MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 1069–1081 doi: 10.1111/1755-0998.12125

The potential of distance-based thresholds and character-based DNA barcoding for defining problematic taxonomic entities by CO1 and ND1

T. BERGMANN,*[1] J. RACH,*[1] S. DAMM,* R. DESALLE,† B. SCHIERWATER*† and H. HADRYS*†
*ITZ Ecology & Evolution, TiHo Hannover, Bünteweg 17d, D-30559 Hannover, Germany, †American Museum of Natural History, The Sackler Institute for Comparative Genomics, New York, NY 10024, USA

Abstract

The mitochondrial CO1 gene (cytochrome *c* oxidase I) is a widely accepted metazoan barcode region. In insects, the mitochondrial NADH dehydrogenase subunit 1 (ND1) gene region has proved to be another suitable marker especially for the identification of lower level taxonomic entities such as populations and sister species. To evaluate the potential of distance-based thresholds and character-based DNA barcoding for the identification of problematic species-rich taxa, both markers, CO1 and ND1, were used as test parameters in odonates. We sequenced and compared gene fragments of CO1 and ND1 for 271 odonate individuals representing 51 species, 22 genera and eight families. Our data suggests that (i) the combination of the CO1 and ND1 fragment forms a better identifier than a single region alone; and (ii) the character-based approach provides higher resolution than the distance-based method in Odonata especially in closely related taxonomic entities.

Keywords: character-based DNA barcoding, CO1, combined, insects, ND1, Odonata

Received 6 July 2012; accepted 9 April 2013

Introduction

The identification success of organisms through DNA barcodes primarily depends on the choice of the genetic marker. The main criteria for an appropriate barcoding marker include high interspecific divergence and low intraspecific variability to facilitate the accurate assignment of organisms to a taxonomic group. In addition, as DNA barcoding is a large-scale approach, sequences should be easy to obtain. Mitochondrial protein-coding genes seem to meet the above criteria best for several reasons: (i) high copy numbers per cell (Hoy 2003; Avise 2004) generally enhance PCR amplification (Lin & Danforth 2004); (ii) the haploid character allows the direct sequencing of PCR products (Saccone *et al.* 1999; Hurst & Jiggins 2005); (iii) the lack of introns, rare occurrence of indels (Hebert *et al.* 2003a) and low recombination rate ease the alignment; and (iv) the lack of proofreading mechanisms leads to higher evolutionary rates than in nuclear genes (Hoy 2003).

The Consortium for the Barcode of Life (CBoL) has agreed on the use of a 648-base pair fragment at the 5′ end of the mitochondrial cytochrome *c* oxidase subunit 1 gene region (CO1) as default DNA barcode region for vertebrates, insects and as many other animal groups as possible. As it was first promoted as suitable DNA barcoding marker for many animal groups by Hebert *et al.* (2003b), CO1 has been successfully used for obtaining reliable DNA barcodes and for a broad range of animal groups, such as arthropods (Will & Rubinoff 2004; Monaghan *et al.* 2005; Hajibabaei *et al.* 2006; Smith *et al.* 2006; Witt *et al.* 2006; Ekrem *et al.* 2007; Foley *et al.* 2007), birds (Hebert *et al.* 2004a; Yoo *et al.* 2006; Kerr *et al.* 2007), fishes (Ward *et al.* 2005) and mammals (Clare *et al.* 2007; Dawnay *et al.* 2007). In some animal groups, however, CO1 has failed to deliver reliable DNA barcodes. In cnidarians and sponges, for example, CO1 divergences are extraordinarily low compared with bilaterian animals (Shearer *et al.* 2002; Park *et al.* 2007). On the other hand, in aves, gastropods and amphibians, interspecific variation and also intraspecific variation in CO1 are very high (Remigio & Hebert 2003; Hebert *et al.* 2004b). In 449 dipteran species, the identification success through CO1 'barcodes' was low due to substantial overlaps in inter- and intraspecific divergences (Meier *et al.* 2006). Moreover, it was shown that the vast majority of nucleotide substitutions within the CO1 fragment occur at the third codon position, which might lead to

Correspondence: Heike Hadrys, Fax: +49-511-953-8584;
E-mail: heike.hadrys@ecolevol.de

[1]These authors contributed equally to this work.

rapid saturation (Lin & Danforth 2004; Vences *et al.* 2005).

Animal mitochondrial genomes usually possess 13 protein-coding genes, showing different rates and patterns of nucleotide substitution within and between taxonomic groups (Saccone *et al.* 1999). While the COI gene has proved to be extremely useful in DNA barcoding, other gene regions have potential too. The mitochondrial ND1 (NADH dehydrogenase 1) gene region, for example, showed better performance than CO1 in resolving phylogenetic relationships especially in insects such as in aphids (Lin & Danforth 2004), in Hawaiian drosophilids (Baker & DeSalle 1997) and odonates (Hadrys *et al.* 2006; Dijkstra *et al.* 2007; Rach *et al.* 2008). In mammals, the estimated variability in ND1 is slightly higher than in CO1 (Saccone *et al.* 1999).

Besides the selection of a suitable genetic marker, another critical point for the utility of DNA barcodes is the choice of method for analysing the sequence data. Here, distance-based analysis of standardized DNA barcodes has been the preferred analytical tool as originally introduced by Hebert *et al.* (2003a). The Barcode of Life Data System (Ratnasingham & Hebert 2007) is the most prominent workbench for the acquisition, storage, analysis and publication of DNA barcode records. The identification system of BOLD aligns the query sequence to the global reference alignment through a hidden Markov model of the CO1 protein (Eddy 1998), followed by a linear search of the reference library. Based on the general patterns of sequence variation, the identification system in BOLD delivers species identification if the query sequence shows a tight match, <1% distance, to a reference sequence. The majority of distance matrix analyses are based on a neighbour joining (NJ) algorithm, with a Kimura 2-parameter (K2P) correction (see for instance Borisenko *et al.* 2008; Casiraghi *et al.* 2010; Hebert *et al.* 2004b; Shearer & Coffroth 2008; Ward *et al.* 2005; Wong & Hanner 2008). While this approach is working for many applications, in other studies it has been shown that the translation of diagnostic sequence information into distance thresholds through application of NJ and K2P might be a major obstacle. Here, overlaps in inter- and intraspecific variation hinders species identification (Meyer & Paulay 2005; Meier *et al.* 2006; Wiemers & Fiedler 2007). In theory, the barcoding gap as defined by Hebert *et al.* (2004b) is based on the assumption that differences between species are significantly higher compared to differences within species. When this assumption is met, a barcoding gap can be a useful indicator for the identification of species by application of distance thresholds. Hebert *et al.* (2004b) propose a 10× threshold of the mean intraspecific variation for the group under study (Hebert *et al.* 2004b). But this threshold has fallen short on its promise to be used as guideline for species characterization. Meyer & Paulay (2005), for example, indicated through comparing their gastropod data and Hebert *et al.*'s bird data set (Hebert *et al.* 2004b) that no simple formula based on intraspecific variation will yield a robust threshold to minimize error across groups. One reason for failure stated by Meyer and Paulay was the underestimation of intraspecific variation because of low sample sizes (sample per species) and scale (regional vs. global). Another reason involves using substantially undersampled true sister species pairs and thus causes an overestimation of interspecific divergence.

In cases where CO1 might not be suitable for barcoding, the application of a character-based DNA barcode approach can be a solution. As a method that translates sequence information into diagnostic characters, it can be applied to identify and discriminate species especially when the interspecific variation is substantially low or when a 'barcoding gap' does not exist (DeSalle 2006, 2007; Wiemers & Fiedler 2007; Rach *et al.* 2008; Waugh *et al.* 2008).

In several case studies like Rach *et al.* (2008) on odonates and others (Damm *et al.* 2010b; Nicolalde-Morejon *et al.* 2010; Yassin *et al.* 2010; Reid *et al.* 2011), it has been shown that specific DNA sequence characters could be identified for genera, species, populations and conservation units by means of the Character Attribute Organization System (CAOS) algorithm (Sarkar *et al.* 2002b, 2008; Bergmann *et al.* 2009). In addition, Damm *et al.* (2010b) demonstrated that a character-based barcode can be implemented into a classical taxonomic framework to identify new species by integrating multiple sources of data. In that study, two mtDNA barcode markers CO1 and ND1 were combined with morphological, ecological and biogeographic data sets unmasking two cryptic odonate species.

In the present study, we evaluate the benefits of using character-based barcodes and/or distance-based thresholds when dealing with species with overlapping inter- and intraspecific sequence divergences. We employ CO1 and ND1 for both, the character-based and the barcode gap, approaches to DNA barcoding of 271 individual samples from 51 closely and distantly related odonate species.

