Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers

Michael T Monaghan, Michael Balke, T. Ryan Gregory and Alfried P Vogler

Phil. Trans. R. Soc. B 2005 **360**, 1925-1933
doi: 10.1098/rstb.2005.1724

References

This article cites 33 articles, 6 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1925.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1925.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. B* go to: **http://rstb.royalsocietypublishing.org/subscriptions**

This journal is © 2005 The Royal Society

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2005) **360**, 1925–1933
doi:10.1098/rstb.2005.1724
Published online 12 September 2005

DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers

Michael T. Monaghan[1,2], Michael Balke[1,†], T. Ryan Gregory[3] and Alfried P. Vogler[1,2,*]

[1]*Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK*
[2]*Division of Biology, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK*
[3]*Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada N1G 2W1*

DNA barcoding has been successfully implemented in the identification of previously described species, and in the process has revealed several cryptic species. It has been noted that such methods could also greatly assist in the discovery and delineation of undescribed species in poorly studied groups, although to date the feasibility of such an approach has not been examined explicitly. Here, we investigate the possibility of using short mitochondrial and nuclear DNA sequences to delimit putative species in groups lacking an existing taxonomic framework. We focussed on poorly known tropical water beetles (Coleoptera: Dytiscidae, Hydrophilidae) from Madagascar and dung beetles (Scarabaeidae) in the genus *Canthon* from the Neotropics. Mitochondrial DNA sequence variation proved to be highly structured, with >95% of the observed variation existing between discrete sets of very closely related genotypes. Sequence variation in nuclear 28S rRNA among the same individuals was lower by at least an order of magnitude, but 16 different genotypes were found in water beetles and 12 genotypes in *Canthon*, differing from each other by a minimum of two base pairs. The distribution of these 28S rRNA genotypes in individuals exactly matched the distribution of mtDNA clusters, suggesting that mtDNA patterns were not misleading because of introgression. Moreover, in a few cases where sequence information was available in GenBank for morphologically defined species of *Canthon*, these matched some of the DNA-based clusters. These findings demonstrate that clusters of close relatives can be identified readily in the sequence variation obtained in field collected samples, and that these clusters are likely to correspond to either previously described or unknown species. The results suggest that DNA-assisted taxonomy will not require more than a short fragment of mtDNA to provide a largely accurate picture of species boundaries in these groups. Applied on a large scale, this DNA-based approach could greatly improve the rate of species discovery in the large assemblages of insects that remain undescribed.

Keywords: taxonomy; large subunit ribosomal RNA; Madagascar; DNA barcoding; *cox1*; COI

1. INTRODUCTION

A reliable and accessible classification of species is fundamental to research in ecology, evolutionary biology, biodiversity and conservation biology. While *ca* 1.5 million species have been described to date, this represents only a fraction of the actual diversity on Earth (Tudge 2000; Wilson 2003). Owing to the constant threat of biodiversity loss, there is an increasingly urgent need to accelerate the pace of species discovery and taxonomic databasing (Godfray 2002). Even the routine identification of known species can be difficult, often requiring highly specialized knowledge and representing a limiting factor in ecological studies and biodiversity inventories. In response, recent proposals have called for a more prominent role of efficient DNA-based methods in the delineation and identification of species (Blaxter 2004; Floyd *et al.* 2002; Hebert *et al.* 2003*a*; Tautz *et al.* 2003). Reactions to such proposals have ranged widely, from strongly supportive (Janzen 2004; Proudlove & Wood 2003; Stoeckle 2003) to vigorously opposed (Lipscomb *et al.* 2003; Seberg *et al.* 2003; Wheeler 2004; Will & Rubinoff 2004). The use of DNA-based methods for the delineation and discovery of new species, and thus their broader role in taxonomy, represents an especially contentious issue in this regard. Unfortunately, much of this debate has remained rhetorical, with limited empirical assessment of the benefits and limitations of a DNA-assisted programme of species discovery.

The objective of any method of species delineation, including DNA-based approaches, is to identify reproductively isolated groups of organisms that warrant classification as distinct species. It is widely acknowledged, and reflected in the Linnaean taxonomic system, that living organisms fall into largely discrete groupings recognizable by differences in morphology or other traits. It is then the role of taxonomy to define and name these groupings.

* Author for correspondence (apv@nhm.ac.uk).
† Present address: Zoologische Staatssammlung, Muenchhausenstrasse 21, 81247 Munich, Germany.

One contribution of 18 to a Theme Issue 'DNA barcoding of life'.

© 2005 The Royal Society

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

However, their recognition may be difficult because diagnostic traits are lacking, or species divergences are very small. Hence, even for cases of biologically distinct species, their delineation will depend on the thoroughness of study and the interpretation of complex, sometimes variable traits. Further, the accuracy of species delineation depends on the degree of sampling, as local variation may affect the conclusions about population separation (Davis & Nixon 1992).

To date, DNA barcoding studies have focussed on the identification of pre-defined species (e.g. Hebert *et al.* 2003*a*, 2004*b*; Hogg & Hebert 2004; Vences *et al.* 2005; Smith *et al.* 2005, etc.), and have yet to address the issue of species delineation *per se*. However, where more than a single individual per species has been sequenced, a minimum threshold of approximately one-tenth of the average *p*-distance found between well-established species in a lineage has been interpreted as intra-specific variation, while greater divergences are thought to indicate misidentifications of specimens or overlooked cryptic species (Hebert *et al.* 2004*a*). These cut-off values roughly correspond to maximum intra-specific divergences in mtDNA of 1–2%, and at the upper bound of this range may include several geographically defined 'phylogroups' (Avise & Walker 1999).

