THE ROYAL SOCIETY FirstCite® e-publishing

Received 29 July 2002
Accepted 30 September 2002
Published online

Biological identifications through DNA barcodes

Paul D. N. Hebert*, Alina Cywinska, Shelley L. Ball and Jeremy R. deWaard

Department of Zoology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

Although much biological research depends upon species diagnoses, taxonomic expertise is collapsing. We are convinced that the sole prospect for a sustainable identification capability lies in the construction of systems that employ DNA sequences as taxon 'barcodes'. We establish that the mitochondrial gene cytochrome *c* oxidase I (COI) can serve as the core of a global bioidentification system for animals. First, we demonstrate that COI profiles, derived from the low-density sampling of higher taxonomic categories, ordinarily assign newly analysed taxa to the appropriate phylum or order. Second, we demonstrate that species-level assignments can be obtained by creating comprehensive COI profiles. A model COI profile, based upon the analysis of a single individual from each of 200 closely allied species of lepidopterans, was 100% successful in correctly identifying subsequent specimens. When fully developed, a COI identification system will provide a reliable, cost-effective and accessible solution to the current problem of species identification. Its assembly will also generate important new insights into the diversification of life and the rules of molecular evolution.

Keywords: molecular taxonomy; mitochondrial DNA; animals; insects; sequence diversity; evolution

1. INTRODUCTION

The diversity of life underpins all biological studies, but it is also a harsh burden. Whereas physicists deal with a cosmos assembled from 12 fundamental particles, biologists confront a planet populated by millions of species. Their discrimination is no easy task. In fact, since few taxonomists can critically identify more than 0.01% of the estimated 10–15 million species (Hammond 1992; Hawksworth & Kalin-Arroyo 1995), a community of 15 000 taxonomists will be required, in perpetuity, to identify life if our reliance on morphological diagnosis is to be sustained. Moreover, this approach to the task of routine species identification has four significant limitations. First, both phenotypic plasticity and genetic variability in the characters employed for species recognition can lead to incorrect identifications. Second, this approach overlooks morphologically cryptic taxa, which are common in many groups (Knowlton 1993; Jarman & Elliott 2000). Third, since morphological keys are often effective only for a particular life stage or gender, many individuals cannot be identified. Finally, although modern interactive versions represent a major advance, the use of keys often demands such a high level of expertise that misdiagnoses are common.

The limitations inherent in morphology-based identification systems and the dwindling pool of taxonomists signal the need for a new approach to taxon recognition. Microgenomic identification systems, which permit life's discrimination through the analysis of a small segment of the genome, represent one extremely promising approach to the diagnosis of biological diversity. This concept has already gained broad acceptance among those working with the least morphologically tractable groups, such as viruses, bacteria and protists (Nanney 1982; Pace 1997; Allander *et al.* 2001; Hamels *et al.* 2001). However, the problems inherent in morphological taxonomy are general enough to merit the extension of this approach to all life. In fact, there are a growing number of cases in which DNA-based identification systems have been applied to higher organisms (Brown *et al.* 1999; Bucklin *et al.* 1999; Trewick 2000; Vincent *et al.* 2000).

Genomic approaches to taxon diagnosis exploit diversity among DNA sequences to identify organisms (Kurtzman 1994; Wilson 1995). In a very real sense, these sequences can be viewed as genetic 'barcodes' that are embedded in every cell. When one considers the discrimination of life's diversity from a combinatorial perspective, it is a modest problem. The Universal Product Codes, used to identify retail products, employ 10 alternate numerals at 11 positions to generate 100 billion unique identifiers. Genomic barcodes have only four alternate nucleotides at each position, but the string of sites available for inspection is huge. The survey of just 15 of these nucleotide positions creates the possibility of 4^{15} (1 billion) codes, 100 times the number that would be required to discriminate life if each taxon was uniquely branded. However, the survey of nucleotide diversity needs to be more comprehensive because functional constraints hold some nucleotide positions constant and intraspecific diversity exists at others. The impact of functional constraints can be reduced by focusing on a protein-coding gene, given that most shifts at the third nucleotide position of codons are weakly constrained by selection because of their four-fold degeneracy. Hence, by examining any stretch of 45 nucleotides, one gains access to 15 sites weakly affected by selection and, therefore, 1 billion possible identification labels. In practice, there is no need to constrain analysis to such short stretches of DNA because sequence information is easily obtained for DNA fragments hundreds of base pairs (bp) long. This ability to inspect longer sequences is significant, given two other biological considerations. First, nucleotide composition at third-position sites is often

*Author for correspondence (phebert@uoguelph.ca).

02PB0653.1

strongly biased (A–T in arthropods, C–G in chordates), reducing information content. However, even if the A–T or C–G proportion reached 1, the inspection of just 90 bp would recover the prospect of 1 billion alternatives ($2^{30} = 4^{15}$). The second constraint derives from the limited use of this potential information capacity, since most nucleotide positions are constant in comparisons of closely related species. However, given a modest rate (e.g. 2% per Myr) of sequence change, one expects 12 diagnostic nucleotide differences in a 600 bp comparison of species with just a million year history of reproductive isolation. As both the fossil record and prior molecular analyses suggest that most species persist for millions of years, the likelihood of taxon diagnosis is high. However, there is no simple formula that can predict the length of sequence that must be analysed to ensure species diagnosis, because rates of molecular evolution vary between different segments of the genome and across taxa. Obviously, the analysis of rapidly evolving gene regions or taxa will aid the diagnosis of lineages with brief histories of reproductive isolation, while the reverse will be true for rate-decelerated genes or species.

