Molecular taxonomy in pholcid spiders (Pholcidae, Araneae): evaluation of species identification methods using CO1 and 16S rRNA

JONAS J. ASTRIN, BERNHARD A. HUBER, BERNHARD MISOF & CORNELYA F. C. KLÜTSCH

Accepted: 15 April 2006
doi: 10.1111/j.1463-6409.2006.00239.x

Astrin, J. J., Huber, B. A., Misof, B. & Klütsch, C. F. C. (2006). Molecular taxonomy in pholcid spiders (Pholcidae, Araneae): evaluation of species identification methods using CO1 and 16S rRNA. — *Zoologica Scripta*, *35*, 441–457.

The identification of species using molecular characters is a promising approach in alpha taxonomy and in any discipline depending on reliable assignment of specimens. Previous studies have shown the feasibility of the method, but considerable controversy persists. In this study, we use pholcid spiders in an effort to address two main issues. First, we evaluate and calibrate molecular species (re-)identification within a closely related group of organisms by using specimens that are morphologically unambiguously either conspecific or not. Species limits hypothesized a priori based on morphology were almost universally reconstructed by both mitochondrial markers used. Second, we focus on species identification methodology in a morphology-calibrated scenario, i.e. on how to assess the quality of a dataset and of the method used to obtain distance estimates (e.g. choice of markers, alignment strategy, type of distance data). We develop a number of statistical estimators permitting the measurement and communication of the clarity of species boundaries in a dataset and discuss their benefits and drawbacks. We propose that box plots rather than histograms are the superior tool for graphically illustrating taxonomic signal and that the median is a more appropriate measure of central tendency than the mean. Applying the suggested tools to our data, we propose that in molecular species identification, indel-related alignment uncertainties may often be even advantageous (by accentuating taxonomy-relevant information) and we conclude that — at least for our dataset — 16S is better suited to taxonomy than CO1.

Jonas J. Astrin, Bernhard A. Huber, Bernhard Misof, Cornelya F. C. Klütsch, Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany. Correspondence: J. Astrin, E-mail: j.astrin.zfmk@uni-bonn.de and B. Huber, E-mail: b.huber.zfmk@uni-bonn.de

Introduction

The knowledge gaps in our taxonomic system, the shortage of taxonomists and the resulting handicap to biodiversity management and conservation have been called the 'taxonomic impediment' by the Convention on Biological Diversity (cf. Workshop on 'Removing the Taxonomic Impediment' 1998, Darwin, Australia; http://www.biodiv.org/programmes/cross-cutting/taxonomy/darwin-declaration.asp). To alleviate this impediment, several potential solutions are being actively debated. If not through political or sociological measures, most of these suggestions focus either on changing the way in which taxonomic information is organized (e.g. Winker 1999; Godfray 2002; Wilson 2003, 2004; but see Thiele & Yeates 2002; Scoble 2004), on changing the nomenclatural system (Godfray 2002; Minelli 2003; Tautz *et al.* 2003; but see Seberg *et al.* 2003; Knapp *et al.* 2004; Lughadha 2004), or on changing the priority of characters used for taxonomic research.

Here, our concern is with the last point. Molecular genetic information is rapidly gaining support as an ample source of easily quantifiable, discrete taxonomic characters that can often be homologized over a wide range of taxa and allow rapid standardized analysis even by nonspecialists (among many others: Langor & Sperling 1995; Palumbi & Cipriano 1998; Townson *et al.* 1999; Puorto *et al.* 2001; Westheide & Hass-Cordes 2001; Bond & Sierwald 2003; Hebert *et al.* 2004a; Johnson *et al.* 2004; Paquin & Hedin 2004; López-Legentil & Turon 2005; Markmann & Tautz 2005; Page *et al.* 2005; Vences *et al.* 2005a; Monaghan *et al.* 2006; Smith *et al.* 2006).

Previously, genetic techniques in taxonomy were largely limited to groups characterized by simple or extremely small-scale morphology, to specific developmental stages, to parts of organisms (e.g. root parts) or to mixed samples (e.g. dietary or combined host/endoparasite samples). Recently, however, it has been proposed that these methods be applied to all branches of the tree of life (Tautz *et al*. 2002, 2003) or at least to all animal life (Hebert *et al*. 2003a,b). The former proposal refers to a system in which species should be routinely described and identified by the additional means of DNA sequence analysis; the latter centres mostly on standardized species re-identification through DNA 'barcodes'. Establishing molecular characters as a standard taxonomic tool equivalent to, or even superior to, morphological data has been met with harsh criticism (e.g. Dunn 2003; Lipscomb *et al*. 2003; Seberg *et al*. 2003; Scotland *et al*. 2003; Sperling 2003; Will & Rubinoff 2004).

While much effort has been given to weighing the general 'pros and cons' of DNA taxonomy ('DNA taxonomy' is used here in a technical sense only, as a taxonomic practice which employs DNA sequence analysis as a tool equivalent to, for example, microscopic scrutiny in morphology; but see Tautz *et al*. 2003), there is need for more empirical work calibrating DNA against morphology within specific, manageable groups to achieve adequate species and population sampling (cf. Funk 1999; Funk & Omland 2003; Moritz & Cicero 2004).

Our study focuses on Pholcidae, a group of mostly small, long-legged and inconspicuously coloured web-weaving spiders. They represent one of the most species-rich taxa among haplogyne spiders (Araneomorphae: Haplogynae), with a wide, mostly tropical distribution.

Alpha taxonomy in spiders is commonly approached morphologically, relying in particular on copulatory organs (Huber 2004). Spider genitalia usually present little intraspecific but conspicuous interspecific variation (e.g. Eberhard 1985; Eberhard *et al*. 1998) and are generally considered valuable species-level diagnostic characters able to robustly identify even closely related species. Due to these features, spiders provide good model organisms to test molecularly identified species against solid morphological evidence. Under certain conditions (cf. Barrett & Hebert 2005; but see Prendini 2005), spider taxonomy can itself require molecular approaches (e.g. Bond 2004; Paquin & Hedin 2004). However, in the present study, we chose exclusively morphologically unambiguous cases of close relatives in order to examine whether mitochondrial DNA is able to recover undoubted morphological species. Therefore, we sequenced and analysed the cytochrome *c* oxidase subunit 1 (CO1) and ribosomal large subunit (16S) genes, the two currently most widely applied mitochondrial markers in molecular taxonomy and close areas of research (using both, e.g. Cognato & Vogler 2001; Therriault *et al*. 2004; Ayoub *et al*. 2005; Page *et al*. 2005; Steinke *et al*. 2005; Vences *et al*. 2005a).

We also address specific methodological questions of DNA taxonomy. All of these focus on a scenario of species re-identification. We present and discuss our use of new methods to assess a dataset collectively for its quality, i.e. for the separation or overlap of inter- and intraspecific distances. Through these methods we provide a scaffold that facilitates ascertaining which mode of obtaining distance values delivers the most information (marker choice, alignment strategy, type of distance data) for taxonomy in general or for a particular dataset only. For the latter, even in groups that are morphologically difficult or have received little study, a pilot study can be carried out using some of the least ambiguous species available, thereby giving incentives on which methodology to apply to the group under study or indicating whether a molecular approach might work at all. Species description methodology proper is not touched on here, but useful hints can be gained from the discussed methods. In species description, it is very useful to know whether there is or is not clearly discontinuous variation in a targeted character system. Furthermore, box plots permit the identification and reassessment of conspicuous cases. The following questions are addressed:

1 How much do mitochondrial markers contribute to the re-identification of species, measured by their recovery of morphological species?

2 How can the performance of a specific marker or analytic procedure be measured quantitatively against extrinsic (e.g. morphological) evidence?

3 Which is the best way to simultaneously visualize distances both within and between species? Usually, histograms are applied (Dalebout *et al*. 1998, 2002; Hebert *et al*. 2003b, 2004b; Barrett & Hebert 2005; Vences *et al*. 2005a,b), but these have to be drawn in separate graphics, otherwise they are in danger of suffering from data interference.

4 For the above-mentioned purposes, can we assume a normal distribution and use estimators as the arithmetic mean?

5 What is the impact of the alignment procedure and is it advantageous to include indels for questions of DNA taxonomy? It has been argued that indel-featuring markers are less appropriate for DNA taxonomy because of the increased effort in sequence alignment (Hebert *et al*. 2003a; Seberg *et al*. 2003; Prendini 2005). However, the effect of using alignments which have received little or no manual adjustment has never been tested in this context.

6 Regarding the identification of pholcid species, how do 16S and CO1 perform compared to each other? In zoology, empirical studies that test whether a specific marker could be more suitable for DNA taxonomy than another are rare (but see e.g. Steinke *et al*. 2005; Vences *et al*. 2005a,b).

