DNA barcodes distinguish species of tropical Lepidoptera

Mehrdad Hajibabaei*†, Daniel H. Janzen†‡, John M. Burns§, Winnie Hallwachs‡, and Paul D. N. Hebert*

*Department of Integrative Biology, University of Guelph, Guelph, ON, Canada N1G 2W1; ‡Department of Biology, University of Pennsylvania, Philadelphia, PA 19104; and §Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0127

Contributed by Daniel H. Janzen, December 6, 2005

Although central to much biological research, the identification of species is often difficult. The use of DNA barcodes, short DNA sequences from a standardized region of the genome, has recently been proposed as a tool to facilitate species identification and discovery. However, the effectiveness of DNA barcoding for identifying specimens in species-rich tropical biotas is unknown. Here we show that cytochrome *c* oxidase I DNA barcodes effectively discriminate among species in three Lepidoptera families from Area de Conservación Guanacaste in northwestern Costa Rica. We found that 97.9% of the 521 species recognized by prior taxonomic work possess distinctive cytochrome *c* oxidase I barcodes and that the few instances of interspecific sequence overlap involve very similar species. We also found two or more barcode clusters within each of 13 supposedly single species. Covariation between these clusters and morphological and/or ecological traits indicates overlooked species complexes. If these results are general, DNA barcoding will significantly aid species identification and discovery in tropical settings.

Area de Conservación Guanacaste | cytochrome *c* oxidase I | Hesperiidae | Sphingidae | Saturniidae

Identification systems based on DNA have the potential to facilitate both the identification of known species and the discovery of new ones (1–3). DNA barcoding is based on the premise that sequence diversity within a short, standardized segment of the genome can provide a "biological barcode" that enables identifications at the species level (2, 4). Earlier studies have shown that sequence diversity in a 648-bp region near the 5′ end of the cytochrome *c* oxidase I (COI) mitochondrial gene can resolve >95% of the species in test assemblages of birds (5), fishes (6), and Lepidoptera (2, 7). The few cases of taxonomic ambiguity were within a complex of morphologically similar species. However, prior studies have not evaluated the performance of DNA barcoding in settings where species richness is particularly high.

Here we test the effectiveness of DNA barcoding for the identification and discovery of species of Lepidoptera in the species-rich fauna of Area de Conservación Guanacaste (ACG) in northwestern Costa Rica. ACG is an intensively inventoried 115,000-hectare block of interdigitated tropical dry forest, rain forest, and cloud forest (7–12). We ask whether COI barcodes provide sufficient resolution to identify specimens of the sympatric (or fine-scale parapatric) and morphologically identifiable species in three families of Lepidoptera–Hesperiidae (skipper butterflies), Sphingidae (sphinx moths), and Saturniidae (wild silk moths). Because this fauna has been much studied taxonomically for at least two centuries, it provides a template against which to test the accuracy of DNA barcoding.

Results and Discussion

We obtained COI sequences from 4,260 adults reared from wild-caught caterpillars (see *Materials and Methods* for details) that represent 521 (71%) of the morphologically defined species of hesperiids, sphingids, and saturniids known from ACG (Fig. 1). An average of eight barcode sequences per species was obtained, and just 11% of the taxa were represented by a single individual (Fig. 2). We found that 97.9% of the 521 species were unambiguously distinguishable from all other species because their barcode sequences formed distinct, nonoverlapping clusters in a neighbor-joining (NJ) analysis (13) (Fig. 3 and Figs. 5–7, which are published as supporting information on the PNAS web site). Given the strong evidence for monophyly of each of these three families, it is unlikely that the root of each barcode tree lies within any individual taxon studied here. The species clusters showed an average bootstrap support of 98% (results not shown), reflecting the fact that sequence divergences were generally much greater between species than within them. Congeneric species showed average divergences of 4.58%, 4.41%, and 6.02% (Hesperiidae, Sphingidae, and Saturniidae, respectively), whereas average within-species divergences were 0.17%, 0.43%, and 0.46% for the same three families (Fig. 1). As we note later, these intraspecific values are inflated by a few cases of deep sequence divergence within a morphologically defined species, some of which reflect clusters of overlooked species.

