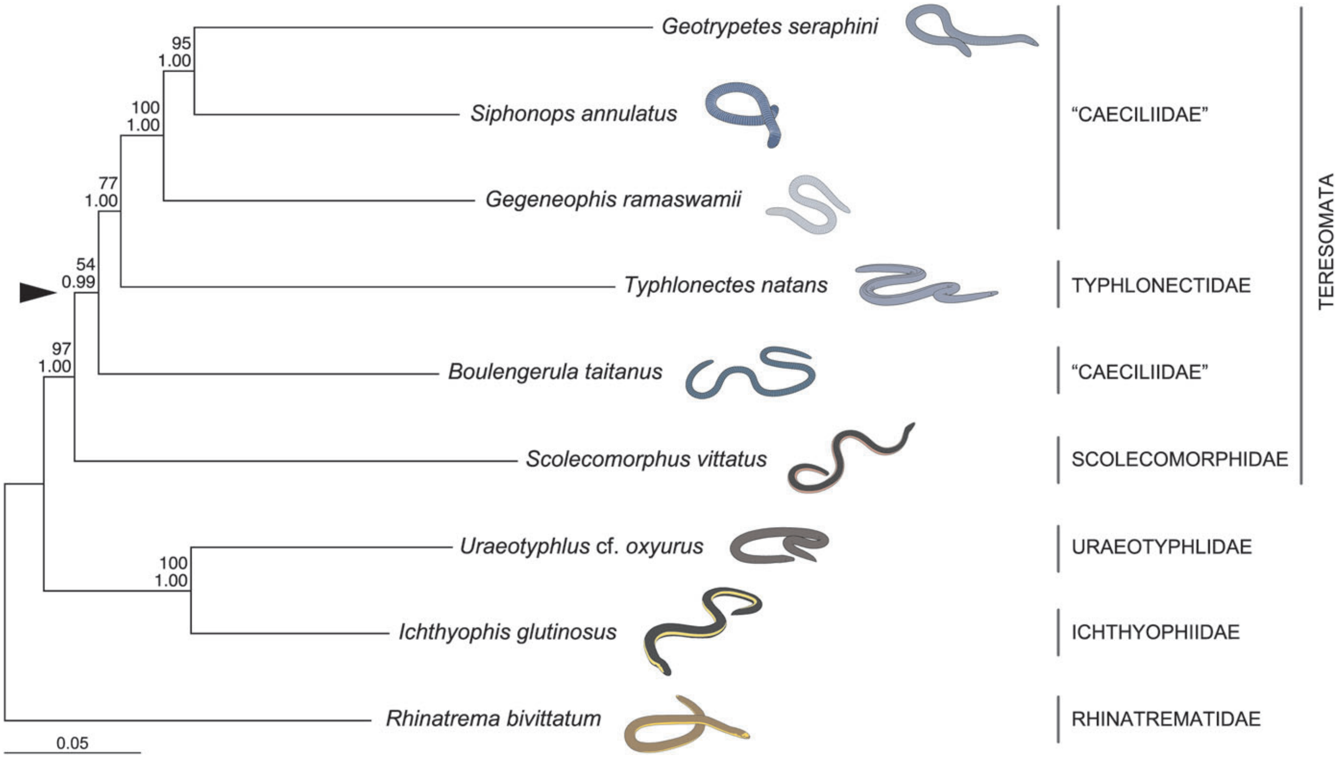


FIGURE 1. ML phylogram for 9 species of caecilian amphibians inferred from our combined mt and nuclear *rag1* data (see text). Numbers above branches represent support for internal branches from ML (RAxML bootstrap proportions; upper value), and BI (posterior probabilities; lower value). Arrowhead indicates the most weakly supported internal branch. Scale bar is in substitutions/site.

Phylogenetic Information and Evolutionary Rates of Data Partitions

Total phylogenetic information about the underlying tree, that is after the information for each partition has been scaled by the relative rate to make it comparable, and evolutionary rates are plotted in Figure 2. Both vary quite widely across the partitions. Substitution rates of partitions RAG1 and CO1 are relatively slower than those of all other mt partitions (Fig. 2), in agreement with previous studies that have indicated the slow evolution of nuclear *rag1* (Groth and Barrowclough 1999; San Mauro et al. 2004) and mt *cox1* (Russo et al. 1996; Zardoya and Meyer 1996; Lopez et al. 1997; San Mauro et al. 2004), this latter one particularly at amino acid level, or after exclusion of third-codon positions, as in our study. Mueller (2006) corroborated that *cox1*, together with the other cytochrome oxidase genes (*cox2*, *cox3*, and *cob*), possesses slow evolutionary rates at amino acid level, and also noted that they also have the fastest rates of all mt genes at nucleotide level (including all codon positions), indicating a relatively higher number of (mainly synonymous) substitutions occurring at the third-codon position of these genes. The rate of evolution of third-codon positions of mt protein-coding genes (partition 3rdPOS) is over 100-fold faster compared with those of all other partitions analyzed (Fig. 2), which agrees with previous studies that reported the faster

TABLE 3. Log-likelihoods and *P* values of PB and AU test for 5 alternative topologies

Alternative topologies	-ln *L*[a]	*P* (PB)	*P* (AU)
1. (*Rbi*,((*Igl*,*Uox*),(*Svi*,(*Bta*,(*Tna*,(*Gra*,(*San*,*Gse*)))))))[b]	55,519.475	–	0.636
2. (*Rbi*,((*Igl*,*Uox*),(*Bta*,(*Svi*,(*Tna*,(*Gra*,(*San*,*Gse*)))))))	55,520.766	<0.001	0.505
3. (*Rbi*,((*Igl*,*Uox*),((*Svi*,*Bta*),(*Tna*,(*Gra*,(*San*,*Gse*))))))	55,528.145	<0.001	0.114
4. (*Rbi*,((*Igl*,*Uox*),(*Bta*,(*Svi*,(*Tna*,(*San*,(*Gra*,*Gse*)))))))[c]	55,534.802	<0.001	0.047
5. (*Rbi*,((*Igl*,*Uox*),(*Bta*,(*Tna*,(*Svi*,(*Gse*,(*Gra*,*San*)))))))[d]	55,548.841	<0.001	0.010

Bta, *Boulengerula taitanus*; *Gra*, *Gegeneophis ramaswamii*; *Gse*, *Geotrypetes seraphini*; *Igl*, *Ichthyophis glutinosus*; *Rbi*, *Rhinatrema bivittatum*; *San*, *Siphonops annulatus*; *Svi*, *Scolecomorphus vittatus*; *Tna*, *Typhlonectes natans*; *Uox*, *Uraeotyphlus* cf. *oxyurus*.

[a] As calculated by PAML.

[b] Unconstrained tree (Fig. 1), Roelants et al. (2007).

[c] Wilkinson et al. (2003).

[d] Frost et al. (2006).

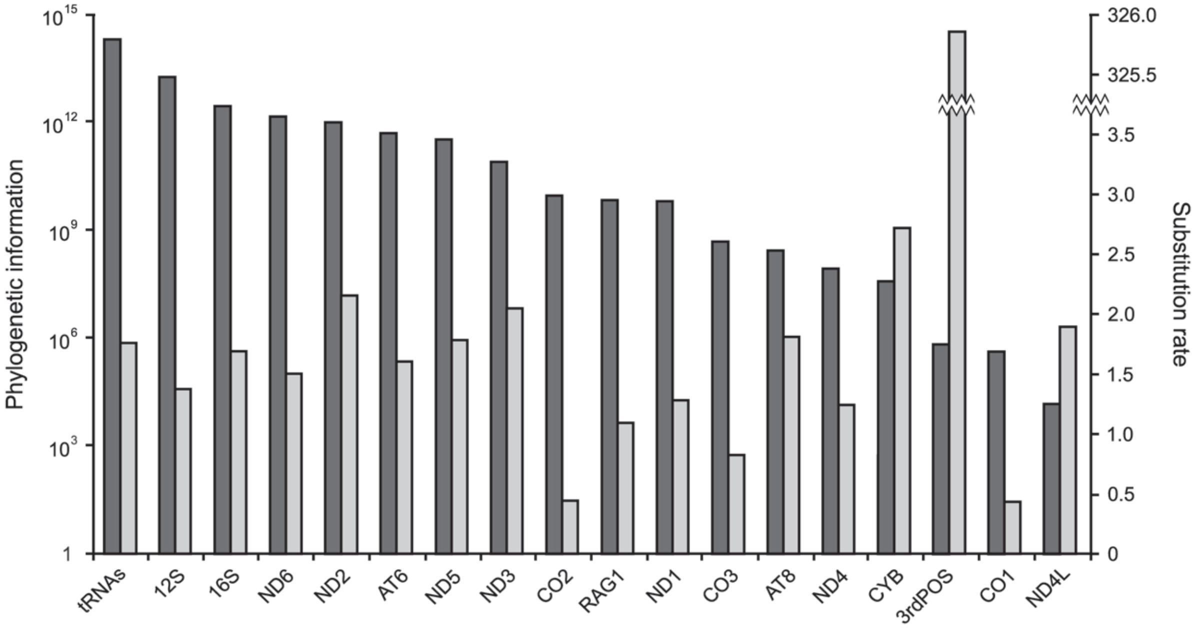


FIGURE 2. Total phylogenetic information per site (dark gray bars; left) and substitution rate per site (light gray bars, right) of each single mt and nuclear *rag1* data partition. Left *y*-axis is on a log scale. Substitution rate is measured as ML tree length.

evolutionary rates of third-codon positions with respect to first and second positions (Irwin et al. 1991; Li and Graur 1991; Johnson and Sorenson 1998; Rodríguez-Trelles et al. 2002) and our finding of saturation. This extremely fast substitution rate is the main reason why we have separately considered all mt third-codon positions (combined) as a single partition for the phylogenetic information analyses of this study.