Methods

Sample collection, processing and sequencing

Tissue samples of 271 individuals representing 51 species, 22 genera and eight families from Europe and Africa (Table S1, Supporting information) were collected during 2001 and 2006 by noninvasive sampling (Hadrys *et al.* 2005) and stored in 70% or 98% ethanol prior to

DNA extraction. Table 1 lists the analysed species and individuals per species.

DNA was extracted using a standard phenol–chloroform method (Hadrys *et al.* 1992). The universal primers LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TAAACTTCAGGGTGACCAAAAAA TCA-3′) were used to amplify the 'Folmer (CO1) fragment' (Folmer *et al.* 1994), and the primer pair P850 (fw), 5′-TTCAAACCGGTGTAAGCCAGG-3′ and P851 (rev) 5′-TAGAATTAGAAGATCAACCAGC-3′ was used to amplify a fragment containing a 5′partial fragment of 16S $tRNA^{Leu}$ and a 3′ partial fragment of the NADH dehydrogenase 1. PCR amplifications were carried out in 25 μL reaction mixture containing 2.5 μL of 10 × *Taq* DNA polymerase buffer (Bioline/Invitrogen), 2.5 mM $MgCl_2$, 0.1 mM dNTPs, 7.5 pM each primer and 0.5 U *Taq* DNA polymerase (either Invitrogen or Bioline). In cases of no immediate amplification success, 0.2 M Trehalose was added to the regular PCR mix (Spiess *et al.* 2004; Hajibabaei *et al.* 2005). Amplification conditions were as follows: initial denaturing at 95 °C 2 min, 30 cycles of 30-s denaturing at 95 °C, 30-s annealing at 48 °C (ND1)/ 50 °C (CO1), 1-min extension at 72 °C, followed by a final extension of 6 min at 72 °C. Amplified products were sequenced either on a MegaBACE 500 sequencer using the DYEnamic ET Dye Terminator Cycle Sequencing kit (Amersham Bioscience) or on an ABI PRISM™ 310 Genetic Analyzer using ABI BigDye® Terminator v1.1 (Applied Biosystems). Sequences were assembled and edited using SEQMANII (v. 5.03; DNASTAR, Inc.). All new sequences were deposited in GenBank (CO1 KC912199-KC912405; ND1 KC912406 - KC912523). In addition, sequences from previous publications of our research (Damm *et al.* 2010a; Damm& Hadrys; Damm *et al.* 2010b; Dijkstra *et al.* 2007; Groeneveld *et al.* 2007; Rach *et al.* 2008) were included in our data sets (see Table 1 for details). The complete CO1 and ND1 data sets used in this manuscript are deposited in the CAOS-Library of the CAOS-Workbench website (http://bol.uvm.edu/caos-workbench/).

Sequences were aligned using MUSCLE (Edgar 2004). The CO1 alignment was trimmed to obtain sequences of uniform lengths of 541 bp. The ND1 alignment revealed indels at the beginning of the amplified fragment in most samples. The ND1 alignment was first trimmed to 436 bp. Afterwards, a second alignment for ND1 was created, containing only the ND1 gene fragment for which no indels were observed. Here, the sequences were shortened to an unambiguous alignable core region of 316 bp.

DNA barcode analyses

For distance-based threshold analyses, mean distances of CO1 and ND1 sequences within and among species were calculated using the Kimura 2-parameter (K2P) substitution model in MEGA 3.1 (Kumar *et al.* 2004). Mean intraspecific as well as lowest and highest mean interspecific K2P distance values for all species are shown in Table 1.

For character-based barcode analyses, each data set of CO1 and ND1 was first aligned with the G-INS-I setting of the Mafft software (Katoh *et al.* 2005), and the alignments were converted into the NEXUS file format with SeaView version 4 (Gouy *et al.* 2010). A maximum-likelihood (ML) tree was created for each data set using RAxML (Stamatakis 2006) with 100 bootstraps. The ML trees served as guide trees for CAOS (Character Attribute Organization System) analyses. Each tree file and the corresponding NEXUS file were saved as one file using MacClade 4 v. 4.06. (Maddison & Maddison 2000) and processed with the CAOS-Analyzer. The CAOS-Analyzer, which can be run on a web server (http://bol.uvm.edu/caos-workbench/) or as a command line program, identifies diagnostic characters, termed 'characteristic attributes', for all clades at each branching node within the given guide tree (Sarkar *et al.* 2002a,b, 2008; Bergmann *et al.* 2009). To produce the character-based barcodes, the output files of the CAOS-Analyzer were run through the CAOS-Barcoder. From the reference barcode created by the CAOS-Barcoder, we selected by eye twenty-nine species-specific simple '*pure*' characteristic attributes (shared by all members of a clade and absent from the other clades descending from the same node) for CO1 and ND1 as a representative example for a character-based barcode (Figs 1 and 2).

For the identification of diagnostic characters for geographical entities, nodes *within* species clusters of the original NJ trees were considered. Numbers of pure characteristic attributes for geographical entities or populations within species were obtained for both data sets.

The CAOS-Classifier assigns query sequences to its closest match by comparing diagnostic characters of reference sequences with the query. To test the accuracy of query assignments to reference data sets by the CAOS-Classifier, a leave-one-out test was performed with the CO1 (234 sequences) and ND1 (266 sequences) data sets. Each sequence within the reference data set was singled out from the data set, it was then used as a query to that data set, and an identification was made. This procedure was accomplished for each taxonomic unit in the study.

We devised a second test of the robustness of character-based diagnostics for the classification of query sequences that involved creating random substitution data sets based on the real data sets. These simulated data sets were then run through the CAOS-Classifier. For both genes, CO1 and ND1, we created 100 random substitution data sets with a 1% nucleotide exchange ratio and 100 random substitution data sets with 5%

Table 1 Mean intra- and interspecific divergences of ND1 and CO1 from 51 odonate species; the source of the sequence is shown for ND1 and CO1; mean intra- and interspecific divergences (Kimura 2-parameter distances) are given in%; lowest and highest interspecific distance values for each species are shown