We counted how often the maximum sequence divergence among individuals of a species exceeded the minimum sequence divergence from another species. These situations, which may confound barcode-based taxonomic assignments, were encountered in 18 species (Fig. 4). Five of these cases involved cryptic species assemblages, which showed exceptionally high levels of within-species divergence (see below). Two of the 18 cases involved very low, but consistent, barcode differences that enabled a specimen to be assigned accurately to its cluster on the NJ tree (Fig. 5). However, another 11 species (2.1% of the total 521) were not separable from one or two other species because a pair or triplet had overlapping barcodes, producing a mixed-species cluster in the NJ tree (Fig. 5). Notably, within-cluster sequence divergence was always <1% (Table 1 and Figs. 4 and 5). No cases of this type were detected in the Saturniidae or Sphingidae, and those in the Hesperiidae involved morphologically similar congeners. For example, three species in the skipper genus *Phocides* formed a mixed-species cluster, as did two species of *Polyctor* (Table 1). Cases of barcode overlap might signal very recent speciation or hybridization (14). No biological information suggests the latter cause, and the species involved are morphologically and ecologically distinct taxa that are either sympatric or fine-scale parapatric in ACG. The key result is that COI barcodes identify all but 2.1% of the species in our test assemblage, and cases of incomplete resolution involve pairs or triplets of closely allied congeners.

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviations: COI, cytochrome *c* oxidase I; ACG, Area de Conservación Guanacaste; NJ, neighbor joining.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ275861–DQ276849, DQ291161–DQ291758, and DQ291759–DQ293943).

†To whom correspondence may be addressed. E-mail: mhajibab@uoguelph.ca or djanzen@sas.upenn.edu.

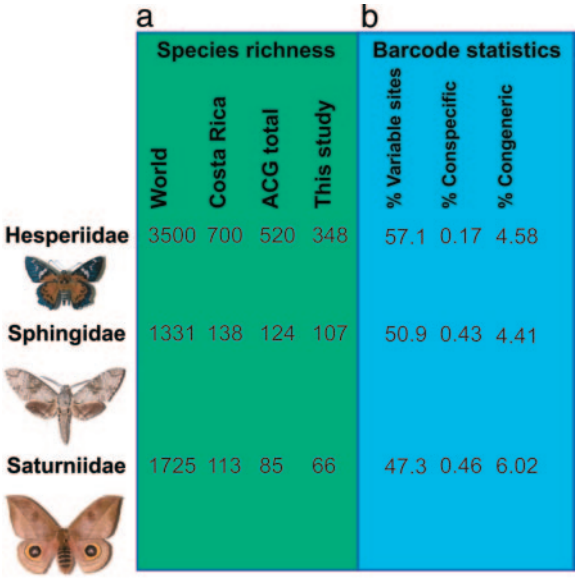

Fig. 1. Lepidoptera families used in this study and their barcode sequence statistics. (*a*) Species used in this study versus the total known for ACG, Costa Rica, and the world (http://janzen.sas.upenn.edu and I. Kitching, personal communication). (*b*) Comparison of COI barcode variation and measures of conspecific and congeneric genetic distances in the three families. Distances are calculated by using the Kimura two-parameter model of base substitution (23).

In contrast to cases of low divergence between taxa, we encountered 13 species whose barcodes were separated into 2–10 distinct clusters in the NJ trees (Figs. 4–7). We ruled out sequencing, methodological, or databasing errors. Furthermore, there were no polymorphic sites or stop codons in any of the barcode sequences, which would have signaled the inadvertent sequencing of a pseudogene in some individuals and a mtDNA sequence in others of the same species (15). These clusters remained distinct as sample sizes increased, supporting the conclusion that each of these 13 species includes distinct COI lineages rather than scattered sequence variation. We also found morphological, ecological, and/or microgeographic differences correlated with the barcode clusters (7) in each of these species. These cases are not discussed in detail here, but *Automeris zugana* provides a representative example. In this saturniid moth, members of one barcode cluster come from ACG dry forest, and members of the other two clusters come from parapatric rain forest, with one rain forest cluster occurring at an elevation of 400–600 m and the other at 600–800 m. Dry forest

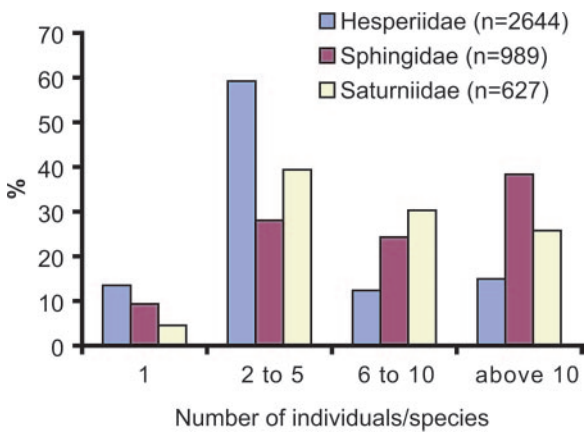

Fig. 2. Number of barcode sequences for each species. The total number of individuals analyzed per family is shown in parentheses.