The phylogenetic information scores, on a per-site basis and after correcting for relative rate of evolution, reveal that the most informative single partitions for the given phylogeny are those for the tRNA genes (1.889×10^{14}), *rrnS* (1.740×10^{13}), and *rrnL* (2.696×10^{12}) (Fig. 2). The phylogenetic performance of these genes is well known, and they (particularly ribosomal genes) have long been used to infer phylogenetic relationships of many diverse organisms spanning a wide range of divergence times (Mindell and Honeycutt 1990; Kumazawa and Nishida 1993; Cummings et al. 1995; Miya and Nishida 2000; Cummings and Meyer 2005; Mueller 2006). Among the protein-coding genes, *nad6* (1.373×10^{12}) and *nad2* (9.447×10^{11}) have the highest information scores (Fig. 2). *Nad2* had already been indicated as good or adequate molecular marker for divergences over 300 million years ago by previous studies on vertebrates (Russo et al. 1996; Zardoya and Meyer 1996; Miya and Nishida 2000; Mueller 2006). In contrast, *nad6* has usually been recovered as a potentially poor (or medium at the most) phylogenetic marker (Zardoya and Meyer 1996; Miya and Nishida 2000; Mueller 2006; but see Russo et al. 1996), with most studies indicating its high variability or rate heterogeneity as probable causes eroding phylogenetic signal. Additionally, the fact that *nad6* encodes on the light strand of the mt DNA and has different base composition biases (Reyes et al. 1998) has led to this gene being routinely excluded from most phylogenetic studies using complete mt genome sequences. One of the main reasons why some of our results on mt protein-coding genes are different from those of previous studies (apart from obvious differences in employed taxa) may be related with the fact that, in our study, mt protein-coding genes are examined to the exclusion of third-codon positions (which are combined and analyzed altogether as a single partition), thus likely reducing the phylogenetic noise associated with multiple substitutions at a given position. In fact, the per-site phylogenetic information score of third-codon positions of mt protein-coding genes (6.751×10^{5}) is among the lowest of all partitions analysed (Fig. 2), and this is probably related to their relatively fast rate of evolution (see above) and the age of caecilian diversification (over 200 million years for the oldest splits; San Mauro et al. 2005; Roelants et al. 2007; see Fig. 3). The partition with the lowest information score is that for *nad4L* (1.453×10^{4}), in full agreement with most previous studies (Russo et al. 1996; Zardoya and Meyer 1996; Miya and Nishida 2000; Mueller 2006) that have indicated the low phylogenetic performance of this gene.

From the first study, much of caecilian molecular phylogenetics has focused on exclusive or majority use of mt *rrnS* (12S) and *rrnL* (16S) fragments with variable success at different levels of divergence (Hedges et al. 1993; Gower et al. 2002, 2005; Wilkinson et al. 2002, 2003). Our results show that alignments made from sequences of these entire genes are among the best partitions for resolving relationships among the major lineages of caecilians included here but, because the results are context

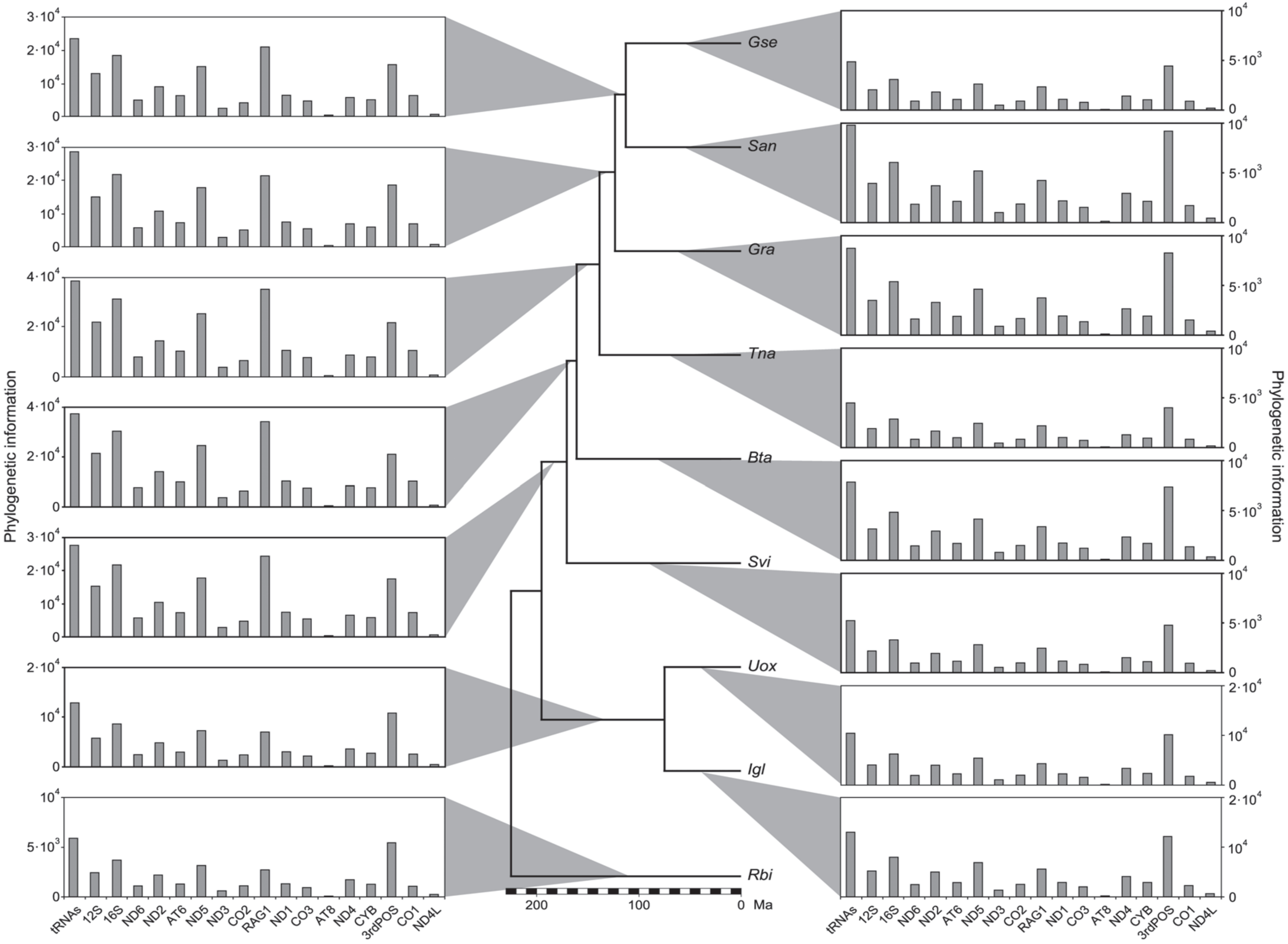


FIGURE 3. Phylogenetic information content of single data partitions estimated per branch of our caecilian tree (see Fig. 1), as mapped onto the timetree of Roelants et al. (2007). *Bta*, *Boulengerula taitanus*; *Gra*, *Gegeneophis ramaswamii*; *Gse*, *Geotrypetes seraphini*; *Igl*, *Ichthyophis glutinosus*; *Rbi*, *Rhinatrema bivittatum*; *San*, *Siphonops annulatus*; *Svi*, *Scolecomorphus vittatus*; *Tna*, *Typhlonectes natans*; *Uox*, *Uraeotyphlus* cf. *oxyurus*.

specific, they do not allow conclusions as to their relative utility in resolving more recent divergences.