Species	No. of individuals	ND1 sequence	CO1 sequence	Mean intraspecific divergence (%) ND1 | CO1	Mean interspecific divergence (range; %) ND1 | CO1
A. Anisoptera					
Aeshna cyanea	4	New	New	0 | 0	5.5–30.4 | 7.7–23.4
Aeshna grandis	1	1	New	—	4.8–29.2 | 8.6–24.7
Aeshna mixta	2	New	New	0 | 0.2	4.8–28.9 | 7.5–25.7
Aeshna rileyi	2	1	New	0 | 0	6.6–29.4 | 8.7–25.0
Anaciaeschna triangulifera	1	1	New	—	5.6–31.6 | 7.5–23.7
Anax ephippiger	10	1	New	0.2 | 0.7	7.6–30.3 | 7.4–26.1
Anax imperator	11	1	New	0.3 | 0.2	2.5–30.0 | 5.8–24.1
Anax speratus	6	1	New	0 | 0	2.5–29.6 | 5.8–24.0
Brachytron pratense	2	1	New	0 | 0	8.0–33.0 | 10.1–26.3
Gynacantha usambarica	9	1	—	0 | —	8.9–31.3 | n/c
Gynacantha villosa	1	1	New	—	9.8–31.1 | 10.8–25.6
Paragomphus geneii	5	1	—	0.9 | —	17.9–35.6 | n/c
Crocothemis erythraea	7	1	New	0.3 | 1.0	18.2–37.4 | 14.6–25.1
Crocothemis sanguinolenta	6	New	New	0 | 0.8	16.1–33.0 | 13.2–27.4
Nesciothemis farinosum	5	New	New	0.6 | 0.3	12.2–31.9 | 13.6–25.7
Orthetrum brachiale	3	1	New	0 | 0.1	8.1–31.2 | 7.1–23.5
Orthetrum chrysostigma	4	1; New	New	0.2 | 0.5	6.0–30.3 | 5.9–24.2
Orthetrum coerulescens	9	1; New	New	0.2 | 0.1	7.4–30.1 | 10.4–23.1
Orthetrum julia falsum	10	1; New	New	0.2 | 0.5	6.0–30.3 | 5.9–23.6
Orthetrum trinacria	5	1	New	0 | 0	11.2–33.1 | 12.0–24.6
Sympetrum sanguineum	2	New	—	0 | —	14.1–32.1 | n/c
Trithemis annulata	3	1	3	0.2 | 0.1	7.9–34.8 | 8.2–23.8
Trithemis arteriosa	1	4	New	—	8.2–32.4 | 9.1–24.1
Trithemis donaldsoni	5	New	New	0.4 | 0.2	12.3–31.9 | 11.8–21.6
Trithemis furva	3	2	3	0.2 | 1.4	9.1–35.2 | 9.7–23.9
Trithemis grouti	2	2	New	1.0 | 0.2	7.9–37.1 | 1.1–25.5
Trithemis hecate	5	1	—	0 | —	13.7–39.2 | n/c
Trithemis kirbyi	4	1; New	New	0.8 | 0.7	15.4–37.3 | 11.3–23.1
Trithemis morrisoni	5	2	3	2.4 | 0.5	4.7–34.8 | 5.0–24.4
Trithemis nuptialis	2	2	3	0 | 0	2.7–34.9 | 1.1–24.8
Trithemis palustris	4	2	3	0.4 | 0.3	4.7–36.7 | 5.0–24.2
Trithemis stictica	7	2; 3	3	0.1 | 0.1	2.7–36.3 | 2.8–24.4
B. Zygoptera					
Calopteryx haemorrhoidales	12	New	—	0.2 | —	15.8–35.5 | n/c
Calopteryx splendens	4	1	—	0 | —	15.8–36.0 | n/c
Platcypha auripes	2	1	New	0.3 | 0	12.0–39.2 | 10.3–25.7
Platcypha caligata	6	1	New	0.3 | 0.2	12.0–36.0 | 10.3–24.8
Ceriagrion tenellum	5	New	New	0 | 0.1	13.4–31.4 | 17.8–23.9
Enallagma cyanthigerum	5	New	New	0 | 0.1	15.1–32.0 | 12.8–23.8
Ischnura graellsii	5	New	New	0 | 0.1	15.1–31.8 | 7.9–25.9
Ischnura senegalensis	5	—	New	— | 0	n/c | 7.9–21.0
Leptagrion elongatum	1	New	New	—	13.4–31.3 | 16.7–25.1
Pseudagrion acaciae	4	1	New	0 | 0.2	0(14.4)–36.0 | 0.6–23.4
Pseudagrion bicoerulans	15	1; New	New	**4.3 | 4.2**	13.1–30.7 | 15.8–23.4
Pseudagrion kersteni	11	1; 6; New	New	1.1 | 1.1	13.1–31.2 | 15.8–26.4
Pseudagrion massaicum	13	1; New	New	0.6 | 0.7	14.4–37.6 | 13.6–25.1
Pseudagrion niloticum	6	1; New	New	0 | 0.7	0(14.4)–36.0 | 0.6–23.7
Teinobasis alluaudi	6	New	New	0.4 | 0.3	15.8–29.6 | 16.9–27.9
Chlorocnemis abbotti	8	New	New	0.3 | 0.2	16.3–35.5 | 18.2–27.1
Coryphagrion grandis	14	5; New	New	2.6 | 2.4	16.7–30.1 | 18.0–24.2
Mecistogaster asticta	1	New	New	—	12.8–35.3 | 13.9–27.9
Mecistogaster martinezi	2	New	New	0 | 0	12.8–37.1 | 13.9–24.8

Bold indicates exceptional high values.
(1) Rach *et al.* (2008).
(2) Damm *et al.* (2010a).
(3) Damm *et al.* (2010b).
(4) Damm & Hadrys (2012).
(5) Groeneveld *et al.* (2007).
(6) Dijkstra *et al.* (2007).

Species	16	19	25	34	76	79	97	103	145	169	214	223	256	259	262	274	277	298	307	310	313	325	346	361	391	409	434	438	469
Aeshna cyanea	A	A	A	A	T	T	T	A	A	A	A	A	A	T	A	C	T	T	T	A	T	T	T	T	T	A	A	T	C
Aeshna grandis	A	A	T	A	T	T	T	T	A	A	T	A	A	C	T	T	A	A	C	A	A	T	C	T	T	G	A	T	T
Aeshna mixta	A	A	A	A	C	T	T	A	A	A	T	T	A	T	A	T	A	A	T	A	G	C	T	A	T	G	A	T/C	A
Aeshna rileyi	A	A	A	A	T	T	T	A	T	G	A	A	T	T	A	T	A	A	C	T	A	T	T	T	T	A	A	T	T
Anaciaeschna triangulifera	A	A	A	A	T	A	T	A	A	G	A	A	G	A	A	T	A	A	T	T	G	T	T	C	T	G	A	T	A
Anax ephippiger	A	A	A	A	T	T	T	T	A	A	C/T	T/C	A	A	A	T	A/G	A	T	A/T	A	C	T	A	T	G	A	A	T
Anax imperator	A	A	A	A	T	T	T	A	A	A	T	C	T	A	A	A	T	A	T	A	A	T	T	T	T	G	A	A	C
Anax speratus	A	A	A	A	T	C	T	A	C	G	T	T	A	A	A	A	T	A	T	A	A	C	T	T	T	A	A	A	T
Brachytron pratense	A	A	A	A	T	A	T	T	A	A	A	A	A	A	A	T	T	A	T	A	A	T	T	A	A	G	A	T	A
Gynacantha villosa	A	A	A	A	C	C	C	A	C	A	T	T	G	T	A	T	A	T	T	A	A	T	T	T	T	G	A	A	A
Crocothemis erythreae	A	T	A	A	T	A	T	A	A	A	T	C	A	A	A	A	C	A	T	G	G/A	C	T	A	T	G	T	T	A
Crocothemis sanguinolenta	A	A	A	T	T	A	T	A	T	A	T	T	G/T	T	T	A	A	A	T	C	G	T	T	T	A	A	T	A	A
Nesciothemis farinosum	C	T	T	A	T	A	T	A	T	A	T	A	A	A	A	C	T	T	T	A	A	T	C	C	T	G	A	C	C
Orthetrum brachiale	T	T	T	A	T	A	C	T	A	G	T	T	A	A	A	T	A	A	T	A	A	T	T	T	A	G	T	T	T
Orthetrum chrysostigma	T	T	T	A	T	A	T	A	A	A	T	T/C	G	A	G/A	T	G	A	T	A	A	T	T	T	A	G	T	T	T
Orthetrum coerulescens	T	A	C	A	T	A	T	T	A	A	A	T	G	A	G	A	A	A	T	A	A	T	C	T	A	A	A	A	T
Orthetrum julia falsum	T	T	T	A	C	A	T	T	A	A	T	C	G	A	A	A	T	A	T	A	A	T	T	A	A/G	G	T	T	T
Orthetrum trinacria	T	T	T	A	T	T	T	T	T	G	T	T	A	A	A	A	T	C	T	T	G	T	T	T	A	A	A	T	A
Trithemis annulata	A	A	A	A	T	A	T	T	C	A	T	T	G	A	T	A	T	T	T	T	A	T	T	T	A	A	A	T	A
Trithemis arteriosa	A	A	T	A	T	A	T	T	T	A	T	T	A	A	A	A	T	T	T	C	A	T	C	T	A	A	A	T	A
Trithemis donaldsonii	A	A	C	A	C	A	T	A	T	A	A	T	G	A	A	A	T	A	C	A	A/G	T	C	C	A	A	A	T	A
Trithemis furva	G	A	A	A	C	A	T	T	C/A	A	A	C	T	T	A	A	A	T	T	T	T	T	T	A	A	A	T	T	A
Trithemis grouti	A	A	A	A	T	A	T	T	T	A	A	T	T	T	A	A	T	A	T	T	G	T	T	A	T	A	T	C	A
Trithemis kirbyi	A	A	A	A	T	A	T	A	T	A	A	T	G	T	A	A	T	A	T	A	G	T	T	A	A	A	T	T	A
Trithemis morrisoni	G	A	A	A	T	A	C	T	T	G	A	T	T	G	A	C	T	T	T	T	G	T	T	A	A	A	C	T	A
Trithemis nuptialis	A	A	C	A	T	A	T	T	T	A	A	T	T	T	A	A	C	A	T	C	G	T	T	A	T	A	A	T	A
Trithemis palustris	A	A	T/C	A	T	A	T	T	T	G	A	T	T	A	A	T	T	T	T	T	T	T/C	T	A	A	A	C	T	A
Trithemis stictica	A	A	C/T	A	T	A	C	T	T	G	A	T	T	T	A	A	T	A	T	T	G	T	T	A	T	A	A	T	A
Platcypha auripes	A	A	A	C	C	C	T	T	A	A	A	C	T	T	C	A	T	A	C	T	A	T	T	T	T	A	A	A	A
Platcypha caligata	A	A	T	A	T	C	T	A	A	A	G	C	G	A	A	A	T	A	T	A	A	T	T	C	C	A	A	G	A
Ceriagrion tennelum	A	T	A	A	T	T	T	T	T	A	C	A	G	T	T	T	A	T	T	A	A	C	C	A	T	A	C	T	A
Enallagma cyanthigerum	A	A	T	A	T	A	T	T	A	G	A	C	A	C	G	T	A	A	T	T	A	T	C	A	A	A	T	T	A
Ischnura graellsii	A	G	T	A	C	T	T	A	A	A	C	T	G	A	A	A	A	A	C	T	A	T	C	C	A	G	T	T	A
Ischnura senegalensis	A	A	T	G	T	C	T	A	A	A	C	T	A	A	A	A	T	A	T	T	A	C	C	C	C	G	C	T	A
Leptagrion elongatum	A	T	A	A	T	T	T	A	T	A	A	A	A	C	A	A	T	T	T	A	A	T	C	C	C	A	A	T	A
Pseudagrion acaciae	G	C	A	A	C	A	C	A	T	A	T	A	T	T	T	T	T	A	C	A	G	T	T	A	A	A	A	C	T
Pseudagrion bicoerulans	A	T	A	A	C	T	C/T	T/A	A	A	C/T	T	T	C	A	A	T	T/C	T	A	G	C	C/T	T	T	A	A	T	A
Pseudagrion kersteni	A	T	A	A	C	T	C	T/A	A	A	A	G	G	T	A	A	A	T	C	A	G/A	T	T	C/T	A	A	A	A	A
Pseudagrion massaicum	A	G	A	A	C	C	C	A	T	A	C	A	G	A	A	A	T	A	T	A	A	C	T	A	T/G	A	A	C	T
Pseudagrion niloticum	G	C	A	A	C	A	C	A	T	A	T	A	T	T	T	T	T	A	C	A	G	T	T	A	A	A	A	C	T
Teinobasis alluaudi	T	A	A	G	T	A	C	A	A	A	T	T	A	T	G	A	A	T	C	A	A	T	T	C	A	A	T	T	T
Chlorocnemis abbotti	A	A	A	A	T	T	T	T	A	G	C	T	A	A	T	T	T	C	C	C/T	A	C	T	C	T	A	A	T	A
Coryphagrion grandis	A	A	T	A	C	C	C/T	A	T	A	C	C	A	C	T	T/C	T/A	A	T	A	A	T	C	C/A	A	A	C	C/T	A
Mecistogaster asticta	C	A	T	A	C	A	C	T	T	G	T	A	G	C	T	A	T	A	C	A	A	T	C	C	T	A	A	T	A
Mecistogaster martinezi	C	A	T	A	C	A	C	T	T	A	T	A	A	A	T	T	T	A	T	A	C	T	T	T	C	A	A	T	A