Although there has never been an effort to implement a microgenomic identification system on a large scale, enough work has been done to indicate key design elements. It is clear that the mitochondrial genome of animals is a better target for analysis than the nuclear genome because of its lack of introns, its limited exposure to recombination and its haploid mode of inheritance (Saccone *et al.* 1999). Robust primers also enable the routine recovery of specific segments of the mitochondrial genome (Folmer *et al.* 1994; Simmons & Weller 2001). Past phylogenetic work has often focused on mitochondrial genes encoding ribosomal (12S, 16S) DNA, but their use in broad taxonomic analyses is constrained by the prevalence of insertions and deletions (indels) that greatly complicate sequence alignments (Doyle & Gaut 2000). The 13 protein-coding genes in the animal mitochondrial genome are better targets because indels are rare since most lead to a shift in the reading frame. There is no compelling *a priori* reason to focus analysis on a specific gene, but the cytochrome *c* oxidase I gene (COI) does have two important advantages. First, the universal primers for this gene are very robust, enabling recovery of its 5′ end from representatives of most, if not all, animal phyla (Folmer *et al.* 1994; Zhang & Hewitt 1997). Second, COI appears to possess a greater range of phylogenetic signal than any other mitochondrial gene. In common with other protein-coding genes, its third-position nucleotides show a high incidence of base substitutions, leading to a rate of molecular evolution that is about three times greater than that of 12S or 16S rDNA (Knowlton & Weigt 1998). In fact, the evolution of this gene is rapid enough to allow the discrimination of not only closely allied species, but also phylogeographic groups within a single species (Cox & Hebert 2001; Wares & Cunningham 2001). Although COI may be matched by other mitochondrial genes in resolving such cases of recent divergence, this gene is more likely to provide deeper phylogenetic insights than alternatives such as cytochrome *b* (Simmons & Weller 2001) because changes in its amino-acid sequence occur more slowly than those in this, or any other, mitochondrial gene (Lynch & Jarrell 1993). As a result, by examining amino-acid substitutions, it may be possible to assign any unidentified organism to a higher taxonomic group (e.g. phylum, order), before examining nucleotide substitutions to determine its species identity.

This study evaluates the potential of COI as a taxonomic tool. We first created a COI 'profile' for the seven most diverse animal phyla, based on the analysis of 100 representative species, and subsequently showed that this baseline information assigned 96% of newly analysed taxa to their proper phylum. We then examined the class Hexapoda, selected because it represents the greatest concentration of biodiversity on the planet (Novotny *et al.* 2002). We created a COI 'profile' for eight of the most diverse orders of insects, based on a single representative from each of 100 different families, and showed that this 'profile' assigned each of 50 newly analysed taxa to its correct order. Finally, we tested the ability of COI sequences to correctly identify species of lepidopterans, a group targeted for analysis because sequence divergences are low among families in this order. As such, the lepidopterans provide a challenging case for species diagnosis, especially since this is one of the most speciose orders of insects. This test, which involved creating a COI 'profile' for 200 closely allied species and subsequently using it to assign 150 newly analysed individuals to species, was 100% successful in identification.

2. MATERIAL AND METHODS

(a) *Sequences*

Approximately one-quarter of the COI sequences (172 out of 655) used in this study were obtained from GenBank. The rest were obtained by preparing a 30 µl total DNA extract from small tissue samples using the Isoquick (Orca Research Inc, Bothell, WA, 1997) protocol. The primer pair LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TAAACTTCAGGGTGACCAAAAAATCA-3′) was subsequently used to amplify a 658 bp fragment of the COI gene (Folmer *et al.* 1994). Each PCR contained 5 µl of 10× PCR buffer, pH 8.3 (10 mM of Tris–HCl, pH 8.3; 1.5 mM of $MgCl_2$; and 50 mM of KCl; 0.01% NP-40), 35 µl of distilled water, 200 µM of each dNTP, 1 unit of Taq polymerase, 0.3 µM of each primer and 1–4 µl of DNA template. The PCR thermal regime consisted of one cycle of 1 min at 94 °C; five cycles of 1 min at 94 °C, 1.5 min at 45 °C and 1.5 min at 72 °C; 35 cycles of 1 min at 94 °C, 1.5 min at 50 °C and 1 min at 72 °C and a final cycle of 5 min at 72 °C. Each PCR product was subsequently gel purified using the Qiaex II kit (Qiagen) and sequenced in one direction on an ABI 377 automated sequencer (Applied Biosystems) using the Big Dye v. 3 sequencing kit. All sequences obtained in this study have been submitted to GenBank; their accession numbers are provided in Electronic Appendices A–C, available on The Royal Society's Publications Web site.

(b) *COI profiles*

We created three COI profiles: one for the seven dominant phyla of animals, another for eight of the largest orders of insects and the last for 200 closely allied species of lepidopterans. These profiles were designed to provide an overview of COI diversity within each taxonomic assemblage and were subsequently used as the basis for identifications to the phylum, ordinal or species level by determining the sequence congruence between each 'unknown' taxon and the species included in a particular profile.

The phylum profile included 100 COI sequences, all obtained from GenBank (see electronic Appendix A available on The Royal Society's Publications Web site). To ensure broad taxonomic coverage, each sequence was derived from a different family and representatives were included from all available classes. Ten sequences were obtained for each of the five phyla (Annelida, Chordata, Echinodermata, Nematoda, Platyhelminthes) that include 5000–50 000 species, while 25 sequences were collected for each of the phyla (Arthropoda, Mollusca) with more than 100 000 species.

The ordinal profile was created by obtaining a COI sequence from a single representative of each of 100 insect families (see electronic Appendix B available on The Royal Society's Publications Web site). The four most diverse orders (more than 100 000 described species) of insects were selected for analysis, together with four additional orders chosen randomly from among the 15 insect orders (Gaston & Hudson 1994) with medium diversity (1000–15 000 described species). Between ten and 25 families were examined for each of the four most diverse orders (Coleoptera, Diptera, Hymenoptera, Lepidoptera), while four to ten families were examined for the other orders (Blattaria, Ephemeroptera, Orthoptera, Plecoptera).

