Syst. Biol. 58(3):298–311, 2009

DOI:10.1093/sysbio/syp027
Advance Access publication on July 1, 2009

Accelerated Species Inventory on Madagascar Using Coalescent-Based Models of Species Delineation

MICHAEL T. MONAGHAN[1,2,3,*], RUTH WILD[1,2], MIRANDA ELLIOT[1,2], TOMOCHIKA FUJISAWA[2], MICHAEL BALKE[1,4], DAEGAN J.G. INWARD[1], DAVID C. LEES[1], RAVO RANAIVOSOLO[5], PAUL EGGLETON[1], TIMOTHY G. BARRACLOUGH[2], AND ALFRIED P. VOGLER[1,2]

[1]*Entomology Department, Natural History Museum, London SW7 5BD, UK;*
[2]*Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK;*
[3]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), 12587 Berlin, Germany;*
[4]*Zoologische Staatssammlung, Muenchhausenstrasse 21, 81247 Munich, Germany; and*
[5]*Department of Animal Biology, University of Antananarivo, 101 Antananarivo, Madagascar;*
**Correspondence to be sent to: Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), 12587 Berlin, Germany; E-mail: monaghan@igb-berlin.de.*

Abstract.—High-throughput DNA sequencing has the potential to accelerate species discovery if it is able to recognize evolutionary entities from sequence data that are comparable to species. The general mixed Yule-coalescent (GMYC) model estimates the species boundary from DNA surveys by identifying independently evolving lineages as a transition from coalescent to speciation branching patterns on a phylogenetic tree. Applied here to 12 families from 4 orders of insects in Madagascar, we used the model to delineate 370 putative species from mitochondrial DNA sequence variation among 1614 individuals. These were compared with data from the nuclear genome and morphological identification and found to be highly congruent (98% and 94%). We developed a modified GMYC that allows for a variable transition from coalescent to speciation among lineages. This revised model increased the congruence with morphology (97%), suggesting that a variable threshold better reflects the clustering of sequence data into biological species. Local endemism was pronounced in all 5 insect groups. Most species (60–91%) and haplotypes (88–99%) were found at only 1 of the 5 study sites (40–1000 km apart). This pronounced endemism resulted in a 37% increase in species numbers using diagnostic nucleotides in a population aggregation analysis. Sample sizes between 7 and 10 individuals represented a threshold above which there was minimal increase in genetic diversity, broadly agreeing with coalescent theory and other empirical studies. Our results from >1.4 Mb of empirical data suggest that the GMYC model captures species boundaries comparable to those from traditional methods without the need for prior hypotheses of population coherence. This provides a method of species discovery and biodiversity assessment using single-locus data from mixed or environmental samples while building a globally available taxonomic database for future identifications. [Biodiversity; coalescent; DNA barcoding; DNA taxonomy; endemism; GMYC; Madagascar; turnover.]

A large proportion of biological diversity remains undescribed within the present system of species classification (Brown and Lomolino 1998; Wilson 2003). This "taxonomic impediment" (Hoagland 1996) greatly hinders progress in the increasingly important task of characterizing the origin, functional role, and fate of species on Earth. Species-rich groups, such as tropical insects, present a particular challenge to taxonomy, but their overwhelming contribution to eukaryotic diversity and ecological function (Groombridge 1992; Godfray et al. 1999; Novotny et al. 2006) demands accelerated methods of species discovery and identification. DNA-based procedures may overcome some of these problems by providing a readily assessed character system (Blaxter and Floyd 2003; Tautz et al. 2003; DeSalle et al. 2005). However, the value of DNA sequences as a tool in taxonomy requires that genetic variation in nature delineates discrete groups that correspond principally to species-level taxa. If confirmed, sequence-based species delimitation could allow rapid biodiversity assessment in critical geographical areas or poorly known taxa (Janzen 2004; Markmann and Tautz 2005; Smith et al. 2005).

Attempts to apply DNA to biodiversity surveys and taxonomy have met with 3 problems. First is the widely held concern that DNA sequence data, in particular short mitochondrial DNA (mtDNA) fragments, are not suited to capture the species category in the way that complex morphological characters can (Moritz and Cicero 2004; Prendini 2005; Will et al. 2005). But regardless of the genetic marker employed, we expect any locus to diverge between species over time and generate a pattern of identifiable clusters where reproductively isolated and independently evolving species exist (Mallet 1995; Barraclough et al. 2003; Fontaneto et al. 2007; Papadapoulou et al. 2008). A second concern is the potentially misleading effect of incomplete lineage sorting or introgression. The process of lineage sorting is stochastic and divergence at a single marker can lag well behind the divergence of lineages into species (Hudson and Coyne 2002). MtDNA introgression across species boundaries has been reported (Funk and Omland 2003), but the frequency of this problem is not well known and may be inflated because of a research focus on recent or problematic species pairs. Taken together, 1 approach is to treat single locus data as a hypothesis to be tested with other data, such as additional genetic loci, geographical distribution, or morphological characters (e.g., DeSalle et al. 2005; Knowles and Carstens 2007; Shaffer and Thompson 2007). A third and potentially far more important problem with applying DNA to taxonomy is that genetic cluster concepts, specifically the quantitative boundaries between them, are difficult to implement where the total variation within and

among the lineages of interest is incompletely known. In practice, these have often used threshold values of genetic divergence for species discrimination, notably the controversial 10× rule in animals (see Hebert and Gregory 2005) and the "97% rule" in prokaryotes (Forney et al. 2004). Not surprisingly, the application of global p-distance criteria against established taxonomic names (i.e., DNA barcoding) has shown mixed success (Hebert et al. 2004; Moritz and Cicero 2004; Meyer and Paulay 2005; Meier et al. 2006). This is partly because of the necessary assumption that extant species within a clade are of a similar age and that the rate of evolutionary change within the gene region of choice is uniform. Other methods (e.g., DeSalle et al. 2005; Rach et al. 2008) require initial assignments of minimum groupings based on morphology, ecology, or geographical location. Such groupings may not be readily available for unknown, mixed, or bulk environmental samples, which nonetheless need methods of determining meaningful groups (e.g., Fraser et al. 2009).

If species are to be delimited from DNA sequences alone, the recognition of a "species boundary" from observed DNA sequence variation constitutes the critical step. The general mixed Yule-coalescent (GMYC) model (Pons et al. 2006; Fontaneto et al. 2007) attempts to identify this boundary as a shift in branching rates on a tree that contains multiple species and populations. Branching patterns within the genetic clusters reflect neutral coalescent processes occurring within species (Kingman 1982), whereas branching among genetic clusters reflects the timing of speciation events (Yule 1924). The GMYC exploits the predicted difference in branching rate under the 2 modes of lineage evolution, assessing the point of highest likelihood of the transition (Pons et al. 2006; Fontaneto et al. 2007). Independently evolving lineages are recovered as putative species, more distinct than predicted if the entire sample derived from a single species without genetic isolation. The GMYC procedure provides a potential means of detecting species from single-locus sequence data, but its use in taxonomy requires tests of congruence with additional genetic loci as well as traditionally recognized taxonomic species.

The GMYC procedure calculates a single threshold value for each input tree and has been applied successfully to selected clades (Pons et al. 2006; Ahrens et al. 2007; Fontaneto et al. 2007). However, as larger data sets encompass a greater number of lineages, a single threshold approach may not reflect the variety of divergence levels among taxa. In an attempt to overcome this, here we develop a multiple threshold method that allows the depth of the coalescent–speciation transition to vary along individual branches of a phylogenetic tree.

The insects of Madagascar exemplify the challenge faced by biologists studying a highly diverse, poorly described, and critically endangered fauna (Paulian and Viette 2003; Hanski et al. 2007). Sequence-based species identification has been attempted for a few Malagasy groups (Smith et al. 2005; Vences et al. 2005) and Madagascar is well known for the prevalence of endemic lineages of closely related species (e.g., Lees and Minet 2003; Yoder et al. 2003; Monaghan, Gattolliat et al. 2005; Orsini et al. 2007). Uncertainty about the nature of species-level entities may be problematic within these lineages (Hey et al. 2003), but it is precisely these groups that are most urgently in need of study (e.g., Monaghan et al. 2006).

Here, we apply the GMYC analysis to 5 taxa (12 families from 4 orders) of paleopteran and holometabolous insects newly sampled in Madagascar. These groups fulfil a wide range of ecological roles (predators, herbivores, detritivores, decomposers), occupy both terrestrial and aquatic habitats, and presumably have differing dispersal abilities. We delineated putative species using both the existing single-threshold approach and a newly developed multiple-threshold model, where the speciation–coalescent transition was allowed to vary across the phylogenetic tree. We then tested the validity of the GMYC approach to taxonomy by comparing these with nuclear (ribosomal RNA [rRNA]) markers and with traditionally recognized morphological species. The latter were identified where known or, in the absence of formal species descriptions, were approximated by separating divergent morphospecies as recognized by taxonomic experts of each group. As an alternative, we explored a character-based diagnosis for the same data using population aggregation analysis (PAA) of nucleotide data (Davis and Nixon 1992; Wiens and Penkrot 2002; DeSalle et al. 2005). We measured haplotype and species turnover among sites to evaluate how well our sampling effort captured genetic diversity. Our results indicate that the GMYC model captures species boundaries that are comparable to those from traditional methods and demonstrate the potential of single-locus, DNA-based procedures to accelerate species discovery and biodiversity assessments, even in highly endemic and species-rich groups of tropical insects.