Materials and methods

Taxon sampling and vouchering

Specimens were acquired from a wide range of localities, mostly centred on the Neotropics. Material was fixed in near-absolute ethanol and stored in the same medium at –20 °C. We sampled 113 pholcid specimens, representing 52 morphospecies and 17 genera. These resulted in a total of 179 sequences. Additionally, 36 pholcid sequences were retrieved from GenBank, all originating from the same study (Bruvo-Madaric *et al*. 2005). Their accession numbers and specimen information are listed in Table 2. Altogether, 112 CO1 and 103 16S sequences were analysed from 61 species and 23 genera.

Roughly half of the species and almost two thirds of all individuals belonged to one of three genera: *Mesabolivar* González-Sponga, 1998, *Metagonia* Simon, 1893 and *Pholcus* Walckenaer, 1805. Within these genera, seven species were represented by four or more individuals from identical or different populations, or from both. Detailed specimen information and collecting data are listed in Table 1. All sequences have been submitted to GenBank (16S: DQ667748–DQ667836; CO1: DQ667854–DQ667943).

Voucher specimens are deposited at the ZFMK (Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany). Likewise, total genomic DNA can be accessed at the ZFMK under the specified DNA voucher numbers.

DNA extraction, amplification and sequencing

Total genomic DNA was extracted either from single whole individuals or from whole prosomata or opisthosomata, using the Nucleo Spin Tissue extraction kit (Macherey-Nagel, Dueren, Germany), following the manufacturer's protocol.

A single set of primers was used for each gene. For the 16S rRNA (LSU) gene: 16s1471-mod: 5′-GCCTGTTTAW-CAAAAACAT-3′ (Crandall & Fitzpatrick 1996; modification: Ch. D. Schubart, unpubl.) and 16sbr-H-mod: 5′-CCG-GTYTGAACTCARATCAYGT-3′ (Palumbi *et al*. 2002; modification: Ch. D. Schubart, unpubl.). The given primer set encloses a fragment of 420–460 bp (aligned length: 504 bp), localized close to the 3′ end of the 16S ribosomal RNA gene. For the CO1 gene: C1-J-1751-SPID: 5′-GAGCTCCTGATATAGCTTTTCC-3′ (Hedin & Maddison 2001), together with C1-N-2191: 5′-CCCGGTAAAAT-TAAAATATAAACTTC-3′ (reviewed in Simon *et al*. 1994). This primer set amplifies 440 bp from the 5′ section of the cytochrome *c* oxidase subunit 1 gene. It is nested within a fragment that has been proposed as a single standardized marker for DNA barcoding studies (Hebert *et al*. 2003a).

PCR reaction mixes (50 µL) contained 125 nmol $MgCl_2$, 5 µL 10× PCR-buffer, 25 pmol of forward and reverse primer each, 5 pmol dNTPs, 1.75 Units of *Taq* polymerase, and 5 µL total DNA template (undiluted). The lab chemicals were purchased from Sigma-Aldrich (Steinheim, Germany).

Thermal cycling for CO1 was performed on a GeneAmp PCR System 2700 (Applied Biosystems, Foster City, CA, USA), for 16S on a TGradient block (Biometra, Göttingen, Germany). The CO1 program encompassed a single cycle set of 30 repeats. Each cycle included 20 s denaturation at 94 °C, 20 s annealing at 48 °C and 40 s extension at 72 °C. The 16S program consisted of two cycle sets, which together constitute an unorthodox mixture of a 'Touch Down' and a 'Step Up' routine (Palumbi 1996). First cycle set (7–9 repeats): 30 s denaturation at 94 °C, 30 s annealing at 55 °C (–1 °C per cycle) and 50 s extension at 72 °C. Second cycle set (23 repeats): 30 s denaturation at 94 °C, 30 s annealing at 50 °C and 50 s extension at 72 °C.

Problematic cases were dealt with by adding 0.4% formamide or by ramping up the annealing temperature. Due to high interspecific genetic variation, some taxa included in this study were impossible to amplify using the employed primers.

Sequence runs were carried out on an ABI Prism 377 sequencer (Applied Biosystems). We used the primers C1-N-2191 (for CO1) and 16s1471-mod (for 16S) for single-stranded sequencing. In cases of suboptimal signal, the second strand was sequenced or the first strand re-sequenced.

Data analysis

Chromatograms were checked by eye. Aligned, truncated sequence length was 312 bp for CO1 and 287 bp for 16S. The reduction resulted from the need to accommodate GenBank haplotypes and shorter, new sequences and because no terminal gaps (coded as missing) were allowed. This was in order to avoid any bias that would otherwise have resulted for the distance calculations. Although a usual and useful practice in phylogeny, coding terminal gaps or other sections of the alignment as missing characters would be detrimental to taxonomy. Gaps required to account for indels in the 16S alignment were treated as fifth character states and as individual characters in order to accommodate indels as evolutionary events (see Giribet & Wheeler 1999; Hancock & Vogler 2000).

Sequence alignment was done in MUSCLE (Multiple Sequence Comparison by Log-Expectation) ver. 3.52 (Edgar 2004a,b), using default parameters and the refine option in an additional run. Apart from delivering reliable alignments, MUSCLE conveys the advantage of processing even large datasets especially fast.

The resulting alignments were checked by eye using BioEdit ver. 7.0.4.1 (Hall 1999). However, not much time was assigned to this task in order to comply with the notion that molecular species identification should follow a straightforward alignment routine. The few very obvious cases of misalignment were changed manually. Final alignments are available as Supplementary Material with the online version of this paper and from the authors.

PAUP* ver. 4.0b10 (Swofford 1998) was used for uncorrected or *p*-distance transformations, for evaluating base

Table 1 List of specimen data and voucher numbers. All specimen vouchers and DNA vouchers were deposited at the ZFMK (Zoologisches Forschungsmuseum Alexander Koenig), Bonn, Germany. 'x' means that the respective sequence was analysed.