The discovery of new, cryptic species from existing, morphologically indiscriminate groups using DNA is neither controversial nor novel (Knowlton 1993), and potential taxonomic revisions inspired by DNA barcoding results have been typically left to experts to resolve on the basis of morphology, behaviour and other features (Hebert *et al.* 2004*b*). It is to be expected that a successful global DNA barcoding program would provide a comprehensive barcoding inventory for a majority of described taxa in the foreseeable future, facilitating the systematic discovery of cryptic species. However, with 85% or more of species still unknown to science, a much greater challenge lies in the potential application of DNA-based methods to the discovery and delineation of new species in poorly characterized taxa.

To explore the utility of DNA-based approaches to species recognition in poorly known groups, we evaluated patterns of variation in both mitochondrial and nuclear genes in a broad sample of water beetles collected from Madagascar. These samples exhibited an unknown level of species diversity within the families Dytiscidae and Hydrophilidae. Using the same methods, we also examined specimens from a single genus of Neotropical dung beetles (*Canthon*). Specimens were collected from various localities in the Neotropics, comprising an unknown number of species in a group that is acknowledged to present difficulties for morphological discrimination. The analysis shows that sequences cluster into cohesive, well-differentiated groups, and identical groups are recovered by both nuclear and mitochondrial markers. Based on multiple lines of evidence, these DNA-based clusters are taken to represent putative species boundaries and could assist with the assembly of a framework for the taxonomy of poorly studied lineages.

2. METHODS

(a) *Field sampling, selection of specimens and DNA sequencing*

Water beetles (Dytiscidae and Hydrophilidae) were collected at five sites in the North and central parts of Madagascar as part of a survey of insect biodiversity in 2004 (Monaghan *et al.*, unpublished). Specimens were collected by sieving through small stream pools, ponds, and packs of leaf litter, and were sorted under a dissecting microscope (10×) into externally distinct morphological groups. Between two and five individuals from each group and each locality were selected for DNA analysis, as a representative sample of the variation of this group. This initial morphological treatment was superficial and was intended to maximize the disparity included in the subset of samples used for sequencing. Specimens of *Canthon* were obtained using baited pitfall traps from locations in Belize, French Guyana, Ecuador and Costa Rica between 1997 and 2001 (Inward 2003). Additional samples were collected from Belize in 2004 (L. Powell, MSc, Imperial College London, 2004). They were assigned to the genus *Canthon* based on a phylogenetic analysis combining them with unpublished sequences for most major groups of Scarabaeinae that also included five species of *Canthon*: *C. doesburgi*, *C. indigaceus*, *C luteicollis*, *C. smaragdulus* and *C. viridis*, plus the closely related *Scybalocanthon pygidialis* (GenBank accessions: AY131633-7 for 28S, AY131814-7 for *cox1*, plus AY131673 and AY131849 for *Scybalocanthon*).

Genomic DNA extraction was performed using Wizard SV 96-well plates (Promega, UK). For both groups, a *ca* 700 bp fragment of 28S rRNA was amplified using primers FF and DD (Inward 2003). Fragments of *cox1* were amplified with primers Pat and Jerry (Simon *et al.* 1994) for *Canthon* (800 bp) or with LCO1490 and HCO2198 (Folmer *et al.* 1994) for water beetles (660 bp). Sequencing was performed in both directions using a BigDye v. 2.1 terminator reaction with the same primers used for PCR. Sequences were analysed on an ABI3730 automated sequencer and forward and reverse strands were assembled in Sequencher software. The 28S fragment was length-variable and was aligned separately for the two datasets using BlastAlign (Belshaw & Katzourakis 2005). *Cox1* was not length-variable for either group.

(b) *Tree construction*

Parsimony trees were obtained with PAUP v. 4b10 (Swofford 2002), with gap characters treated as a 'fifth character state', and branch length optimized under accelerated transformation. Heuristic searches were performed using TBR branch swapping and 100 replicates. We performed 1000 random addition replicates saving only a single tree in each case. Because the dataset contained many identical or very similar haplotypes, a large number of trees were found, one of which was selected arbitrarily for further analysis. To calculate Bremer Support (Bremer 1994), constraint files for parsimony searches enforcing the absence of the focal nodes were produced with TreeRot v. 2.0a. Bremer Support values of 0 indicate unresolved nodes which would be collapsed in a strict consensus of all shortest trees. Trees were rooted with sequences from related Carabidae taken from GenBank in the case of water beetles, and using a sequence from the *Canthon* dataset generated here for rooting the 28S tree. The single species of *Scybalocanthon* was used as the outgroup to root the tree of *Canthon*.

(c) *Variation in cox1*

We examined *cox1* variation within and among clusters of sequences (see §3) using analysis of molecular variance