Materials and Methods

Sampling

Insect communities were studied in 5 areas of northern and eastern Madagascar, located among national parks (NPs) and special reserves (SRs) and in 1 unprotected area: Montagne d'Ambre NP and adjacent Foret d'Ambre SR; Ankarana SR; Antsaba-Galoko (unprotected); Mantadia-Andasibe NP and adjacent Analamazaotra SR; and Ranomafana NP (Supplementary Fig. S1; all supplemental figures are available online at http://www.sysbio.oxfordjournals.org). Over a 5-d period at each site, we collected mayflies (Ephemeroptera: Baetidae and Oligoneuridae), termites (Blattodea: Termitoidae), butterflies (Lepidoptera: Hesperiidae, Papilionoidea, Pieridae, Nymphalidae, Lycaenidae, and Riodinidae), dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae), and water beetles (Coleoptera: Dytiscidae and Hydrophilidae). We used hand nets, modified Surber samplers and kitchen sieves (mayflies, water beetles), fruit traps and hand nets (butterflies), and flight intercept, dung, and carrion traps (dung beetles).

Dead wood, soil, and termite mounds were inspected using trowels and machetes (termites). Samples were preserved in 100% ethanol in the field, returned to the laboratory and sorted to morphospecies by the authors and by experienced taxonomists under a dissecting microscope (10×). Where possible, 5 individuals from each morphospecies and each locality were selected for DNA analysis as a representative sample of the genetic variation. For a subset of species, more individuals were analysed in order to examine the effects of sample size on estimates of genetic diversity. A list of individuals from which DNA was extracted is provided as supplementary information (Supplementary Table S1). Specimens are deposited in the research collection of the Entomology Department, Natural History Museum, London.

DNA Extraction, Sequencing, and Multiple Alignment

We performed nondestructive DNA extraction to facilitate taxonomic identification after molecular analysis. DNA was extracted using Wizard SV extraction kits (Promega, UK). Whole specimens were soaked overnight in extraction buffer with proteinase K at 55 °C. Specimens were then removed from the buffer, washed in 70% ethanol, and either dried and mounted (beetles, butterflies) or placed in 70% ethanol (mayflies, termites) for long-term storage. Larger beetle specimens were split between the head and the thorax to expose tissue and re-glued after drying. Butterfly wings were removed prior to extraction. Gene regions from 1 nuclear rRNA marker and 1 mitochondrial protein-coding marker were amplified using standard primers for each taxonomic group (Supplementary Table S2). Because the GMYC uses a coalescent approach, we used the most polymorphic mtDNA marker (Morando et al. 2003) that could be readily amplified for each taxonomic group based on previous experience in our laboratory. Nuclear rRNA markers were chosen based on previous examination of a subset of dung beetles and water beetles (Monaghan, Balke et al. 2005), where unique 28S rRNA genotypes corresponded to single mtDNA clusters.

Polymerase chain reaction products were cleaned using MultiScreen HTS 96-well plates (Millipore, Watford, UK) and sequenced using a BigDye version 1.1 (Warrington, UK) reaction and electrophoresis on an ABI-3730. Forward and reverse sequencing reads were assembled and edited using the STARS platform (Chan and Ventress 2001) with phred/phrap base calling and assembly.

STARS produced consensus sequences when both forward and reverse reads have a phred score ≥ 40 (corresponding to an error probability of <0.0001). Single sequencing reads, available when the reverse complement failed a phred score, were assessed manually using Sequencher version 4.6 (Gene Codes Corporation, Ann Arbor, MI). In many cases, low-quality sequence at 5′ or 3′ ends could be trimmed and the score of a failed sequence could be raised to the point that assembly and editing were possible as normal. All sequences were deposited to GenBank and accession numbers are provided in Supplementary Table S1. Matrices have been submitted to TreeBASE, submission number SN4377.

Each nuclear and mitochondrial data set was subjected to multiple alignment using ClustalW (http://align.genome.jp) with an IUB weight matrix and default gap penalties (open = 15, extension = 6.66). Ingroup length variation in rRNA (Table 1) was mostly limited to 1- and 2-bp indels. In mtDNA, only *cox2* was length variable (Table 1) due to the presence of 6- or 9-bp indels. These were unambiguously aligned by eye by minimizing amino acid changes and avoiding stop codons. The *cox1* and *cob* markers were not length variable in any samples. Gaps in the termite matrix were coded as a fifth character state in parsimony analysis. Tree length and the number of parsimony-informative sites (S_i) were determined with maximum parsimony searches, conducted using TNT (Goloboff et al. 2004) with 10 ratchet iterations, 10 cycles of tree drifting, and 3 rounds of tree fusing for each of 200 random addition sequence starting trees.

Branch Length Estimation and the Single Threshold GMYC Model

The GMYC approach (Pons et al. 2006; Fontaneto et al. 2007) uses branching rates that are estimated from evo-

TABLE 1. Ingroup nuclear and mitochondrial sequence variation

Taxon	28S rRNA							mtDNA						
	n	bp	K_n	S	S_i	$\%S_i$	L	n	bp	K_{mt}	S	S_i	$\%S_i$	L
Mayflies	90	624 (616–621)	34	134	114	18	282	274[a]	357	109	207	204	57	2159
Termites	80	695 (665–689)	14	68	67	9	100	166[b]	697	74	346	318	46	1122
Butterflies	137	708 (632–667)	44	159	140	20	379	390[a]	357	206	208	200	56	2953
Dung beetles	162	654 (643–649)	34	64	61	9	123	133[c]	801	106	344	325	41	2111
Water beetles	492	745 (600–656)	88	299	258	35	1041	521[c]	658	347	310	294	45	5921

Notes: Genetic variation was measured at nuclear (28S rRNA) and mitochondrial (*cob*, *cox2*, or *cox1*) gene regions for all ingroup specimens of the 5 target groups; Gaps were scored as a fifth character state for these calculations. Abbreviations: n = number of sequences; bp = aligned matrix length including outgroup (minimum and maximum ingroup sequence length given for rRNA); K = haplotypes; S = polymorphic sites; S_i = parsimony-informative sites; % S_i = percentage of sites that were parsimony informative; L = maximum parsimony tree length for ingroup data sets.

[a] *cob*.

[b] *cox2* (length variation 598–697 bp).

[c] *cox1*.

lutionary models and so is potentially influenced by the methods used, particularly the assumptions required to make evolution clocklike. Thus, we compared 3 methods of estimating branch lengths: 2 versions of a relaxed lognormal clock (Drummond et al. 2006) and a simple clock-constrained maximum likelihood search (see below). The 2 relaxed lognormal clock analyses were conducted with Bayesian analysis as implemented in BEAST version 1.4.7 (Drummond and Rambaut 2007). Branch lengths were estimated once using a coalescent prior and once using a Yule prior. Because the GMYC uses a coalescent as the null model to explain branching patterns (Pons et al. 2006), we viewed the use of a coalescent prior in BEAST as the more conservative option, that is, more likely to not recognize a coalescent–speciation transition. The third analysis was a computationally simpler approach whereby branch lengths were estimated under maximum likelihood using a GTR model and uniform molecular clock in PAUP* version 4.0b10 (Swofford 2002). This model was chosen for comparison with more parameter-rich models under the relaxed clock approach (see below). Each analysis used a UPGMA starting tree and all model parameters were estimated from the data.

Duplicate haplotypes were removed from the matrix using Collapse 1.2 (Posada 2004). Extensive preliminary analyses showed that a GTR + I + Γ model with unlinked codon positions was preferred in the relaxed clock analyses. MCMC chains were run for 20–30 M generations and only terminated after all parameters of the evolution model reached an estimated sample size (ESS) >100 after burnin. In some cases, ESS was improved by increasing the burnin from the default of 10% to values of 50–60%. After termination, the MCMC output was analyzed using TreeAnnotater version 1.4.7 (http://beast.bio.ed.ac.uk). We used all trees after the burnin, a posterior probability limit of 0.5, targeting the maximum clade credibility tree, and keeping the target node heights. For the uniform clock, all collapsed nodes from the maximum likelihood tree were resolved with zero-length branches arbitrarily in TreeEdit (Rambaut and Charleston 2002). Lineage-through-time plots were drawn using the *ltt.plot* command from the ape library (Paradis et al. 2004) in R (R Core Development Team 2005).

The single-threshold GMYC model was optimized onto the output trees from all 3 branch length estimates for each of the 5 mtDNA data sets. Optimization was performed using a newly created script, GMYC, available as a tool within the SPLITS package for R (available from http://r-forge.r-project.org/projects/splits/). For the likelihood branch lengths in PAUP, in addition to uniform-clock constraints, we explored the effect of normalizing branch lengths with penalized likelihood or nonparametric rate smoothing (r8s; Sanderson 2003). Although these 2 methods produced only slightly different patterns of branching, differences were toward the base of the tree, and there was no effect on the GMYC model fit or the number of entities (data not shown).