The species profile was based upon COI data for a single individual from each of the 200 commonest lepidopteran species from a site near Guelph, Ontario, Canada (electronic Appendix C available on The Royal Society's Publications Web site). This profile examined members of just three allied superfamilies (Geometroidea, Noctuoidea, Sphingoidea) to determine the lower limits of COI divergence in an assemblage of closely related species. The Noctuoidea included members of three families (Arctiidae, Noctuidae, Notodontidae) while the others included representatives from just a single family each (Geometridae and Sphingidae, respectively).

(c) *Test taxa*

Additional sequences were collected to test the ability of each profile to assign newly analysed species to a taxonomic category (electronic Appendices A–C). COI sequences were obtained from 55 'test' taxa to assess the success of the phylum profile in assigning newly analysed species to a phylum. These 'test' taxa included five representatives from each of the five 'small' phyla, and 15 representatives from both the Mollusca and the Arthropoda. When possible, the 'test' taxa belonged to families that were not included in the phylum profile. A similar approach was employed to test the ability of the ordinal profile to classify newly analysed insects to an order. Fifty new taxa were examined, including between one and five representatives from each small order and five to ten representatives from each of the four large orders. When possible, the 'test' taxa belonged to families or genera that were not included in the ordinal profile. A test of the species profile required a slightly different approach, as identifications were only possible for species represented in it. As a result, sequences were obtained from another 150 individuals belonging to the species included in this profile.

(d) *Data analysis*

Sequences were aligned in the SeqApp 1.9 sequence editor. They were subsequently reduced to 669 bp for the phylum analysis, 624 bp for the ordinal analysis and 617 bp for the species-level analysis. Analyses at the ordinal and phylum levels examined amino-acid divergences, using Poisson corrected *p*-distances to reduce the impacts of homoplasy. For the species-level analysis, nucleotide-sequence divergences were calculated using the Kimura-two-parameter (K2P) model, the best metric when distances are low (Nei & Kumar 2000) as in this study.

Neighbour-joining (NJ) analysis, implemented in MEGA2.1 (Kumar *et al.* 2001), was employed to both examine relationships among taxa in the profiles and for the subsequent classification of 'test' taxa because of its strong track record in the analysis of large species assemblages (Kumar & Gadagkar 2000). This approach has the additional advantage of generating results much more quickly than alternatives. The NJ profiles for both the orders and the phyla possessed 100 terminal nodes, each representing a species from a different family, while the species NJ profile had 200 nodes, each representing a different lepidopteran species. A member of the primitive insect order Thysanura (family Lepismatidae) was used as the outgroup for the insect profile, while single members of three primitive lepidopteran families were employed as the outgroup for the species profile.

Each of the three (phylum, order, species) NJ profiles was subsequently used as a classification engine, by re-running the analysis with the repeated addition of a single 'test' taxon to the dataset. Following each analysis, the 'test' species was assigned membership of the same taxonomic group as its nearest neighbouring node. For example, in the ordinal analysis, a 'test' taxon was identified as a member of the order Lepidoptera if it grouped most closely with any of the 24 lepidopteran families included in the profile. The success of classification was quantified for both the phylum and the ordinal analyses by determining the proportion of 'test' taxa assigned to the proper phylum/order. In the case of species, a stricter criterion was employed. A 'test' taxon was recognized as being correctly identified only if its sequence grouped most closely with the single representative of its species in the profile.

Multidimensional scaling (MDS), implemented in SYSTAT 8.0, was employed to provide a graphical summary of the species-level results because of the very large number of taxa. MDS explores similarity relationships in Euclidean space and has the advantage of permitting genetically intermediate taxa to remain spatially intermediate, rather than forcing them to cluster into a pseudogroup as in hierarchical methods (Lessa 1990). In the present case, a similarity matrix was constructed by treating every position in the alignment as a separate character and ambiguous nucleotides as missing characters. The sequence information was coded using dummy variables (A = 1, G = 2, C = 3, T = 4). However, as noted earlier, a NJ profile was also constructed for the lepidopteran sequences using K2P distances and this is provided in electronic Appendix D available on The Royal Society's Publications Web site.

3. RESULTS

(a) *Taxon profiles*

Each of the 100 species included in the phylum and ordinal profiles possessed a different amino-acid sequence at COI. The phylum profile showed good resolution of the major taxonomic groups (figure 1). Monophyletic assemblages were recovered for three phyla (Annelida, Echinodermata, Platyhelminthes) and the chordate lineages formed a cohesive group. Members of the Nematoda were separated into three groups, but each corresponded to one of the three subclasses that comprise this phylum. Twenty-three out of the 25 arthropods formed a monophyletic group, but the sole representatives of two crustacean classes (Cephalocarida, Maxillopoda) fell outside this group. Twelve out of the 25 molluscan lin-