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

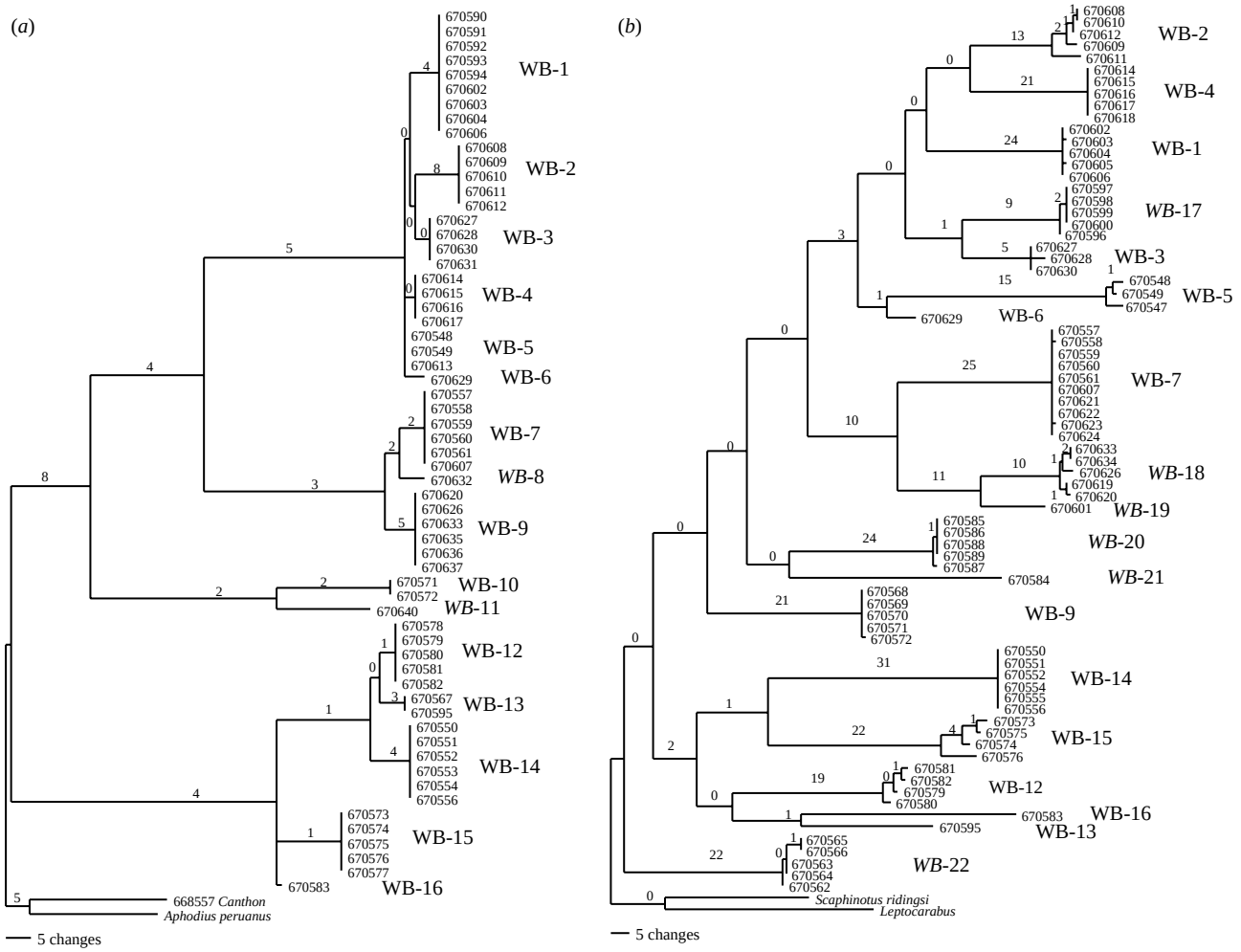

Figure 1. Parsimony trees for water beetles using 28S (*a*) and *cox1* (*b*) sequences. Cluster names are given to the right of groups. Italics in 28S denote groups for which no *cox1* data were available, and vice versa. Bremer Support values (see table 2) are reported above branches.

(AMOVA) (Excoffier *et al.* 1992) of pairwise differences as implemented in ARLEQUIN v.2.000 (Schneider *et al.* 2000). We used a two-level hierarchical analysis to partition total *cox1* variation into within-cluster and among-cluster covariance components. Individuals were included into a single *cox1* cluster if they exhibited identical sequences in the 28S gene (below) or, if no 28S sequence was available for a specimen, the *cox1* sequence grouped within these clusters. The fixation index calculated among groups (analogous to a population genetics F_{ST}) was tested for significance using ARLEQUIN v. 2.000. Groups with only one representative *cox1* sequence (e.g. WB-6, WB-11, see figure 1) were excluded from analysis because within-group variation could not be measured.

3. RESULTS

A total of 75 and 71 individuals were included in the analysis of water beetles and *Canthon*, respectively. Sequencing of 28S rRNA was successful for 63 and 62 specimens, and aligned matrices contained 699 and 746 characters in the respective groups. In the water beetle dataset, we detected 16 different 28S genotypes. Four genotypes were present in only a single individual, and the remaining occurred in groups ranging in size from two to nine individuals (figure 1*a*). Sequences differed from one another by a minimum of two nucleotides, e.g. an AC insertion separated WB-4 and WB-5 (figure 1*a*). DNA sequencing for *cox1* revealed 42 haplotypes. Parsimony tree searches uncovered 14 clusters of similar *cox1* sequences, plus five isolated sequences without close relatives ('singletons') (figure 1*b*). Results for *Canthon* were very similar. There were 12 different 28S genotypes and all but two were represented by >1 individual (figure 2*a*). The 46 *cox1* haplotypes grouped into 12 clusters, with two singletons (figure 2*b*).

The clustering of *cox1* sequences in the parsimony trees showed complete congruence with the 28S genotypes for both the water beetle and *Canthon* datasets. Closely related *cox1* haplotypes all exhibited the same 28S genotype and none of the groups defined by 28S genotypes were polyphyletic in the *cox1* tree (figures 1 and 2). It was not possible to judge incongruence in the 12 water beetles for which the 28S sequencing had failed (figure 1*b*). Equally, the *Canthon* dataset included missing sequences in both the 28S and *cox1* datasets, although there was perfect congruence for the 50 individuals sequenced for both genes (called 'core terminals', below).