Multiple-Threshold GMYC Model

We extended the single-threshold optimization, described above, to a multiple-threshold method. Starting with a set of most recent common ancestor (MRCA) nodes given by an arbitrary threshold time, the multiple-threshold implementation obtains new possible sets of MRCAs by dividing 1 cluster into 2 (Fig. 1a) or by fusing 2 clusters to 1 (Fig. 1b). From among this set, the MRCAs with the maximum improvement in likelihood are chosen as a new starting set. This process is continued until no likelihood value is improved by the renewal. The number of parameters of the multiple-threshold model depends on how many separate thresholds are needed to specify the final set of MRCAs. Scaling parameters (p) and branching rates (λ) are assumed to be identical among clusters within each threshold group in order

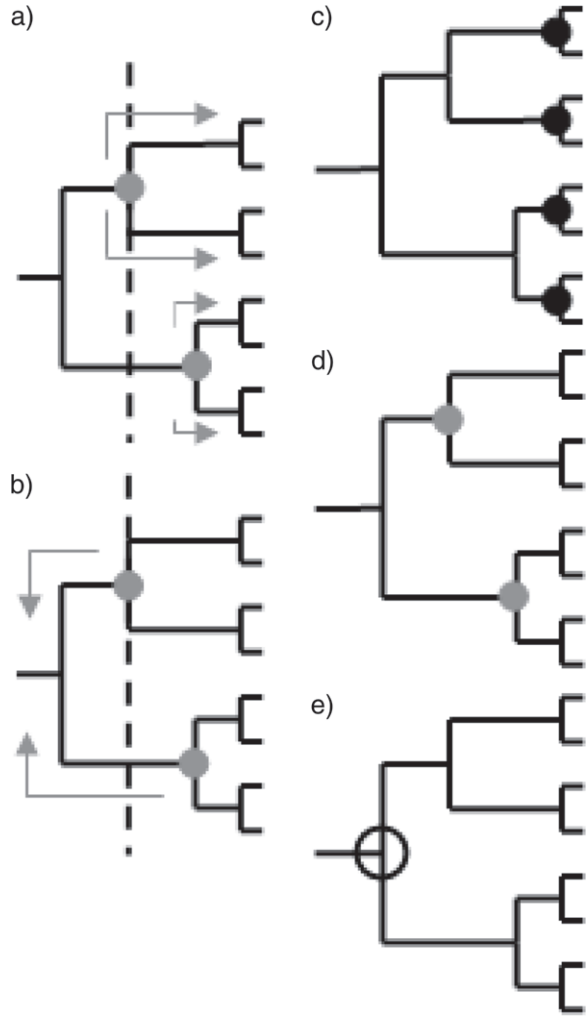

FIGURE 1. Illustrative view of optimization using the multiple threshold GMYC aproach. A single, median-age threshold is established for the whole tree in a first iteration and then undergoes iterative division (a) and fusion (b) of MRCA nodes. Threshold nodes can then be shifted toward the tips (c), remain the same (d), or be shifted toward the base (e) depending on which yields the highest likelihood value.

to prevent a large increase in the number of parameters. Degrees of freedom for comparing the single- and multiple-threshold models are determined by the number of additional thresholds, with each additional threshold increasing the degrees of freedom by 3 (1 for the threshold age, 1 for the scaling parameter, and 1 for the branching rate associated with that threshold group).

The starting point of this process can be selected arbitrarily; however, the optimization process is often attracted to local optima if only 1 starting point is used (data not shown). Therefore, the following recursive approach was used to consider multiple starting points for the fusing/dividing algorithm. First, the age of the node with the median age is chosen as the first starting threshold. The fusing–dividing optimization algorithm is then performed, storing the maximum likelihood solution found. Next, 2 new starting thresholds are chosen as the median age of the nodes older and younger than the initial node. The fusing–dividing algorithm is repeated for each of these. If the maximum likelihood solution starting at either the older or younger node is not greater than the first solution, then the first solution is kept as the global maximum likelihood solution. In contrast, if new thresholds lead to an improvement in the likelihood score, then the process is repeated recursively. The process is repeated until no improvement of likelihood is obtained. A function implementing this approach is provided in the GMYC package.

Character-Based Delineation of Subgroups within Clusters

PAA (Davis and Nixon 1992) delineates groups under the criterion of having 1 or more diagnostic traits (i.e., nucleotides) that are absent from other groups. The procedure requires the delineation of a priori populations in order to test them for the presence of diagnostic nucleotide changes. Here, we used entities from the GMYC model analysis as prior hypotheses of species and sought to establish subgroups within these entities that could be delineated by geography (sensu DeSalle et al. 2005). Such a character-based group delineation requires at least 4 individuals, 2 of which are found in the same study site and are therefore a diagnosable subgroup and 2 of which are found in 1 or more different sites. The analysis could be applied to 53 putative GMYC species (see results).

Nuclear rRNA and Morphological Species

Membership of individuals within GMYC clusters was evaluated for conflict with nuclear rRNA genotypes and morphological groups. Because mtDNA is expected to coalesce at up to 4 times the rate of nuclear DNA, distinct GMYC clusters that shared a single rRNA genotype were not considered to be in conflict. Where a single mtDNA cluster subsumed 2 or more nuclear clusters or where mitochondrial and nuclear markers placed an individual in alternative groups, these were both interpreted as cases of conflict. All sequenced individuals were examined morphologically using standard diagnostic characters for each taxonomic group. Species were named where descriptions were available or otherwise sorted into morphological groups by workers experienced with the taxonomy and routine identification of each group in question (Supplementary Table S1).

Genetic and Species Diversity

In order to assess the amount of genetic variation captured by our sample sizes, we estimated genetic diversity within each species for which multiple sequences were available. We measured nucleotide polymorphism as the average proportion of nucleotide differences per site (π) and with the Watterson estimator of the mutation parameter calculated with silent sites (θ_W). Under neutral evolution (i.e., mutation-drift equilibrium), θ and π should estimate the same parameter and we tested for significant departures from neutrality using Tajima's D (Tajima 1989). There are a number of tests for neutrality, with varying levels of power depending on the modes of departure (Depaulis et al. 2003), but owing to small sample size in most species here and the problem of significance of multiple tests, we limited our analysis to D. We calculated D and values of θ_W and its variance assuming no recombination for mtDNA. These values were calculated from the aligned mtDNA matrices using DNAsp version 4.0 (Rozas et al. 2003). Species diversity was tabulated from GMYC estimates from the multiple-threshold model analysis (see results). Richness estimates also were calculated using a Chao (1984) estimator implemented in EstimateS (Colwell 2005). The value derives from an extrapolation of diversity based on the frequency of encountering taxa with only 1 or 2 representatives.

RESULTS

Species Delineation with the GMYC Model

We analyzed 1656 individuals for a total of 2444 sequences (aligned bp: 684 046 rRNA; 802 759 mtDNA). This produced 214 rRNA genotypes and 860 mtDNA haplotypes (Table 1). Nucleotide polymorphism ($\%S_i$) in 28S rRNA ranged from 9% to 35%, whereas the same values ranged from 45% to 57% for mtDNA (Table 1). Mitochondrial polymorphism was higher for mayflies and butterflies (*cob* locus) than for termites (*cox2*), water beetles, and dung beetles (*cox1*; Table 1). The number of parsimony-informative sites was 1.1–5.3 times greater for mtDNA than rRNA, and the resulting phylogenetic trees (Supplementary Figs. S2–S6) had 6–17 times the number of steps under maximum parsimony (Table 1). Tree length was positively correlated with the number of haplotypes (Pearson $R = 0.97$) but was independent of gene region.

The GMYC approach was applied to trees reconstructed with a relaxed lognormal coalescent, a relaxed lognormal Yule, and a uniform clock. Bayes factors indicated that a coalescent prior was a better fit than a Yule prior for the relaxed lognormal clock estimates in each of the 5 data sets (Supplementary Table S3).