Fig. 1 Character-based DNA barcodes for 45 odonate species based on CO1 sequences; unique combinations of character states at 29 nucleotide positions for each species are shown; grey shaded cells show two different bases at the particular nucleotide position within a species.

nucleotide exchange ratio. The substitution included the random selection of either an 'A', 'T', 'C', 'G', '-', '?' or an 'N' at a random position within the sequences. Each of the 100 random matrices contained all 234 sequences for CO1 and all 266 sequences for ND1.

At last to test the accuracy of the CAOS-Classifier and BOLD, both platforms were confronted with our data set of 234 odonate CO1 sequences.

Results

Distance-based thresholds

Interspecific distances. The mean interspecific K2P distances ranged from 0.6% to 27.9% within CO1 and 2.5% and 39.2% within ND1 sequences. The lowest distance values were observed between *Pseudagrion acaciae* and *Pseuadgrion niloticum*, with no difference in ND1 and only 0.6% divergence in CO1. The pairwise distances between CO1 sequences of these two species differed between 0.37% and 0.76%. Very low mean CO1 distances were also observed between *Trithemis nuptialis* and *Trithemis grouti* (1.1%) and between *T. nuptialis* and *T. stictica* (2.8%). With respect to ND1, lowest mean interspecific K2P distances in ND1 were observed between *Anax imperator* and *Anax speratus* (2.5%) and *T. nuptialis* and *T. stictica* (2.7%).

In rare cases, distances between samples of congeneric species were higher than between samples from different higher taxa. For example, the mean interspecific distance of CO1 sequences between the libellulids

Species	48	60	66	69	72	75	81	84	87	90	93	102	105	114	123	129	132	144	153	157	165	171	177	183	226	243	288	298	304
Aeshna cyanea	A	T	T	T	A	A	A	T	A	G	G	T	A	T	T	T	T	A	T	G	T	T	T	G	T	C	A	T	A
Aeshna grandis	A	T	T	T	A	A	G	T	A	G	G	T	A	T	T	T	T	A	T	G	T	T	T	G	C	T	A	T	A
Aeshna mixta	A	T	T	T	A	C	G	T	A	A	A	T	A	T	T	T	T	T	T	G	T	T	T	G	T	T	A	T	A
Aeshna rileyi	A	T	T	T	A	T	A	T	A	A	A	A	A	T	T	T	A	T	A	G	T	T	T	G	T	T	G	T	T
Anaciaeshna triangulifera	A	A	T	A	A	T	A	T	A	A	A	T	A	T	A	G	A	A	G	G	T	T	T	A	T	T	A	T	T
Anax ephippiger	A	A	T	A	A	T	A	T	A	A	G	T	A	T	T	T	T	T	G	G	T	T	T	G	T	T	A	C	T
Anax imperator	A	T	A	A	A	T	G	T	A	A	A	T	A	T	T	T	A	T	T	G	T	T	T	A	T	T	A	T	T
Anax speratus	A	T	A	A	A	C	A	T	G	G	A	T	A	T	T	T	A	T	T	G	T	T	T	A	T	G	A	T	T
Brachytron pratense	A	A	T	T	A	A	G	T	A	A	G	T	A	T	T	T	C	A	T	G	A	T	T	A	T	C	A	C	A
Gynacantha usambarica	A	T	T	A	G	A	A	C	A	A	A	T	A	T	T	T	T	T	A	G	T	A	T	G	T	T	G	T	T
Gynacantha villosa	A	T	T	T	G	T	A	T	T	G	G	T	A	C	T	T	T	G	A	G	C	A	T	G	T	T	G	T	T
Paragomphus genei	A	T	A	A	G	C/T	A	T	A	G	G	T	A	C	T	C/T	T	T	T	G	C	T	C	G	T	T	G/A	T	A
Crocothemis erythreae	G	G	A	T	T	C	T	T	C	T	G	T	A	T	T	T	C	G	G	G	A	T	T	A	T	T	A/G	T	T
Crocothemis sanguinolenta	A	A	T	G	A	C	G	T	A	T	A	A	A	T	T	T	T	G	A	G	T	T	T	A	T	T	T	T	G
Nesciothemis farinosum	A	A/G	T	A	T	T	G	C	A	G	A	C	A	T	T	A	T	A	T	A	T	T	A	G	T	T	G	T	G/A
Orthetrum brachiale	G	T	A	T	T	C	A	C	A	A	A	T	A	T	T	T	A	T	A	G	T	T	T	A	T	T	A	T	T
Orthetrum chrysostigma	A	T	A	T	T	T	A	T	A	A	A	T	A	T	T	T	A	T	T	G	T	T	C	G	T	T	A	T	T
Orthetrum coerulescens	A	T	A	T	T	T	G	T	A	A	G	T	A	T	T	T	T	T	T	A	C	T	C	A	T	T	A/G	T	A
Orthetrum julia falsum	A	T	A	T	T	T	A	T	A	G	A	T	A	C	T	T/C	A	C	T	A	T	T	T	A	C	T	A	T	T
Orthetrum trinacria	A	T	T	T	T	T	A	T	A	G	A	C	A	T	T	T	A	A	A	G	T	T	T	A	A	A	A	T	T
Sympetrum sanguineum	A	C	T	G	T	T	A	T	A	G	A	T	G	T	T	T	A	C	A	G	C	T	T	A	T	T	A	T	T
Trithemis annulata	A	T	T	C	A	T	A	A	G	A	A	A	A	T	T	T	A	G	A	G	C	A	T	A	T	T	A	T	T
Trithemis arteriosa	A	A	T	T	A	T	A	A	A	A	A	A	A	T	T	T	A	A	A	A	T	T	T	G	T	A	A	T	T
Trithemis donaldsonii	A	A	T	T	A	T	A	A	A	T	A/G	A	A	T	A	T	A	T	A	G	T	T	A	A	T	T	G/A	T	T
Trithemis furva	A	T	T	C	A	C	A	A	A	T	A	A	A	T	T	G	C	T	T	A	C	T	T	A	C	C	A	T	G
Trithemis grouti	G	A	T	T	A	A	A	A	A	A	A	A	G	T	T	T	A	T	T	A	A	G	T	G	T	T	A	T	C
Trithemis hecate	A	A	T	T	A	T	A	A	G	T	A	A	A	T	T	A	A	T	G	G	C	T	A	G	T	T	A	T	T
Trithemis kirbyi	A	A	T	G	A	A	A	A	G	T	A	A	A	T	G	T	T	T	T	G	C	T	A	T	T/C	A	A	T	T
Trithemis morrisoni	A	T	T	G	A	T	A	A	A	A	A	A	A	T	T	T	A	T	T	G	A	A	T	A	T	T	A	T	T
Trithemis nuptialis	A	T	T	G	A	A	A	A	A	A	A	G	A	T	T	T	A	T	T	A	A	A	T	A	T	T	A	T	T
Trithemis palustris	A	T	T	G	A/G	T	A	A	A	A	A	A	A	T	T	T	A	T	T	G	A	G	T	A	C	C	A	T	T
Trithemis stictica	G	T	T	T	A	A	A	A	A	A	A	G	A	T	T	C	A	T	T	A	A	A	C	A	T	T	G	T	T
Calopteryx haemorrhoidales	A	A	T	T	T	C	G	T	G	G	G	T	A	A	T	C	T	A/G	T	A	T	T	C	A	T	T	C	C	A
Calopteryx splendens	G	G	T	C	T	C	G	T	A	A	A	T	A	G	C	C	T	C	T	A	T	T	T	A	A	T	G	T	T
Platycypha auripes	A	T	T	T	A	T	A	C	A	G	A	T	T	A	A	T	T	T	T	C	A	T	C	A	T	T	T	T	A
Platycypha caligata	A	T	T	T	G	C	A	T	A	A	A	T	T	A	T	A	A	A	G	T	T	T	C	A	T/C	T	T	T	A
Ceriagrion tennelum	G	T	T	A	A	T	G	A	G	A	A	T	A	A	T	T	T	A	A	A	T	T	A	G	T	T	T	T	G
Enallagma cyanthigerum	A	T	T	A	T	T	A	A	A	A	G	T	A	G	T	T	A	T	A	T	T	T	A	A	T	T	A	T	T
Ischnura graellsii	A	T	A	A	T	T	A	A	A	A	A	T	T	A	T	C	A	T	T	T	T	T	T	G	T	T	G	T	T
Leptagrion elongatum	A	T	A	T	G	T	A	A	G	G	G	T	A	G	T	T	T	T	A	A	T	T	T	G	T	T	A	T	T
Pseudagrion acaciae	G	A	T	T	T	T	A	C	A	T	G	T	A	A	C	T	T	G	T	T	C	T	T	G	A	C	A	T	T
Pseudagrion bicoerulans	A	A	G/A	C/T	A/G	C/T	G/A	T	A	T	G	T	A	A	C	A/C	C	T		T		A	C		T		A	T	
Pseudagrion kersteni	A	G	A	C	A	C	A	T	A	T	G	T	A	A	T	T/C	A	T	A	T	C	A	T	A	T	T	G/A	T	T
Pseudagrion massaicum	A	A	T	T	T	C	A	T	A	T	A	T	A	G	T	T	A	G	T	T	T	G	T	A	T	C	G	T	C/T
Pseudagrion niloticum	G	A	T	T	T	T	A	C	A	T	G	T	A	A	C	T	T	G	T	T	C	T	T	G	A	C	A	T	T
Teinobasis alluaudi	A	C	T	T	T	T	A	T	A	G	A	T	A	A	T	T	A	T	A	G	T	T	A	G	T	T	A	T	G
Chlorocnemis abbotti	A	T	G	T	T	T	T	A	A	A	G	T	A	G	T	G	T	T	G	A	A	T	G	G	T	T	T	T	G
Coryphagrion grandis	A	T	T	A	A	C	A	T/C	A	A	A	T	A	A	T	A	A	T/C	G	A/G	T	T	A/G	A/G	T	T	A	T	T
Mecistogaster asticta	A	T	G	T	G	T	G	T	A	A	A	T	A	G	T	G	T	C	A	A	T	T	G	A	T	T	A	T	T
Mecistogaster martinezi	A	T	G	T	G	T	A	T	G	G	G	T	A	A	C	T	T	C	G	A	T	T	G	A	T	T	G	T	T