Phylogenetic Information Per Branch

Figure 3 shows information scores estimated per branch of the unrooted caecilian tree plotted against Roelants et al.'s (2007) ultrametric timetree. In general, information scores of all partitions are lower in terminal than in internal branches, particularly those spanning a time depth of 75–196 million years. As for the partition totals (Fig. 2), the most informative partition in all branches is that for the tRNA genes (Fig. 3). The rank order of partition information is not constant across branches, and the relative performance of slow-evolving *rag1* and fast-evolving mt third-codon positions changes more markedly between internal and terminal branches (*rag1* performing better in internal branches, mt third-codon positions performing better in terminal branches; Fig. 3). We conducted a factorial (2-way) analysis of variance to assess variations in log-transformed phylogenetic information between terminal and internal branches (main effect "branch type"), and between slow-evolving *rag1*, fast-evolving mt third-codon positions, and all other partitions (main effect "gene rate"). Both main effects are highly significant ($F_{1,264} = 26.146$ for "branch type"; $F_{2,264} = 15.986$ for "gene rate"; $P < 0.001$ in both cases), indicating that information scores are significantly higher in internal than in terminal branches, and that, in general, fast- and slow-evolving partitions perform better than all other partitions (taken together), although apparently in different parts of the tree (mt third-codon positions perform better in terminal branches). The interaction of the 2 main effects was not significant ($F_{2,264} = 0.723$; $P = 0.486$). The reason why terminal branches have in general or for the most part less information is elusive, but likely related to the fact that the information estimated is really about branch lengths and these are less constrained for the terminal than for the internal branches.

Combining Information of Mt Data Partitions: Assessing Mitogenomic Information

Overall phylogenetic information for the complete mt genome can be determined from the information matrices of the partitions. Although the phylogenetic information, being the determinant of the information matrix, is not additive, the information matrices can simply be added together and then the determinant taken. For partitions with different relative rates, the information matrices first have to be made comparable, as described above, before being summed. The alternative of estimating phylogenetic information from the concatenation of the partitions is expected to be potentially misleading because of the averaging of substitution model parameters for the concatenated data. To explore this, phylogenetic information was estimated directly from concatenated data sets (PROTS-NO3, PROTS-ALL, and mtGENOME-NO3; Table 2) and compared with the combined phylogenetic information scores for the component partitions. The results show that there is a notable variation in information scores between those data sets averaging phylogenetic information and those adding up information (Fig. 4). For example, phylogenetic information for PROTS-ALL (1.073×10^{72}) is higher than the combined phylogenetic information for PROTS-NO3 plus 3rdPOS (1.095×10^{70}) despite being based on the same set of sequence characters. Similarly, phylogenetic information of mtGENOME-NO3 (9.506×10^{71}) is more than the combined information scores of all single mt partitions excluding third-codon positions (2.241×10^{71}).

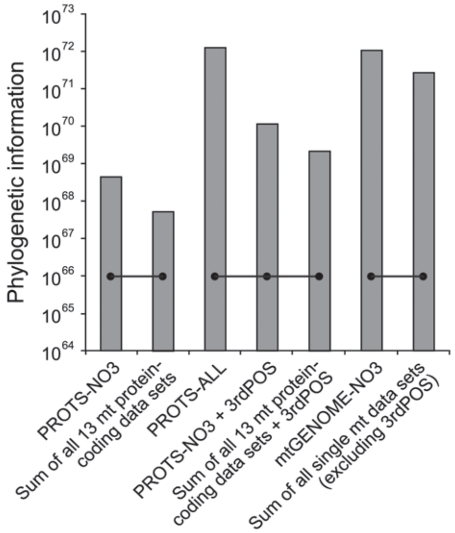


FIGURE 4. Phylogenetic information scores for composite mt data sets (total information for the partition). Columns linked by horizontal bars are based on the same set of sequence characters. *Y*-axis is on a log scale.

Our results demonstrate the substantial impact that concatenation and consequent substitution model misspecification can have for estimates of phylogenetic information: all these results show that misspecification leads to overestimating how informative the data are and so false confidence in the topology. In general, it is better to combine information scores estimated separately for partitions with differing best fit models of sequence evolution than to estimate scores from concatenated data. This raises the possibility that further subdivision of our partitions (e.g., first and second codon postions and stem and loop regions of ribosomal genes) would alter our assessments of phylogenetic information.

Experimental Design and Caecilian Systematics

As indicated above, a point of disagreement among recent molecular studies (Wilkinson et al. 2003; Frost et al. 2006; Roelants et al. 2007) and the greatest uncertainty in our caecilian phylogeny (both from ML bootstrap scores and AU tests of alternative topologies) involves the relationships among *Scolecomorphus*, *Boulengerula*, and other teresomatans (Fig. 1 and Table 3).

We calculated the Fisher information of the branch separating *Scolecomorphus* and *Boulengerula* to identify positions in our caecilian tree at which a hypothetical taxon can be added so as to best increase phylogenetic information for the branch resolving these relationships. The increase in phylogenetic information is strongly inversely correlated ($R^2 > 0.980$; $F_{1,10} > 489.634$; $P < 0.001$ in all cases) with the distance between the controversial branch and the position at which the hypothetical taxon is added (Fig. 5).

We used analysis of covariance (distance as covariate) to assess variation in the log-transformed increase of phylogenetic information and planned comparisons to examine contrasts between adding the hypothetical taxon to specific branches. The greatest increase in phylogenetic information (significantly higher than those in all other branches; $F_{1,153} = 639.285$; $P < 0.001$) occurs when the hypothetical taxon joins internal branch 1 neighboring the controversial internal branch (phylogenetic information going higher than 3.2×10^5) (Fig. 5b). Unfortunately, it seems unlikely that known extant caecilian diversity (Wilkinson and Nussbaum 2006) includes any lineage that would join branch 1. We consider it likely that most, if not all, other extant caecilians would join our tree individually on the terminal branches. Of the terminal branches, significant increases in information ($F_{1,153}$=172.809; $P < 0.001$) occur with the addition of the hypothetical taxon to the *Scolecomorphus* branch (also going higher than 3.2×10^5), followed by the *Boulengerula* and *Rhinatrema* branches (Fig. 5b). When the hypothetical taxon is added to any other terminal branch, the increase in phylogenetic information is not