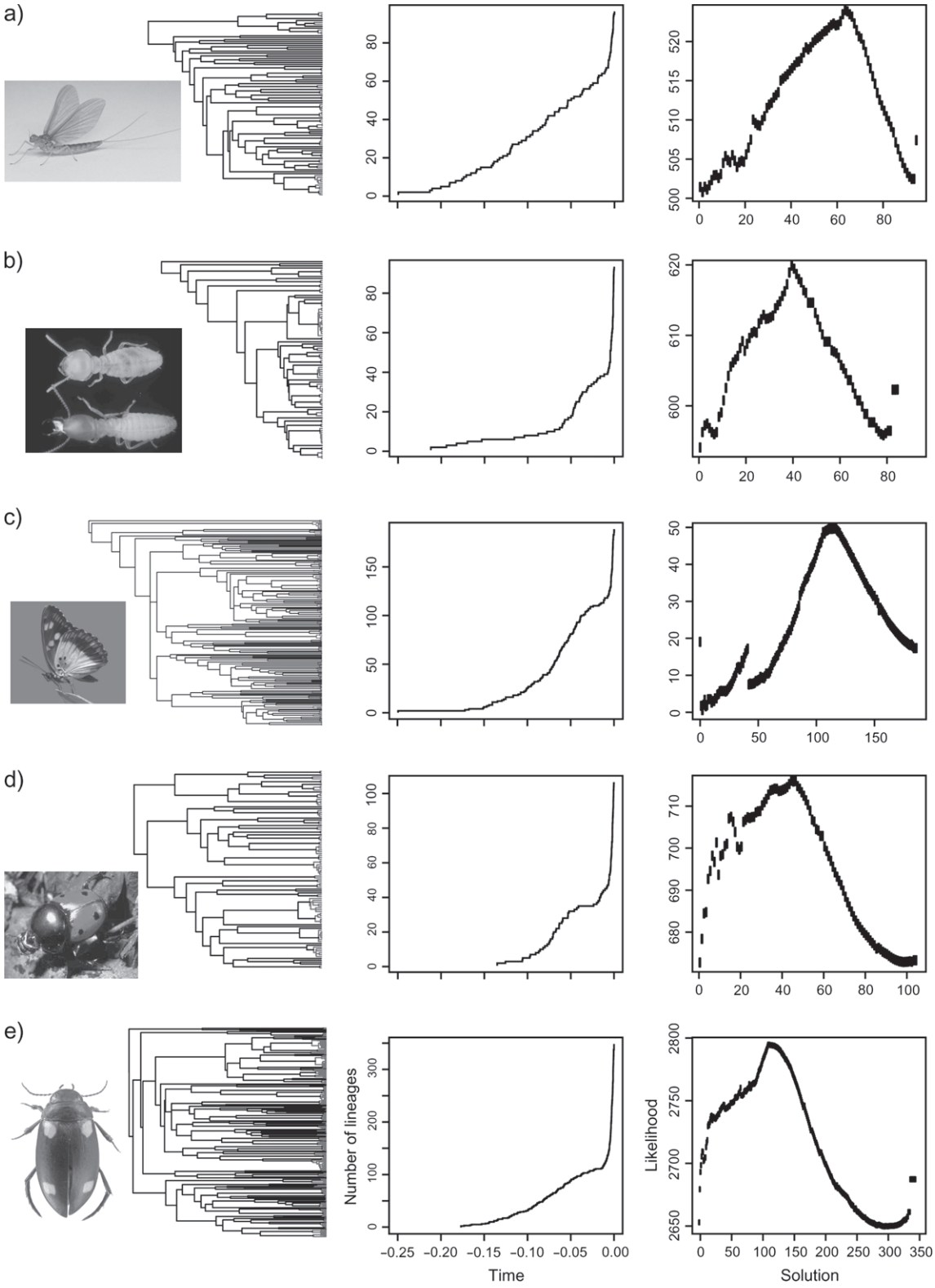

FIGURE 2. Fit of the GMYC model to mtDNA data from the 5 taxonomic groups. Phylogenetic trees (left, here with branch lengths from a uniform GTR model, see text), lineage-through-time plots (center), and single-threshold likelihood solutions to the GMYC model (right) for (a) mayflies, (b) termites, (c) butterflies, (d) dung beetles, and (e) water beetles. The shift in banching rate was modeled as a speciation–coalescent transition (see text, Table 2).

TABLE 2. Lineage branching patterns fit to single- and multiple-threshold variants of the GMYC model

	Model	T	N_{GMYC}	(CI)	λ_1	λ_2	p_1	p_2	L_0	L_{GMYC}	LR
Mayflies	Single	0.0187	64	(57–69)	226.1211	0.8441	1.3246	0.0029	477.054	510.758	67.408*
	Multiple	—	57	(57–59)	23.9580	0.9384	1.0573	1.2640		509.513	64.917*
Termites	Single	0.0112	23	(23–26)	337.1556	1.3755	1.5935	0.4089	445.062	460.120	30.118*
	Multiple	—	24	(24–27)	469.1736	1.3686	0.2832	1.5942		460.344	30.565*
Butterflies	Single	0.0198	129	(127–134)	168.4488	16.9121	0.0692	0.2394	1109.264	1173.555	128.583*
	Multiple	—	125	(125–127)	50.7482	20.8548	0.8722	0.6330		1172.018	125.509*
Dung beetles	Single	0.0506	42	(30–43)	36.5441	0.3909	1.6243	0.7743	527.952	538.922	21.939*
	Multiple	—	43	(40–48)	40.8988	0.4356	0.7056	1.5919		539.358	22.812*
Water beetles	Single	0.0188	112	(108–116)	158.0237	18.0475	0.7418	0.5751	2526.476	2630.186	207.442*
	Multiple	—	112	(112–114)	125.2721	28.2144	0.6321	0.6235		2631.118	209.425*

Notes: All parameters are based on analysis of relaxed lognormal branch lengths using a coalescent prior (see text). Model outputs include the threshold genetic distance from the branch tips where transition occurred (T, presented for single-threshold models), the number of putative species as the sum of sequence clusters and singletons (N_{GMYC}), and confidence intervals (CI) as solutions within 2 log-likelihood units of the maximum likelihood. Parameters fit to the data during optimization include branching rates (λ_1 and λ_2) and scaling parameters (p_1 and p_2) between and within sequence clusters, respectively. Likelihoods are presented for null (L_0) and GMYC (L_{GMYC}) models, where null likelihoods are the same for single and multiple threshold model comparisons. Significance of the likelihood ratio (LR) was evaluated using a chi-square test with 3 degrees of freedom to compare GMYC and null models.
*$p < 0.001$.

Nonetheless, we analyzed both results, along with the third branch length estimate (uniform clock-constrained maximum likelihood) in order to examine the sensitivity of the model to these estimates. In each case, lineage-through-time plots exhibited a pronounced increase in branching rates at the tips of the linearized trees (Fig. 2). The mixed model quantified this transition with a higher likelihood than the null model of uniform (coalescent) branching rates in all 5 taxonomic groups (Table 2, see also Supplementary Table S4).

Using the single-threshold GMYC, the depth (T) from the branch tips at which the speciation–coalescent transition occurred ranged from 0.051 to 0.011 substitutions per site with the relaxed lognormal coalescent prior (Table 2). This resulted in 370 putative species (hereafter referred to as GMYC species), 178 as distinct clusters of haplotypes ($K_{mt} > 1$) and 192 as singletons. The total ranged from 345 to 388 based on model solutions at 2 log-likelihood units from the maximum (Table 2), equivalent to 95% confidence intervals (Edwards 1972). The number of species delineated using clock-constrained maximum likelihood was very similar to the number using the relaxed lognormal coalescent prior (Fig. 3). The relaxed lognormal Yule prior resulted in a greater number of species than the other methods except in the water beetles (Fig. 3); however, we noted already that the Yule model was an inferior fit to the data matrix (Supplementary Table S1) and do not consider it further.

The multiple-threshold model produced a total of 15 changes to the depth of the Yule coalescent transition across the 5 data sets when compared with the single-threshold analysis. Eleven of these resulted in a new threshold that was shifted toward the base, whereas 4 were shifted toward the tips (Supplementary Figs. S2–S6). The resulting changes to the extent of GMYC groups led to the delineation of 358 GMYC species. All changes fell within the confidence interval of the single-threshold model with the exception of the butterflies (Table 2). Changes to the extent of groups are further presented in a comparison with morphological groups below.

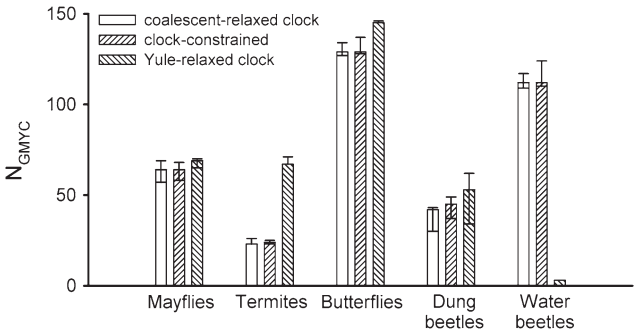

FIGURE 3. Putative species estimates obtained using 3 different branch length estimators. The number of species is reported for the single-threshold GMYC model using branch lengths calculated 3 different ways: with a relaxed lognormal clock using a coalescent (constant size) tree prior, a GTR uniform clock, and a relaxed lognormal clock using a Yule (speciation) tree prior.

Character-Based Delineation and Distribution of Haplotypes

Character-based species delineation (Wiens and Penkrot 2002; Sites and Marshall 2003) might recover further diagnosable groupings under a phylogenetic species concept if fixed nucleotide differences are encountered that are unique to a geographic location. From the 52 GMYC groups that met the minimum requirement for applicability of the test, 19 additional groups could be diagnosed by nucleotide polymorphisms that were exclusive to all individuals in a locality (Table 3). This increase in the number of diagnosable groups (approximately 37%) indicates a high degree of geographical isolation of haplotypes. Throughout the entire data set, only 39 haplotypes were found in more than 1 study site (Table 3). This was fairly consistent among mayflies, termites, dung beetles, and water

TABLE 3. Malagasy endemism and phylogenetic species within GMYC species

	N_{PAA}		K_{mt} shared	P_{end}	P_{Mad}	γ_{Mad}	Chao (1984) estimate (CI)	
Mayflies	1	(1)	4	0.85	0.80	71[a]	92	(74–143)
Termites	4	(8)	3	0.60	0.44	54[b]	32	(26–62)
Butterflies	5	(15)	24	0.75	0.39	317[c]	166	(143–212)
Dung beetles	2	(4)	1	0.91	0.20	215[d]	52	(47–69)
Water beetles	7	(24)	7	0.82	0.39	309[e]	135	(121–166)

Abbreviations: N_{PAA} = additional species delineated within clusters by population aggregation analysis (see text), with the maximum possible number of cases presented in parentheses; K_{mt} shared = number of haplotypes shared among study sites; P_{end} = proportion of GMYC species found at only 1 site; P_{Mad} = estimated proportion of the Malagasy fauna sampled in this study as GMYC species (multiple threshold results from Table 2) based on current estimates of richness for all of Madagascar (γ_{Mad}). Summary of data from Supplementary Figures S2–S6. Nonparametric Chao (1984) estimates (see Supplementary Fig. S6) are reported with confidence intervals (CI) in parentheses.
[a]Elouard et al. (2003).
[b]Eggleton and Davies (2003).
[c]Davis et al. (2002).
[d]Lees and Minet (2003).
[e]Bergsten J. (personal communication).

beetles. Each group had 1–4% of haplotypes that were shared among at least 2 study sites. Butterflies were the exception, with 24 shared haplotypes, or 12% (Table 3).