Fig. 2 Character-based DNA barcodes for 50 odonate species based on ND1 sequences; unique combinations of character states at 29 nucleotide positions for each species are shown; dashed cells indicate the occurrence of three or four character states within a species; grey shaded cells show two different bases at the particular nucleotide position within a species.

Crocothemis erythreae and *Crocothemis sanguinolenta* was 16.5%, while 14.9% divergence was observed between *C. erythreae* and the aeshnid *Aeshna mixta*, but only 14.6% between *C. erythreae* (suborder Anisoptera) and *Ischnura senegalensis* (suborder Zygoptera). The mean K2P distance of ND1 sequences between the two *Crocothemis* species, *C. erythreae* and *C. sanguinolenta*, was 23.2%, while distances between *C. erythreae* and all eleven species of the family Aeshnidae were lower (19.4%–23.1%). Another example was observed for *Pseudagrion massaicum* (suborder Zygoptera) and the two congeneric species *Pseudagrion kersteni* and *Pseudagrion bicoerulans*. Here, the interspecific K2P distances in CO1 were higher (21.2%/20.8%) than those between *P. massaicum* and all three *Anax* species (suborder Anisoptera; 18.5%–18.6%). The ND1 fragment revealed a lower mean K2P distance value between *P. massaicum* and *Ischnura graellsi* (20.8%) than between *P. massaicum* and *P. bicoerulans* (21.4%).

Intraspecific distances. The mean intraspecific K2P distances ranged from 0% to 4.2% in CO1 and 0% to 4.3% in ND1. For six of 45 species, only one sample was analysed and intraspecific divergences could not be calculated. The highest values were observed for *Pseudagrion bicoerulans* (CO1:4.2%; ND1 4.3%). Here, all four analysed populations form distinct clusters. High intraspecific distances of at least 1% within one fragment were also detected for *Coryphagrion grandis* (CO1/ND1: 2.6%), *Pseudagrion kersteni* (CO1/ND1: 1.1%), *Trithemis furva* (CO1: 1.4%), *Crocothemis erythreae* (1%) and *Trithemis grouti* (1%). Intraspecific distances of more than 0.5% either

within ND1 or CO1 were observed for further eight species (see Table 1).

Character-based DNA barcodes

Diagnostic characters for species. A core sequence of 29 nucleotide positions of the CO1 fragment showed the highest number of diagnostic characters for groups at the important nodes and exhibited diagnostic characters for very closely related species (Fig. 1). The character states at the chosen nucleotide positions revealed unique base compositions – character-based DNA barcodes – for 43 of the 45 species. No diagnostic characters were found for differentiating specimens of *Pseudagrion niloticum* from those of *Pseuagrion acaciae*.

Similar to the CO1 sequences, a core region of 29 nucleotide positions of the ND1 fragment was selected (Fig. 2). Of the 29 nucleotide positions, 23 were used previously as character-based DNA barcodes in dragonflies. As the 5′ end of the sequences were trimmed by 142 bp, the numbers of nucleotide positions changed and six positions were additionally included. 48 of 50 species revealed unique combinations of character states at the 29 nucleotide positions. Again, no diagnostic characters were found to distinguish *P. acaciae* and *P. niloticum*.

Table 2 lists the numbers of *pure* diagnostic characters for sister species pairs. The lowest number of diagnostic characters within the CO1 fragment was found for *Trithemis nuptialis* and *Trithemis grouti*, which differed by five nucleotide positions. The ND1 fragment revealed 21 *pure* diagnostic characters for this sister species pair. Very low numbers of diagnostic characters within the ND1 fragment were found for *A. imperator* and *A. speratus* (six diagnostic characters) and *Trithemis stictica* and *T. nuptials* (eight diagnostic characters). The CO1 fragment exhibited 29 diagnostic characters for the differentiation of *A. imperator* and *A. speratus* and 17 for *T. stictica* and *T. grouti*. For all other pairs of sister species, at least 16 diagnostic characters within the CO1 or ND1 fragment have been found.