Taxon	Voucher #	DNA Voucher #	CO1	16S	Collecting Data
Artema atlanta Walckenaer, 1837	pb05-G101	DNA05-JA119	x	x	EGYPT, Cairo i. 2002 (H. El-Hennawy)
Carapoia paraguaensis González-Sponga, 1998	pb05-V37	DNA05-JA97	x	x	VENEZUELA, km 44 from El Dorado xii. 2002 (B.A. Huber)
Carapoia ubatuba, Huber, 2005	pb05-B2	DNA05-JA102	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Carapoia ubatuba, Huber, 2005	pb05-B2	DNA05-JA9	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Coryssocnemis simla Huber, 2000	pb05-G21	DNA05-JA107	x	x	TRINIDAD, Arima-Blanchisseuse Road. iii. 2002 (Starr & Sewlal)
Coryssocnemis simla Huber, 2000	pb05-V50	DNA05-JA100	x	x	VENEZUELA, Cascada Chorro xii. 2002 (B.A. Huber)
Kaliana yuruani Huber, 2000	pb05-V53	DNA05-JA101	x	x	VENEZUELA, km 109 from El Dorado xii. 2002 (B.A. Huber)
Mecolaesthus arima Huber, 2000	pb05-G22	DNA05-JA108	—	x	TRINIDAD, Arena Forest iv. 2002 (Starr & Sewlal)
Mecolaesthus longissimus Simon, 1893	pb05-V12	DNA05-JA89	x	x	VENEZUELA, Tovar xii. 2002 (B.A. Huber)
Mesabolivar aurantiacus (Mello-Leitão, 1930)	pb05-G1	DNA05-JA25	x	x	TRINIDAD, Arena Forest vi. 2002 (Starr & Sewlal)
Mesabolivar brasiliensis (Moenkhaus, 1898)	pb05-B16	DNA05-JA51	x	—	BRAZIL, PN Cantareira xii. 2003 (B.A. Huber)
Mesabolivar brasiliensis (Moenkhaus, 1898)	pb05-G2	DNA05-JA26	x	—	BRAZIL, São Paulo, P.E. Cantareira vi. 2001 (Pinto-Rocha & Rheims)
Mesabolivar cyaneomaculatus (Keyserling, 1891)	pb05-B21	DNA05-JA52	x	—	BRAZIL, Est. Alto da Serra xii. 2003 (B.A. Huber)
Mesabolivar cyaneotaeniatus (Keyserling, 1891)	pb05-B32	DNA05-JA57	x	x	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar cyaneotaeniatus (Keyserling, 1891)	pb05-B32	DNA05-JA58	x	x	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar cyaneotaeniatus (Keyserling, 1891)	pb05-B33	DNA05-JA18	—	x	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar cyaneotaeniatus (Keyserling, 1891)	pb05-B33	DNA05-JA59	x	x	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar cyaneotaeniatus (Keyserling, 1891)	pb05-B34	DNA05-JA60	x	x	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar eberhardi Huber, 2000	pb05-V35	DNA05-JA93	x	x	VENEZUELA, Canaima near Salto Ara xii. 2002 (B.A. Huber)
Mesabolivar eberhardi Huber, 2000	pb05-V49	DNA05-JA99	—	x	VENEZUELA, Cascada Chorro xii. 2002 (B.A. Huber)
Mesabolivar eberhardi Huber, 2000	pb05-V5	DNA05-JA34	x	x	VENEZUELA, PN Ávila xii. 2002 (B.A. Huber)
Mesabolivar eberhardi Huber, 2000	pb05-V62	DNA05-JA35	x	—	VENEZUELA, Yacambú xii. 2002 (B.A. Huber)
Mesabolivar luteus Huber, 2000	pb05-G6	DNA05-JA41	x	x	BRAZIL, M. Gerais, Cotas Altas, S. Caraça iv. 2002 (A.J. Santos)
Mesabolivar sp. 1	pb05-B5	DNA05-JA29	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Mesabolivar sp. 2	pb05-B15	DNA05-JA14	—	x	BRAZIL, PN Cantareira xii. 2003 (B.A. Huber)
Mesabolivar sp. 2	pb05-B15	DNA05-JA50	x	x	BRAZIL, PN Cantareira xii. 2003 (B.A. Huber)
Mesabolivar sp. 3	pb05-B36	DNA05-JA19	x	x	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar sp. 3	pb05-B41	DNA05-JA32	x	x	BRAZIL, Hotel Fazenda Colina Verde xii. 2003 (B.A. Huber)
Mesabolivar sp. 4	pb05-V14	DNA05-JA21	x	—	VENEZUELA, km 126 from El Dorado xii. 2002 (B.A. Huber)
Mesabolivar sp. 4	pb05-V17	DNA05-JA90	x	—	VENEZUELA, Bolívar, 'km 118'-village xii. 2002 (B.A. Huber)
Mesabolivar sp. 5	pb05-B4	DNA05-JA10	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Mesabolivar sp. 5	pb05-B4	DNA05-JA28	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Mesabolivar sp. 6	pb05-B23	DNA05-JA16	—	x	BRAZIL, Est. Alto da Serra xii. 2003 (B.A. Huber)
Mesabolivar sp. 6	pb05-B23	DNA05-JA53	x	x	BRAZIL, Est. Alto da Serra xii. 2003 (B.A. Huber)
Mesabolivar sp. 7	pb05-B14	DNA05-JA49	x	—	BRAZIL, PN Cantareira xii. 2003 (B.A. Huber)
Mesabolivar togatus (Keyserling, 1891)	pb05-B29	DNA05-JA54	x	—	BRAZIL, Sâo Paulo, Zoo xii. 2003 (B.A. Huber)
Mesabolivar togatus (Keyserling, 1891)	pb05-G4	DNA05-JA62	x	—	BRAZIL, Minas Gerais, Taisbeiras iv. 2002 (A.J. Santos)
Metagonia belize Gertsch, 1986	pb05-J140a	DNA05-JA118	—	x	BELIZE, Cockcomb Basin xi. 2003 (J.J. Astrin)
Metagonia belize Gertsch, 1986	pb05-J140a	DNA05-JA38	—	x	BELIZE, Cockcomb Basin xi. 2003 (J.J. Astrin)
Metagonia belize Gertsch, 1986	pb05-J140a	DNA05-JA78	—	x	BELIZE, Cockcomb Basin xi. 2003 (J.J. Astrin)
Metagonia belize Gertsch, 1986	pb05-J140a	DNA05-JA79	—	x	BELIZE, Cockcomb Basin xi. 2003 (J.J. Astrin)
Metagonia conica Simon, 1893	pb05-V8	DNA05-JA86	x	—	VENEZUELA, Tovar xii. 2002 (B.A. Huber)
Metagonia delicata (O. Pickard-Cambridge, 1895)	pb05-J140	DNA05-JA117	—	x	BELIZE, Cockcomb Basin xi. 2003 (J.J. Astrin)
Metagonia mariguitarensis (González-Sponga, 1998)	pb05-V43	DNA05-JA23	x	x	VENEZUELA, Mariguitar xii. 2002 (B.A. Huber)
Metagonia mariguitarensis (González-Sponga, 1998)	pb05-V43	DNA05-JA94	x	x	VENEZUELA, Mariguitar xii. 2002 (B.A. Huber)
Metagonia paranapiacaba Huber, Rheims, Brescovit, 2005	pb05-B22	DNA05-JA15	x	x	BRAZIL, Est. Alto da Serra xii. 2003 (B.A. Huber)
Metagonia sp. 1	pb05-B1	DNA05-JA27	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Metagonia sp. 2	pb05-B6	DNA05-JA11	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Metagonia sp. 3	pb05-B12	DNA05-JA12	x	x	BRAZIL, PN Cantareira xii. 2003 (B.A. Huber)
Metagonia sp. 4	pb05-V54	DNA05-JA4	x	—	VENEZUELA, km 109 from El Dorado xii. 2002 (B.A. Huber)
Metagonia sp. 4	pb05-V54	DNA05-JA8	x	x	VENEZUELA, km 109 from El Dorado xii. 2002 (B.A. Huber)
Metagonia sp. 4	pb05-V54	DNA05-JA95	x	x	VENEZUELA, km 109 from El Dorado xii. 2002 (B.A. Huber)
Metagonia sp. 5	pb05-V63	DNA05-JA24	x	—	VENEZUELA, Yacambú xii. 2002 (B.A. Huber)
Metagonia sp. 6	pb05-V2	DNA05-JA84	x	x	VENEZUELA, Canaima, village forest xii. 2002 (B.A. Huber)
Metagonia sp. 6	pb05-V2	DNA05-JA85	x	x	VENEZUELA, Canaima, village forest xii. 2002 (B.A. Huber)
Metagonia sp. 6	pb05-V2	DNA05-JA115	x	x	VENEZUELA, Canaima, village forest xii. 2002 (B.A. Huber)
Metagonia sp. 6	pb05-V2	DNA05-JA2	—	x	VENEZUELA, Canaima, village forest xii. 2002 (B.A. Huber)