Sequences making up the *cox1* clusters were very similar to each other, but very different from other clusters. Based on pairwise differences in AMOVA, within-group variation accounted for only 4.1% of the total variation in the dataset in *Canthon*, and only 2.5% in water beetles (table 1). Absolute divergence

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

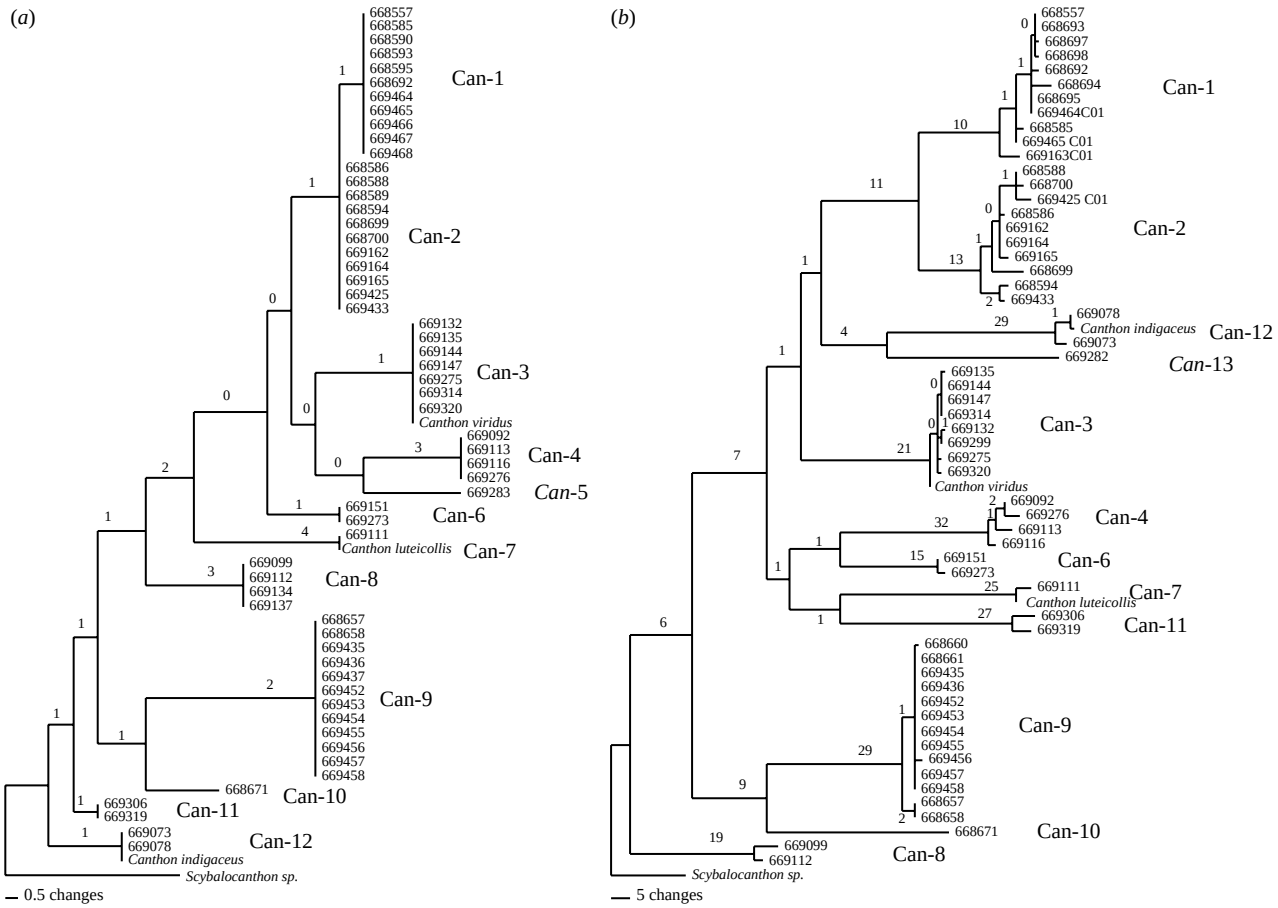

Figure 2. Parsimony trees for *Canthon*, as in figure 1.

Table 1. *Cox1* variation among and within clusters (figures 1 and 2) measured with AMOVA (Excoffier *et al.* 1994). (*$p<0.001$.)

	source of variation					
	among clusters			within clusters		
	df	Var	% variation	df	var	% variation
Canthon	9	1632.451	95.69*	40	68.089	4.31
water beetles	14	3130.381	97.49*	56	68.400	2.51

(uncorrected *p*-distance) of sequences within clusters ranged from 0 to 2%, whereas the mean divergence between clusters was 10 and 19% for *Canthon* and water beetles, respectively (table 2). These patterns of divergence are similar to those reported for intra- versus inter-species comparisons in other barcoding studies (Hebert *et al.* 2003*b*). The findings also appear to indicate that clusters in *Canthon* were more closely related to one another than were the clusters in the more diverse sample of water beetles.