Diagnostic characters identifying geographical clusters or flagging of populations with diagnostics. We also use the DNA barcoding information to group specimens within distinct species according to geographical origin to test for diagnosis of these groups as potential novel species. This process has been called 'flagging' (Goldstein & DeSalle 2011), where flagging refers to the process of designating populations as potential species worthy of further anatomical, behavioural or other work to determine species existence. Species showing distinct geographical clusters are listed in Table 2, and the number of diagnostic characters for each of the geographical clusters is given. For the two German populations of *Orthetrum coreulescens*, one diagnostic character each was found within the CO1 and ND1 sequences to distinguish them from the Italian population. Five diagnostic characters within the CO1 and one within the ND1 fragment could differentiate the two Namibian populations of *Orthetrum julia falsum*. The *Trithemis furva* sample from South Africa shows different character states when comparing it to the two Ethiopian samples (nine nucleotide positions within CO1 and one within ND1). For *Pseudagrion bicoerulans*, distinct clusters for all four populations were observed. Here, the lowest numbers of diagnostic characters were found for the two Kenyan populations from Mount Kenya and Mount Elgon (CO1: 2; ND1: 4). The third Kenyan population and the Tanzanian population differed from the others by at least 14 nucleotide positions within the CO1 and 16 positions within the ND1

Table 2 Number of *pure* diagnostic characters identified within the CO1 and ND1 sequences for five sister species pairs. Number of pure diagnostics characters identified for populations or geographical groups of five odonate species. For further explanations, see text

		No. of pure diagnostics CO1/ND1
Sister species pairs		
Anax imperator	*Anax speratus*	29/6
Aeshna cyanea	*Aeshna mixta*	35/19
Trithemis nuptialis	*Trithemis grouti*	5/21
Trithemis stictica	*Trithemis grouti*	17/20
Trithemis stictica	*Trithemis nuptialis*	16/8
Populations		
P. bicoerulans, Mt. Elgon, Kenya	*P. bicoerulans*, Mt. Kenya	2/4
O. julia falsum, Waterberg, Namibia	*O. julia falsum*, Tsauchab, Namibia	5/1
C. grandis, Kenya	*C. grandis*, Tanzania	17/11
O. coerulescens, Germany	*O. coerulescens*, Italy	1/1
T. furva, South Africa	*T. furva*, Ethiopia	9/1

fragment. For *Coryphagrion grandis*, two distinct clusters were detected, one comprised all three Kenyan and the other all three Tanzanian populations. The clusters revealed *pure* diagnostic characters at 17 nucleotide positions within CO1 and 11 within ND1.

Leave-one-out test. To test the validity of the CAOS-Classifier for assigning queries to the correct species, 234 Odonata reference data sets for CO1 were created all leaving out one of the 234 sequences. For 227 of the 234 left-out sequences, the best hit was at the same species level with an identity of 98.34–100% (Table S2, Supporting information). For the seven remaining query sequences, *Mecistogaster asticta*, *Leptagrion elongatum*, *Gynacantha villosa*, *Aeshna grandis*, *Anaciaeschna triangulifera*, *Aeshna mixta and Trithemis arteriosa*, the best hits were between 82.62% and 90.20%. All of these queries belong to species with specimen sizes of $n = 1$, and only for *A. mixta*, the number of specimen sequences was $n = 2$. Three queries (*M. asticta*, *L. elongatum* and *T. arteriosa*) were matched with their closest relative in the data set, while the remaining four queries were assigned to the wrong species.

Two hundred and sixty-six Odonata reference data sets for ND1 were created all leaving out one of the 266 sequences. For 260 of the 266 left-out sequences, the best hit was at the same species level with an identity of 98.73–100% (Table S3, Supporting information). For the six remaining query sequences, *Aeshna grandis*, *Anaciaeschna triangulifera*, *Gynacantha villosa*, *Mecistogaster asticta*, *Leptagrion elongatum* and *Trithemis arteriosa*, the best hits were between 78.80% and 90.49%. All of these queries belong to species with specimen sizes of $n = 1$. Two queries (*M. asticta* and *T. arteriosa*) were matched with their closest relative in the data set, while the remaining four queries were assigned to the wrong species.

Random substitution test. To test the robustness of diagnostic characters for species identification, we created randomly generated sequences and challenged the CAOS-Classifier with these sequences. The average score of correct species assignments for 100 randomly substituted data sets was evaluated (Table S4, Supporting information). For the CO1 data sets with 1% substitution ratio, we observed an average score of 233 correct assignments of 234 (99.5%). Increasing the substitution ratio to 5% led to a reduction in correct assignments to 225 of 234 (96.1%). For the ND1 data sets with 1% substitution ratio, we observed an average score of 249 correct assignments of 266 (93.8%). Increasing the substitution ratio to 5% led to a reduction in correct assignments to 237 of 266 (89.1%).

CAOS-Classifier vs. BOLD. All 234 CO1 odonate sequences were tested on the CAOS-Classifier and BOLD (Table S5, Supporting information). Using the reference barcodes for these sequences, all 234 queries were correctly assigned by the CAOS-Classifier to the species they belong to. For BOLD, 131 of 234 were assigned to a species with an identity of 97.39%–100%. The remaining 103 queries showed no match. Interestingly, three specimens we identified as *Pseudagrion acaciae* were identified as *Pseudagrion niloticum* (99.43%–99.63%) by BOLD. Of the five specimens we identified as *Enallagma cyathigerum*, only one was identified as *E. cyathigerum* (99.81%) and the remaining as *Coenagrion hastulatum* (99.81%). All five specimens we identified as *Ischnura senegalensis* were identified as *Pseudagrion abyssinica* (100%). All five specimens we identified as *Ischnura graellsii* were identified as *Ischnura elegans* (99.80%–100%). All five specimens we identified as *Trithemis donaldsonii* were identified as *Trithemis aconita* (99.63%–100%). All three specimens we identified as *Orthetrum brachiale* were identified as *Orthetrum stemmale* (99.81%–100%). Of the four specimens we identified as *Orthetrum chrysostigma* three were identified as *Orthetrum julia* (100%). All three specimens we identified as *Aeshna rileyi* were identified as *Aeshna subpupillata* (99.63%).

Discussion

The value and utility of DNA barcoding decisively depend on the trade-off between investments in marker isolation and identification and the resolution of these markers to unambiguously distinguish between species or related taxonomic units. This study of 51 odonate species suggests that the employment of two combined genetic markers substantially enhances DNA barcoding in this insect order and possibly many other animal groups.

CO1 vs. ND1 vs. CO1/ND1

The main criterion for an efficient DNA-based identification system is the straightforward acquisition of comparative informative sequences. In this study, the CO1 and ND1 sequences were obtained from most species by using a single primer pair each. This is cost- and time-saving because all PCRs are carried out under the same conditions and no optimization is required. However, in some cases, the amplification of mitochondrial genes for all species of a particular animal group using one or two sets of universal primers can be a challenge due to high substitution rates. Besides, mitochondrial-like sequences frequently occur in the nuclear genome, which can complicate PCR amplification and sequencing of authentic mitochondrial genes (Zhang & Hewitt 1996; Behura 2007). In our study, putative pseudogenes of the CO1 gene region have been observed for at least five of 51

species. For ND1, more than one pseudogene fragment was amplified only in one case. However, for all 51 species, at least one sequence was obtained and could be utilized as a DNA barcode.

Although both markers used in this study are of mitochondrial origin, and therefore inherited jointly, their substitution patterns within and between taxonomic entities differ substantially. For example, only six *pure* characteristic attributes were observed within the ND1 fragment to differentiate the sister species *Anax imperator* and *Anax speratus,* while the corresponding CO1 sequences revealed the high number of 29 *pure* diagnostic characters. The species *Trithemis stictica* and *Trithemis grouti* differed by 20 diagnostic characters within ND1 but by 17 within their CO1 fragments. In contrast, for *Trithemis nuptialis* and *T. grouti,* the ND1 sequences exhibited 21 *pure* characteristic attributes, while the CO1 sequences revealed only five. The complementarity of the two fragments was also observed when diagnostics for populations below the species level were examined. For example, the two Kenyan populations of *Pseudagrion bicoerulans* from Mount Kenya and Mount Elgon differed by four diagnostic characters within the ND1 and only two within the CO1 fragment. The CAOS analysis of the CO1 sequences revealed five *pure* diagnostic characters for the discrimination of the two Namibian populations of *Orthetrum julia falsum* and nine for the South African and the Ethiopian populations of *Trithemis furva*. In both cases, only one *pure* characteristic attribute was found within the ND1 gene region. Thus, it cannot be predicted which fragment reveals the better information but both together do the job of identifying populations nicely.

In summary, both, ND1 and CO1, are suitable DNA barcoding markers and deliver reliable character-based DNA barcodes for the vast majority of species. However, neither one alone could resolve all species. It was shown that combining both markers is highly beneficial for discriminating species, in particular sister species as well as geographical entities. It cannot be predicted which marker delivers the higher degree of information in which species. This *per se* suggests that both markers should be used in these cases.