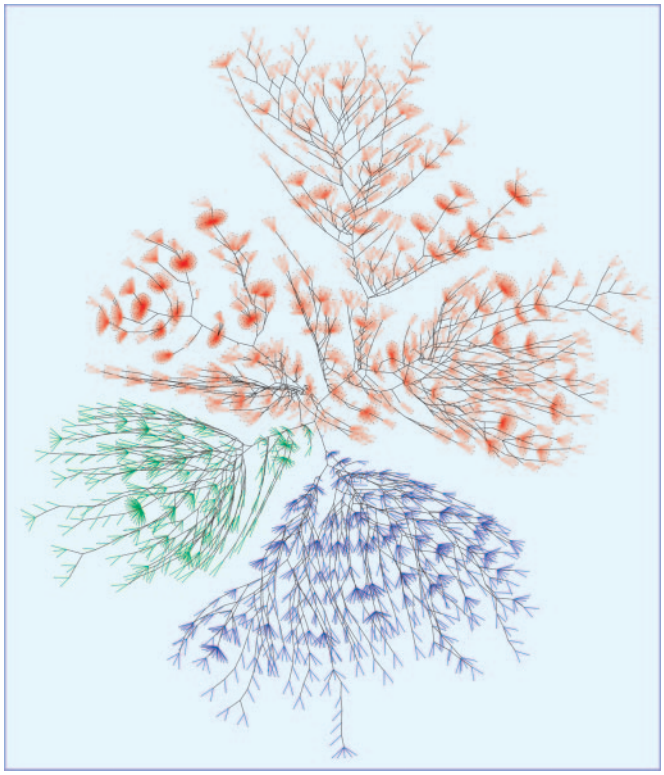

Fig. 3. A tree representation of COI barcodes for 4,260 individuals of three Lepidoptera families showing the clear separation of branches leading to individuals of Hesperiidae, Sphingidae, and Saturniidae (red, blue, and green, respectively). The original tree was made by using the NJ algorithm (13) based on Kimura two-parameter distances (23). To make this tree we modified the NJ tree by collapsing all of the branches with zero length, which creates the star-shape clusters in the tree (see Figs. 5–7 for detailed NJ trees for each family).

specimens are paler than their rain forest counterparts, and the members of one rain forest cluster are smaller than the other. Moreover, male genitalia differ as much between clusters as they commonly do between known species of *Automeris*. Covariation between barcode clusters and both ecological and morphological traits provides strong evidence that *A. zugana* in ACG is actually a complex of three species (16). Evidence of similar trait covariation was found in 12 other species displaying deeply divergent barcode clusters (Fig. 4). The discovery of possible cryptic species within some currently recognized taxa did not detract from the accuracy of identifications because barcode clusters for a single morphological species were always positioned together in the NJ tree. However, their treatment in our analyses as a single species has inflated estimates of within-species variation. When we treated the 13 barcode clusters that displayed covariation in morphology and/or ecology as distinct species, within-species barcode divergences were substantially reduced (Table 2).

Levels of intraspecific barcode variation differed considerably among the other species that we examined (Tables 3 and 4, which are published as supporting information on the PNAS web site). Most species showed low variation, whereas some showed sequence diversity rivaling that found between very similar species. However, because we found no evidence of morphological or life history covariation with the barcode variants within these taxa, we here regard each of them as single species. Some of these cases may simply represent unusually high levels of intraspecific variation, perhaps reflecting merged phylogeographic variants or retained ancestral polymorphisms. In other cases, further study may reveal additional overlooked species (7, 17). The search for such cryptic