Table 1 *Continued*

Taxon	Voucher #	DNA Voucher #	CO1	16S	Collecting Data
Metagonia sp. 6	pb05-V2	DNA05-JA3	x	x	VENEZUELA, Canaima, village forest xii. 2002 (B.A. Huber)
Metagonia sp. 6	pb05-V2	DNA05-JA96	x	x	VENEZUELA, Canaima, village forest xii. 2002 (B.A. Huber)
Micropholcus fauroti (Simon, 1887)	pb05-G38	DNA05-JA66	x	x	CUBA, La Habana, indoors iii. 2001 (A. Pérez Gonzalez)
Ninetine gen. sp. indet.	pb05-V1	DNA05-JA83	x	—	VENEZUELA, Canaima, Salto Sapo xii. 2002 (B.A. Huber)
Pholcophora americana Banks, 1896	pb05-G89	DNA05-JA45	x	x	USA, CA, Mono County, Inyo Forest vi. 2003 (Paquin & Duperre)
Pholcophora americana Banks, 1896	pb05-G91	DNA05-JA111	x	x	USA, OR, Josephine Co., Siskiyou Forest ix. 2003 (Paquin & Wytrykush)
Pholcophora americana Banks, 1896	pb05-G91	DNA05-JA75	—	x	USA, OR, Josephine Co., Siskiyou Forest ix. 2003 (Paquin & Wytrykush)
Pholcophora americana Banks, 1896	pb05-G92	DNA05-JA46	x	x	USA, MT, Missoula Co., Lolo Forest ix. 2003 (Paquin & Wytrykush)
Pholcus opilionoides (Schrank, 1781)	pb05-G58	DNA05-OJ1	x	—	AUSTRIA, Reitpoidl near Linz, around bldg. vii. 2000 (B.A. Huber)
Pholcus opilionoides (Schrank, 1781)	pb05-G58	DNA05-OJ21	x	—	AUSTRIA, Reitpoidl near Linz, around bldg. vii. 2000 (B.A. Huber)
Pholcus opilionoides (Schrank, 1781)	pb05-G61	DNA05-JA1	x	x	AUSTRIA, between Frauenstein & Ramsau viii. 2003 (B.A. Huber)
Pholcus opilionoides (Schrank, 1781)	pb05-G61	DNA05-JA6	x	x	AUSTRIA, between Frauenstein & Ramsau viii. 2003 (B.A. Huber)
Pholcus opilionoides (Schrank, 1781)	pb05-G61	DNA05-JA7	x	x	AUSTRIA, between Frauenstein & Ramsau viii. 2003 (B.A. Huber)
Pholcus opilionoides (Schrank, 1781)	pb05-J280	DNA05-JA39	x	x	AUSTRIA, Burgenland, Zeiler Berg ix. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-B26	DNA05-JA30	x	x	BRAZIL, Est. Alto da Serra xii. 2003 (B.A. Huber)
Pholcus phalangioides (Fuesslin, 1775)	pb05-G50	DNA05-OJ9	—	x	BRAZIL, Rio de Janeiro, Itatiaia vi. 2001 (H.F. Japyassú)
Pholcus phalangioides (Fuesslin, 1775)	pb05-G52	DNA05-JA69	x	x	USA, SC, Pickens Co., Clemson University iii. 2001 (W. Reeves)
Pholcus phalangioides (Fuesslin, 1775)	pb05-J17	DNA05-JA37	—	x	MADEIRA/PORTUGAL, São Vicente, Laranjal ii. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-J17	DNA05-OJ11 & 22	x	x	MADEIRA/PORTUGAL, São Vicente, Laranjal ii. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-J17	DNA05-OJ12	x	—	MADEIRA/PORTUGAL, São Vicente, Laranjal ii. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-J17	DNA05-OJ23	—	x	MADEIRA/PORTUGAL, São Vicente, Laranjal ii. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-J17	DNA05-OJ30	—	x	MADEIRA/PORTUGAL, São Vicente, Laranjal ii. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-J271	DNA05-JA64	x	x	SPAIN, Teulada, Moraira, indoors xii. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-G100a	DNA05-JA43	x	x	GERMANY, Bonn, ZFMK, basement vii. 2004 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-G100	DNA05-OJ34	x	x	GERMANY, Bonn, ZFMK, basement x. 2003 (J.J. Astrin)
Pholcus phalangioides (Fuesslin, 1775)	pb05-G55	DNA05-OJ10	x	x	GERMANY, Bonn, ZFMK, basement ii. 2002 (B.A. Huber)
Pholcus manueli Gertsch, 1937	pb05-G25	DNA05-JA42	x	x	USA, PA, ~20 mi NE Stroudsburg vii. 2000 (B.A. Huber)
Pholcus manueli Gertsch, 1937	pb05-G25	DNA05-JA63	x	x	USA, PA, ~20 mi NE Stroudsburg vii. 2000 (B.A. Huber)
Pholcus manueli Gertsch, 1937	pb05-G63	DNA05-OJ29	x	—	USA, PA, ~20 mi NE Stroudsburg vii. 2000 (B.A. Huber)
Pholcus manueli Gertsch, 1937	pb05-G63	DNA05-OJ31	x	—	USA, PA, ~20 mi NE Stroudsburg vii. 2000 (B.A. Huber)
Pholcus manueli Gertsch, 1937	pb05-G63	DNA05-OJ35	x	—	USA, PA, ~20 mi NE Stroudsburg vii. 2000 (B.A. Huber)
Pholcus manueli Gertsch, 1937	pb05-G28	DNA05-JA65	x	x	USA, PA, NW Scranton, Slumber Vly. vi. 2000 (B.A. Huber)
Pholcus sp. 1	pb05-G44	DNA05-JA68	x	—	CAPE VERDE, Fogo, Cueva de Gaucho i. 1999 (P. Oromi)
Physocyclus dugesi Simon, 1893	pb05-G71	DNA05-JA71	x	—	COSTA RICA, San Pedro de Montes de Oca v. 2002 (Peretti & Eberhard)
Physocyclus globosus (Taczanowski, 1874)	pb05-G60	DNA05-OJ28	x	x	COMORO ISLANDS, Anjouan, Pomoni v. 2003 (Jocqué & Spiegel)
Physocyclus globosus (Taczanowski, 1874)	pb05-G60	DNA05-OJ36	—	x	COMORO ISLANDS, Anjouan, Pomoni v. 2003 (Jocqué & Spiegel)
Physocyclus globosus (Taczanowski, 1874)	pb05-G77	DNA05-JA44	—	x	COMORO ISLANDS, Anjouan, Hombo v. 2003 (Jocqué & Spiegel)
Physocyclus globosus (Taczanowski, 1874)	pb05-J93	DNA05-JA76	—	x	GUATEMALA, Lívingston, indoors xi. 2003 (J.J. Astrin)
Physocyclus globosus (Taczanowski, 1874)	pb05-J93	DNA05-JA77	—	x	GUATEMALA, Lívingston, indoors xi. 2003 (J.J. Astrin)
Physocyclus globosus (Taczanowski, 1874)	pb05-V41	DNA05-JA36	—	x	VENEZUELA, Mariguitar xii. 2002 (B.A. Huber)
Physocyclus sp.	pb05-G97	DNA05-JA5	—	x	MEXICO, Hidalgo, PN Meztitlán iii. 2003 (A. Peretti)
Physocyclus sp.	pb05-G97	DNA05-JA70	x	x	MEXICO, Hidalgo, PN Meztitlán iii. 2003 (A. Peretti)
Priscula binghamae (Chamberlin, 1916)	pb05-J253	DNA05-JA114	x	x	BOLIVIA, La Paz, indoors iv. 2003 (J.J. Astrin)
Priscula binghamae (Chamberlin, 1916)	pb05-J253	DNA05-JA82	x	x	BOLIVIA, La Paz, indoors iv. 2003 (J.J. Astrin)
Priscula sp.	pb05-V24	DNA05-JA22	x	x	VENEZUELA, Lara, near Cueva Guacharo xii. 2002 (B.A. Huber)
Priscula sp.	pb05-V24	DNA05-JA91	x	x	VENEZUELA, Lara, near Cueva Guacharo xii. 2002 (B.A. Huber)
Priscula venezuelana Simon, 1893	pb05-V28	DNA05-JA92	x	x	VENEZUELA, Aragua, PN Pittier, Rancho Grande xii. 2002 (B.A. Huber)
Psilochorus simoni (Berland, 1911)	pb05-G99	DNA05-JA106	x	—	GERMANY, Bonn, indoors ii. 2005 (B.A. Huber)
Psilochorus sp.	pb05-G16	DNA05-JA105	—	x	USA, UT, Wayne Co., 2 mi N Bicknell vii. 2003 (P. Paquin *et al.*)
Stenosfemuraia sp.	pb05-V10	DNA05-JA20	x	x	VENEZUELA, Tovar xii. 2002 (B.A. Huber)
Stenosfemuraia sp.	pb05-V10	DNA05-JA87	x	x	VENEZUELA, Tovar xii. 2002 (B.A. Huber)
Systenita prasina Simon, 1893	pb05-V11	DNA05-JA88	x	—	VENEZUELA, Tovar xii. 2002 (B.A. Huber)
Systenita prasina Simon, 1893	pb05-V45	DNA05-JA98	x	—	VENEZUELA, Cerro Picacho xii. 2002 (B.A. Huber)
Tupigea sp. 1	pb05-B3	DNA05-JA48	x	x	BRAZIL, Fazenda Angelim xii. 2003 (B.A. Huber)
Tupigea sp. 2	pb05-B13	DNA05-JA13	x	x	BRAZIL, PN Cantareira xii. 2003 (B.A. Huber)
Tupigea sp. 3	pb05-B45	DNA05-JA61	x	x	BRAZIL, Hotel Fazenda Colina Verde xii. 2003 (B.A. Huber)

Table 2 List of accesssion numbers for the sequences retrieved from GenBank (www.ncbi.nih.gov).