Congruence with Morphological Species and Nuclear rRNA

Morphological analysis delineated 348 species among the sequenced specimens (Table 4, Supplementary Table S1). These belonged to 104 genera (Supplementary Table S1), many of which are endemic to Madagascar and some of which have pantropical or global distributions. The mayfly species delineated include 5 genera of Baetidae (*Afroptilum*, *Dicentroptilum*, *Herbrosus*, *Guloptiloides*, and *Xyrodromeus*) that together comprise a single endemic lineage (Monaghan, Gattolliat et al. 2005) with uncertain generic status and several undescribed species (Gattolliat and Sartori 2000; Elouard et al. 2003; Gattolliat 2004). Within termites, we sampled *Coarctotermes*, *Capritermes*, and several other undescribed endemic genera of Kalotermitidae and Nasutitermitinae (Inward et al. 2007). The butterfly genera *Heteropsis*, *Perrotia*, and *Strabena* are particularly species-rich groups in Madagascar (Lees and Minet 2003), as are the dung beetle genera *Aleiantus*, *Sphaerocanthon*, *Arachnodes*, *Apotolamprus*, and *Helictopleuris* (e.g., Orsini et al. 2007) and water beetles in *Anacaena*, *Copelatus*, and *Madaglymbus* (Shaverdo et al. 2008).

Morphological diagnosis was highly congruent with GMYC species membership for all groups (Table 4, Supplementary Figs. S2–S6). The overall increase of approximately 6% (from 348 morphospecies to 370 genetic species) resulted from changes to the extent of groups. Rather than a simple subdivision of some morphospecies, there were 29 instances of subdivision and 7 of fusion (Table 4). The majority of cases involved the subdivision of 1 morphospecies into 2 GMYC species (Supplementary Figs. S2–S6). The increase was consistent across taxa, but varied in magnitude. Mayflies increased by 20%, whereas there were 1–8% more GMYC

TABLE 4. Congruence of morphological and nuclear genotype assignment to GMYC species

	Morphology							28S rRNA N_{GMYC} subsumed				
	N_{morph}	Subdivided		Fused		Net change		$K_n > 1$		Total	= 2	> 2
Mayflies	53	11	(5)	0	(1)	0.21	(0.08)	1[a]	(1)	4	2	1
Termites	22	3	(3)	2	(1)	0.05	(0.10)	0	(0)	6	2	2
Butterflies	128	1	(0)	0	(3)	0.01	(0.02)	0	(1)[b]	27	6	3
Dung beetles	39	3	(4)	0		0.08	(0.10)	0	(0)	13	7	3
Water beetles	106	11	(9)	5	(4)	0.05	(0.05)	2[c]	(0)	29	9	2
Total	**348**	**29**	**(21)**	**7**	**(9)**			**3**	**(2)**		**(26)**	**(11)**

Notes: N_{morph} = number of morphological groups tabulated using standard diagnostic characters for each group. For instances of subdivided or fused morphospecies, single-threshold GMYC results are presented, followed by results from the multiple-threshold model in parentheses. GMYC delineation increased (divided) or reduced (fused) the number of groups (summarized here from Supplementary Figs. S2–S6). There was a net proportional increase (+) in all taxa. The multiple threshold model reduced the number of divisions and increased the number of fusions. In 3 cases, multiple 28S rRNA genotypes (K_n) were found within single GMYC species; reduced to 2 cases with the multiple-threshold model. When more than 1 GMYC species was subsumed within a single 28S genotype, the majority were pairs (=2) of closely related species, with large groups (> 2).
[a]*Ellassoneuria sp.* 5 had 2 different 28S genotypes.
[b]*Perrotia gala* and *Perrotia kikideli* were fused by multiple-threshold GMYC but had different 28S genotypes.
[c]*Madaglymbus* sp. 1 and *Madaglymbus* sp. 2 had different 28S genotypes and were 1 GMYC species under a single-threshold model, but 2 species under the multiple-threshold model. The same was true for *Madaglymbus* sp. 3 and *Madaglymbus* sp. 4.

species in the other taxa (Table 4). There was not a single instance of conflict, whereby morphological traits and GMYC clustering placed 1 individual in alternative groups.

The multiple-threshold model produced fewer GMYC species, resulting in an approximately 3% increase over morphology (compared with 6% above). An intriguing result was that allowing the Yule coalescent transition to vary across the tree produced 8 fewer subdivisions of morphospecies compared with a single transition (Table 4). The ability to detect a variable threshold appears to be more congruent to morphological assignment. As noted earlier, the majority of changes resulted from an increased depth of the point of Yule coalescent transition.

Sequence data were sufficient to permit the direct comparison of mitochondrial and nuclear genetic data for a subset of GMYC species (248, or approximately 65% of those examined). Seventy-nine GMYC species were subsumed within 38 rRNA genotypes, resulting from the lower levels of nuclear variation relative to mtDNA. In the majority of cases, 2 GMYC groups shared a single rRNA genotype, with fewer cases of multiple mixed Yule coalescent groups sharing a single rRNA genotype (Table 4). The latter were members of highly species-rich groups. For example, 6 species of *Heteropsis*, 6 species of *Strabena* (both butterflies), and 9 species of *Anacaena* water beetles each shared an identical rRNA genotype. There were 146 single-threshold GMYC groups with a 28S rRNA genotype that was unique to the group and invariable in all members.

In only 3 instances (i.e., 1.2% of cases) was the assignment of single-threshold GMYC species to rRNA genotypes not straightforward (Supplementary Fig. S7). In each case, a single-threshold species had 2 different 28S genotypes. Two of these were *Madaglymbus* water beetles. The 2 morphospecies *Madaglymbus* sp. 1 and *Madaglymbus* sp. 2 also had different 28S genotypes (1-bp difference) but were recognized as a single GMYC species. *Madaglymbus* sp. 3 and *Madaglymbus* sp. 4 were also separate morphospecies, but 1 individual was placed in alternative groups by the 2 genetic markers (Supplementary Fig. S6). Interestingly, morphology was congruent with mtDNA and rRNA was congruent with geography, there being 1 phylogenetic species in Montage d'Ambre and 1 in Andasibe according to diagnostic rRNA nucleotides. The third case was the mayfly *Elassoneuria* sp. 1. It was recovered as a single GMYC group and single morphospecies despite having 2 28S genotypes (Supplementary Fig. S7). Under the multiple-threshold model, there were 5 instances of single GMYC species having 2 nuclear genotypes. The recognition of *Madaglymbus* sp. 1 and *Madaglymbus* sp. 2 as distinct GMYC species increased agreement between mtDNA and rRNA groupings (above). In contrast, the fusing of 6 butterfly species to 3 by the multiple-threshold model (Table 4) meant that 3 additional GMYC groups had 2 different 28S genotypes. In each case, 28S genotypes were unique to morphospecies.

Genetic Diversity and Sampling

The degree of subdivision and distribution of genetic variation was further assessed within the 239 GMYC groups with more than 1 individual. This included 168 entities with multiple haplotypes (1121 individuals in total). In 60 groups with fewer than 6 individuals, each genotype was unique, but the 1:1 relationship of individuals:haplotypes appeared to reach a threshold between 6 and 10 individuals (Fig. 4a). The only exception was the water beetle *Anacaena* sp. 4, with all 15 individuals having a unique *cox1* haplotype. Only 4 GMYC groups had >10 mtDNA haplotypes, despite there being 31 groups with >10 individuals sequenced (Supplementary Table S1).

None of the 239 GMYC groups were found to depart significantly from a neutral distribution using Tajima's *D* (i.e., θ_W and π should estimate genetic polymorphism within species equally), which suggests that selective sweeps and range expansions did not feature in the distribution of variation. Nucleotide variation within

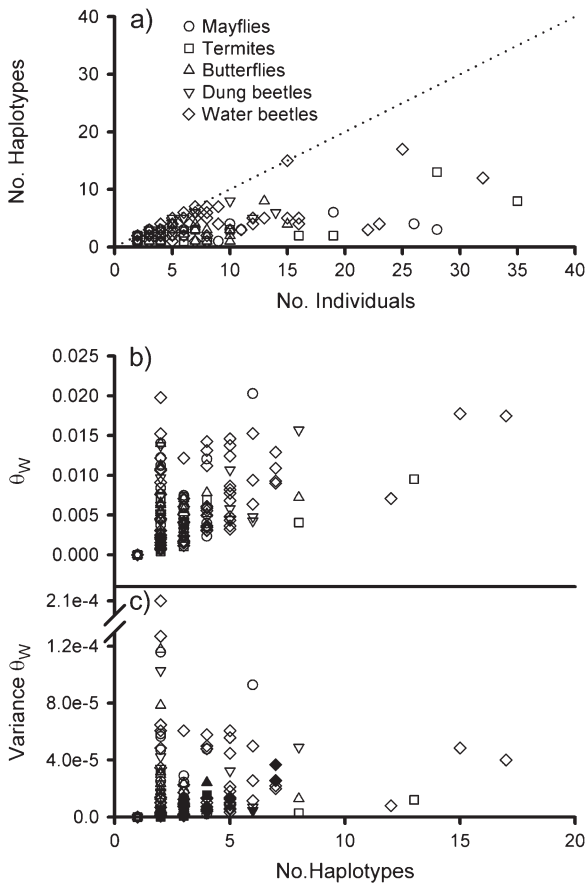

FIGURE 4. Genetic diversity estimates for 239 GMYC species. The relationship between the number of individuals and haplotypes (a) appeared to be saturated after 7 individuals, except in 1 species of *Anacaena* water beetles, where 15 individuals each had a unique haplotype. Relationships between the number of haplotypes and (b) genetic diversity (θ_W) and (c) its variance appeared to have similar thresholds after 7 individuals.