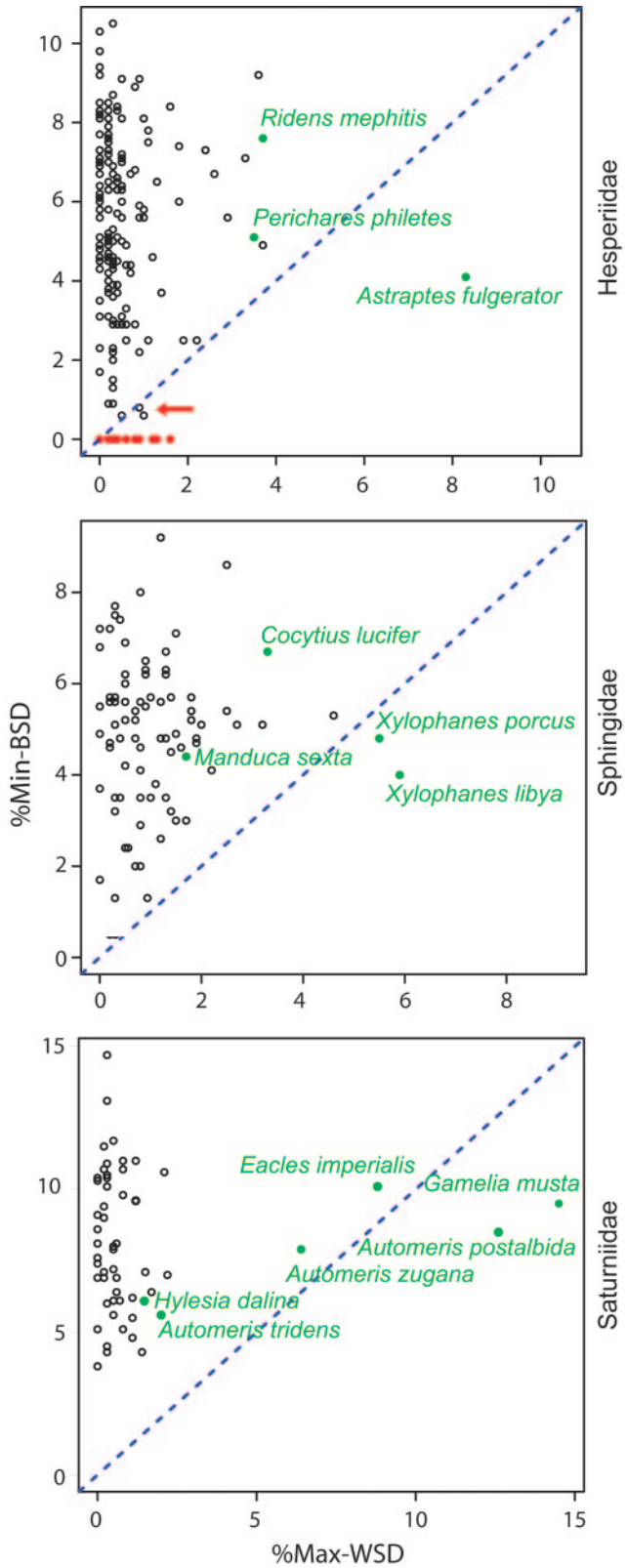

Fig. 4. Patterns of COI divergence for the 315 species of Lepidoptera from ACG that were represented by three or more individuals. Minimum between-species divergence (Min-BSD) is plotted against maximum within-species divergence (Max-WSD) for each morphologically defined species. Points above the diagonal represent cases where species identification is straightforward. Species with overlapping COI barcodes (and therefore very low Min-BSD values) are shown as red dots. Species with two or more distinct barcode clusters showing covariation in biological traits are shown as green dots (see text for details). In two species of Hesperiidae, indicated by an arrow, Min-BSD values are slightly (≈10%) smaller than Max-WSD values; but individuals of these species are not confused in the NJ analysis because their clusters do not overlap (13) (see Fig. 5). Distances are calculated by using the Kimura two-parameter model of base substitution (23).

Table 1. Species with overlapping COI barcodes

Genus	Species	% COI similarity*
Hesperiidae		
Cobalus	*fidicula* (11), *virbius* (5)	99.67
Neoxeniades	Burns01 (14), *luda* (8)	99.62
Phocides	*belus* (5), *pigmalion* (13), Warren01 (6)	99.72
Polyctor	*cleta* (23), *polyctor* (19)	99.74
Saliana	*fusta* (11), *triangularis* (7)	99.84

Number of individuals per morphological species is shown in parentheses.
*Mean between species similarity based on the Kimura two-parameter model of base substitution (23).

taxa would logically begin with those species that show the greatest within-species barcode divergences.