Comparing character-based barcoding and distance-based thresholds

The majority of DNA barcoding studies have focused on the distance-based approach for analysing DNA barcodes (Hebert *et al.* 2003a). The accuracy of this method depends on the discrepancy between intra- and interspecific values – the 'barcoding gap' (Meyer & Paulay 2005). In odonates, high intra- and low interspecific variability has been observed leading to the conclusion that distance-based methods are ill-suited for DNA barcoding in this insect order (Rach *et al.* 2008). Our data confirm these findings. High mean intraspecific K2P distances of more than 1% are observed for four of 50 species in the ND1 and for five of 45 species in the CO1 fragment. The highest intraspecific distance values are seen in *P. bicoerulans* (ND1: 4.3%; CO1: 4.2%), and an rapid speciation in this species has been suspected as in former studies (Hadrys *et al.* 2006; Dijkstra *et al.* 2007). In contrast, in some cases, the distance values between sister species are extraordinarily low. For example, the mean K2P divergence of ND1 among *A. speratus* and *A. imperator* (2.5%) is lower than the observed mean intraspecific distance in *C. grandis* (2.6%). The mean interspecific distance between *T. nuptialis* and *T. stictica* is only slightly higher (2.7%). The CO1 distance between *T. nuptialis* and *T. grouti* is only 1.1% and is exceeded by the mean intraspecific CO1 distances in four species. Although we examined only a small part of the worldwide dragonfly diversity, we assume that cases of overlapping intra- and interspecific distances are prevalent.

The two examples of the genera *Crocothemis* and *Pseudagrion* indicate that due to overlapping distance values between congenerics and members of different higher taxa, incorrect assignments might occur when a critical species is missing in the DNA barcode database. Here, we suggest that the character-based approach for DNA barcoding is a powerful complement to the currently used distance-based methods. Cut-offs for species boundaries are needless, and diagnostic characters can be easily identified at different taxonomic levels by means of the CAOS algorithm.

Diagnostic characters for geographical clusters; flagging of populations with diagnostics

The ND1 and COI sequences can also be examined for diagnostics within distinct geographical clusters of individuals. There are two purposes for searching for such diagnostics. First, the diagnostics can be used to identify populations of origin for unidentified specimens. Such diagnostics can then be used in ecological monitoring studies where samples are hard to identify to population. Second, if diagnostics do exist, then these populations can be flagged for future, integrated taxonomic studies (DeSalle 2006; Rubinoff 2006) that might result in species descriptions for these diagnosable populations (Goldstein *et al.* 2000).

Hence, we have detected diagnostic markers for these populations for use in ecological monitoring studies that might be useful as bio-indicators. In addition, we suggest that further taxonomic study using integrated taxonomic approaches (Rubinoff *et al.* 2006) should be applied to

these populations to determine whether taxonomic revision of these entities is needed.

Leave-one-out test

Testing the assignment of new sequences to a reference database by the CAOS-Classifier showed that in most cases the correct species was assigned by the program. However, in both test groups, CO1 and ND1, some queries were assigned to a species that was not its closest match. In 4 of 234 cases, we observed it for CO1, and in 4 of 266 cases, we observed it for ND1. For *A. grandis*, the closest possible match was *A. rileyi*, but *T. nuptialis* was selected by the CO1 test set. After reviewing the decision tree of the Classifier, we located the source of the problem. At one point in the classification, the query was compared with two groups, one including two specimens of *A. rileyi* and a second including 237 specimen of different species. The first group having only two specimens showed only one diagnostic character in comparison with the second group with 237 specimens and 192 diagnostic characters. As the diagnostic character was truly unique for *A. rileyi* while 24 diagnostic characters were shared between the query and the second group, the classification returned an incorrect diagnosis. While we never observed this misclassification with query sequences sharing at least one close member in the reference data set, the assignment of truly unique or new sequences by the Classifier can be suboptimal if at some point of the decision tree a group of few specimens is compared with a group of many.

Random substitution test

Our random substitution test showed that even with substitution ratios of one to five per cent, the CAOS-Classifier in most cases assigns the query to the correct species. This demonstrates that even when sequences of new, undocumented populations are entered or sequencing errors are present in the query sequence, a mostly accurate result is presented.

While the accuracy for CO1 was above 99% at 1% substitution ratio and above 95% at 5% substitution ratio, the results for ND1 were slightly lower. With ND1 we observed around 94% correct assignments at 1% substitution ratio and 89% at 5% substitution ratio. The explanation for this bias of accuracy between CO1 and ND1 is the difference in number of nucleotide positions used for each gene. While the CO1 data sets included 541 characters, only 316 characters were used within the ND1 data sets. Considering that ND1 is shorter by 225 characters (almost 42%) compared with CO1, the accuracy of the CAOS-Classifier is still high. This not only highlights the potential of ND1 as a barcode marker for insects but also validates character-based identification tools as a means for classification. We expect even better results when compound characters are added as diagnostics in addition to simple pure characters that are currently used.

Comparison of the CAOS-Classifier and BOLD

All test sequences for CO1 were correctly assigned by the CAOS-Classifier to the corresponding species. This result shows that when queries are tested that have at least one representative species sequence within the reference library, an accurate match is identified by the CAOS-Classifier. When we tested the same sequences with BOLD, only 131 of 234 sequences were assigned to a species. In all fairness, we have to mention that at this point, our CO1 sequences were not submitted to BOLD, and the identification was performed using only Odonata data that were included by other researchers. Nevertheless, it also shows that even BOLD has problems with the assignment of sequences when no closely related reference data are available to the program. Of the 131 sequences that were assigned to species, 100 shared the same species as predicted by the CAOS-Classifier. The remaining 31 query sequences were assigned either to a closely related species (22 times) or to a different species than we had assumed (nine times). The first observation can be explained by insufficient data in the BOLD library and a strong similarity of sequences between closely related species. In the second observation, the four specimens we had assigned to *Enallagma cyathigerum* were assigned by BOLD to *Coenagrion hastulatum* with 99.81% identity. Five specimens that we assigned to *Ischnura senegalensis* were assigned by BOLD to *Pseudagrion abyssinica* with 100% identity. Especially in the last case, we can only assume that either the other researchers who added the reference sequences to BOLD made an error in specimen identification or alternatively we have made identification errors. The scenario that both species share the same sequences could be possible but is unlikely.

Conclusions

In this study, we have used 271 odonate samples belonging to 51 species. The analyses of the genetic data reveal that odonates are a challenging test-bed for DNA barcoding. The employment of two combined genetic markers highly enhances the identification of organisms through DNA sequences, even if both markers are of mitochondrial origin. The number of diagnostic characters for the discrimination of taxonomic groups increases substantially with the use of two genetic markers in odonates. The

acquisition of an additional marker is not necessarily cost-intensive, but can become a *conditio sine qua non* for many closely related species. A database containing reliable DNA barcodes of as many species as possible highly enhances the discovery of yet unknown species or speciation processes and can be of priceless value for fast biodiversity assessment. It is also clear from this study that diagnostic characters for geographical clusters of specimens are valuable 'flags' for long-time monitoring, speciation studies, conservation management and identification of larval stages.

Acknowledgements

We thank all our collaborators who helped collecting tissue samples in the field, for example, Adolfo Cordero, Viola Claunitzer, Klaas Douwe Djikstra, Frederico A. A. Lencioni, Frank Suhling. Jessica Rach was supported by a fellowship of the Evangelische Studienstiftung, and Sandra Damm was financed by the Federal Government Research Program (BMBF) BIOTA South (S08). Tjard Bergmann was financed by the German Academic Exchange Service (DAAD) and the H. Wilhelm Schaumann Stiftung. RD thanks the Sackler Family and Korein Family Foundations as well as the Lewis and Dorothy Cullman Program in Molecular Systematics at the American Museum of Natural History. Heike Hadrys thanks the German Science Foundation (DFG-HA1947/6-1). This manuscript benefitted greatly from insightful comments from the Subject Editor and two anonymous reviewers.

References

Avise JC (2004) *Molecular Markers, Natural History, and Evolution*, 2nd edn. Sinauer Associates, Sunderland, Massachusetts.

Baker RH, DeSalle R (1997) Multiple sources of character information and the phylogeny of *Hawaiian drosophilids*. *Systematic Biology*, **46**, 654–673.

Behura SK (2007) Analysis of nuclear copies of mitochondrial sequences in honeybee (Apis mellifera) genome. *Molecular Biology and Evolution*, **24**, 1492–1505.

Bergmann T, Hadrys H, Breves G, Schierwater B (2009) Character-based DNA barcoding: a superior tool for species classification. *Berliner Und Munchener Tierarztliche Wochenschrift*, **122**, 446–450.

Borisenko AV, Lim BK, Ivanova NV, Hanner RH, Hebert PDN (2008) DNA barcoding in surveys of small mammal communities: a field study in Suriname. *Molecular Ecology Resources*, **8**, 471–479.

Casiraghi M, Labra M, Ferri E, Galimberti A, De Mattia F (2010) DNA barcoding: a six-question tour to improve users' awareness about the method. *Briefings in Bioinformatics*, **11**, 440–453.