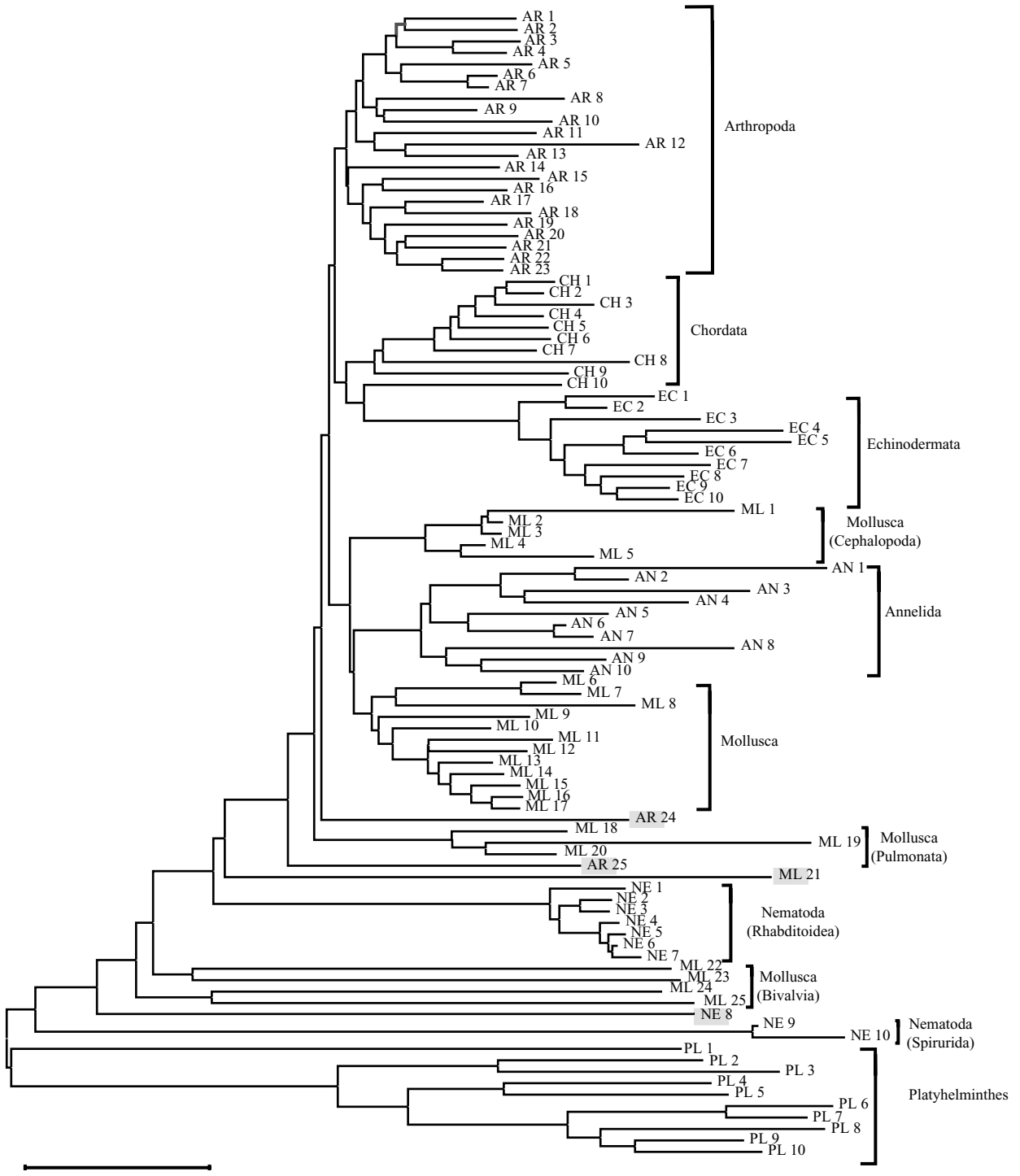

Figure 1. NJ analysis of Poisson corrected *p*-distances based on the analysis of 223 amino acids of the COI gene in 100 taxa belonging to seven animal phyla. The taxa in the grey boxes represent outliers. AR 24 and AR 25 are the sole representatives of the arthropod classes Cephalocarida and Maxillopoda, respectively. ML 21 is a member of the molluscan class Bivalvia, while NE 8 is the sole member of the nematode subclass Enoplia. Scale bar, 0.1.

eages formed a monophyletic assemblage allied to the annelids, but the others were separated into groups that showed marked genetic divergence. One group consisted solely of cephalopods, a second was largely pulmonates and the rest were bivalves. It is worth emphasizing that these outlying COI sequences always showed considerable amino-acid divergence from sequences possessed by other taxonomic groups. As such, the rate acceleration in these lineages generated novel COI amino-acid sequences rather than secondary convergence on the amino-acid arrays of other groups.

The ordinal profile showed high cohesion of taxonomic groups with seven out of the eight orders forming monophyletic assemblages (figure 2). The sole exception was the Coleoptera whose members were partitioned into three groups. Two of these groups included 21 families belonging to the very diverse suborder Polyphaga, while the other group included four families belonging to the