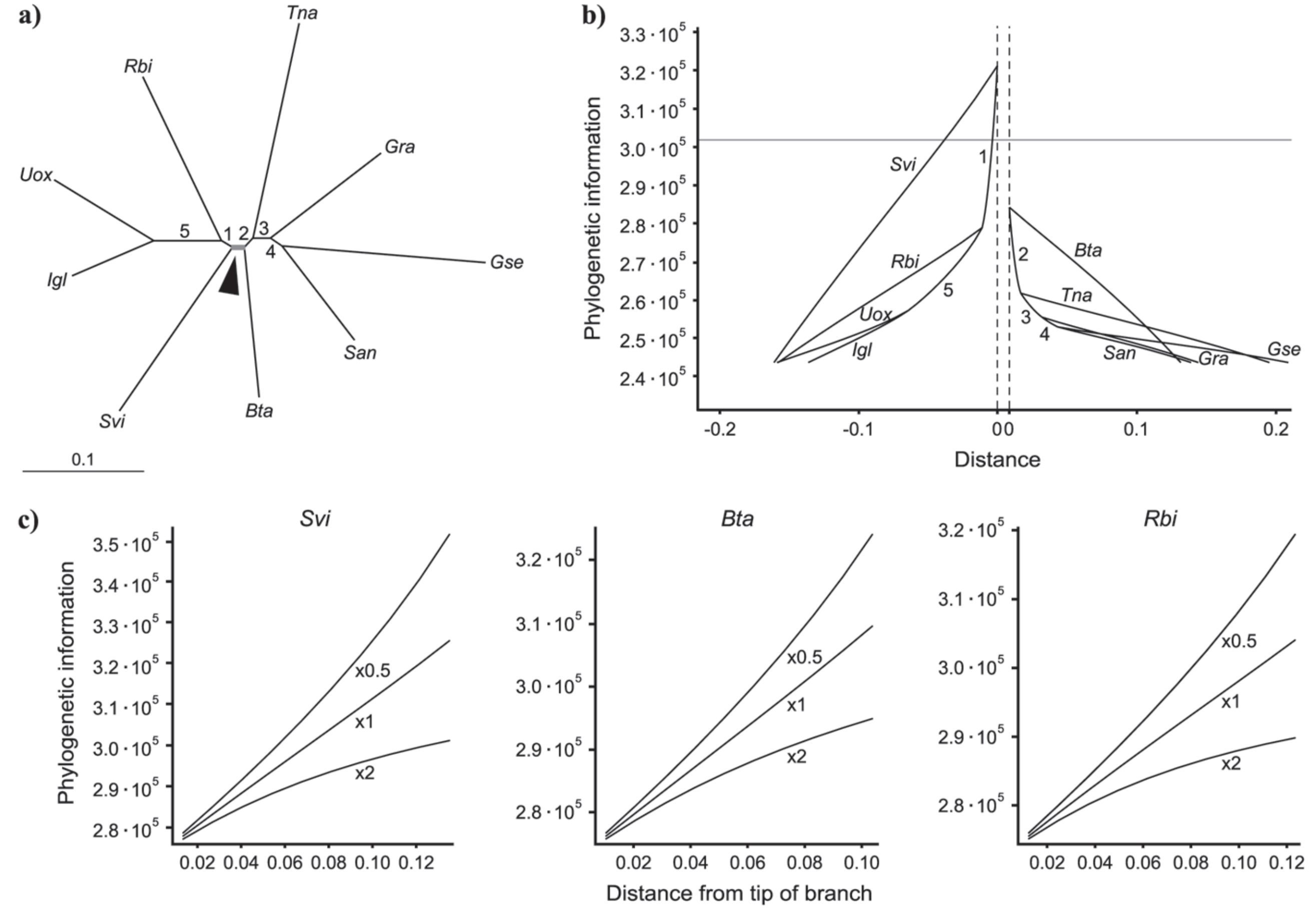


FIGURE 5. Changes in phylogenetic information for the most weakly supported internal branch of our caecilian tree (see Fig. 1) when a new, hypothetical taxon is added to different parts of the tree. a) Our ML caecilian phylogeny, indicating the most weakly supported internal branch (arrowhead). Scale bar is in substitutions/site. b) Increase in phylogenetic information of most weakly supported internal branch plotted against the distance from that branch at which the hypothetical taxon has been added. Terminal branches are labeled with the name of the taxon at the tip of the branch; other internal branches are labeled following branch numbers in (a). Vertical dashed lines denote the boundaries of the controversial branch. Horizontal gray line indicates the increase in phylogenetic information of the most weakly supported internal branch without increasing taxon sampling but instead increasing character sampling by 1300 bp (sequence data of the same nature as already sequenced) for each of the 9 original taxa. c) Phylogenetic information for the 3 most informative terminal branch additions of a new hypothetical taxon when the length of the branch joining the new taxon is variably: equal to the mean of the adjacent branch (×1), half that length (×0.5), and twice that length (×2). X-axis in (b) and (c) are absolute values (substitutions/site) corresponding to branch lengths as given in scale in (a). *Bta*, *Boulengerula taitanus*; *Gra*, *Gegeneophis ramaswamii*; *Gse*, *Geotrypetes seraphini*; *Igl*, *Ichthyophis glutinosus*; *Rbi*, *Rhinatrema bivittatum*; *San*, *Siphonops annulatus*; *Svi*, *Scolecomorphus vittatus*; *Tna*, *Typhlonectes natans*; *Uox*, *Uraeotyphlus* cf. *oxyurus*.

significant. The increase in phylogenetic information is inversely related to the branch length of the hypothetical taxon (Fig. 5c). The horizontal gray line in Figure 5b indicates the expected increase in phylogenetic information of the controversial branch obtained by adding 1300 bp of sequence data (of the same kind—mitogenomic + *rag1*) for each of the 9 included taxa (without any additional hypothetical taxa), simulating the effect of sequencing an additional, hypothetical gene for each of our current taxa. Sequencing 1300 bp for 9 taxa represents approximately the same amount of total sequencing effort (in terms of total bp sequenced) as sequencing our final combined data (11 867 bp) for a single additional taxon. This gives quantitative insight into the relative merits of sampling more characters versus more taxa.

These results combined with background knowledge of caecilian diversity and phylogeny provide guidance for future sampling to provide compelling resolution of the relationships of *Scolecomorphus, Boulengerula*, and other teresomatans, and the potential paraphyly of the Caecillidae with respect to the Scolecomorphidae. Caecilians with the greatest chance of increasing phylogenetic accuracy in this part of our tree are any of those that would join the 1) *Scolecomorphus*, 2) *Boulengerula*, and, less intuitively, 3) *Rhinatrema* branches. Addition of a single taxon to any other terminal branch is predicted to result in a much smaller increase in phylogenetic information. According to recent studies (Frost et al. 2006; Wilkinson and Nussbaum 2006; Roelants et al. 2007), extant caecilians that would join our tree at these 3 most promising terminal branches are 1) the 2 unsampled species of *Scolecomorphus* and the 3 species of *Crotaphatrema*, 2) the 6 unsampled species of *Boulengerula* and 2 species of *Herpele*, and 3) the 8 species of *Epicrionops*. At least some of these (*Crotaphatrema*, *Herpele* and *Boulengerula boulengeri, Epicrionops*) appear to join the terminal branches of our phylogeny proximal to the controversial branch (Gower et al. 2002; Wilkinson et al. 2003; Frost et al. 2006; Loader et al. 2007) offering additional hope that a compelling resolution of this controversy is attainable using *rag1* and complete mt genomes or a suite of the most informative genes.

Given limited resources, generating whole mitogenomic and *rag1* data for 1 or more of the identified priority additional taxa joining the *Scolecomorphus* branch may be a better (more efficient) strategy than sequencing a similar amount of additional nucleotides spread across the current taxon sampling. Moreover, obtaining whole mitogenomic data provide information, such as gene order, that may provide additional evidence of phylogenetic relationships (Rokas and Holland 2000; San Mauro et al. 2006).

Concluding Remarks

Goldman's (1998) method offers a powerful tool for experimental design in molecular phylogenetics that has yet to receive much attention. Data for caecilian amphibians illustrate how this method can be used to provide quantitative comparisons of the phylogenetic information content of different genes or other data partitions across an entire tree or per branch. This comparison can be used to identify the most informative markers for the phylogenetic question at hand and to predict the impact of additional data in the form of new characters and/or taxa. The latter offers a coherent framework for determining whether it is most efficient to add more characters or more taxa or a combination of both. Although cheap, high-throughput sequencing might make careful choice of molecular markers less important in future, the design of polymerase chain reaction (PCR) primers and optimization of PCR conditions will remain a rate-limiting step in many molecular systematic studies, so there will still be a significant cost to adding markers. Providing additional taxa will be dependent on the availability of tissue samples for the organisms, which often involves directed and time-consuming fieldwork, so taxon choice will certainly remain an important problem.

Most of our results regarding the informativeness of different markers confirm insights from previous studies, such as the utility of mt ribosomal and transfer RNA genes and the poor performance of *nad4L* for inferring deeper divergences (Mindell and Honeycutt 1990; Kumazawa and Nishida 1993; Cummings et al. 1995; Zardoya and Meyer 1996; Groth and Barrowclough 1999; Miya and Nishida 2000; Cummings and Meyer 2005; Mueller 2006). Importantly, Goldman's method takes into account the specific phylogenetic context when assessing the informativeness of different markers.

Our results are also consistent with the widely held intuition regarding the greater informativeness of additional taxa that have short branches, and that join the tree closer to controversial internal branches (Goldman 1998; Geuten et al. 2007). We find the quantitative support provided by Goldman's method for these intuitions to be reassuring, and the potential for less intuitive insights to be exciting.

Although not comprehensive (e.g., we did not consider additions of multiple hypothetical taxa), our investigations of sampling in caecilian molecular phylogenetics are highly illustrative. Further assessment of Goldman's method should benefit from empirical tests of the specific predictions we have made. It is important to always bear in mind that the results produced by Goldman's method are context specific, but it might be the case that our results are more broadly extendable to other phylogenetic questions concerning similarly deep divergences.

SUPPLEMENTARY MATERIAL

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

FUNDING

This work received financial support from grants of the Ministry of Science and Innovation of Spain

(CGL2004-00401, and MEC/Fulbright postdoctoral fellowship 2007-0448), the Natural Environment Research Council (GST/02/832) and the Biotechnology and Biological Sciences Research Council (40/G18385) of the United Kingdom, the European Molecular Biology Laboratory, and the European Commission's Research Infrastructure Action via the SYNTHESYS Project.

Acknowledgements

We thank Salvador Carranza, Rob Cruickshank, Mario García-París, Adrian Paterson, David Posada, Jack Sullivan, and 2 anonymous reviewers for insightful comments on earlier versions of the manuscript. M.W. gratefully acknowledges the assistance of Anton Espira in the field, the National Museums of Kenya for support and loans, and the Kenya Wildlife Service for collection and export permits. D.J.G. and M.W. thank Jeannot and Odette (Camp Patawa) for their hospitality while completing some of the final stages of their contribution to this study.

References

Akaike H. 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov B.N., Csaki F., editors. Second international symposium of information theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.

Atkinson A.C., Donev A.N. 1992. Optimum experimental designs. London: Oxford University Press. p. 352.

Buckley T.R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst. Biol. 51:509–523.

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

Cotton J.A., Wilkinson M. 2008. Quantifying the potential utility of phylogenetic characters. Taxon. 57:131–136.

Cummings M.P., Meyer A. 2005. Magic bullets and golden rules: data sampling in molecular phylogenetics. Zoology. 108:329–336.

Cummings M.P., Otto S.P., Wakeley J. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12:814–822.

Curole J.P., Kocher T.D. 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. Trends Ecol. Evol. 14:394–398.

Duellman W.E., Trueb L. 1986. Biology of amphibians. New York: McGraw-Hill. p. 670.

Edwards A.W.F. 1972. Likelihood. Cambridge (UK): Cambridge University Press. p. 144–160.

Efron B. 1985. Bootstrap confidence intervals for a class of parametric problems. Biometrika. 72:45–58.