Clare EL, Lim BK, Engstrom MD, Eger JL, Hebert PDN (2007) DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes*, **7**, 184–190.

Damm S, Hadrys H (2012) A dragonfly in the desert: genetic pathways of the widespread *Trithemis arteriosa* (Odonata: Libellulidae) suggest male-biased dispersal. *Organisms Diversity & Evolution*, **12**, 267–279.

Damm S, Dijkstra KDB, Hadrys H (2010a) Red drifters and dark residents: the phylogeny and ecology of a Plio-Pleistocene dragonfly radiation reflects Africa's changing environment (Odonata, Libellulidae, Trithemis). *Molecular Phylogenetics and Evolution*, **54**, 870–882.

Damm S, Schierwater B, Hadrys H (2010b) An integrative approach to species discovery in odonates: from character-based DNA barcoding to ecology. *Molecular Ecology*, **19**, 3881–3893.

Dawnay N, Ogden R, McEwing R, Carvalho GR, Thorpe RS (2007) Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International*, **173**, 1–6.

DeSalle R (2006) Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. *Conservation Biology: The Journal of the Society for Conservation Biology*, **20**, 1545–1547.

DeSalle R (2007) Phenetic and DNA taxonomy; a comment on Waugh. *Bioessays*, **29**, 1289–1290.

Dijkstra K-DB, Groeneveld LF, Clausnitzer V, Hadrys H (2007) The Pseudagrion split: molecular phylogeny confirms the morphological and ecological dichotomy of Africa's most diverse genus of Odonata (Coenagrionidae). *International Journal of Odonatology*, **10**, 31–41.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Ekrem T, Willassen E, Stur E (2007) A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, **43**, 530–542.

Foley DH, Wilkerson RC, Cooper RD, Volovsek ME, Bryan JH (2007) A molecular phylogeny of *Anopheles annulipes* (Diptera: Culicidae) sensu lato: the most species-rich anopheline complex. *Molecular Phylogenetics and Evolution*, **43**, 283–297.

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. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Goldstein PZ, DeSalle R (2011) Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays*, **33**, 135–147.

Goldstein PZ, DeSalle R, Amato G, Vogler AP (2000) Conservation genetics at the species boundary. *Conservation Biology*, **14**, 120–131.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**, 221–224.

Groeneveld LF, Clausnitzer V, Hadrys H (2007) Convergent evolution of gigantism in damselflies of Africa and South America? Evidence from nuclear and mitochondrial sequence data. *Molecular Phylogenetics and Evolution*, **42**, 339–346.

Hadrys H, Balick M, Schierwater B (1992) Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Molecular Ecology*, **1**, 55–63.

Hadrys H, Schroth W, Schierwater B, Streit B, Fincke OM (2005) Tree hole odonates as environmental monitors: non-invasive isolation of polymorphic microsatellites from the neotropical damselfly *Megaloprepus caerulatus*. *Conservation Genetics*, **6**, 481–483.

Hadrys H, Clausnitzer V, Groeneveld LV (2006) The present role and future promise of conservation genetics for forest Odonates. In: *Forests and Dragonflies* (ed. Rivera A), pp. 279–299. Pensoft Publishers Sofia-Moscow, Pontevedra, Spain.

Hajibabaei M, deWaard JR, Ivanova NV *et al*. (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1959–1967.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PD (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 968–971.

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. *Proceedings. Biological Sciences/ The Royal Society*, **270**, 313–321.

Hebert PD, Ratnasingham S, deWaard JR (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological sciences/ The Royal Society*, **270**(Suppl 1), S96–S99.

Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14812–14817.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b) Identification of birds through DNA barcodes. *PLoS Biology*, **2**, 1657–1663.

Hoy M (2003) *Insect Molecular Genetics*, 2nd edn. Academic Press, San Diego, California.

Hurst GD, Jiggins FM (2005) Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings. Biological sciences/ The Royal Society*, **272**, 1525–1534.

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.

Kerr KCR, Stoeckle MY, Dove CJ *et al.* (2007) Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes*, **7**, 535–543.

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.

Lin CP, Danforth BN (2004) How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. *Molecular Phylogenetics and Evolution*, **30**, 686–702.

Maddison D, Maddison W (2000) *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, Massachusetts.

Meier R, Shiyang K, Vaidya G, Ng PK (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, **55**, 715–728.

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.

Monaghan MT, Balke M, Gregory TR, Vogler AP (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1925–1933.

Nicolalde-Morejon F, Vergara-Silva F, Gonzalez-Astorga J, Stevenson DW (2010) Character-based, population-level DNA barcoding in Mexican species of Zamia L. (Zamiaceae: Cycadales). *Mitochondrial DNA*, **21**, 51–59.

Park MH, Sim CJ, Baek J, Min GS (2007) Identification of genes suitable for DNA barcoding of morphologically indistinguishable Korean Halichondriidae sponges. *Molecules and Cells*, **23**, 220–227.

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings. Biological sciences/ The Royal Society*, **275**, 237–247.

Ratnasingham S, Hebert PD (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Reid BN, Le M, McCord WP *et al.* (2011) Comparing and combining distance-based and character-based approaches for barcoding turtles. *Molecular Ecology Resources*, **11**, 956–967.

Remigio EA, Hebert PD (2003) Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Molecular Phylogenetics and Evolution*, **29**, 641–647.

Rubinoff D (2006) DNA barcoding evolves into the familiar. *Conservation Biology : The Journal of the Society for Conservation Biology*, **20**, 1548–1549.

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *The Journal of Heredity*, **97**, 581–594.

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, **238**, 195–209.

Sarkar IN, Planet PJ, Bael TE *et al.* (2002a) Characteristic attributes in cancer microarrays. *Journal of Biomedical Informatics*, **35**, 111–122.

Sarkar IN, Thornton JW, Planet PJ *et al.* (2002b) An automated phylogenetic key for classifying homeoboxes. *Molecular Phylogenetics and Evolution*, **24**, 388–399.

Sarkar IN, Planet PJ, DeSalle R (2008) CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources*, **8**, 1256–1259.

Shearer TL, Coffroth MA (2008) Barcoding corals: limited by interspecific divergence, not intraspecific variation. *Molecular Ecology Resources*, **8**, 247–255.

Shearer TL, Van Oppen MJ, Romano SL, Worheide G (2002) Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). *Molecular Ecology*, **11**, 2475–2487.

Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PD (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3657–3662.

Spiess AN, Mueller N, Ivell R (2004) Trehalose is a potent PCR enhancer: lowering of DNA melting temperature and thermal stabilization of taq polymerase by the disaccharide trehalose. *Clinical Chemistry*, **50**, 1256–1259.

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

Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1859–1868.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1847–1857.

Waugh J, Huynen L, Millar C, Lambert D (2008) DNA barcoding of animal species-response to DeSalle. *Bioessays*, **30**, 92–93.

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, **4**, 8.

Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics–The International Journal of the Willi Hennig Society*, **20**, 47–55.

Witt JD, Threloff DL, Hebert PD (2006) DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology*, **15**, 3073–3082.

Wong EHK, Hanner RH (2008) DNA barcoding detects market substitution in North American seafood. *Food Research International*, **41**, 828–837.

Yassin A, Markow TA, Narechania A, O'Grady PM, DeSalle R (2010) The genus Drosophila as a model for testing tree- and character-based methods of species identification using DNA barcoding. *Molecular Phylogenetics and Evolution*, **57**, 509–517.

Yoo HS, Eah JY, Kim JS *et al.* (2006) DNA barcoding Korean birds. *Molecules and Cells*, **22**, 323–327.

Zhang DX, Hewitt GM (1996) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends in Ecology & Evolution*, **11**, 247–251.

H.H. and B.S. initiated and designed the project. T.B. and J.R. performed the research. T.B. and R.D. wrote and performed the bioinformatics. S.D. provided additional barcode information. T.B., J.R., R.D. and H.H. wrote the manuscript.

Data Accessibility

DNA sequences: CO1 KC912199 – KC912405; ND1 KC912406 – KC912523.

The reference matrices and sequence data were added to the CAOS-Library:

Website: http://bol.uvm.edu/caos-workbench/

CAOS-Library: http://bol.uvm.edu/caos-workbench/librarystart.php.

To access and download Odonata data sets, enter 'Odonata' as search value with 'Overview' as search category. Select both data sets 'Odonata1' and 'Odonata2'

and choose 'download selected data' in the bottom top-down menu to download the data sets.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 SpeciesList: List of all samples used in this study.

Table S2 CO1-L1O: Results for CO1 leave one out test.

Table S3 ND1-L1O: Results for ND1 leave one out test.

Table S4 Random_substitution_test: Results for the random substitution test.

Table S5 CAOS-BOLD: The 234 sequences of the CO1 data set were tested on the CAOS-Classifier and BOLD.