The monophyly of the *cox1* clusters was highly supported. When Bremer Support was considered separately for three categories of node levels (tip nodes within a cluster, nodes immediately subtending a cluster, and nodes defining basal relationships between the clusters), the majority of total tree support was derived from nodes immediately below (i.e. defining) the clusters (88% of total Bremer Support for water beetles, 80% for *Canthon*; table 3). Tip nodes within clusters, and basal nodes had low support

Table 2. Mean uncorrected *p*-distances for *cox1* within and among clusters.
(Singletons were included in the among-cluster calculation, but within-cluster *p*-distance could not be measured.)

	among clusters		within clusters	
	mean	range	mean	range
Canthon	0.115	0.099–0.125	0.005	0–0.010
water beetles	0.162	0.140–0.190	0.005	0–0.018

(figures 1 and 2, table 3). To test for congruence in phylogenetic signal, a simultaneous analysis of *cox1* and 28S datasets was conducted for *Canthon*. Because some individuals were successfully sequenced for only one fragment, we either combined all terminals in a single 'supermatrix' (*cox1*+28S all terminals, $n=71$), or removed all terminals which were not complete for either one of the two gene partitions (core terminals,

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

Table 3. Parsimony analysis on the four datasets produced trees with minimal length and homoplasy values as indicated. (Bremer Support was calculated for three categories of nodes, corresponding to those near the tips within a cluster (tips), the nodes immediately below a cluster (sub-cluster) and the nodes defining basal relationships between the clusters (basal). The number given is the sum of the nodal Bremer Support values for this category in the entire tree, and numbers in parentheses give the number of nodes assigned to each category (many of them collapsed near the tips because sequences are identical or very similar). Total Bremer Support refers to the sum of the values for the entire tree. For *Canthon*, a combined analysis of *cox1* and 28S datasets was conducted, either combining all terminals in a single 'supermatrix' (all terminals), or removing all terminals which were not complete for either one of the two gene partitions (core terminals, $n=50$).)

data	Bremer support (no. nodes)				no. terminals	variable positions	no. steps	CI	RI
	total	tips	sub-cluster	basal					
water beetles									
cox1	284	17 (41)	249 (14)	18 (20)	77	255	1155	35	85
28S	62	0 (32)	28 (14)	34 (16)	63	157	332	76	93
Canthon (all terminals)									
cox1	296	13 (39)	240 (10)	42 (10)	61	287	705	52	88
28S	24	0 (39)	17 (10)	7 (10)	62	34	66	68	94
cox1+28S	92	7 (48)	68 (10)	17 (10)	71	321	784	53	89
Canthon (core terminals)									
cox1	260	10 (27)	214 (10)	36 (11)	50	284	695	52	85
28S	24	0 (27)	16 (10)	8 (11)	50	34	66	68	94
cox1+28S	277	12 (27)	224 (10)	41 (11)	50	318	774	53	86

$n=50$; table 3). Total tree support was much higher in the analysis of core terminals as compared to when all individuals were included in the combined analysis (figure 3, table 3). However, the incongruence length difference was minimal, indicating that the drop is not due to conflict between both markers but the reduced discriminatory power of the dataset once a large number of missing entries is included in the data matrix. In all cases, total tree support was higher for *cox1* than 28S for all analyses, regardless of whether all *Canthon* or only core individuals were used in the calculation (table 3), presumably due to the larger number of character changes in the former.

4. DISCUSSION

(a) *The partitioning of genetic variation*

The most striking result of the DNA analysis was the strong clustering of the sequence variation, with comparably large distances between groups of closely related sequences. In addition, these clusters showed remarkably high levels of nodal support for their monophyly according to *cox1* and 28S genes. Support in the combined analysis was even higher and is essentially the sum of the individual partitions, showing a high degree of congruence for the two markers. Nodes defining the clusters included >80% of the total Bremer Support provided by the datasets, although they specify only one-third or less of the total number of nodes in the tree. Variation in the 28S nuclear gene, while showing overall fewer character changes, was also strongly clustered. Genotypes were shared by many individuals and were separated by a minimum of a single base pair (bp) in *Canthon* and a minimum 2 bp insertion differentiating two water beetle genotypes (e.g. WB-4 and WB-5). For the *cox1* variation, >95% of variation occurred among these different clusters, with only *ca* 2.5–4.5% of the variation within these groups. Remarkably, the observed pattern of clustering was very similar in the two groups of beetles, even though they are composed of members of two different suborders Adephaga and Polyphaga, and obtained from different parts of the world (Madagascar and the Neotropics).

A further key finding of this study is that the nuclear and mitochondrial gene data were completely congruent for both samples of beetles. Individuals were grouped into clusters in the exact same way whether based on the *cox1* or 28S genotypes. *Cox1* sequences were more variable than 28S, but multiple *cox1* haplotypes in a cluster were monophyletic with respect to a single 28S genotype. For the water beetles this may be biased by the fact that these were field-samples from a given locality, raising uncertainty as to whether sister or even closely related taxa co-occur and were collected. For *Canthon*, by contrast, a single lineage was deliberately chosen from a larger sample of dung beetle communities (unpublished), and a wider sampling range covered, in order to increase the probability of sampling sister taxa. Notably, even in the case of the very closely related Can-1 and Can-2 clusters, where only a 2 bp insertion segregated 28S genotypes, the *cox1* phylogenetic analysis was completely congruent with separation into two distinct groups.

(b) *What is the nature of the clusters?*

Several lines of evidence suggest that the clusters identified in this study represent distinct species, rather than any other level of hierarchical organization. Phenetic sequence divergence in mtDNA within these groups never exceeded 2% and usually was much lower, whereas divergence between the clusters was often greater by more than an order of magnitude. This is in general agreement with empirical levels of divergence found between species in phylogeographic analyses (Avise & Walker 1999) and barcoding studies (Hebert *et al.* 2003*b*). For *Canthon*, the existing molecular phylogenetic and taxonomic framework