Efron B., Hinkley D.V. 1978. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika. 65:457–487.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Frost D.R., Grant T., Faivovich J., Bain R.H., Haas A., Haddad C.F.B., de Sá R.O., Channing A., Wilkinson M., Donnellan S.C., Raxworthy C.J., Campbell J.A., Blotto B.L., Moler P., Drewes R.C., Nussbaum R.A., Lynch J.D., Green D.M., Wheeler W.C. 2006. The amphibian tree of life. Bull. Am. Mus. Nat. Hist. 297:1–370.

Gauthier O., Lapointe F.-J. 2007. Seeing the trees for the network: consensus, information content, and superphylogenies. Syst. Biol. 56:345–355.

Geuten K., Massingham T., Darius P., Smets E., Goldman N. 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56:609–622.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. Lond. B. 265:1779–1786.

Goldman N., Anderson J.P., Rodrigo A.G. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Gower D.J., Bahir M., Mapatuna Y., Pethiyagoda R., Raheem D., Wilkinson M. 2005. Molecular phylogenetics of Sri Lankan Ichthyophis (Amphibia: Gymnophiona: Ichthyophiidae), with discovery of a cryptic species. Raffles Bull. Zool. (Suppl 12):153–161.

Gower D.J., Kupfer A., Oommen O.V., Himstedt W., Nussbaum R.A., Loader S.P., Presswell B., Müller H., Krishna S.B., Boistel R., Wilkinson M. 2002. A molecular phylogeny of ichthyophiid caecilians (Amphibia: Gymnophiona: Ichthyophiidae): out of India or out of South East Asia? Proc. R. Soc. Lond. B. 269:1563–1569.

Graybeal A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst. Biol. 43:174–193.

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.

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.

Hardman M., Hardman L.M. 2006. Comparison of the phylogenetic performance of neodermatan mitochondrial protein-coding genes. Zool. Scr. 35:655–665.

Hedges S.B., Nussbaum R.A., Maxson L.R. 1993. Caecilian phylogeny and biogeography inferred from mitochondrial DNA sequences of the 12SrRNA and 16S rRNA genes (Amphibia: Gymnophiona). Herpetol. Monogr. 7:64–76.

Hedtke S.M., Townsend T.M., Hillis D.M. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522–529.

Hillis D.M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigatior bias. Syst. Biol. 47:3–8.

Hofacker I.L. 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31:3429–3431.

Hofacker I.L., Fontana, W., Stadler P.F., Bonhoeffer L.S., Tacker M., Schuster P. 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. 125:167–188.

Holm S. 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6:65–70.

Huelsenbeck J.P., Hillis D.M., Jones R. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. In: Ferarris J.D., Palumbi S.R., editors. Molecular zoology: advances, strategies, and protocols. New York: Wiley-Liss. p. 19–45.

Huelsenbeck J.P., Ronquist F.R. 2001. MRBAYES: bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Huelsenbeck J.P., Ronquist F.R., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294:2310–2314.

Irwin D.M., Kocher T.D., Wilson A.C. 1991. Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. 32:128–144.

Johnson K.P., Sorenson M.D. 1998. Comparing molecular evolution in two mitochondrial protein coding genes (cytochrome b and ND2) in the dabbling ducks (Tribe: Anatini). Mol. Phylogenet. Evol. 10:82–94.

Kim J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

Kim J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47:43–60.

Kumazawa Y., Nishida M. 1993. Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. J. Mol. Evol. 37:380–398.

Li C., Lu G., Orti G. 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst. Biol. 57:519–539.

Li W.-H., Graur D. 1991. Fundamentals of molecular evolution. Sunderland (MA): Sinauer. p. 284.

Loader S.P., Pisani D., Cotton J.A., Gower D.J., Day J.J., Wilkinson M. 2007. Relative time scales reveal multiple origins of parallel disjunct distributions of African caecilian amphibians. Biol. Lett. 3:505–508.

Lopez J.V., Culver M., Stephens J.C., Johnson W.E., O'Brien S.J. 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Mol. Biol. Evol. 14:277–286.

Massingham T., Goldman N. 2000. EDIBLE: experimental design and information calculations in phylogenetics. Bioinformatics. 16: 294–295.

McGuire J.A., Witt C.C., Altshuler D.L., Remsen J.V. Jr. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Syst. Biol. 56:837–856.

Mindell D.P., Honeycutt R.L. 1990. Ribosomal RNA in vertebrates: evolution and phylogenetic applications. Annu. Rev. Ecol. Syst. 21: 541–566.

Miya M., Nishida M. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. Mol. Phylogenet. Evol. 17:437–455.

Mueller R.L. 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst. Biol. 55:289–300.

Nussbaum R.A. 1977. Rhinatrematidae: a new family of caecilians (Amphibia: Gymnophiona). Occas. Pap. Mus. Zool. Univ. Mich. 682:1–30.

Nussbaum R.A. 1979. The taxonomic status of the caecilian genus *Uraeotyphlus* Peters. Occas. Pap. Mus. Zool. Univ. Mich. 687: 1–20.

Nussbaum R.A., Wilkinson M. 1989. On the classification and phylogeny of caecilians (Amphibia: Gymnophiona), a critical review. Herpetol. Monogr. 3:1–42.

Nylander J.A.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J.L. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Nylander J.A.A., Wilgenbusch J.C., Warren D.L., Swofford D.L. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 24:581–583.

Poe S., Swofford D.L. 1999. Taxon sampling revisited. Nature. 398: 299–300.

Pollock D.D., Bruno W.J. 2000. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. 17:1854–1858.

Pollock D.D., Zwickl D.J., McGuire J.A., Hillis D.M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.

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

Rannala B., Huelsenbeck J.P., Yang Z., Nielsen R. 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47:702–710.

Reeves J.H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17–31.

Reyes A., Gissi C., Pesole G., Saccone C. 1998. Asymetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. 15:957–966.

Rodríguez F., Oliver J.F., Marín A., Medina J.R. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142: 485–501.

Rodríguez-Trelles F., Alarcón L., Fontdevila A. 2002. Molecular evolution and phylogeny of the *buzzatii* complex (*Drosophila repleta* group): a maximum-likelihood approach. Mol. Biol. Evol. 17: 1112–1122.

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. U.S.A. 104:887–892.

Rokas A., Carroll S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.

Rokas A., Holland P.W.H. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15:454–459.

Ronquist F. 1996. Matrix representation of trees, redundancy, and weighting. Syst. Biol. 45:247–253.

Ronquist F., Huelsenbeck J.P. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572–1574.

Rosenberg M.S., Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. U.S.A. 98:10751–10756.

Russo C.A.M., Takezaki N., Nei M. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13:525–536.

Sambrook J., Fritsch E.F., Maniatis T. 1989. Molecular cloning. A laboratory manual. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. p. E.3–E.4.

San Mauro D., Gower, D.J., Oommen O.V., Wilkinson M., Zardoya R. 2004. Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. Mol. Phylogenet. Evol. 33:413–427.

San Mauro D., Gower D.J., Zardoya R., Wilkinson M. 2006. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. Mol. Biol. Evol. 23:227–234.

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.

Schwarz G. 1978. Estimating the dimensions of a model. Ann. Stat. 6:461–464.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17: 1246–1247.

Soltis D.E., Albert V.A., Savolainen V., Hilu K., Qiu Y.L., Chase M.W., Farris J.S., Stefanovic S., Rice D.W., Palmer J.D., Soltis P.S. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.

Springer M.S., DeBry R.W., Douady C.J., Amrine H.M., Madsen O., deJong W.W., Stanhope M.J. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol. Biol. Evol. 18:132–143.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Stamatakis A., Blagojevic F., Nikolopoulos D., Antonopoulos C. 2007. Exploring new search algorithms and hardware for phylogenetics: RAxML meets the IBM cell. J. VLSI Signal Process. 48:271–286.

StatSoft Inc. 2001. STATISTICA (data analysis software system). Version 6. StatSoft. Available from: URL http://www.statsoft.com.

Strimmer K., Rambaut A. 2001. Inferring confidence sets of possible misspecified gene trees. Proc. R. Soc. Lond. B 269:137–142.

Swofford D.L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sunderland (MA): Sinauer Associates, Inc.

Taylor E.H. 1968. The caecilians of the world: a taxonomic analysis. Lawrence (KS): University of Kansas Press. p. 848.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin J., Higgins D.G. 1997. The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.

Thorley J.L., Wilkinson M., Charleston M.A. 1998. The information content of consensus trees. In: Rizzi A., Vichi M., Bock H.-H., editors. Advances in data science and classification. Berlin (Germany): Springer. p. 91–98.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Townsend J.P., López-Giráldez F., Friedman R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67:437–447.

Wägele J.W., Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol. Biol. 7:147.

Wilkinson M. 1992. The phylogenetic position of the Rhinatrematidae (Amphibia: Gymnophiona): evidence from the larval lateral line system. Amphib.-Reptil. 13:74–79.