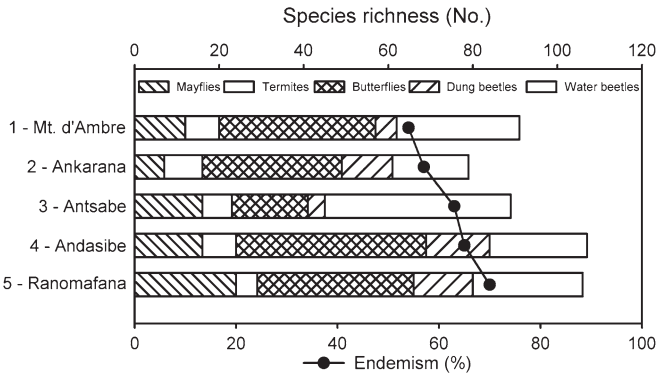

FIGURE 5. Species richness and endemism at each of the study sites in Madagascar. Total species richness is presented as the sum of single-threshold GMYC species for each taxonomic group. Endemism is summed across all taxa; for taxon-specific endemism, see Table 3 (P_{end}).

species ranged from $\theta_W = 0.000$ to 0.021 and was unrelated to sample size, as either haplotypes (K_{mt}, Pearson $R = 0.47$, Fig. 4b) or sequences (n, Pearson $R = 0.01$, data not shown). Variance decreased with increasing sample size and maximum values were not observed past 6 haplotypes (Fig. 4c).

The prevalence of geographical subgroups with diagnostic mtDNA characters (above) suggests an important role for range size in determining genetic variation. Although there was a significantly positive correlation between θ_W and the number of study sites in which a species occurred, the relationship was weak ($R^2 = 0.023, F_{1,237} = 5.6060, P = 0.018$). In fact, only 1 of the 16 highest θ_W values came from a cluster containing >1 diagnosable units, thus θ_W and its variance did not appear to be indicative of GMYC groups that contained diagnostic haplotypes. Taken together, greater sampling intensity did not appear to be the main factor determining genetic diversity, particularly as it was partitioned into GMYC groups. Equally, the diagnosis of phylogenetic species was not the result of greater genetic diversity per se, but resulted from the way diversity is partitioned among localities for those widespread species.

Richness and Endemism

The relative contribution to total species richness by the 5 target groups was fairly uniform across study sites (Fig. 5). Comparison of the more conservative, multiple-threshold GMYC species numbers with published and expert estimates indicates that our samples comprised 20–80% of the known fauna for all of Madagascar (Table 3). Nonparametric estimates (Chao 1984) of richness suggest that we undersampled species richness (Table 3); however, the utility of the estimator is probably limited by the fact that not all sampled individuals were successfully sequenced. The resulting increase in the number of singletons and doubletons will influence the shape of the sampling curve (Supplementary Fig. S7). Local endemism was pronounced in all taxonomic groups, with approximately 60–91% of species being found in only 1 of the 5 study sites (Table 3). Endemism was highest in dung beetles and lowest for termites. There was some effect of spatial scale on these results because study sites closer together were faunistically more similar according to a Mantel test (faunal dissimilarity and geographic distance; $Z = 47070, P = 0.048$). However, when measured for study sites rather than for taxa, each site had a fairly similar level of endemism, ranging from approximately 55% to 70% (Fig. 5).

DISCUSSION

Success of the GMYC Approach

Sequence-based approaches to group delimitation and identification have enormous potential to accelerate studies of biodiversity. This is particularly true where there is little or no means to apply morphological or ecological criteria (Blaxter 2004). Unfortunately, recent approaches to the use of DNA for taxonomy deal only with part of this task: either they attempt an identification based on a nearest match or they use arbitrary thresholds of sequence divergence to establish "molecular operational taxonomic units" (MOTUs). The former has been shown to be problematic (e.g., Meyer and Paulay 2005; Meier et al. 2006; Elias et al. 2007), in particular, if the reference database is incomplete (Paquin and Hedin 2004), and MOTUs may correspond to any hierarchical level, but not necessarily to the species category (Floyd et al. 2002; Blaxter 2004; Forney et al. 2004). This is a major short coming of existing approaches, as they may miss what is arguably the critical hierarchical level in the analysis—the species (Vogler and Monaghan 2007).

The key finding here was the ability to delineate species effectively within mixed DNA samples by applying an evolutionary model designed to capture the signature of independently evolving groups from single-locus sequence data, with minimal conflict with the traditional procedures of species recognition from morphology. The pattern was consistent across a broad diversity of tropical insects, spanning a wide range of evolutionary history and ecological diversity. This is somewhat surprising given the expectations that closely related species of endemic lineages such as those on Madagascar would be especially problematic for a mtDNA-based approach (Hey et al. 2003; Shaffer and Thompson 2007).

An important advance to previous applications of the GMYC model was the novel, multiple-threshold approach that allows for a variable transition from coalescent to speciation across different clades. The closer congruence to morphological species, although relatively minor in total change, suggests that variable rates of evolution and degrees of divergence can be effectively modeled and identified. Future investigation of morphological characters may result in a subdivision of groups in line with those delineated by the model,

whereas further molecular markers may either refute these groups or may confirm their validity as "cryptic" species not recognizable based on external characters. An advantage of any DNA-based system is that the data are readily available for further analysis, for example, with alternative models or approaches.

The 28S marker played an important role in the study by providing confirmation that the GMYC groups were congruent with the nuclear genome. This advances the argument for the use of nuclear markers in taxonomy (e.g., Monaghan, Balke et al. 2005, Sonnenberg et al. 2007) and allays concerns over the frequent gene flow of mtDNA (Funk and Omland 2003) for these species. In nearly 60% of all examined cases, mitochondrial and nuclear clusters were congruent. For nearly all the remaining groups, insufficient variation in the nuclear marker led to frequent lumping of 2 or more mtDNA GMYC species into a single genotype. In this case, gene flow between close relatives may remain undetected, although the high degree of morphological congruence here suggests that this also is minimal. For the 3 cases in which single species had 2 nuclear genotypes (5 under the multiple threshold model), 28S was congruent with strict morphological analysis. Only once was an individual misassigned by 1 or the other marker. This was due to a 1-bp difference in the 28S genotype. Given the extent of sampling here, even with stringent phred scores ($>$40), 68 bp in the data set were likely to be miscalled. Thus, this level of inaccuracy (1 incongruent individual from nearly 1500) is well within analytical error. This is another important consideration, in that although 28S served to establish the high degree of congruence, it is too conservative to establish species-level membership in these taxa.

Taken together, our results support the use of mtDNA as an effective means of species delineation. Mitochondrial incongruence, such as may occur through hybridization, was limited to 1 individual from nearly 1500 across nearly 350 morphospecies. The lack of a departure from neutrality in all species studied suggested that no bottlenecks or selective sweeps (see Bazin et al. 2006) had an effect on the patterns of variation in mtDNA. This conclusion is based on small sample sizes for many species, although 44 species with $>$5 individuals and 10 with $>$15 individuals represent very similar sample sizes to those used in explicit studies of neutrality (e.g., Baudry et al. 2006).

Other recently developed methods for species delineation rely on analysis of multiple unlinked loci (e.g., Knowles and Carstens 2007; Shaffer and Thompson 2007). Although multiple-marker approaches are expected to be more powerful than single locus approaches in most circumstances, universal primers for nuclear markers that are sufficiently variable for species delineation are not yet widely available for many groups. Improvements to technology and shifts in research foci may change this, but most of the present emphasis is placed on large-scale surveys of single-locus variation. Examples include *cox1* in animals (Hebert et al. 2004) and 16S in prokaryotes (Forney et al. 2004).

Sampling

It is important to consider the effects that our sampling scheme may have on the interpretation of patterns. The critical question is whether haplotypes were fully sampled from the lineages in question, regardless of whether we consider them to belong to the same species delineated here or to new unsampled species. Based on our overall results, we sampled approximately 1 haplotype for every 1.7 individuals. From our actual sample sizes, ranging from 1 to 35 individuals per species, we could estimate that 7–10 individuals (approximately 4–6 haplotypes) represented a threshold above which genetic diversity increased in smaller amounts. From neutral coalescent theory, the probability of sampling the MRCA of a coalescent group is calculated as $(n - 1)/(n + 1)$, where n is the number of individuals (Saunders et al. 1984; Nordborg 2001). Thus, the threshold sample sizes here had a probability of approximately 0.75–0.82 of sampling the MRCA. To reach a theoretical probability of 0.90 requires 19 individuals; 0.95 requires 39. These are probably not realistic sample sizes with existing sequencing methods, and more fundamentally because many species in any ecological survey are rare. Fortunately, these probabilities are conservative because the calculation is based on demographic assumptions that are probably not met in natural populations (see Nordborg 2001; Morando et al. 2003). When the criteria are not met, the necessary sample size usually becomes smaller, so the probability is higher than our calculations above but difficult to estimate. Morando et al. (2003), using phylogenetic bootstrap support and nested clade analysis to delineate 13 species of *Liolaemus* lizards, concluded that sample sizes of 5–10 individuals per locality should be adequate in studies of species delineation. Our empirical data used to delineate more than 350 species suggest that similar sample sizes are also appropriate for the GMYC approach.

The above limitations to sampling of haplotypes require a better understanding of how samples sizes affect the ability of the model to extract a coalescent–speciation transition. Perhaps the most critical question is how many singletons the data set can tolerate. In our data set, singletons accounted for up to 60% of delineated GMYC species (Supplementary Table S4), yet the model performed very well. But there must certainly be a point at which a transition from coalescent to speciation is missing, for example, in a typical phylogenetic tree of species-level taxa. The model is amenable to studies using coalescent simulation (e.g., Papadapoulou et al. 2008) and could also be evaluated by random removal of terminals from empirical phylogenetic trees.