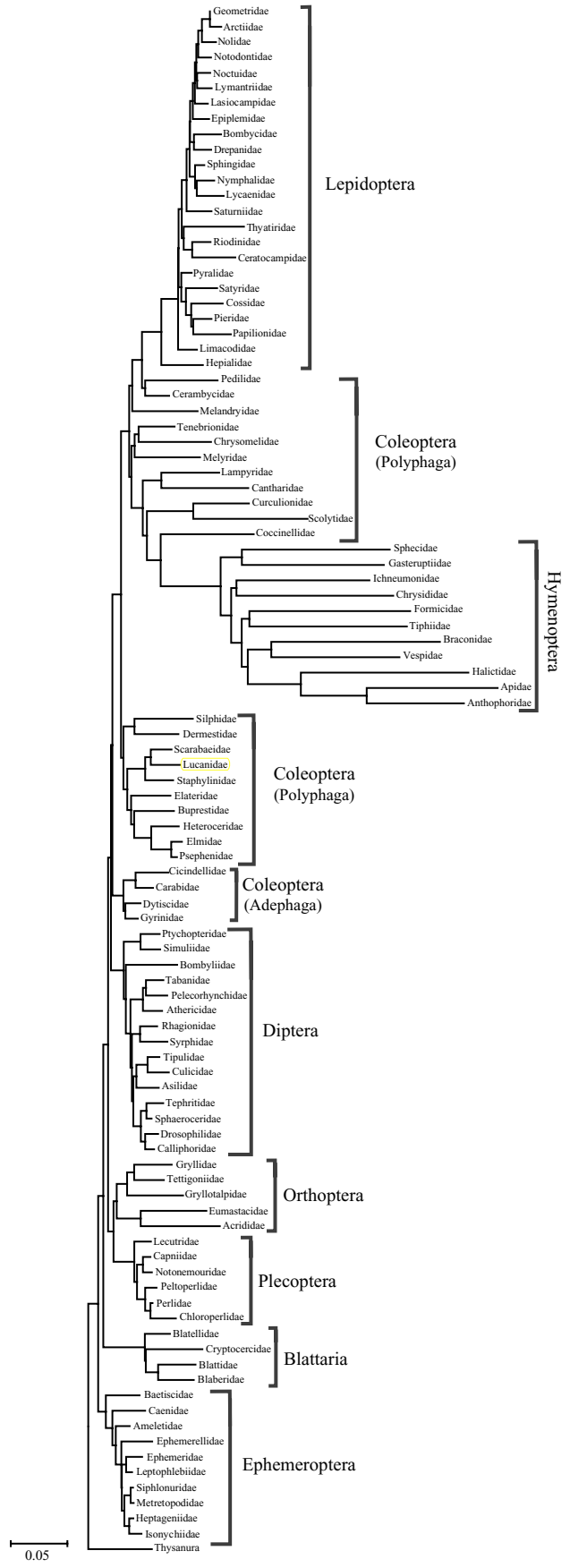

Figure 2. NJ analysis of Poisson corrected *p*-distances based on the analysis of 208 amino acids in 100 taxa belonging to eight insect orders. Scale bar, 0.05.

Table 1. Mean Poisson corrected *p*-distances (*d*) for 208 amino acids of COI in 100 insect families belonging to eight orders.
(*n* indicates the number of families analysed in each order. G/C content is also reported.)

order	*n*	*d*	s.e.	G/C (%)
Hymenoptera	11	0.320	0.028	27.7
Coleoptera	25	0.125	0.015	35.6
Orthoptera	5	0.119	0.019	35.4
Blattaria	4	0.076	0.014	35.7
Diptera	15	0.055	0.011	34.1
Lepidoptera	24	0.054	0.009	31.0
Ephemeroptera	10	0.036	0.008	40.5
Plecoptera	6	0.031	0.008	39.5

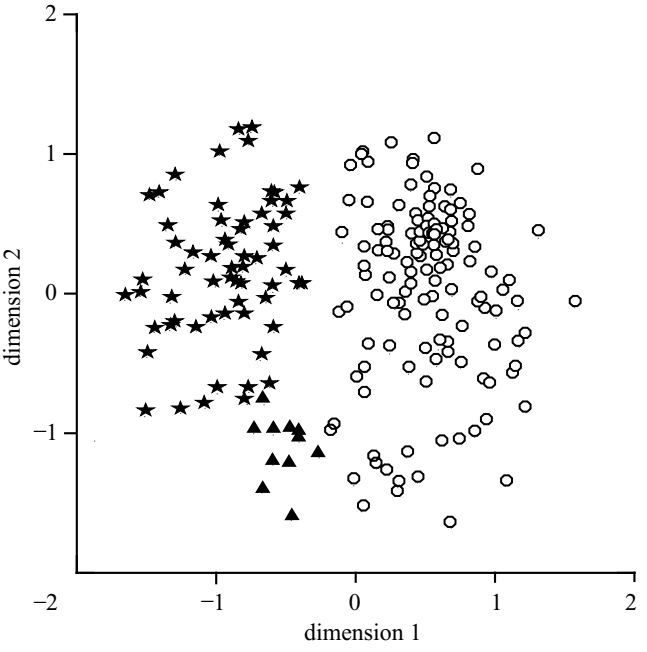

Figure 3. Multidimensional scaling of Euclidian distances among the COI genes from 200 lepidopteran species belonging to three superfamilies: Geometroidea (stars); Sphingoidea (triangles) and Noctuoidea (circles).

suborder Adephaga. Among the four major orders, species of Diptera and Lepidoptera showed much less variation in their amino-acid sequences than did the Hymenoptera, while the Coleoptera showed an intermediate level of divergence (table 1).

Each of the 200 lepidopterans included in the species profile possessed a distinct COI sequence. Moreover, a MDS plot showed that species belonging to each of the three superfamilies fell into distinct clusters (figure 3), signalling their genetic divergence. A detailed inspection of the NJ tree (electronic Appendix D) revealed further evidence for the clustering of taxonomically allied species. For example, 23 genera were represented by two species, and these formed monophyletic pairs in 18 cases. Similarly, five out of the six genera represented by three species formed monophyletic assemblages.

Table 2. Percentage success in classifying species to membership of a particular taxonomic group based upon sequence variation at COI.
(*n* indicates the number of taxa that were classified using each taxon 'profile'.)

taxon	target group	*n*	% success
kingdom Animalia	7 phyla	55	96.4
class Hexapoda	8 orders	50	100
order Lepidoptera	200 species	150	100

(b) *Testing taxonomic assignments*

Fifty-three out of the 55 'test' species (96.4%) were assigned to the correct phylum in the analyses at this level (table 2). The exceptions were a polychaete annelid that grouped most closely with a mollusc and a bivalve that grouped with one of the arthropod outliers. However, in both cases, there was substantial sequence divergence (13% and 25%, respectively) between the test taxon and the lineage in the profile that was most similar to it. Identification success at the ordinal level was 100% as all 50 insect species were assigned to the correct order. Moreover, when a 'test' species belonged to a family represented in the ordinal profile, it typically grouped most closely with it. Identification success at the species level was also 100%, as each of the 150 'test' individuals clustered most closely with its conspecific in the profile. The sequences in the species profile were subsequently merged with those from the 'test' taxa to allow a more detailed examination of the factors enabling successful classification (electronic Appendix E available on The Royal Society's Publications Web site). MDS analysis (figure 4) showed that 'test' taxa were always either genetically identical to or most closely associated with their conspecific in the profile. Examination of the genetic distance matrix quantified this fact, showing that divergences between conspecific individuals were always small, the family values averaging 0.25% (table 3). By contrast, sequence divergences between species were much greater, averaging 6.8% for congeneric taxa and higher for more distantly related taxa (table 3). A few species pairs showed lower values, but only four out of the 19 900 pairwise comparisons showed divergences that were less than 3%. Figure 4*a* shows one of these cases, involving two species of *Hypoprepia*, but, even in these situations, there were no shared sequences between taxa.