Wilkinson M. 1996. The heart and aortic arches of rhinatrematid caecilians (Amphibia: Gymnophiona). Zoomorphology. 105:277–295.

Wilkinson M. 1997. Characters, congruence and quality: a study of neuroanatomical and traditional data in caecilian phylogeny. Biol. Rev. 72:423–470.

Wilkinson M., Cotton J.A., Thorley J.L. 2004. The information content of trees and their matrix representations. Syst. Biol. 53:989–1001.

Wilkinson M., Loader S.P., Gower D.J., Sheps J.A., Cohen B.L. 2003. Phylogenetic relationships of African caecilians (Amphibia: Gymnophiona): insights from mitochondrial rRNA gene sequences. Afr. J. Herpetol. 52:83–92.

Wilkinson M., Nussbaum R.A. 1996. On the phylogenetic position of the Uraeotyphlidae (Amphibia: Gymnophiona). Copeia. 1996: 550–562.

Wilkinson M., Nussbaum R.A. 1997. Comparative morphology and evolution of the lungless caecilian *Atretochoana eiselti* (Taylor) (Amphibia: Gymnophiona: Typhlonectidae). Biol. J. Linn. Soc. 62:39–109.

Wilkinson M., Nussbaum R.A. 1999. Evolutionary relationships of the lungless caecilian *Atretochoana eiselti* (Amphibia: Gymnophiona: Typhlonectidae). Zool. J. Linn. Soc. 126:191–223.

Wilkinson M., Nussbaum R.A. 2006. Caecilian phylogeny and classification. In: Exbrayat J.-M., editor. Reproductive biology and phylogeny of Gymnophiona (Caecilians). Enfield (NH): Science Publishers. p. 39–78.

Wilkinson M., Sheps J.A., Oommen O.V., Cohen B.L. 2002. Phylogenetic relationships of Indian caecilians (Amphibia: Gymnophiona) inferred from mitochondrial rRNA gene sequences. Mol. Phylogenet. Evol. 23:401–407.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z., Goldman N., Friday A. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. 34:384–399.

Zardoya R., Meyer A. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol. Biol. Evol. 13:933–942.

Zwickl D.J., Hillis D.M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.

Received 12 May 2008; reviews returned 17 July 2008; accepted 29 June 2009
Associate Editor: Adrian Paterson

ONLINE APPENDIX 1

DISTINCT STRUCTURAL FEATURES OF THE MT GENOMES OF *BOULENGERULA TAITANUS* AND *GEOTRYPETES SERAPHINI*

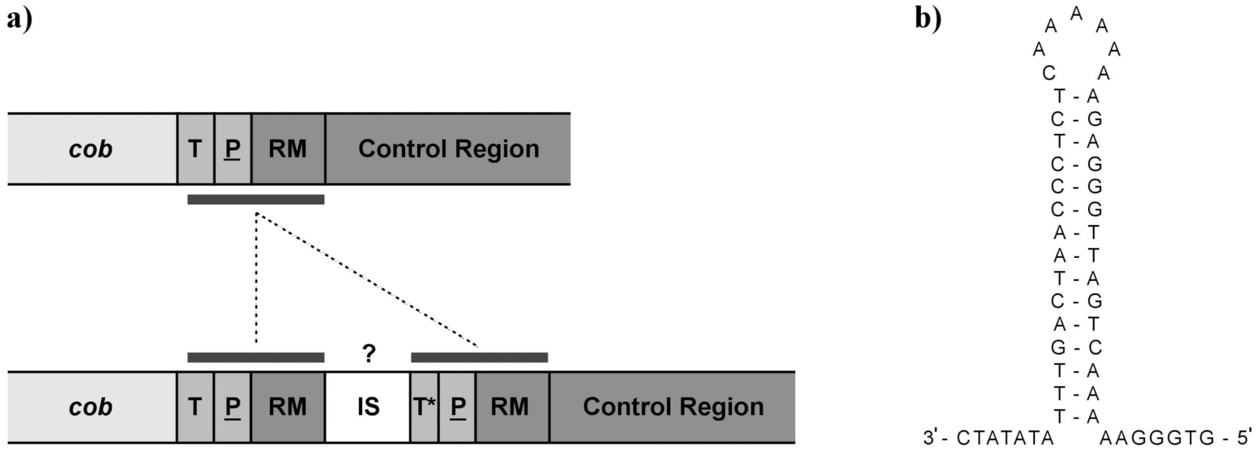
Mt genome of *Boulengerula taitanus*

The mt genome of *Boulengerula taitanus* conforms to the vertebrate consensus mt gene arrangement (Jameson et al., 2003), but possesses two copies of the 3'-half portion of *trnT*, *trnP*, and a 50 bp-long motif of the 5'-end of the control region, with the two paralogous regions separated by a 61-bp-long non-coding spacer (online Appendix 1 Figure 1a). The two paralogous regions are virtually identical in nucleotide sequence, and are most parsimoniously interpreted as the result of a single tandem duplication. Given that duplicated mtDNA sequences are probably lost rapidly under strong selective pressure to constrain mt genome (Wolstenholme, 1992), the high sequence conservation in this case might be the result of a relatively recent duplication event and/or concerted evolution following duplication (Kumazawa et al., 1996). The nature of the intergenic spacer is less clear, but it is most parsimoniously interpreted as resulting from the same duplication event, and represents a downstream portion of the control region that has undergone more rapid evolution.

Duplications in this mt region have been reported for other vertebrates, such as the amphisbaenian *Bipes biporus* (Macey et al., 1998) and the scincomorph lizard *Cordylus warreni* (Kumazawa, 2004), providing further evidence that duplications are more likely to occur in close proximity to (or involving) replication origins (e.g., Moritz and Brown, 1987; Kumazawa et al., 1998; Mindell et al., 1998).

Mt genome of *Geotrypetes seraphini*

The mt genome organization of *Geotrypetes seraphini* also conforms to the vertebrate consensus mt gene arrangement (Jameson et al., 2003), but it lacks the origin of light-strand replication (O_L) at its typical position. Instead, only five nucleotides occur between *trnN* and *trnC*, and no stem-loop structure can be identified. However, the mt genome does possess a long (301 bp) non-coding intergenic spacer between *trnI* and *trnQ*, which includes a 37 bp-long region that can be folded into a stem-loop structure (online Appendix 1 Figure 1b). This stem-loop has a nucleotide sequence that is fairly dissimilar to those reported for O_Ls of other caecilians (Zardoya and Meyer, 2000; San Mauro et al., 2004, 2006), and it lacks some functional motifs interpreted as necessary for light strand replication in human and mouse O_Ls (Brennicke and Clayton, 1981; Wong and Clayton, 1985; Hixson et al., 1986). However, this stem-loop structure in *G. seraphini* might represent a novel O_L, co-opted for the function of light strand replication, similar to reports of some tRNA genes in other animal species (Clayton, 1982; Clary and Wolstenholme, 1985). A long intergenic spacer between *trnI* and *trnQ* (corresponding to a duplicated control region, ψ*trnP*, and *trnL1*) has been reported in the mt genome of the snake *Dinodon semicarinatus* (Kumazawa et al., 1998). Unlike *G. seraphini*, this snake's mt genome also possesses a typical functional O_L, and sequence similarity between the *G. seraphini* spacer and any part of the mtDNA control region of the same animal is low, and BLAST searches (Altschul et al., 1990) produced no close matches.



ONLINE APPENDIX 1 – FIGURE 1. Distinct structural features of the newly determined mt genomes of *Boulengerula taitanus* and *Geotrypetes seraphini*. (a) Region between *cob* and the control region in *B. taitanus* (below), with interpretation of partial duplication from ancestral genome arrangement (above). tRNA genes are abbreviated with the corresponding one-letter amino acid code, with gene (*trnP*) encoded by the light-strand underlined. IS, intergenic spacer; RM, repeated motif probably representing duplication of 5'-end of the control region; T*, truncated *trnT* (27 bp shorter than in other caecilians). (b) Proposed secondary structure of the stem-loop region found between *trnI* and *trnQ* in mt genome of *G. seraphini*.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Brennicke, A., and D. A. Clayton. 1981. Nucleotide assignment of alkalisensitive sites in mouse mitochondrial DNA. J. Biol. Chem. 256:10613–10617.

Clary, D. O., and D. R. Wolstenholme. 1985. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization and genetic code. J. Mol. Evol. 22:252-271.

Clayton, D. A. 1982. Replication of animal mitochondrial DNA. Cell 28:693-705.

Hixson, J. E., T. W. Wong, and D. A. Clayton. 1986. Both the conserved stem-loop and divergent 5'-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. J. Biol. Chem. 261:2384-2390.

Jameson, D., A. P. Gibson, C. Hudelot, and P. G. Higgs. 2003. OGRe: a relational database for comparative analyses of mitochondrial genomes. Nucleic Acids Res. 31:202-206.