A second sampling consideration is that of geographical scale, particularly with respect to conclusions about species-level groups using geography as an alternative criterion (Davis and Nixon 1992; Hedin and Wood 2002; DeSalle et al. 2005). In our study, application of a character-based concept increased the number of species substantially because of the pronounced local

endemism of haplotypes. This number of phylogenetic species diagnosable within GMYC groups may be reduced if spatially intermediate populations contain a combination of what were diagnostic haplotypes. The higher number of GMYC groups relative to morphological entities we observed also could merely reflect unrecognized isolation-by-distance phenomena within morphological species. Indeed, minimum distances here ranged from 40 (Ankarana-Montagne d'Ambre) to 280 km (Ranomafana-Andasibe). Many intermediate populations and (sister) species within clades may be missing (Sperling 2004; Prendini 2005). If they hold additional haplotypes of intermediate divergence, these could weaken the distinction between discrete clusters, potentially resulting in fewer GMYC groups. However, many of the species we studied here were widespread and genetic variation of those groups did not increase with a greater number of genotypes or study sites sampled (beyond an initial increase from sampling the basic diversity). This is in agreement with other studies of widespread geographic samples (Hebert and Gregory 2005; Pons et al., 2006) where mtDNA patterns were retained under broad geographical sampling. Only a denser sampling of study sites throughout the island will reveal the total number of groups that have been missed with our sampling regime.

Madagascar Biodiversity

Using GMYC groups as the basis for species numbers thus provides the possibility for assessing the biodiversity of Malagasy insects. An important finding is that total species richness from the 5 study sites was between 20% and 80% of existing estimates for Madagascar as a whole. These values were even higher after applying the Chao estimator to correct for incomplete sampling in the 5-d period. Although these comparisons are made only on the basis of species numbers and are slightly inflated compared with morphologically recognized species (3% in the GMYC), it is clear that we have either sampled a large proportion of total diversity or the current estimates are too low. Based on the high turnover and local endemism, we might conclude that estimates for Madagascar as a whole are too low. However, recent extinctions may have reduced the total richness. For example, estimates of dung beetle extinction found a relatively recent loss of some 50% of the known species based on a combined evaluation of museum specimens and intense field sampling (Hanski et al. 2007). An important task for further studies is to link the GMYC groups to those species already named. This would provide access to the relevant historical records and facilitate studies similar to the work of Hanski et al. (2007) in other taxonomic groups.

A second finding was the pronounced degree of species turnover between study sites. Taken at face value, our data suggest a very high level of local endemism in Madagascar, affecting all target taxa despite the broadly different ecological roles and evolutionary histories. This was despite what appeared to be a greater degree of dispersal in butterflies, which had 3 times the proportion of shared haplotypes. Although the broad geographical scale of endemism cannot be fully addressed with these data, it is noteworthy that high turnover was fairly uniform over distances of 40–1000 km, with only weak correlation of species similarity with geographic distance. Similar conclusions of local endemism were obtained using a MOTU survey of ants in northeastern Madagascar (Smith et al. 2005) and using detailed range extrapolations in dung beetles (Hanski et al. 2007). The outstanding question remaining is the scale of species turnover across a broader sample of biodiversity and the link with topography and ecozones (Wilme et al. 2006), which will determine the total number of insect species on Madagascar (Paulian and Viette 2003).

Conclusions

An important step in biodiversity assessment is the delineation of groups of biological significance, typically species. Whereas group delineation is often the most time-consuming and subjective step, the GMYC approach provides an objective and rapid means to delineate species from DNA sequences alone. Our finding that GMYC clusters correspond closely to traditionally defined species greatly enhances the biological relevance of the groupings produced, as any "collateral" information (Janzen et al. 2005) can be associated with evolutionarily coherent entities rather than similarity-based operational units. This is particularly true where samples will be collected in bulk or where there is little or no means to apply morphological or ecological criteria (Blaxter 2004; Forney et al. 2004; Bass et al. 2007). These data are immediately available for the assessment of macroecological patterns, can be applied to all individuals present in a sample regardless of sex or life stage, and can immediately recognize samples from widely separated localities as belonging to the same or different groups. As sequencing technology and computational tools improve, the generation of large amounts of DNA data from diverse assemblages within environmental samples (e.g., Rusch et al. 2007) may be the fastest avenue for species discovery in the future. An important task remains the linkage of sequence diversity to species already described in order to integrate the 90% of species awaiting discovery (Wilson 2003) with our existing knowledge of biodiversity.

Supplementary Material

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

Funding

Research was funded by the Biotechnology and Biological Sciences Research Council, UK (BBS/B/04358).

ACKNOWLEDGEMENTS

Field work would not have been possible without the generous support of staff at the Madagascar Institute for the Conservation of Tropical Ecosystems, especially Benjamin Andriamihaja, Liva Ravelonarivo, Benjy Randrianambina, and Jean-Marcel Rakotoarison. Sampling was conducted with the help of Doug Ottke, Roger Andriamparany, Kelly Inward, and Pierre Razafindraire. In Ranomafana, we thank Anigo Summer-Nielson, Jean-Philippe Puyravaud, and Patricia Wright. In London, Julia Llewellyn-Hughes, Claire Griffin, Andie Hall, Lisa Smith, Jo Ingle, Anna Papadapoulou, Sandra Sterman, Alex Martin, and Sylvia Fabrizi supported each stage of the laboratory analysis. Johannes Bergsten, Jean-Luc Gattolliat, and Olivier Montreuil provided expertise with specimen identifications. Data analysis was assisted by Ben Isambert, Mafalda Lopez da Silva, Eleri Randall, Aline Moore, Gail Bartlett, Tom Conner, Toby Hunt, and Paul van Dyk. Comments from Anna Papadapoulou helped to improve an earlier draft. We thank Marshal Hedin and 3 anonymous reviewers for comments on the manuscript. Access to protected areas and permission to sample (Permit No. 175 MINENVEF/SG/DGEF/DPB/SCBLF) and export (Permit No. 354N-EV11/MG04) specimens for research was granted by the Madagascar Ministry of the Environment, Water and Forests and the National Association for the Management of Protected Areas.

REFERENCES

Ahrens D., Monaghan M.T., Vogler A.P. 2007. DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). Mol. Phylogenet. Evol. 44:436–449.

Barraclough T.G., Birky C.W., Burt A. 2003. Diversification in sexual and asexual organisms. Evolution. 57:2166–2172.

Bass D., Richards T.A., Matthai L., Marsh V., Cavalier-Smith T. 2007. DNA evidence for global dispersal and probable endemicity of protozoa. BMC Evol. Biol. 7:162.

Baudry E., Derome N., Huet M., Veuille M. 2006. Contrasted polymorphism patterns in a large sample of populations from the evolutionary genetics model *Drosophila simulans*. Genetics. 173:759–767.

Bazin E., Glemin S., Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science. 312:570–572.

Blaxter M.L. 2004. The promise of a DNA taxonomy. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359:669–679.

Blaxter M.L., Floyd R. 2003. Molecular taxonomics for biodiversity surveys: already a reality. Trends. Ecol. Evol. 18:268–269.

Brown J.H., Lomolino M.V. 1998. Biogeography. 2nd ed. Sunderland (MA): Sinauer.

Chan M.S., Ventress N. 2001. STARS: sequence typing analysis and retrieval system. Oxford: University of Oxford. Available from: http://neelix.molbiol.ox.ac.uk:8080/userweb/ mchan/stars/.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scand. Stat. Theory Appl. 11:265–270.

Colwell R.K. 2005. EstimateS: statistical estimation of species richness and shared species from samples. Storrs (CT): University of Connecticut. Available from: http://purl.oclc.org/estimates.

Davis A.L.V., Scholtz C.H., Philips T.K. 2002. Historical biogeography of scarabaeine dung beetles. J. Biogeogr. 29:1217–1256.

Davis J.I., Nixon K.C. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. Syst. Biol. 41:421–435.

Depaulis F., Mousset S., Veuille M. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. J. Mol. Evol. 57:S190–S200.

DeSalle R., Egan M.G., Siddall M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1905–1916.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88

Drummond A.J., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Edwards A.W.F. 1972. Likelihood. London: John Hopkins University Press.

Eggleton P., Davies R. 2003. Isoptera, termites. In: Goodman S.M., Benstead J.P., editors. The natural history of Madagascar. Chicago (IL): University of Chicago. p. 654–660.

Elias M., Hill R.I., Willmott K.R., Dasmahapatra K.K., Brower A.V.Z., Mallet J., Jiggins C.D. 2007. Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc. R. Soc. Lond. B. Biol. Sci. 274:2881–2889.

Elouard J.-M., Gattolliat J.-L., Sartori M. 2003. Ephemeroptera, mayflies. In: Goodman S.M., Benstead J.P., editors. The natural history of Madagascar. Chicago (IL): University of Chicago. p. 639–645.

Floyd R., Abebe E., Papert A., Blaxter M. 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. 11:839–850.

Fontaneto D., Herniou E.A., Boschetti C., Caprioli M., Melone G., Ricci C., Barraclough T.G. 2007. Independently evolving species in asexual bdelloid rotifers. PLoS Biol. 5:914–921.

Forney L.J., Zhou X., Brown C.J. 2004. Molecular microbial ecology: land of the one-eyed king. Curr. Opin. Microbiol. 7:210–220.

Fraser C., Alm E.J., Polz M.F., Spratt B.G., Hanage W.P. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 323:741–746.

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. Syst. 34:397–423.