Taxon	Acc. # CO1	Acc. # 16S	Collecting Data
Artema atlanta	AY560771.1	AY560663.1	USA, Florida, Sarasota date unknown
Ciboneya antraia	AY560794.1	AY560665.1	CUBA, Pinar del Rio, Ceja Francisco vii. 2000
Coryssocnemis simla	AY560773.1	—	TRINIDAD, Arima iii. 2002
Crossopriza lyoni (isolate 1)	AY560774.1	—	S. AFRICA, Kwazulu Natal, Mkuze, bldg. iii. 2001
Crossopriza lyoni (isolate 2)	AY560775.1	AY560667.1	BELGIUM, Antwerp viii. 2002
Crossopriza lyoni (isolate 3)	—	AY560666.1	VENEZUELA, Sucre xi. 2002
Holocnemus pluchei (isolate 1)	AY560776.1	—	AUSTRIA, Vienna viii. 2001
Holocnemus pluchei (isolate 2)	AY560777.1	—	AUSTRIA, Vienna viii. 2001
Kaliana yuruani	—	AY560668.1	VENEZUELA, Bolívar xii. 2002
Mesabolivar aurantiacus (isolate 1)	AY560778.1	AY560669.1	TRINIDAD & TOBAGO, Arena Forest iv. 2002
Mesabolivar aurantiacus (isolate 2)	AY560779.1	AY560670.1	TRINIDAD & TOBAGO, Arena Forest iv. 2002
Mesabolivar brasiliensis	AY560780.1	—	BRAZIL, São Paulo, P.E.Cantareira vi. 2001
Mesabolivar cyaneotaeniatus	AY560781.1	AY560671.1	BRAZIL, Rio de Janeiro, Itatiaia vi. 2001
Metagonia sp. BB-2004a (here: *M.* sp. 3)	AY560783.1	—	BRAZIL, São Paulo, P. E. Cantareira vi. 2001
Metagonia sp. BB-2004b (here: *M.* sp. 7)	AY560784.1	AY560673.1	BRAZIL, Iporanga, São Paulo, Petar vi. 2002
Pholcus opilionoides	—	AY560674.1	AUSTRIA, Upper Austria: Reitpoidl vii. 2000
Pholcus phalangioides (isolate 1)	—	AY560675.1	AUSTRIA, Vienna x. 2000
Pholcus phalangioides (isolate 2)	—	AY560676.1	AUSTRIA, Vienna x. 2000
Pholcus phalangioides (isolate 3)	—	AY560677.1	AUSTRIA, Vienna x. 2000
Pholcus sp. BB-2004 (here: *Ph. manneli*)	AY560786.1	—	USA, PA, ~20 mi NE Stroudsburg
Physocyclus dugesi	AY560787.1	—	COSTA RICA, San Pedro Montes Oca v. 2002
Physocyclus globosus	AY560788.1	—	CUBA, La Habana iii. 2001
Psilochorus itaguyrussu	AY560782.1	AY560672.1	BRAZIL, São Paulo, P.E.Cantareira vi. 2001
Psilochorus simoni	AY560789.1	—	GERMANY, Bonn ii. 2002
Quamtana embuleni	AY560793.1	—	S. AFRICA, Mpumalanga, Badplaas, iii. 2001
Quamtana vidal	AY560792.1	—	S. AFRICA, Kwazulu Natal, Cape Vidal iv. 2001
Spermophora senoculata	AY560791.1	—	USA, NY, New York v. 2000
Trichocyclus sp. BB-2004	AY560772.1	—	AUSTRALIA, W Austr., Gundaring Res. vi. 2001

composition and information content as well as for reconstruction of neighbor-joining (NJ, Saitou & Nei 1987) trees. NJ, often applied in molecular taxonomy (e.g. Dalebout *et al*. 1998; Moon-van der Staay *et al*. 2001; Floyd *et al*. 2002; Hebert *et al*. 2003a, 2004a,b; Blaxter *et al*. 2004; Hogg & Hebert 2004; Paquin & Hedin 2004; Armstrong & Ball 2005; Barrett & Hebert 2005; Janzen *et al*. 2005; López-Legentil & Turon 2005; Markmann & Tautz 2005; Vences *et al*. 2005a; Ward *et al*. 2005; Hajibabaei *et al*. 2006; Smith *et al*. 2006), was chosen to build a species identification tree (distinct from a tree chosen when striving for phylogenetic accuracy). As an exclusively algorithmic, phenetic procedure, NJ is fast at processing large datasets, but produces only a single uncontested tree. It performs less robustly than phylogenetic methods in cases of incomplete lineage sorting, producing a highly resolved tree even in cases of insufficient evidence. This suggests the inappropriateness of NJ for species descriptions, but (through the increased speed and operational simplicity) not necessarily for standardized species identification approaches.

For translation of DNA into amino acid sequences in CO1, the GenBank Invertebrate Mitochondrial translation table was used with BioEdit. For CO1, several models of sequence evolution were implemented. Testing for normal (Gaussian) distribution of distance observations, the Shapiro–Wilk (on partitioned data, due to its bulk) and Kolmogorov–Smirnov tests were implemented in SPSS.

Three new measures to estimate taxon separation (i.e. within-species vs. between-species distances) are proposed here and are used in this study.

Gap Range. A plain Euclidean measure of the interval between the lowest interspecific and the highest intraspecific observation. A negative value indicates the numerical extent of overlap of both categories (intraspecific vs. interspecific distances). The algebraic sign serves only for illustration purposes, since the *p*-distance denotes a proportion of real nucleotides and cannot be negative.

Overlap. The degree of overlap between categories is expressed by the proportion of observations out of the total number of cases in both categories that falls into the *Gap Range*. The *Gap Range* has to be zero or negative for the *Overlap* to be considered, i.e. the different categories have to intersect.

5-95 Range. By making use of the percentiles which delineate the most marginal 5% of the data, it is possible to gain a focus on the section of the statistical (sub)population closest to the other category, while excluding most if not all outliers between the groups. Thus, one will not truncate the dataset in excess of the usual threshold level of significance ($P = 0.05$). Accordingly, by subtracting the 95th percentile of within-species distances from the 5th percentile of between-species distances, the *5-95 Range* can be obtained. A high value for the *5-95 Range* indicates a good group separation for 95% of the data. On the other hand, if it became zero or less, this would imply absence of separation for the 5%-pruned portion of the data (and possibly much more *Overlap* than just 5% of the observations, since this is uncorrelated). Being derived from the connection of two separate categories, the *5-95 Range* does not necessarily correspond to the range comprising the five percent of the dataset with the 'weakest' taxon separation. So, if a particular dataset is characterized by very far-flung outliers, these can lie beyond the *5-95 Range*. Less than 5% of the data would then be comprised in the *5-95 Range*. This possibility can be checked by comparing the respective percentiles to the greatest and smallest observation, precluding an 'invasion' of the 5%-truncated dataset by outliers if

$$P_5 - x_{max1} \geq 0 \leq x_{min2} - P_{95}$$

with P being the respective percentiles, x_{max1} the maximal intraspecific and x_{min2} the minimal interspecific observation.

Results

Information content and base composition

Of all 16S characters, 75.6% were variable, 72.5% were parsimony informative; in CO1, 61.4% were variable and 57.2% informative. On a finer scale, CO1 codon triplet positions were considered independently. More than half of the overall CO1 variation was localized within third codon positions. Only 2% of the positions of these selectively weak 'wobble' bases were constant for all taxa. Analysing the CO1 amino acid sequences, 50% of the translated codons were variable and 45% were informative. An A+T bias existed for CO1 and 16S (Table 3). This was higher in 16S (70.8%) than in CO1 (65.0%), although third codon positions (77.0%) exceeded the A+T bias of 16S.

Table 3 Base composition. Mean values are given together with standard deviation to illustrate degree of statistical dispersion. Since the hypothesis of normal distribution is rejected by the data, robust statistics would be more appropriate. However, the difference would amount to less than a real unit (1 bp) and is hence negligible in this case.

	A	C	G	T	A+T
16S	27.7 (± 3.3)	12.5 (± 1.6)	16.7 (± 2.8)	43.1 (± 2.5)	70.8 (± 4.1)
CO1	21.3 (± 2.7)	14.0 (± 1.3)	20.9 (± 3.0)	43.7 (± 2.3)	65.0 (± 3.5)
CO1, 1st + 2nd positions	19.1 (± 1.6)	18.9 (± 1.1)	22.1 (± 1.8)	39.9 (± 1.6)	59.0 (± 2.3)
CO1, 3rd positions	25.9 (± 6.2)	4.2 (± 2.9)	18.7 (± 6.7)	51.2 (± 5.6)	77.0 (± 8.0)

Uncorrected distances

Using pairwise *p*-distance data, the pholcids studied generally show a high genetic divergence among species in the two mitochondrial genes (see below). Pholcids are more diverse than other spiders studied (interspecific mean for CO1: 19.8% vs. Hebert *et al.* 2003b: 14.4% *p*-distance; Barrett & Hebert 2005: 16.4% K2P-distance). Length variation due to indels (16S) and evidence for considerable base substitution events (CO1, 16S) were encountered.

The Shapiro–Wilk and Kolmogorov–Smirnov tests clearly rejected the hypothesis of normal distribution for distance observations of both markers ($P < 0.0005$).

Distances between individuals were arranged so as to fit into three categories: (1) within species; (2) within genera but between species; (3) between pholcid genera. Haplotype sharing was not encountered among taxa. Within species, it was common (16 haplotypes shared by 48 individuals in CO1; 21 shared by 60 individuals in 16S). Even when sequences were not identical among conspecifics, genetic divergence was mostly much lower than that registered for individuals of the same genus: variation within species ranged from 0.0 to 10.9% (CO1) and from 0.0 to 12.2% (16S), whereas between pairs of congeneric allospecifics it was 8.7–28.5% (CO1) and 15.0–34.1% (16S). Intraspecific medians lie at 0.0 and 0.2% (CO1, 16S), while interspecific ones lie at 19.6 and 24% (see Table 4). This pattern is shown in Fig. 1. Divergences within genera were often as high as between genera, although the latter tended to have higher values (Table 4).

Table 4 Pairwise *p*-distance values (in percent). For each category are listed: median, interquartile range (IQR; interval into which the 'central' 50% of the data fall), smallest and greatest observation ($x_{min.}$–$x_{max.}$) and — for comparability reasons — the arithmetic mean and its standard deviation.