Kumazawa, Y. 2004. Mitochondrial DNA sequences of five squamates: phylogenetic affiliation of snakes. DNA Res. 11:137-144.

Kumazawa, Y., H. Ota, M. Nishida, and T. Ozawa. 1996. Gene rearrangements in snake mitochondrial genomes: highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA cluster. Mol. Biol. Evol. 13:1242-1254.

Kumazawa, Y., H. Ota, M. Nishida, and T. Ozawa. 1998. The complete nucleotide sequence of snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions. Genetics 150:313-329.

Macey, J. R., J. A. Schulte II, A. Larson, and T. J. Papenfuss. 1998. Tandem duplication via light-strand synthesis may provide a precursor for mitochondrial genomic rearrangement. Mol. Biol. Evol. 15:71-75.

Mindell, D. P., M. D. Sorenson, and D. E. Dimcheff. 1998. Multiple independent origins of mitochondrial gene order in birds. Proc. Natl. Acad. Sci. USA 95:10693-10697.

Moritz, C., and W. M. Brown. 1987. Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. Proc. Natl. Acad. Sci. USA 84:7183-7187.

San Mauro, D., D. J. Gower, O. V. Oommen, M. Wilkinson, and R. Zardoya. 2004. Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. Mol. Phylogenet. Evol. 33:413-427.

San Mauro, D., D. J. Gower, R. Zardoya, and M. Wilkinson. 2006. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. Mol. Biol. Evol. 23:227-234.

Wolstenholme, D. R. 1992. Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. 141:173-216.

Wong, T. W., and D. A. Clayton. 1985. In vitro replication of human mitochondrial DNA: accurate initiation at the origin of light-strand synthesis. Cell 42: 951-958.

Zardoya, R., and A. Meyer. 2000. Mitochondrial evidence on the phylogenetic position of Caecilians (Amphibia: Gymnophiona). Genetics 155:765-775.

ONLINE APPENDIX 2

JUSTIFICATION OF THE PARTITIONING SCHEME EMPLOYED IN BAYESIAN AND RAxML ANALYSES

Six alternative partitioning schemes (of 1, 2, 4, 7, 17, and 32 partitions, respectively; see online Appendix 2 Table 1 for a complete description of the partitions) were compared as in recent studies (McGuire et al., 2007; Li et al., 2008). For both maximum likelihood (ML) and Bayesian frameworks, we used the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). The preferred partitioning scheme is that with the minimum observed AIC or BIC. We employed the following equations:

$$\mathrm{AIC}_i = -2 \ln L_i + 2\, k_i$$

$$\mathrm{BIC}_i = -2 \ln L_i + k_i \ln n$$

in which L_i is the likelihood of the model, k_i is the number of parameters in the model i (see online Appendix 2 Table 1), and n is the number of sites (11,867 in our case). The ln L_i is substituted by the harmonic mean log likelihood (HML$_i$) in the Bayesian framework (see McGuire et al., 2007, for a similar approach). When the ratio $n/k_i < 40$, AIC$_\mathrm{c}$ was used instead of AIC to correct for small sample size (Burnham and Anderson, 2002). AIC$_\mathrm{c}$ was calculated as:

$$\mathrm{AIC}_{\mathrm{c}i} = -2 \ln L_i + 2\, k_i + 2\, k_i\, (k_i + 1) / (n - k_i - 1)$$

In the case of the Bayesian framework, standard Bayes factors (BF; Nylander et al., 2004) were also employed, applying Kass and Raftery's (1995) conventions (2 ln BF > 10 are considered to be "very strong" support for a particular partitioning model). BF were calculated as:

$$\mathrm{BF}_i = -\mathrm{HML}_i - (-\mathrm{HML}_{best})$$

Data partition statistics including number parameters of each scheme and model-selection criteria are shown in online Appendix 2 Table 2. Bayes factors are shown in online Appendix 2 Table 3. In the both Bayesian and ML frameworks, all decision criteria preferred the seven-partition scheme (P7), with only one exception (AIC of the ML framework). Thus, this partitioning scheme was employed in our BI and RAxML analyses.

REFERENCES

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle*in* Second international symposium of information theory (B. N. Petrov, and F. Csaki, eds.). Akademiai Kiado, Budapest, Hungary.

Burnham, K., and D. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach, 2nd Edition. Springer-Verlag, New York.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773-795.

Li, C., G. Lu, and G. Orti. 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Syst. Biol. 57:519-539.

McGuire, J. A., C. C. Witt, D. L. Altshuler, and J. V. Remsen Jr. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Syst. Biol. 56:837-856.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47-67.

Schwarz, G. 1978. Estimating the dimensions of a model. Ann. Stat. 6:461-464.

ONLINE APPENDIX 2 – TABLE 1. Alternative data partitions evaluated and substitution models applied to each (RAXML used the GTR+Γ_4 model for any partition). The number of free parameters (*k*) is indicated for each substitution model. Abbreviations are defined as follows: mt , mitochondrial; nu, nuclear; prots, protein-coding genes; pos1, first codon position; pos2, second codon position; pos3, third codon position.

Partition		Bayesian framework		ML framework	
Name	Description	Substitution model	*k*	Substitution model	*k*
P1	All positions	GTR+Γ_4+I	11	GTR+Γ_4	9
		Rate multiplier	0	Rate multiplier	0
Total	1		11		9
P2	mtDNA	GTR+Γ_4+I	11	GTR+Γ_4	9
	nu (RAG1)	GTR+Γ_4	10	GTR+Γ_4	9
		Rate multiplier	2	Rate multiplier	0
Total	2		23		18
P4	mt prots	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt rRNAs	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt tRNAs	GTR+Γ_4	10	GTR+Γ_4	9
	nu RAG1	GTR+Γ_4	10	GTR+Γ_4	9
		Rate multiplier	4	Rate multiplier	0
Total	4		46		36
P7	mt prots pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt prtos pos2	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt rRNAs	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt tRNAs	GTR+Γ_4	10	GTR+Γ_4	9
	nu RAG1pos1	GTR+I	10	GTR+Γ_4	9
	nu RAG1pos2	GTR+I	10	GTR+Γ_4	9
	nu RAG1pos3	GTR+Γ_4	10	GTR+Γ_4	9
		Rate multiplier	7	Rate multiplier	0
Total	7		80		63
P17	mt ATP6	GTR+Γ_4	10	GTR+Γ_4	9
	mt ATP8	HKY+Γ_4	5	GTR+Γ_4	9
	mt COX1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt COX2	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt COX3	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt Cytb	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND2	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND3	GTR+Γ_4	10	GTR+Γ_4	9
	mt ND4L	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND4	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND5	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND6	GTR+Γ_4	10	GTR+Γ_4	9
	mt tRNAs	GTR+Γ_4	10	GTR+Γ_4	9
	mt 12S	GTR+Γ_4	10	GTR+Γ_4	9
	mt 16S	GTR+Γ_4+I	11	GTR+Γ_4	9
	nu RAG1	GTR+Γ_4	10	GTR+Γ_4	9
		Rate multiplier	17	Rate multiplier	0
Total	17		192		153

P32	mt ATP6 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ATP6 pos2	GTR+Γ_4	10	GTR+Γ_4	9
	mt ATP8 pos1	HKY+I	5	GTR+Γ_4	9
	mt ATP8 pos2	F81	3	GTR+Γ_4	9
	mt COX1 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt COX1 pos2	GTR	9	GTR+Γ_4	9
	mt COX2 pos1	GTR+Γ_4	10	GTR+Γ_4	9
	mt COX2 pos2	GTR+I	10	GTR+Γ_4	9
	mt COX3 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt COX3 pos2	HKY+Γ_4+I	6	GTR+Γ_4	9
	mt Cytb pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt Cytb pos2	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND1 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND1 pos2	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND2 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND2 pos2	GTR+I	10	GTR+Γ_4	9
	mt ND3 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND3 pos2	GTR+Γ_4	10	GTR+Γ_4	9
	mt ND4L pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND4L pos2	GTR+Γ_4	10	GTR+Γ_4	9
	mt ND4 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND4 pos2	GTR+Γ_4	10	GTR+Γ_4	9
	mt ND5 pos1	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND5 pos2	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt ND6 pos1	GTR+Γ_4	10	GTR+Γ_4	9
	mt ND6 pos2	GTR+Γ_4	10	GTR+Γ_4	9
	mt 12S	GTR+Γ_4	10	GTR+Γ_4	9
	mt 16S	GTR+Γ_4+I	11	GTR+Γ_4	9
	mt tRNAs	GTR+Γ_4	10	GTR+Γ_4	9
	nu RAG1 pos1	GTR+I	10	GTR+Γ_4	9
	nu RAG1 pos2	GTR+I	10	GTR+Γ_4	9
	nu RAG1 pos3	GTR+Γ_4	10	GTR+Γ_4	9
		Rate multiplier	32	Rate multiplier	0
Total	32		349		288

ONLINE APPENDIX 2 – TABLE 2. Data partition statistics (number of parameters [k_i], likelihood scores [HML$_i$, ln L_i], model-selection criteria [AIC, AIC$_c$, BIC]) for the partitioning schemes evaluated. Values in bold indicate the preferred partitioning scheme according to each model-selection criterion.