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

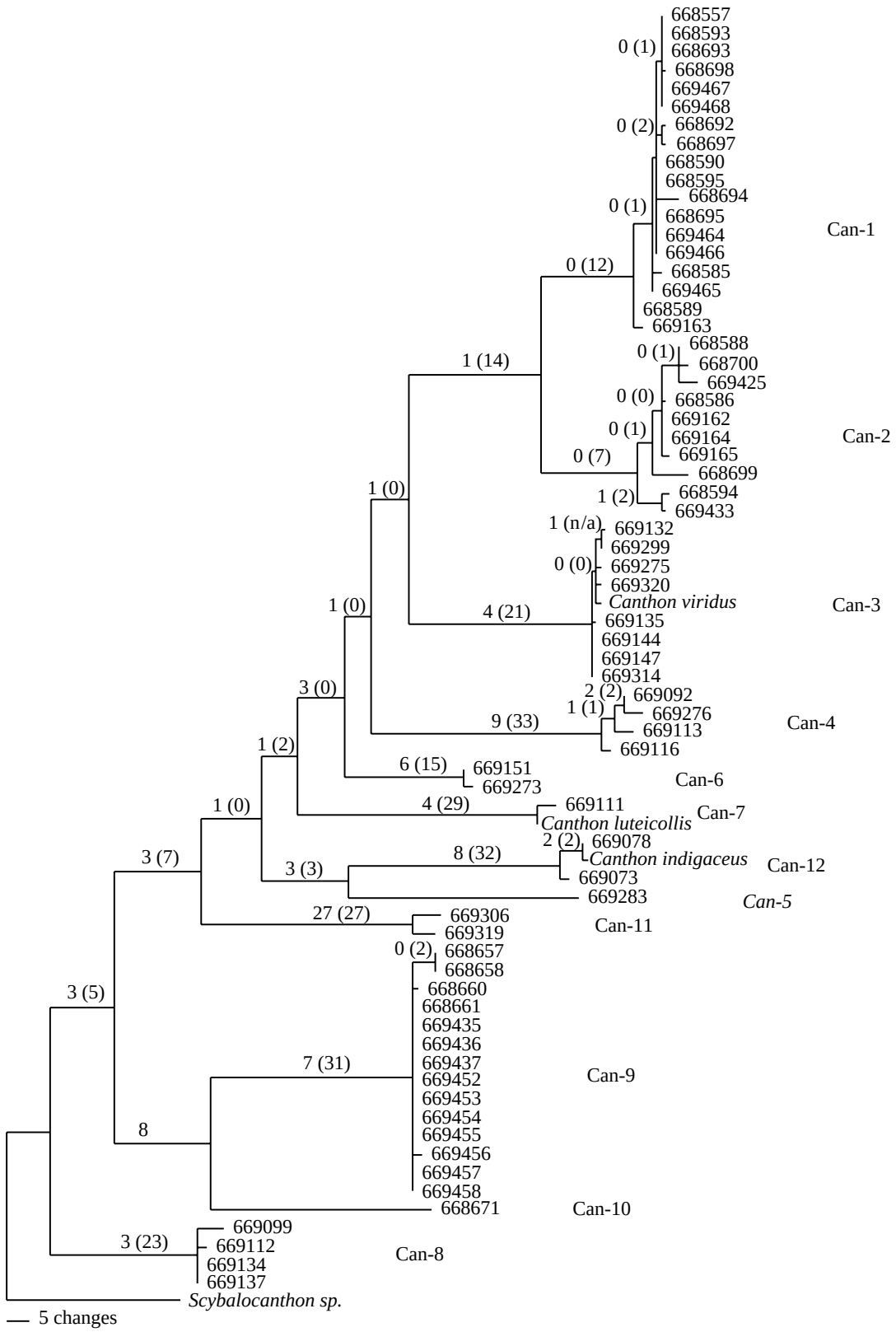

Figure 3. Tree from the combined data matrix (28S-*cox1* all terminals; table 2) for *Canthon*. Bremer Support values are reported above branches for the tree pictured. Values in parentheses are from the analysis of only the core terminals (50 individuals; table 2).

also supports this conclusion, as clusters from our study necessarily represent subgroups below the genus level, and GenBank database entries of various species of *Canthon* correspond to different clusters in our analysis.

Beyond simple comparisons of phenetic divergence, inspection of the phylogenetic trees revealed a striking shift in branch length, long branches leading to subtending nodes and short branches within tip clusters, as seen in other studies of closely related species (Barraclough unpublished; Pons *et al.* unpublished). In addition, the strong support for the sub-cluster nodes (and no other node level) also confirms the unique status of this particular level of

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

hierarchy in the trees. Whether or not this represents the species boundary remains to be investigated further. Ultimately, additional information, such as field studies of the sampled populations and broader genetic surveys including sister species, is required to confirm that the groups defined by these nodes are defining the species. However, the species category does take up a special place in the taxonomic hierarchy as the only 'natural' level of organization of the classificatory system, in contrast to the higher levels, such as genera and families (Cracraft 1983). It is our hypothesis that the transition in branching patterns, and the shift from strong to negligible branch support, represents a genetic signature of this unique level of organization.

(c) *Methodological issues of species delineation from sequence data*

Existing approaches to species delineation from sequence variation alone have been applied mainly to very small organisms, such as prokaryotes or soil nematodes, in which morphological discrimination is difficult or impossible (Floyd *et al.* 2002; Gregory & DeSalle 2005). In the case of nematodes, Molecular Operational Taxonomic Units (MOTUs) have been assigned based solely on sequence divergence (Blaxter 2004). While there may be no better way to classify these organisms to date, it remains unclear how these MOTUs correspond to evolutionarily differentiated groups, and how meaningful they are with respect to species cohesion. While the observation of large inter- and low intra-species variation promises easy identification of described species and the discovery of many cryptic species (Hebert *et al.* 2003*b*; Hebert *et al.* 2004*b*), there is concern regarding variability in the threshold values both between individual sister species pairs and among major lineages (DeSalle *et al.* 2005; Moritz & Cicero 2004).