Gattolliat J.-L. 2004. A distinctive new species of *Xyrodromeus* Lugo-Ortiz & McCafferty (Ephemeroptera: Baetidae) from Madagascar. Zootaxa. 452:1–10.

Gattolliat J.-L., Sartori M. 2000 Contribution to the systematics of the genus *Dabulamanzia* (Ephemeroptera:Baetidae) in Madagascar. Rev. Suisse Zool. 107, 561–577.

Godfray H.C.J., Lewis O.T., Memmott J. 1999. Studying insect diversity in the tropics. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 354:1811–1824.

Goloboff P., Farris S., Nixon K. 2004. TNT (tree analysis using new technology). Cladistics. 20:84.

Groombridge B., editor. 1992. Global biodiversity: status of the earth's living resouces. London: Chapman & Hall.

Hanski I., Koivulehto H., Cameron A., Rahagalala P. 2007. Deforestation and apparent extinctions of endemic forest beetles in Madagascar. Biol. Lett. 3:344–347.

Hebert P.D.N., Gregory T.R. 2005. The promise of DNA barcoding for taxonomy. Syst. Biol. 54:852–859.

Hebert P.D.N., Penton E.H., Burns J.M., Janzen D.H., Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Natl. Acad. Sci. USA. 101:14812–14817.

Hedin M., Wood D.A. 2002. Genealogical exclusivity in geographically proximate populations of *Hypochilus thorelli* Marx (Araneae, Hypochilidae) on the Cumberland Plateau of North America. Mol. Ecol. 11:1975–1988.

Hey J., Waples R.S., Arnold M.L., Butlin R.K., Harrison R.G. 2003. Understanding and confronting species uncertainty in biology and conservation. Trends Ecol. Evol. 18:597–603.

Hoagland K.E. 1996. The taxonomic impediment and the Convention of Biodiversity. Assoc. Syst. Coll. Newslett. 24:61–67.

Hudson R.D., Coyne J.A. 2002. Mathematical consequences of the genealogical species concept. Evolution. 56:1557–1565.

Inward D.J.G., Vogler A.P., Eggleton P. 2007. A comprehensive phylogenetic analysis of termites (Isoptera) illuminates key aspects of their evolutionary biology. Mol. Phylogenet. Evol. 44:953–967.

Janzen D.H. 2004. Now is the time. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359:731–732.

Janzen D.H., Hajibabaei M., Burns J.M., Hallwachs W., Remigio E., Hebert P.D.N. 2005. Wedding biodiversity inventory of a large and

complex Lepidoptera fauna with DNA barcoding. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1835–1845.
Kingman J.F.C. 1982. On the genealogy of large populations. J. Appl. Probab. 19A:27–43.
Knowles L.L., Carstens B.C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–895.
Lees D.C., Minet J. 2003. Lepidoptera: systematics and diversity. In: Goodman S.M., Benstead J.P., editors. The natural history of Madagascar. Chicago (IL): University of Chicago Press. p. 748–761.
Mallet J. 1995. A species definition for the modern synthesis. Trends Ecol. Evol. 10:294–299.
Markmann M., Tautz D. 2005. Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1917–1924.
Meier R., Shiyang K., Vaidya G., Ng P.K.L. 2006. DNA barcoding and taxonomy in diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. 55:715–728.
Meyer C.P., Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3:2229–2238.
Monaghan M.T., Balke M., Gregory T.R., Vogler A.P. 2005. DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1925–1933.
Monaghan M.T., Balke M., Pons J., Vogler A.P. 2006. Beyond barcodes: complex DNA taxonomy of a South Pacific island radiation. Proc. R. Soc. Lond. B. Biol. Sci. 273:887–893.
Monaghan M.T., Gattolliat J.-L., Sartori M., Elouard J.-M., James H., Derleth P., Glaizot O., de Moor F., Vogler A.P. 2005. Trans-oceanic and endemic origins of the small minnow mayflies (Ephemeroptera, Baetidae) of Madagascar. Proc. R. Soc. Lond. B. Biol. Sci. 272:1829–1836.
Morando M., Avila L.J., Sites J.W. (2003). Sampling strategies for delimiting species: genes, individuals, and populations in the *Liolaemus elongatus-kriegi* complex (Squamata: Liolaemidae) in Andean–Patagonian South America. Syst. Biol. 52:159–185.
Moritz C., Cicero C. 2004. DNA barcoding: promise and pitfalls. PLoS Biol. 2:1529–1531.
Nordborg M. 2001. Coalescent theory. In: Balding D.J., Bishop M., Cannings C., editors. Handbook of statistical genetics. Chichester (UK): Wiley. p. 179–212.
Novotny V., Drozd P., Miller S.E., Kulfan M., Janda M., Basset Y., Weiblen G.D. 2006. Why are there so many species of herbivorous insects in tropical rainforests? Science. 313:1115–1118.
Orsini L., Koivulehto H., Hanski I. 2007. Molecular evolution and radiation of dung beetles in Madagascar. Cladistics. 23:145–168.
Papadapoulou A., Bergsten J., Fujisawa T., Monaghan M.T., Barraclough T.G., Vogler A.P. 2008. Speciation and DNA barcodes—testing the effects of dispersal on the formation of discrete sequence clusters. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363:2987–2996.
Paquin P., Hedin M. 2004. The power and perils of 'molecular taxonomy': a case study of eyeless and endangered Cicurina (Araneae: Dictynidae) from Texas caves. Mol. Ecol. 13:3239–3255.
Paradis E., Claude J., Strimmer K. 2004. APE: analyses of phylogenetics evolution in R language. Bioinformatics. 20:289–290.
Paulian R., Viette P. 2003. An introduction to terrestrial and freshwater invertebrates. In: Goodman S.M., Benstead J.P., editors. The natural history of Madagascar. Chicago (IL): University of Chicago Press. p. 503–511.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Posada D. 2004. Collapse: describing haplotypes from sequence alignments. Vigo (Spain): University of Vigo. Available from: http://darwin.uvigo.es/software/collapse.html.
Prendini L. 2005. Comment on "Identifying spiders through DNA barcodes." Can. J. Zool. 83:498–504.
R Core Development Team. 2005. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Rach J., DeSalle R., Sarkar I.N., Schierwater B., Hadrys H. 2008. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proc. R. Soc. Lond. B. Biol. Sci. 275:237–247.
Rambaut A., Charleston M. 2002. TreeEdit. Phylogenetic tree editor version 1.0 alpha 10. Oxford: Oxford University.
Rozas J., Sanchez-DelBarrio J.C., Messeguer X., Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.
Rusch D.B., Halpern A.L., Sutton G., Heidelberg K.B., Williamson S., Yooseph S., Wu D.Y., Eisen J.A., Hoffman J.M., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter J.E., Li K., Kravitz S., Heidelberg J.F., Utterback T., Rogers Y.H., Falcon L.I., Souza V., Bonilla-Rosso G., Eguiarte L.E., Karl D.M., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari M.R., Strausberg R.L., Nealson K., Friedman R., Frazier M., Venter J.C. 2007. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5: 398–431.
Sanderson M.J. 2003. r8s: inferring absolute rates of molecular evolution, divergence times in the absence of a molecular clock. Bioinformatics. 19:301–302.
Saunders I.W., Tavaré S., Watterson G.A. 1984. On the genealogy of nested subsamples from a haploid population. Adv. Appl. Probab. 16:471–491.
Shaffer H.B., Thompson R.C. 2007. Delimiting species in recent radiations. Syst. Biol. 56:896–906.
Shaverdo H., Monaghan M.T., Lees D.C., Ranaivosolo R., Balke M. 2008. A new genus of Malagasy endemic diving beetles and description of a highly unusual species based on morphology and DNA sequence data (Dytiscidae: Copelatinae). Syst. Biodivers. 6:43–51.
Sites J.W., Marshall J.C. 2003. Delimiting species: a renaissance issue in systematic biology. Trends Ecol. Evol. 18:462–470.
Sonnenberg R., Nolte A.W., Tautz D. 2007. An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. Front. Zool. 4:6.
Smith M.A., Fisher B.L., Hebert P.D.N. 2005. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1825–1834.
Sperling F.A.H. 2004. DNA barcoding: deus ex machina. Newsl. Biol. Surv. Canada 22:50–53.
Swofford D.L. 2002. PAUP*: phylogenetic analysis using parsimony. Version 4.0b. Sunderland (MA): Sinauer Associates.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.
Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18:70–74.
Vences M., Thomas M., Bonett R.M., Vieites D.R. 2005. Deciphering amphibian diversity through DNA barcoding: chances and challenges. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1859–1868.
Vogler A.P., Monaghan M.T. 2007. Recent advances in DNA taxonomy. J. Zool. Syst. Evol. Res. 45:1–10.
Wiens J.J., Penkrot T.A. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (Sceloporus). Syst. Biol. 51:69–91.
Will K.W., Mishler B.D., Wheeler Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. Syst. Biol. 54:844–851.
Wilme L., Goodman S.M., Ganzhorn J.U. 2006. Biogeographic evolution of Madagascar's microendemic biota. Science. 312:1063–1065.
Wilson E.O. 2003. The encyclopedia of life. Trends Ecol. Evol. 18: 77–80.
Yoder A.D., Burns M.M., Zehr S., Delefosse T., Veron G., Goodman S.M., Flynn J.J. 2003. Single origin of Malagasy Carnivora from an African ancestor. Nature. 421:734–737.
Yule G.U. 1924. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, FRS. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 213:21–87.

Received 1 April 2008; revised 9 June 2008; accepted 1 February 2009
Associate Editor: Marshal Hedin