Because 97.9% of the 521 species examined in our study were unambiguously identified, it appears that DNA barcoding will be an effective tool for species recognition in tropical settings. This ability of barcoding to deliver species-level identifications (2, 5–7, 18) should allay the concern (19, 20) that short standardized gene sequences would be unable to provide resolution below the level of genus or family. In fact, the few cases of incomplete resolution that we encountered involved morphologically similar congeners. Moreover, if barcoding is used to tally species richness, these cases are more than offset by the revelation of overlooked species as evidenced by the discovery of 13 likely species complexes in our study. We emphasize that even in a group with a well established taxonomy, such as the Costa Rican Lepidoptera examined here, DNA barcoding enables the rapid detection of deep "intraspecific" barcode divergences that often flag overlooked species. Barcoding may also be applied to lesser known groups, where a count of barcode lineages showing deep divergence (e.g., >2%) will provide a preliminary signal of species richness. However, we emphasize that such an application of barcoding is no substitute for full taxonomic analysis, because the coupling of detailed morphological and ecological investigations with barcode results is critical for a final documentation of species richness (7).

Materials and Methods

Specimens. The specimens examined in this study were reared from wild-collected caterpillars by D.H.J., W.H., and a parataxonomist team (http://janzen.sas.upenn.edu) during the last 27 years of biodiversity inventory in ACG (16). All specimens were killed upon eclosion with cyanide or freezing (usually) and were spread and oven-dried in the field. We analyzed multiple individuals from each morphologically defined species when they were available. For <1% of the species, samples from the rearing program were augmented by wild-caught adults from the same site. Further details on each specimen are available at http://janzen.sas.upenn.edu.

We increased sample sizes whenever deep (i.e., >2%) sequence variation was found among members of a single morphologically defined species. These additional specimens were selected from different caterpillar food plants, contrasting caterpillar color patterns, and different but parapatric ecosystems of origin, so as to determine whether these barcode clusters remained distinct or merged to form a single variable assemblage. We also increased sample sizes when individuals of

Table 2. Incidence of within-species barcode clusters showing correlated biological differences and their impact on estimates of intraspecific sequence divergence in three families of Lepidoptera

Family	No. of morphological species	No. of species with barcode clusters*	% species with barcode clusters	Mean barcode divergences,† %		P‡
				Morphological species	Morphological species + barcode-detected species	
Hesperiidae	348	3	0.86	0.17	0.16	0.52
Sphingidae	107	4	3.74	0.43	0.35	0.16
Saturniidae	66	6	9.09	0.46	0.19	0.05

*Number of morphologically defined species composed of two or more barcode clusters with correlated biological differences.
†Distances are calculated by using the Kimura two-parameter model of base substitution (23).
‡Calculated by using a paired *t* test.

different species were found either to share barcode sequences or have sequences that were intermingled.

COI Amplification. DNA was extracted with standard protocols (21) from single legs removed from dried voucher specimens, which are marked with small yellow labels that say "Legs away/for DNA" and are housed in the National Museum of Natural History. We examined 4,260 specimens, including 2,644 individuals from 348 morphologically defined species of skipper butterflies (Hesperiidae), 989 individuals from 107 species of sphinx moths (Sphingidae), and 627 individuals from 66 species of wild silk moths (Saturniidae). We included sequence records for 459 members (and 10 cryptic species) of the *Astraptes fulgerator* complex (Hesperiidae) obtained in an earlier study (7). For ≈80% of the samples, the primers LepF (5′-ATTCAACCAATCATAAAGATATTGG-3′) and LepR (5′-TAAACTTCTGGATGTCCAAAAAATCA-3′) amplified the target 658-bp fragment of COI. In ≈7% of the cases where these primers did not produce a PCR product, we used primer Enh_LepR1 (5′-CTCCWCCAGCAGGATCAAAA-3′) as reverse primer. Combination of this primer and LepF amplifies a 612-bp fragment of COI. Finally, for the 13% of samples that were recalcitrant, most of which were >10 years old, we amplified shorter fragments by using the primer combination MF1 (5′-GCTTTC-CCACGAATAAATAATA-3′)-LepR (407-bp amplicon) and MH-MR1 (5′-CCTGTTCCAGCTCCATTTTC-3′)-LepF (311-bp amplicon). These shorter PCR products either were used alone (as a short DNA barcode) or were concatenated (in the case where both fragments were amplified for a given sample). Sequences were obtained by using either ABI 377 (25% of total sequences, unidirectional read) or ABI 3730 (75% of total sequences, bidirectional read) sequencers (Applied Biosystems).