4. DISCUSSION

This study establishes the feasibility of developing a COI-based identification system for animals-at-large. PCR products were recovered from all species and there was no evidence of the nuclear pseudogenes that have complicated some studies employing degenerate COI primers (Williams & Knowlton 2001). Moreover, the alignment of COI sequences was straightforward, as indels were uncommon, reinforcing the results of earlier work showing the rarity of indels in this gene (Mardulyn & Whitfield 1999). Aside from their ease of acquisition and alignment, the COI sequences possessed, as expected, a high level of diversity.

We demonstrated that differences in COI amino-acid sequences were sufficient to enable the reliable assignment of organisms to higher taxonomic categories. It is worth emphasizing that most newly analysed taxa were placed in the correct order or phylum despite the fact that our profiles were based on a tiny fraction of the member species. For example, our ordinal profile, which was based on just 0.002% of the total species in these orders, led to 100% identification success. The two misidentifications at the phylum level were undoubtedly a consequence of the limited size and diversity of our phylum profile. The misplaced polychaete belonged to an order that was not in the profile, while the misidentified mollusc belonged to a subclass that was represented in the profile by just a single species. Such misidentifications would not occur in profiles that more thoroughly surveyed COI diversity among members of the target assemblage. The general success of COI in recognizing relationships among taxa in these cases is important because it signals that character convergence and horizontal gene transfer (i.e. via retroviruses) have not disrupted the recovery of expected taxon affinities. Moreover, it establishes that the information content of COI is sufficient to enable the placement of organisms in the deepest taxonomic ranks.

The gold standard for any taxonomic system is its ability to deliver accurate species identifications. Our COI species profile was 100% successful in identifying lepidopteran species, and we expect similar results in other groups, since the Lepidoptera are one of the most taxonomically diverse orders of animals and they show low sequence divergences. There is also reason to expect successful diagnosis at other locales, as the species richness of lepidopterans at the study site exceeds that which will be encountered in regional surveys of most animal orders. Higher diversities will be encountered for some orders in tropical settings (Godfray *et al.* 1999), but COI diagnoses should not fail at these sites unless species are unusually young.

COI-based identification systems can also aid the initial delineation of species. For example, inspection of the genetic distance matrix for lepidopterans indicated that divergence values between species are ordinarily greater than 3%. In fact, when this value was employed as a threshold for species diagnosis, it led to the recognition of 196 out of the 200 (98%) species recognized through prior morphological study. The exceptions were four congeneric species pairs that were genetically distinct but showed low (0.6–2.0%) divergences, suggesting their recent origin. The general ease of species diagnosis reveals one of the great values of a DNA-based approach to identification. Newly encountered species will ordinarily signal their presence by their genetic divergence from known members of the assemblage.

The prospect of using a standard COI threshold to guide species diagnosis in situations where prior taxonomic work has been limited is appealing. It is, however, important to validate this approach by determining the thresholds that distinguish species in other geographical regions and taxonomic groups. Thresholds will particularly need to be established for groups with differences in traits, such as generation length or dispersal regime, that are likely to alter rates of molecular evolution or the extent of population subdivision. However, differences in