		Among conspecifics	Among congeners	Between genera
CO1	median (IQR)	0.0 (1.9)	19.6 (4.5)	24.0 (3.2)
	$x_{min.}$–$x_{max.}$	0.0–10.9	8.7–28.5	13.8–32.1
	mean	1.7 (± 0.3)	19.8 (± 0.1)	24.0 (± 0.0)
16S	median (IQR)	0.2 (1.7)	24.0 (7.2)	32.4 (3.8)
	$x_{min.}$–$x_{max.}$	0.0–12.2	15.0–34.1	15.7–44.6
	mean	0.9 (± 0.1)	23.4 (± 0.2)	32.2 (± 0.1)

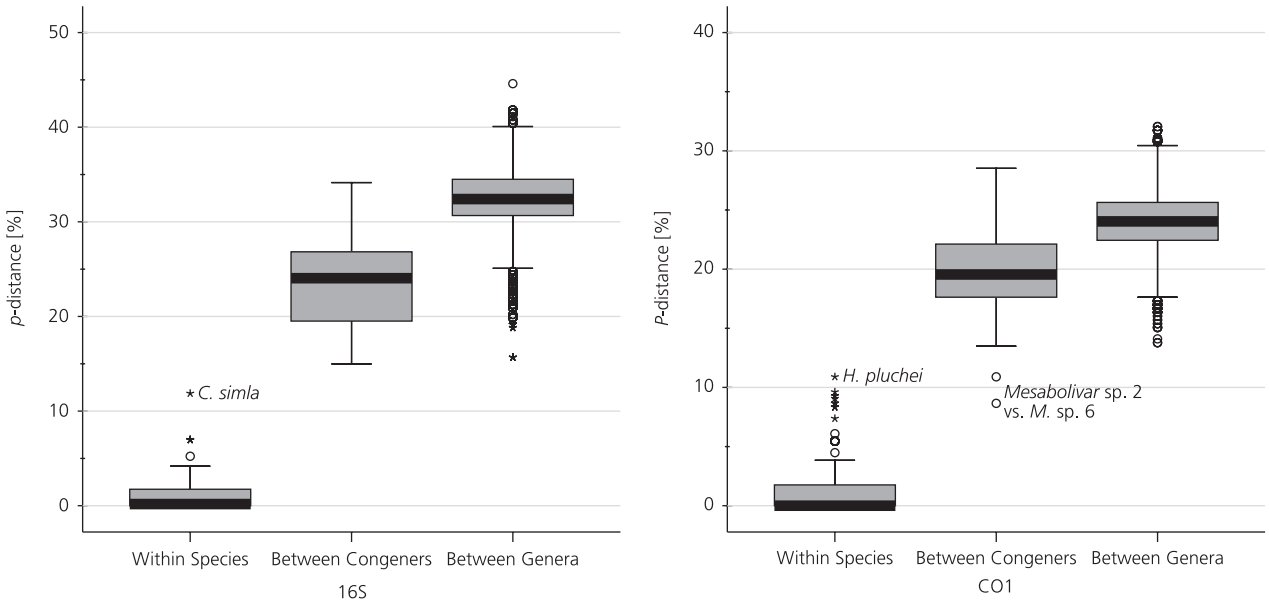

Fig. 1 Box plots of *p*-distances. Boxes indicate interquartile range (IQR: between upper [Q3] and lower [Q1] quartile). Black bar designates median, whiskers indicate values lying within 1.5× the IQR beneath Q1 or 1.5× above Q3. 'Mild' outliers (circles): between 1.5× and 3× IQR; extreme outliers (asterisks): above 3× IQR. Extremes are labelled (see text).

Haplotype variation sometimes appeared to be only loosely correlated with geographical range. In two cases (*Pholcus phalangioides*, *Pholcophora americana*; for a single nucleotide also in *Physocyclus globosus*), haplotypes distinct from others in their population were shared with haplotypes of individuals collected several hundred kilometres away. However, the sampling does not allow in-depth scrutiny of the extent of geographical structuring nor of the population biology of the analysed pholcid taxa.

Clustering of specimens

The topologies of the CO1 and 16S NJ species identification trees shown in Fig. 3 differ considerably from each other. Morphological conspecifics always grouped together — closely in most cases — whereas between species, obvious segregation could be discerned. This segregation was always deeper than within-species separations in 16S and mostly so in CO1 (Fig. 1; see Table 5 for quantitative results).

Splits within species were conspicuous in several cases. In *Coryssocnemis simla* both markers showed a relatively high divergence between individuals from the Lesser Antilles (Trinidad) and northern South America (Venezuela) (*p*-distance 16S: 12.2% [within-species category range: 0.0–12.2%]; CO1: 8.3% [range: 0.0–10.9%]). This is in contrast to the total absence of detectable morphological variation.

The opposite was the case in *Mesabolivar eberhardi*, represented by one specimen each from four geographically separated Venezuelan populations. As usual in this species (Huber 2000), these showed an uncommonly high variation of phenotypic traits. However, molecular analysis revealed only one population to be divergent from a group formed by three other populations (CO1: 8.7% *p*-distance vs. 1.3% within the mentioned group; 16S: 0.3–1.4% within this group, the divergent specimen missing from the 16S dataset).

A split within a screened population appeared in the CO1 tree for *Holocnemus pluchei* (10.9%). This also occurred in *Metagonia* sp. 6 (CO1: 8.7–9.6%; 16S not conspicuous: 3.1–4.2%). In the 16S tree, *Metagonia belize* showed an only slightly marked within-population split (7.0%).

Individuals from separate species were conspicuously close only in the case of *Mesabolivar* sp. 6 with *Mesabolivar* sp. 2 for CO1 (10.9%; category range: 8.7–28.5%). For 16S this was not the case. Morphologically, the two species appear closely related yet clearly distinct.

The implications of the data for pholcid phylogeny will be presented elsewhere, using technically more appropriate methods, additional markers, an adapted sampling, and phylogeny-adjusted alignments (not containing ambiguously aligned sections and including longer sequences by the incorporation of sporadic terminal gaps).

Discussion

Molecules in pholcid species identification

In clustering analysis, morphologically identified conspecifics always proved to be reciprocally monophyletic in the NJ tree (Fig. 3). Thus, molecular markers (genospecies criterion)

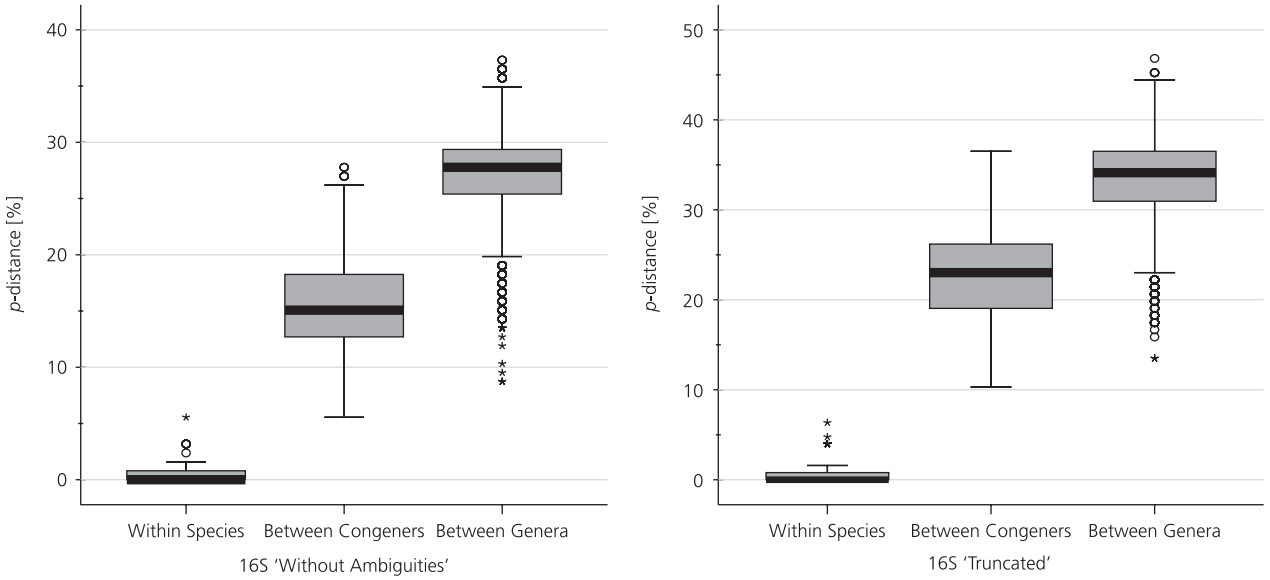

Fig. 2 Box plots of *p*-distances for an alignment from which indel data was removed ('Without Ambiguities') and for a terminally truncated, indel-rich version of the original alignment ('Truncated').

universally recovered individual species hypothesized by morphology (morphospecies criterion). Considering the distance values collectively, it is obvious that levels of intraspecific sequence divergence range much lower than interspecific divergence levels (see Fig. 1). The 16S dataset features a gap between taxonomic categories, whereas in CO1 a slight overlap can be found between these (cf. Table 5). Thus, assignment of specimens based on a 'threshold' value of sequence divergence (Hebert *et al*. 2003a, 2004b; Barret & Hebert 2005) would mostly work for the present scenario. We fully agree on theoretical grounds with Meyer & Paulay (2005) that it would be dangerous to use thresholds as global or pivotal evidence in taxonomy (see also Nielsen & Matz 2006, who stress the importance of the respective mutation rates and effective population sizes in the analysed group).