Sequence Analysis. Sequences were edited to remove ambiguous base calls and primer sequences and were assembled by using SEQUENCHER (Gene Codes, Ann Arbor, MI). Sequences were then aligned by using CLUSTALW (22) software and manually edited. Sequence information was entered in the Barcode of Life Database (BOLD, www.barcodinglife.org) along with an image and collateral information for each voucher specimen. The detailed specimen records and sequence information, including trace files, are available on the BOLD in three project files (Hesperiidae of ACG1, Sphingidae of ACG1, and Saturniidae of ACG1). All sequences have been submitted to GenBank (Table 5, which is published as supporting information on the PNAS web site). Kimura's two-parameter model of base substitution (23) was used to calculate genetic distances in MEGA3 software (24), and NJ trees were produced by using BOLD and MEGA3 software. MEGA3 was used to perform bootstrap analysis on NJ trees (1,000 replicates).

We thank Ed Remigio for generating some of the saturniid barcodes; Stephanie Kirk for assistance with molecular work; Angela Holliss for DNA sequencing; Tanya Dapkey for preparing specimens; Donald Harvey for dissecting hesperiid and saturniid genitalia; ACG parataxonomists for rearing the caterpillars; Sujeevan Ratnasingham for database management; Gregory Singer for assistance with computerized analyses; and Charles Mitter, May Berenbaum, and Naomi Pierce for constructive reviews of the manuscript. This study was supported by the Natural Sciences and Engineering Research Council (Canada) (P.D.N.H.), the Canada Research Chairs program (P.D.N.H.), the Gordon and Betty Moore Foundation (P.D.N.H.), National Science Foundation Grants DEB 0072730 and 0515699 (to D.H.J. and W.H.), the Guanacaste Dry Forest Conservation Fund (D.H.J. and W.H.), ACG (D.H.J. and W.H.), and the National Museum of Natural History Small Grants Program (J.M.B.).

1. Blaxter, M. (2003) *Nature* **421,** 122–124.
2. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. (2003) *Proc. R. Soc. London Ser. B* **270,** 313–321.
3. Hebert, P. D. N., Ratnasingham, S. & deWaard, J. R. (2003) *Proc. R. Soc. London Ser. B* **270,** Suppl. 1, S96–S99.
4. Marshall, E. (2005) *Science* **307,** 1037.
5. Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S. & Francis, C. M. (2004) *PLoS Biol.* **2,** E312.
6. Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N. (2005) *Philos. Trans. R. Soc. London B* **360,** 1847–1857.
7. Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 14812–14817.
8. Burns, J. M. & Janzen, D. H. (2001) *J. Lepid. Soc.* **54,** 15–43.
9. Gauld, I. D. & Janzen, D. H. (2004) *Zool. J. Linn. Soc.* **141,** 297–351.
10. Janzen, D. H. (2004) *J. Appl. Ecol.* **41,** 181–187.
11. Janzen, D. H. (2000) *Biodiversity* **1,** 7–20.
12. Janzen, D. H. (2003) in *Arthropods of Tropical Forests: Spatio-Temporal Dynamics and Resource Use in the Canopy*, eds. Basset, Y., Novotny, V., Miller, S. E. & Kitching, R. L. (Cambridge Univ. Press, Cambridge, U.K.), pp. 369–379.
13. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
14. Sites, J. W. & Marshall, J. C. (2001) *Trends Ecol. Evol.* **18,** 462–470.
15. Bensasson, D., Zhang, D., Hartl, D. L. & Hewitt, G. M. (2001) *Trends Ecol. Evol.* **16,** 314–321.
16. Janzen, D. H., Hajibabaei, M., Burns, J. M., Hallwachs, W., Remigio, E. & Hebert, P. D. N. (2005) *Philos. Trans. R. Soc. London B* **360,** 1835–1845.
17. Saez, A. G. & Lozano, E. (2005) *Nature* **433,** 111.
18. Meyer, C. P. & Paulay, G. (2005) *PLoS Biol.* **3,** E422.
19. Will, K. W. & Rubinoff, D. (2004) *Cladistics* **20,** 47–55.
20. Hurst, G. D. & Jiggins, F. M. (2005) *Proc. R. Soc. London Ser. B* **272,** 1525–1534.
21. Hajibabaei, M., deWaard, J. R., Ivanova, N. V., Ratnasingham, S., Dooh, R. T., Kirk, S. L., Mackie, P. M. & Hebert, P. D. N. (2005) *Philos. Trans. R. Soc. London B* **360,** 1959–1967.
22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
23. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
24. Kumar, S., Tamura, K. & Nei, M. (2004) *Brief Bioinform.* **5,** 150–163.