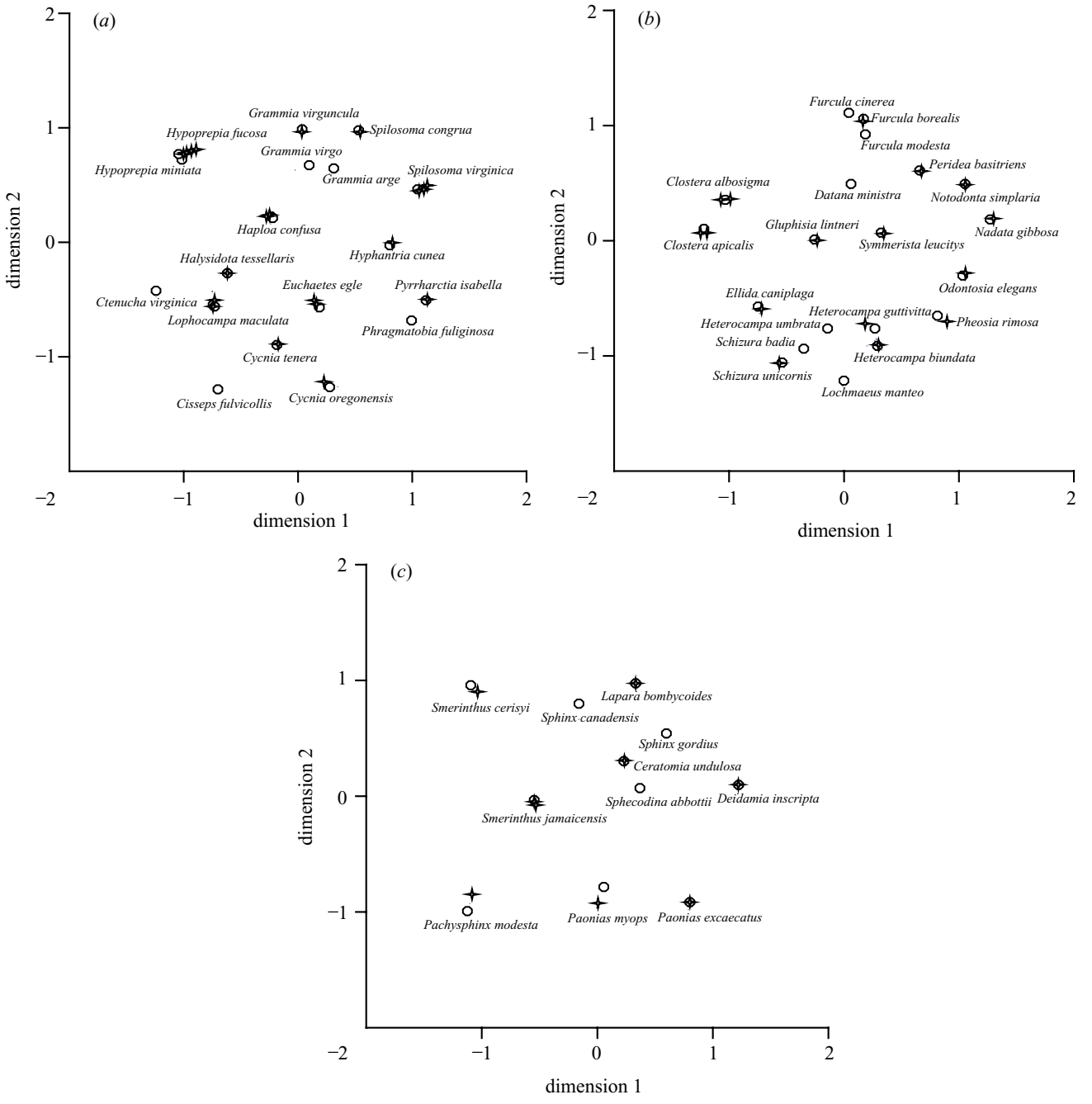

Figure 4. Multidimensional scaling of Euclidian distances among the COI genes from (*a*) 18 species of Arctiidae, (*b*) 20 species of Notodontidae and (*c*) 11 species of Sphingidae. Circles identify the single representatives of each species included in the profile, while the crosses mark the position of 'test' individuals. Profile and test individuals from the same species always grouped together.

thresholds may be smaller than might be expected. For example, different species of vertebrates ordinarily show more than 2% sequence divergence at cytochrome *b* (Avise & Walker 1999), a value close to the 3% COI threshold adopted for lepidopterans in this study.

The likely applicability of a COI identification system to new animal groups and geographical settings suggests the feasibility of creating an identification system for animals-at-large. Certainly, existing primers enable recovery of this gene from most, if not all, animal species and its sequences are divergent enough to enable recognition of all but the youngest species. It is, of course, impossible for any mitochondrially based identification system to resolve fully the complexity of life. Where species boundaries are blurred by hybridization or introgression, supplemental analyses of one or more nuclear genes will be required. Similarly, when species have arisen through polyploidization, determinations of genome size may be needed. While protocols will be required to deal with such complications, a COI-based identification system will undoubtedly provide taxonomic resolution that exceeds that which can be achieved through morphological studies. Moreover, the generation of COI profiles will provide a partial solution to the problem of the thinning ranks of morphological taxonomists by enabling a crystallization of their knowledge before they leave the field. Also, since COI sequences can be obtained from museum specimens without their destruction, it will be possible to regain taxo-

Table 3. Percentage nucleotide sequence divergence (K2P distances) at COI between members of five lepidopteran families at three levels of taxonomic affinity.
(At the species level, *n* indicates the number of species for which two or more individuals were analysed. At a generic level, *n* represents the number of genera with two or more species, while at the family level it indicates the total number of species that were analysed.)

family	*n*	within species	*n*	within genus	*n*	within family
Arctiidae	13	0.33	4	7.0	18	10.0
Geometridae	30	0.23	10	9.1	61	12.5
Noctuidae	42	0.17	12	5.8	90	10.4
Notodontidae	14	0.36	4	5.9	20	12.4
Sphingidae	8	0.17	3	6.4	11	10.5

nomic capability, albeit in a novel format, for groups that currently lack an authority.

We believe that a COI database can be developed within 20 years for the 5–10 million animal species on the planet (Hammond 1992; Novotny *et al.* 2002) for approximately $1 billion, far less than that directed to other major science initiatives such as the Human Genome project or the International Space Station. Moreover, initial efforts could focus on species of economic, medical or academic importance. Data acquisition is now simple enough for individual laboratories to gather, in a single year, COI profiles for 1000 species, a number greater than that in many major taxonomic groups on a continental scale. Once completed, these profiles will be immediately cost-effective in many taxonomic contexts, and innovations in sequencing technology promise future reductions in the cost of DNA-based identifications.

If advanced comprehensively, a COI database could serve as the basis for a global bioidentification system (GBS) for animals. Implementation on this scale will require the establishment of a new genomics database. While GenBank aims for comprehensive coverage of genomic diversity, the GBS database would aim for comprehensive taxonomic coverage of just a single gene. Through web-based delivery, this system could provide easy access to taxonomic information, a particular benefit to developing nations. Its adoption by an organization such as the Global Biodiversity Information Facility or the All Species Foundation would be an important step towards ensuring the longevity lacking in many web-based resources. Once established, this microgenomic identification system will overcome the deficits of morphological approaches to species discrimination: the bounds of intraspecific diversity will be quantifiable, sibling species will be recognizable, taxonomic decisions will be objective and all life stages will be identifiable. Moreover, once complete, the GBS will allow single laboratories to execute taxon diagnoses across the full spectrum of animal life. The creation of the GBS will be a substantial undertaking and will require close alliances between molecular biologists and taxonomists. However, its assembly promises both a revolution in access to basic biological information and a newly detailed view of the origins of biological diversity.

This work was supported by grants from NSERC and the Canada Research Chairs Program to P.D.N.H. Teri Crease, Melania Cristescu, Derek Taylor, Jonathan Witt and two reviewers provided helpful comments on earlier drafts of this manuscript. The authors thank Win Bailey, Klaus Bolte, Steve Burian, Don Klemm, Don Lafontaine, Steve Marshall, Christine Nalepa, Jeff Webb and Jack Zloty for either providing specimens or verifying taxonomic assignments. They also thank Lisa Schieman, Tyler Zemlak, Heather Cole and Angela Holliss for their assistance with the DNA analyses.

REFERENCES

Allander, T., Emerson, S. U., Engle, R. E., Purcell, R. H. & Bukh, J. 2001 A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc. Natl Acad. Sci. USA* **98**, 11 609–11 614.