Bayesian framework				
Partition	k_i	–HML$_i$	AIC	BIC
P1	11	57,024.78	114,071.56	114,152.76
P2	23	56,964.00	113,974.00	114,143.77
P4	46	56,577.20	113,246.40	113,585.95
P7	80	55,872.29	**111,904.58**	**112,495.10**
P17	192	56,214.68	112,813.36	114,230.61
P32	349	55,873.71	112,466.63*	115,021.57
Maximum likelihood framework				
Partition	k_i	–ln L_i	AIC	BIC
P1	9	57,018.00	114,053.99	114,120.43
P2	18	56,833.37	113,702.74	113,835.61
P4	36	56,423.94	112,919.87	113,185.61
P7	63	55,619.94	111,365.88	**111,830.91**
P17	153	55,975.27	112,256.54	113,385.91
P32	288	55,004.97	**110,585.94**	112,711.81

*AIC$_c$

ONLINE APPENDIX 2 – TABLE 3. Bayes factors (BF) for the partitioning schemes evaluated (values above the diagonal). Values below the diagonal are 2 ln BF, with those in bold being highly significant according to Kass and Raftery (1995) conventions (2 ln BF > 10).

	P1	P2	P4	P7	P17	P32
P1	–	60.78	447.58	1,152.49	810.10	1,151.07
P2	8.21	–	386.80	1,091.71	749.32	1,090.29
P4	**12.21**	**11.92**	–	704.91	362.52	703.49
P7	**14.10**	**13.99**	**13.12**	–	342.39	1.42
P17	**13.39**	**13.24**	**11.79**	**11.67**	–	340.97
P32	**14.10**	**13.99**	**13.11**	0.70	**11.66**	–

ONLINE APPENDIX 3

BEST-FITTING SUBSTITUTION MODELS (AND ASSOCIATED ESTIMATED PARAMETERS) USED FOR ASSESSING PHYLOGENETIC INFORMATION CONTENT OF EACH EMPLOYED DATA SET

Data set name (genes included, number of positions)	**Best-fit model**	**Base frequencies**		**Substitution rate matrix (Ti:Tv ratio for HKY models)**		**Γ_4-shape parameter (α)**	**Prop. invariable sites (I)**
AT6 (*atp6* without the 3rd position; 452 bp)	TrN+Γ_4	A = 0.239 C = 0.296	G = 0.114 T = 0.350	A-C = 1.000 A-G = 2.733 A-T = 1.000	C-G = 1.000 C-T = 4.208 G-T = 1.000	0.296	0
AT8 (*atp8* without the 3rd position; 42 bp)	HKY+Γ_4	A = 0.234 C = 0.231	G = 0.090 T = 0.445	1.642		0.311	0
CO1 (*cox1* without the 3rd position; 1,020 bp)	TrN+Γ_4+I	A = 0.229 C = 0.235	G = 0.215 T = 0.321	A-C = 1.000 A-G = 4.196 A-T = 1.000	C-G = 1.000 C-T = 10.854 G-T = 1.000	1.401	0.747
CO2 (*cox2* without the 3rd position; 450 bp)	TVM+Γ_4+I	A = 0.299 C = 0.229	G = 0.171 T = 0.302	A-C = 1.300 A-G = 6.030 A-T = 1.924	C-G = 0.212 C-T = 6.030 G-T = 1.000	1.950	0.593
CO3 (*cox3* without the 3rd position; 522 bp)	TVM+Γ_4+I	A = 0.216 C = 0.246	G = 0.211 T = 0.327	A-C = 4.595 A-G = 13.714 A-T = 4.600	C-G = 0.673 C-T = 13.714 G-T = 1.000	1.126	0.640
CYB (*cob* without the 3rd position; 752 bp)	GTR+Γ_4+I	A = 0.256 C = 0.250	G = 0.167 T = 0.327	A-C = 2.698 A-G = 3.428 A-T = 1.436	C-G = 0.566 C-T = 8.745 G-T = 1.000	0.716	0.579
ND1 (*nad1* without the 3rd position; 612 bp)	GTR+Γ_4+I	A = 0.232 C = 0.281	G = 0.169 T = 0.318	A-C = 0.402 A-G = 2.124 A-T = 0.854	C-G = 0.263 C-T = 3.226 G-T = 1.000	0.630	0.487
ND2 (*nad2* without the 3rd position; 678 bp)	TVM+Γ_4+I	A = 0.279 C = 0.280	G = 0.115 T = 0.326	A-C = 1.644 A-G = 4.346 A-T = 1.411	C-G = 0.453 C-T = 4.346 G-T = 1.000	0.885	0.313
ND3 (*nad3* without the 3rd position; 216 bp)	TVM+Γ_4	A = 0.206 C = 0.295	G = 0.156 T = 0.344	A-C = 31.609 A-G = 95.307 A-T = 28.624	C-G = 6.870 C-T = 95.307 G-T = 1.000	0.279	0
ND4 (*nad4* without the 3rd position; 900 bp)	GTR+Γ_4+I	A = 0.275 C = 0.264	G = 0.135 T = 0.327	A-C = 1.779 A-G = 3.678 A-T = 1.560	C-G = 0.470 C-T = 6.716 G-T = 1.000	1.012	0.413
ND4L (*nad4L* without the 3rd position; 184 bp)	GTR+Γ_4+I	A = 0.222 C = 0.269	G = 0.169 T = 0.340	A-C = 1.111 A-G = 2.608 A-T = 2.360	C-G = 0.614 C-T = 5.151 G-T = 1.000	2.695	0.461
ND5 (*nad5* without the 3rd position; 1,098 bp)	GTR+Γ_4+I	A = 0.289 C = 0.246	G = 0.155 T = 0.311	A-C = 3.100 A-G = 4.051 A-T = 2.170	C-G = 0.706 C-T = 9.505 G-T = 1.000	1.071	0.429
ND6 (*nad6* without the 3rd position; 286 bp)	GTR+Γ_4	A = 0.134 C = 0.142	G = 0.280 T = 0.444	A-C = 0.820 A-G = 6.373 A-T = 2.806	C-G = 1.172 C-T = 2.365 G-T = 1.000	0.538	0
PROTS-NO3 (mt protein-coding genes without the 3rd position; 7,212 bp)	GTR+Γ_4+I	A = 0.249 C = 0.253	G = 0.170 T = 0.328	A-C = 1.889 A-G = 3.594 A-T = 1.800	C-G = 0.493 C-T = 6.515 G-T = 1.000	0.914	0.487
PROTS-ALL (mt protein-coding genes - all positions; 10,818 bp)	GTR+Γ_4+I	A = 0.338 C = 0.242	G = 0.125 T = 0.295	A-C = 1.501 A-G = 3.105 A-T = 2.150	C-G = 0.306 C-T = 11.817 G-T = 1.000	1.022	0.355
3rdPOS (3rd positions of mt protein-coding genes; 3,606 bp)	GTR+Γ_4+I	A = 0.469 C = 0.210	G = 0.059 T = 0.262	A-C = 0.000 A-G = 8.826 A-T = 0.060	C-G = 0.000 C-T = 11.456 G-T = 1.000	0.352	0.007
12S (mt *rrnS*; 699 bp)	GTR+Γ_4	A = 0.327 C = 0.245	G = 0.198 T = 0.231	A-C = 7.324 A-G = 19.659 A-T = 12.774	C-G = 0.000 C-T = 37.794 G-T = 1.000	0.510	0
16S (mt *rrnL*; 1,169 bp)	GTR+Γ_4+I	A = 0.379 C = 0.207	G = 0.175 T = 0.240	A-C = 3.878 A-G = 5.883 A-T = 5.736	C-G = 0.000 C-T = 19.712 G-T = 1.000	0.883	0.306
tRNAs (all mt tRNA genes except *trnF*; 1,278 bp)	TVM+Γ_4	A = 0.317 C = 0.183	G = 0.182 T = 0.318	A-C = 1.673 A-G = 10.638 A-T = 2.136	C-G = 0.000 C-T = 10.638 G-T = 1.000	0.800	0
mtGENOME-NO3 (all single mt data sets combined, excluding 3rd positions; 10,358bp)	GTR+Γ_4+I	A = 0.278 C = 0.237	G = 0.174 T = 0.311	A-C = 2.182 A-G = 5.165 A-T = 2.333	C-G = 0.366 C-T = 8.326 G-T = 1.000	1.031	0.422
RAG1 (nuclear *rag1*; 1,509 bp)	GTR+Γ_4	A = 0.311 C = 0.199	G = 0.227 T = 0.263	A-C = 1.748 A-G = 6.635 A-T = 1.476	C-G = 0.970 C-T = 9.758 G-T = 1.000	0.457	0