In part, the problem of quantitative species delimitation could be overcome by searching for diagnostic character variation (Cracraft 1983), or complex character combinations (DeSalle *et al.* 2005) to define the species limits based on quantitative methods (Sites & Marshall 2003). These tests, which are rooted in the phylogenetic species concept, establish whether *a priori* populations can be 'aggregated' into a single species based on the distribution of characters or tree topology. These methods may not be practical when applied to large-scale species discovery and barcoding studies, where the cohesion of populations is unknown and broad sampling across species' geographic ranges may not be possible.

A possible alternative to aggregation methods is to interpret branch length itself as being suggestive of species boundaries, assuming that the long branches defining the clusters could only have arisen if populations diverged longer than around N_e (effective population size) generations ago (Hudson & Coyne 2002). Appropriate methods for estimating these shifts include Templeton's statistical parsimony analysis that partitions the variation into homoplastic (i.e. long branches) and non-homoplastic (short branches) variation (Templeton 2001). Similarly, it may be possible to statistically differentiate the shifts of lineage branching from interspecific, long branches to intraspecific, short branches using maximum likelihood methods (Barraclough, unpublished; Pons *et al.* submitted). A further approach could be based on population genetics analyses. It is possible to interpret the AMOVA results used to calculate intra- versus inter-cluster variation in a way analogous to *F*-statistics (Wright 1978). In this scenario, $F_{ST} > 0.95$ for both water beetles and *Canthon* datasets, meaning that $>95\%$ of the total genetic variation in the dataset arises from differences among groups. A threshold of $<5\%$ within-group variation seems a reasonable means of minimizing the chance of overlooking distinct taxa. As an example, combining Can-1 and Can-2 into a single group ('Can1-2') and recalculating AMOVA statistics results in a 15.6% value for within-group variation, as opposed to the 2.5% value when these two groups are considered as separate entities (table 1). This demonstrates the stringent clustering of the data, and provides a simple procedure to identify groups that have been grouped incorrectly.

(d) *Conclusions and prospects*

The aim of the present study was to investigate the efficacy of short sequence fragments for use in the discovery, delineation and routine identification of species. The analysis neither strictly constitutes a test of whether DNA can delimit pre-defined species, nor was it an analysis of Type I or Type II errors of species assignment (Quicke 2004). Instead, we used parsimony analysis to simultaneously examine a large number of sequences to assess patterns of variation in nature, and enquired whether this conforms to expectations of clustering at the species level of the biological hierarchy. The results were striking: sequence variation clustered very strongly for both nuclear and mitochondrial markers, and nodes defining these clusters were well supported whereas tip nodes, connecting closely related individuals, were not. Whether all the clusters we identified correspond to pre-existing, named species remains to be tested, and would require the input of specialists experienced in these taxonomic groups. Notably, comparisons with GenBank sequences indicated that at least three of the clusters identified here do indeed correspond to named species of *Canthon*.

The analytical approach will allow these clusters to be delineated objectively and repeatedly by anyone using the sequence data matrix. As a result, DNA data can form the basis of testable taxonomic hypotheses that could be examined with additional types of data in the future. The benefits of this approach are manifold: it provides a rapid division into probable groups of reproductively isolated individuals and generates more direct links to their evolutionary past; it will facilitate the determination of distinctive morphological features (i.e. through a focussed comparison of pre-delineated groups); it would allow the study of these putative species to proceed even while formal description is pending; it would link individuals from the same species collected in different localities or in different studies in a way that arbitrary designations (e.g. '*Canthon* sp.1') do not; and it would immediately

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

provide the data needed for future DNA barcoding identification.

The results of the present study are based on a relatively small number of species, but nonetheless demonstrate the general feasibility of using DNA-based methods in the large-scale delineation and discovery of new species, even in poorly characterized groups. While more research is needed to establish the best approach for species delineation using DNA (e.g. through phylogenetic or coalescent methods, with phenetic barcode 'thresholds', or some combination thereof), it is becoming evident that DNA methods present a promising new means of assessing and identifying biological diversity in some of the most species rich taxa and environments on Earth. There is reason for optimism that, if fully developed and implemented on a broad scale, DNA-based tools such as those examined here may provide the first opportunity for creating a comprehensive inventory of life.

We are grateful to Daegan Inward, Richard Davies and Liz Powell for collecting *Canthon* dung beetles; to David Lees, Ravomiarana Ranaivosolo, Pierre Razafindraire, Roger Andriamparany and Doug Ottke for assistance with water beetle collection; to Ruth Wild and Miranda Elliot for laboratory analysis; and to Silvia Fabrizi for mounting specimens. In Madagascar, we thank York Pareik at King de la Piste and Madame Liva and Benjamin Andriamihaja at MICET (Madagascar Institut pour la Conservation des Ecosystèmes Tropicaux). Tim Barraclough and an anonymous referee provided helpful comments on the manuscript.

REFERENCES

Avise, J. C. & Walker, D. E. 1999 Species realities and numbers in sexual vertebrates: perspectives from an asexually transmitted genome. *Proc. Natl Acad. Sci.* **96**, 992–995. (doi:10.1073/pnas.96.3.992.)

Belshaw, R. & Katzourakis, A. 2005 BlastAlign: a program that uses blast to align problematic nucleotide sequences. *Bioinformatics* **21**, 122–123. (doi:10.1093/bioinformatics/bth459.)

Blaxter, M. L. 2004 The promise of a DNA taxonomy. *Phil. Trans. R. Soc. B* **359**, 669–679. (doi:10.1098/rstb.2003.1447.)

Bremer, K. 1994 Branch support and tree stability. *Cladistics* **10**, 295–304. (doi:10.1111/j.1096-0031.1994.tb00179.x.)