However, based on the results presented here, we disagree with their point that independently of the group of organisms studied, a 'barcoding gap' between interspecific and intraspecific distance values would likely disappear in studies featuring both dense within-species sampling and closely related species. Based on the constrained taxon dispersal in the strongly structured Neotropics (where most of our collecting was conducted), we might also ask whether Meyer & Paulay's (2005) argument — that the presence of 'barcoding gaps' will be unlikely for supra-regional sampling — can be generalized. Often, the most challenging question for a particular taxon will be how to look for conspicuous separation of within- and between-species distances.

Tree-based taxon clustering as well as statistical taxon separation analysis indicate that molecular evidence does coincide with morphological hypotheses. Hence, species identification based on DNA sequence analysis proved to be feasible for the analysed taxa.

Table 5 Quantitative evaluation of taxon separation. '%, *p*-dist.' refers to values obtained from *p*-distances; *5-95 Range* values: parentheses indicate a 5%-truncated dataset 'invaded' by outliers (see text).

	Gap Range [%, *p*-dist.]	*Overlap* [% of observations]	*5-95 Range* [%, *p*-dist.]
CO1	–2.2	1.2	7.3
16S	+2.8	—	13.0
CO1 1st + 2nd positions	–2.0	4.6	(2.4)
CO1 3rd positions	–8.7	0.9	14.6
CO1 amino acid seq.	–3.9	8.5	3.9
CO1 'Shared Cases'	+1.3	—	6.5
16S 'Shared Cases'	+2.8	—	12.6
CO1 +16S combined	+4.2	—	10.9
16S 'Without Ambiguities'	±0.0	—	7.1
16S 'Truncated'	+4.0	—	14.7

Calibrated measurement of taxon separation

In molecular species identification, one of the most relevant aspects in marker choice — and the only criterion for choosing adequate analytic procedures — is the clarity with which the marker or method routinely reconstructs species limits. To obtain a measure for the degree to which all conspecific distances are separated from all congeneric distances, several methods are conceivable when (a) proper species descriptions based on, for example, phenotypic traits exist and are all based

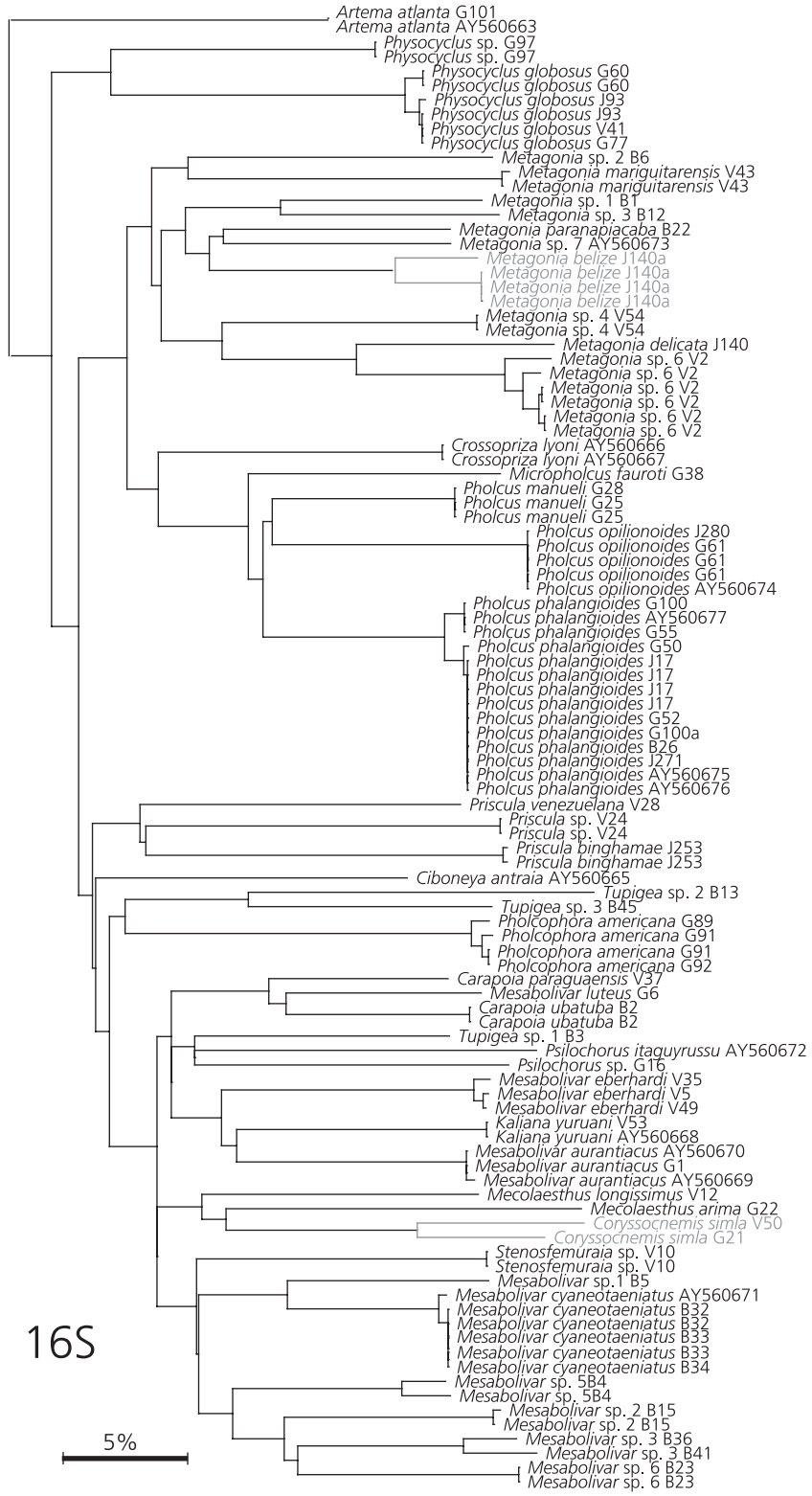

Fig. 3 Neighbor-joining trees of CO1 and 16S. Scale: 5% *p*-distance. Branches and names in grey: species with higher than usual intraspecific variation, offset branches: sister species lying closer to each other than usual. The trees do not aim at phylogenetic accuracy and an outgroup is hence omitted here. Phylogenetic reconstructions will be presented elsewhere.