Avise, J. C. & Walker, D. 1999 Species realities and numbers in sexual vertebrates: perspectives from an asexually transmitted genome. *Proc. Natl Acad. Sci. USA* **96**, 992–995.

Brown, B., Emberson, R. M. & Paterson, A. M. 1999 Mitochondrial COI and II provide useful markers for *Weiseana* (Lepidoptera, Hepialidae) species identification. *Bull. Entomol. Res.* **89**, 287–294.

Bucklin, A., Guarnieri, M., Hill, R. S., Bentley, A. M. & Kaartvedt, S. 1999 Taxonomic and systematic assessment of planktonic copepods using mitochondrial COI sequence variation and competitive, species-specific PCR. *Hydrobiology* **401**, 239–254.

Cox, A. J. & Hebert, P. D. N. 2001 Colonization, extinction and phylogeographic patterning in a freshwater crustacean. *Mol. Ecol.* **10**, 371–386.

Doyle, J. J. & Gaut, B. S. 2000 Evolution of genes and taxa: a primer. *Plant Mol. Biol.* **42**, 1–6.

Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. 1994 DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**, 294–299.

Gaston, K. J. & Hudson, E. 1994 Regional patterns of diversity and estimates of global insect species richness. *Biodivers. Conserv.* **3**, 493–500.

Godfray, H. C. J., Lewis, O. T. & Memmott, J. 1999 Studying insect diversity in the tropics. *Phil. Trans. R. Soc. Lond.* B **354**, 1811–1824. (DOI 10.1098/rstb.1999.0523.)

Hamels, J., Gala, L., Dufour, S., Vannuffel, P., Zammatteo, N. & Remacle, J. 2001 Consensus PCR and microarray for diagnosis of the genus *Staphylococcus*, species, and methicillin resistance. *BioTechniques* **31**, 1364–1372.

Hammond, P. 1992 Species inventory. In *Global biodiversity: status of the earth's living resources* (ed. B. Groombridge), pp. 17–39. London: Chapman & Hall.

Hawksworth, D. L. & Kalin-Arroyo, M. T. 1995 Magnitude and distribution of biodiversity. In *Global biodiversity assessment* (ed. V. H. Heywood), pp. 107–191. Cambridge University Press.

Jarman, S. N. & Elliott, N. G. 2000 DNA evidence for mor-

phological and cryptic Cenozoic speciations in the Anaspididae, 'living fossils' from the Triassic. *J. Evol. Biol.* **13**, 624–633.

Knowlton, N. 1993 Sibling species in the sea. *A. Rev. Ecol. Syst.* **24**, 189–216.

Knowlton, N. & Weigt, L. A. 1998 New dates and new rates for divergence across the Isthmus of Panama. *Proc. R. Soc. Lond.* B **265**, 2257–2263. (DOI 10.1098/rspb.1998.0568.)

Kumar, S. & Gadagkar, S. R. 2000 Efficiency of the neighbour-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. *J. Mol. Evol.* **51**, 544–553.

Kumar, S., Tamura, K., Jacobsen, I. B. & Nei, M. 2001 *MEGA2: molecular evolutionary genetics analysis software*. Tempe, AZ: Arizona State University.

Kurtzman, C. P. 1994 Molecular taxonomy of the yeasts. *Yeast* **10**, 1727–1740.

Lessa, P. 1990 Multidimensional scaling of geographic genetic structure. *Syst. Zool.* **39**, 242–252.

Lynch, M. & Jarrell, P. E. 1993 A method for calibrating molecular clocks and its application to animal mitochondrial DNA. *Genetics* **135**, 1197–1208.

Mardulyn, P. & Whitfield, J. B. 1999 Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera: Braconidae): evidence of a high diversification rate in this group of parasitoids. *Mol. Phylogenet. Evol.* **12**, 282–294.

Nanney, D. L. 1982 Genes and phenes in *Tetrahymena*. *Bioscience* **32**, 783–788.

Nei, M. & Kumar, S. 2000 *Molecular evolution and phylogenetics*. Oxford University Press.

Novotny, V., Baset, Y., Miller, S. E., Weiblen, G. D., Bremer, B., Cizek, L. & Drezel, P. 2002 Low host specificity of herbivorous insects in a tropical forest. *Nature* **416**, 841–845.

Pace, N. R. 1997 A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740.

Saccone, C., DeCarla, G., Gissi, C., Pesole, G. & Reynes, A. 1999 Evolutionary genomics in the Metazoa: the mitochondrial DNA as a model system. *Gene* **238**, 195–210.

Simmons, R. B. & Weller, S. J. 2001 Utility and evolution of cytochrome *b* in insects. *Mol. Phylogenet. Evol.* **20**, 196–210.

Trewick, S. A. 2000 Mitochondrial DNA sequences support allozyme evidence for cryptic radiation of New Zealand *Peripatoides* (Onychophora). *Mol. Ecol.* **9**, 269–282.

Vincent, S., Vian, J. M. & Carlotti, M. P. 2000 Partial sequencing of the cytochrome oxidase-b subunit gene. I. A tool for the identification of European species of blow flies for post mortem interval estimation. *J. Forensic Sci.* **45**, 820–823.

Wares, J. P. & Cunningham, C. W. 2001 Phylogeography and historical ecology of the North Atlantic intertidal. *Evolution* **12**, 2455–2469.

Williams, S. T. & Knowlton, N. 2001 Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp *Alpheus*. *Mol. Biol. Evol.* **18**, 1484–1493.

Wilson, K. H. 1995 Molecular biology as a tool for taxonomy. *Clin. Infect. Dis.* **20**(Suppl.), 192–208.

Zhang, D.-X. & Hewitt, G. M. 1997 Assessment of the universality and utility of a set of conserved mitochondrial primers in insects. *Insect Mol. Biol.* **6**, 143–150.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.

Visit http://www.pubs.royalsoc.ac.uk to see electronic appendices to this paper.