Cracraft, J. 1983 Species concept and speciation analysis. *Curr. Ornithol.* **1**, 159–187.

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

DeSalle, R., Egan, M. G. & Siddall, M. 2005 The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Phil. Trans. R. Soc. B* **360**, 1905–1916. (doi:10.1098/rstb.2005.1722.)

Excoffier, L., Smouse, P. & Quattro, J. 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Floyd, R., Abebe, E., Papert, A. & Blaxter, M. 2002 Molecular barcodes for soil nematode identification. *Mol. Ecol.* **11**, 839–850. (doi:10.1046/j.1365-294X.2002.01485.x.)

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.

Godfray, H. C. J. 2002 Challenges for taxonomy—the discipline will have to reinvent itself if it is to survive and flourish. *Nature* **417**, 17–19. (doi:10.1038/417017a.)

Gregory, T. R. & DeSalle, R. 2005 Comparative genomics in prokaryotes. In *The evolution of the genome* (ed. T. R. Gregory), pp. 585–675. San Diego: Elsevier.

Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. 2003*a* Biological identifications through DNA barcodes. *Proc. R. Soc. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218.)

Hebert, P. D. N., Ratnasingham, S. & DeWaard, J. R. 2003*b* Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. B* **270**(Suppl. 1), S96–S99. (doi:10.1098/rspb.2002.2218.)

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. 2004*a* Ten species in one: DNA barcoding reveals cryptic species in the Neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl Acad. Sci.* **101**, 14 812–14 817. (doi:10.1073/pnas.0406166101.)

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S. & Francis, C. M. 2004*b* Identification of birds through DNA barcodes. *PLOS Biol.* **2**, 1657–1663. (doi:10.1371/journal.pbio.0020312.)

Hogg, I. D. & Hebert, P. D. N. 2004 Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Can. J. Zool.* **82**, 749–754. (doi:10.1139/z04-041.)

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

Inward, D. G. 2003 The evolution of dung beetle assemblages. Ph.D. thesis, Imperial College, London.

Janzen, D. H. 2004 Now is the time. *Phil. Trans. R. Soc. B* **359**, 731–732. (doi:10.1098/rstb.2003.1444.)

Knowlton, N. 1993 Sibling species in the sea. *Trends Ecol. Evol.* **24**, 189–216.

Lipscomb, D., Platnick, N. & Wheeler, Q. 2003 The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends Ecol. Evol.* **18**, 65–68. (doi:10.1016/S0169-5347(02)00060-5.)

Moritz, C. & Cicero, C. 2004 DNA barcoding: promise and pitfalls. *PLOS Biol.* **2**, 1529–1531. (doi:10.1371/journal.pbio.0020354.)

Proudlove, G. & Wood, P. J. 2003 The blind leading the blind: cryptic subterranean species and DNA taxonomy. *Trends Ecol. Evol.* **18**, 272–273. (doi:10.1016/S0169-5347(03)00095-8.)

Quicke, D. J. L. 2004 The world of DNA barcoding and morphology—collision or synergism and what of the future. *Systematist* **23**, 8–12.

Schneider, S. D., Roessli, D. & Excoffier, L. 2000 *ARLEQUIN version 2.000: a software for population genetics data analysis*. Switzerland: Genetics and Biomery Laboratory, University of Geneva.

Seberg, O., Humphries, C. J., Knapp, S., Stevenson, D. W., Petersen, G., Scharff, N. & Andersen, N. M. 2003 Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends Ecol. Evol.* **18**, 63–65. (doi:10.1016/S0169-5347(02)00059-9.)

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. 1994 Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* **87**, 651–701.

Sites, J. W. & Marshall, J. C. 2003 Delimiting species: a Renaissance issue in systematic biology. *Trends Ecol. Evol.* **18**, 462–470. (doi:10.1016/S0169-5347(03)00184-8.)

Smith, M. A., Fisher, B. L. & Hebert, P. D. N. 2005 DNA barcoding for effective biodiversity assessment of a

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011

hyperdiverse arthropod group: the ants of Madagascar. *Phil. Trans. R. Soc. B* **360**, 1825–1834. (doi:10.1098/rstb.2005.1714.)

Stoeckle, M. 2003 Taxonomy, DNA, and the barcode of life. *Bioscience* **53**, 796–797.

Swofford, D. L. 2002 *PAUP**: Phylogenetic analysis using parsimony. Version 4.0b. Sunderland, MA: Sinauer Associates.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74. (doi:10.1016/S0169-5347(02)00041-1.)

Templeton, A. R. 2001 Using phylogeographic analyses of gene trees to test species status and processes. *Mol. Ecol.* **10**, 779–791. (doi:10.1046/j.1365-294x.2001.01199.x.)

Tudge, C. 2000 *The variety of life*. Oxford: Oxford University Press.

Vences, M., Thomas, M., Bonett, R. M. & Vieites, D. R. 2005 Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Phil. Trans. R. Soc. B* **360**. (doi:10.1098/rstb.2005.1717.)

Wheeler, Q. D. 2004 Taxonomic triage and the poverty of phylogeny. *Phil. Trans. R. Soc. B* **359**, 571–583. (doi:10.1098/rstb.2003.1452.)

Will, K. W. & Rubinoff, D. 2004 Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* **20**, 47–55. (doi:10.1111/j.1096-0031.2003.00008.x.)

Wilson, E. O. 2003 The encyclopedia of life. *Trends Ecol. Evol.* **18**, 77–80. (doi:10.1016/S0169-5347(02)00040-X.)

Wright, S. 1978 *Evolution and the genetics of populations*. Chicago: University of Chicago.