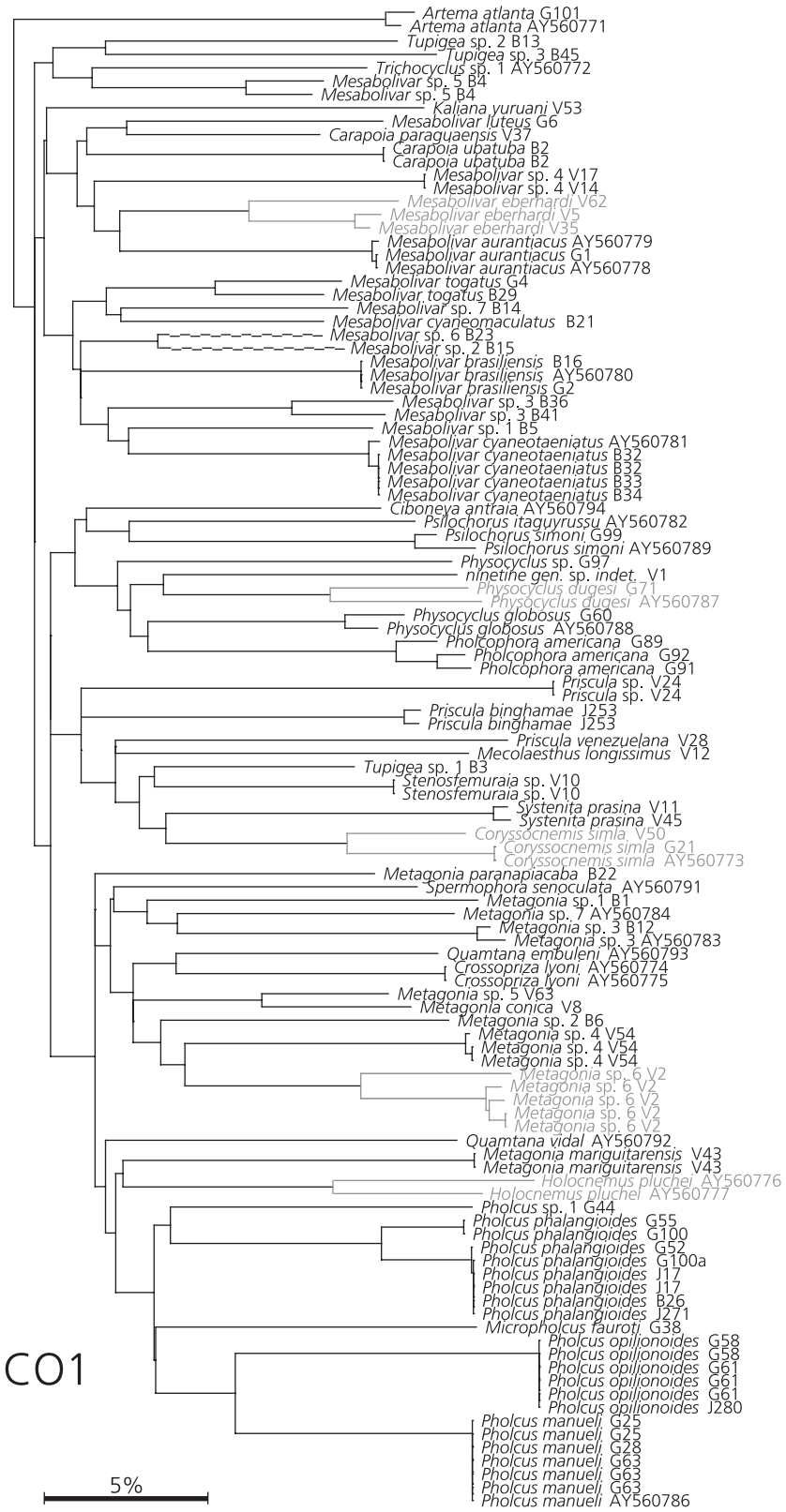

Fig. 3 *Continued.*

on the same concept or criterion, (b) the available material is identified accordingly, and (c) lineage sorting is complete. The three quantitative estimators introduced above and used in Table 5 are here discussed theoretically, together with a graphic representation method useful in molecular taxonomy.

Each of the outlined measures has advantages and drawbacks: the *Gap Range*, although serving as a quick overview tool, is overly influenced by outliers and ignores the actual data distribution between the extremes. The *5-95 Range* is less prone to outliers — at the cost of having to give up those 5% of the data with the worst separation. Using the above-described pigeonholing device, however, one can bring these observations back on a minor, merely qualitative scale, indicating problematic cases when those excluded 5% invade the pruned, main dataset. In contrast, the *Overlap* value can only measure taxon separation if the categories intersect, hence focusing on 'bad' datasets and testing negatively for quality. Of the measures presented here, it is the most labour-intensive to produce since it implies manually sorting the dataset (as long as no software can do this). However, it is also a very valuable estimator, since it will indicate the likelihood of the approach to fail. No measure by itself can adequately describe all important aspects of taxon separation without employing a more complex formula.

We argue that a percentile-based box plot (Tukey 1977; Fig. 1), although not strictly quantifiable as a graphical method, will express taxon separation most comprehensively and clearly. On a 'proto-quantitative' scale, it solves the dilemma of having to choose between the majority of the data and outliers. It offers the crucial advantage of being capable of identifying not only outliers (and hence potentially interesting evolutionary cases) in a simple way, but also contaminants (*sensu* Barnett & Lewis 1994, i.e. cases of inappropriate taxonomy). Thus, the tendency of box plots to put weight on the tails of the distribution is advantageous in the current context. Additionally, box-and-whisker plots appear perfectly suited for DNA taxonomy purposes due to their ability to display several subpopulations both simultaneously and in relation to each other. Different box plots can then be compared as a graphical test criterion for any method, experimental or analytical, used to obtain the tested distances. This is in contrast to the limited versatility that (two-dimensional) histograms show for this task. They have to be either drawn separately (e.g. Hebert *et al.* 2003b, 2004b; Vences *et al.* 2005a) or in one graphic, but then always in a potentially interfering way (see Dalebout *et al.* 1998, 2002; Barrett & Hebert 2005; Vences *et al.* 2005b). We therefore suggest that box plots rather than histograms represent a generally convenient way to depict frequency distributions of distinct subpopulations in DNA taxonomy.

The descriptive statistical methods suggested above are based on robust statistics, i.e. they do not depend on a normal (Gaussian) distribution. The fact that the data used here does not follow a normal distribution implies the necessity of making use of such robust statistics. Thus, it precludes the use of measures connected to assumptions that are valid for normally distributed data only and that would become biased in an asymmetric distribution. Such measures are the arithmetic mean (the truncated mean to a lesser degree), average absolute deviation of the mean, variance, standard deviation, etc. On the other hand, percentiles and derived measures are less prone to outliers by reducing their weight. The 50th percentile (i.e. the statistical median) — as opposed to the mean — is a convenient measure of central tendency when it comes to dealing with skewed or multipeaked distributions. The median can be as easily computed as the mean and it would be pointless in this case to adhere to a measure which has proven to deliver a flawed result. In this study, for instance, mean–median difference of the intraspecific category amounted to five (CO1) and two nucleotides (16S), both values infringing the assumption that mean–median differences are not significant (15.5% and 5.7% of the total range).

The mean has been used frequently in DNA taxonomy (e.g. Dawson & Jacobs 2001; Hebert *et al.* 2003a,b, 2004b; Hogg & Hebert 2004; Armstrong & Ball 2005; Barrett & Hebert 2005; Kress *et al.* 2005; Lorenz *et al.* 2005; Monaghan *et al.* 2005; Page *et al.* 2005; Vences *et al.* 2005b; Ward *et al.* 2005; Hajibabaei *et al.* 2006). This use may in some contexts have been unjustified, because well-suited datasets include a split between conspecific and congeneric distances that precludes a normal distribution. The mode (e.g. Vences *et al.* 2005a) is a good measure for data at a nominal level, but with ordinal data it is surpassed by the median, which will be unique for any given distribution, as opposed to the possibility of several modes.

Can indels bias the alignment in a beneficial way?

As usual in nonpeptide-coding DNA, the 16S rRNA gene presented a high number of indels. These pose alignment difficulties and invoke the possibility of missing positional homology between parts of the alignment. In phylogeny, problematic regions are often removed from the alignment in order to avoid biasing the resulting trees (Wägele 2005; also adopted in taxonomy by Floyd *et al.* 2002). However, this is sometimes seen as a procedure equally prone to bias (Lutzoni *et al.* 2000). As has been pointed out previously (e.g. Blaxter 2004; Schander & Willassen 2005; Steinke *et al.* 2005; but see Wiens & Penkrot 2002; Will & Rubinoff 2004; DeSalle *et al.* 2005; Prendini 2005), a deep phylogenetic signal is not of central importance for molecular (alpha) taxonomy, which focuses on terminal branches. Thus, little gain should be expected from truncating the alignment. Recent work revealed that the largest proportion of genetic variation between closely related individuals has to be attributed to indels,

which 'dominate the process of early divergence' (Britten *et al*. 2003). Hence, including indels should deliver important information about taxon separation, while excluding them is here considered an unjustified loss of data for taxonomy.

In fact, 16S indels in the pholcid dataset usually improved discrimination of species by creating local 'blocks' of similar sequences within ambiguous alignment regions (cf. electronic supplement). Such a phenomenon is not created by a shortcoming in the alignment software but is due rather to evolutionary processes that may not allow linking the data in indel regions. Within these alignment blocks, positional homology is highly probable due to sequence similarity (de Pinna 1991). Between different blocks, positional homology was sometimes dubious. Although prohibiting any phylogenetic inference, such an approach (i.e. allowing a local alignment 'artifact' by grouping closely related specimens even closer together) reveals the taxonomic diagnostic signal more readily.

We see no evidence in the dataset indicating that the described artifact could become detrimental. Even *Coryssocnemis simla*, the extreme among conspecific 16S observations (Fig. 1), reached its high intraspecific variation mostly through individual base substitutions (21 positions) and included only two detectable, hypothetical indels (three characters). On a statistical basis, an alignment containing indel data was superior to one without. We created two alignments of equal length, about half the size of the original alignment. For the first, we removed all regions of dubious homology ('Without Ambiguities') (Fig. 2). The second represented a truncated version of the original alignment ('Truncated'; Fig. 2), stemming from an indel-rich region. Box plots (Fig. 3) and quantitative analysis (Table 5) both indicate a much better taxon separation for the indel-rich dataset.

When using a wide array of very distantly related taxa for an indel-featuring gene, alignment might become virtually impossible, rendering an approach like the one outlined here useless. However, for the cases in which this would occur, i.e. when distinguishing between higher taxa, molecular approaches to taxonomy are most often unnecessary since a superficial morphological screening usually suffices for identification (cf. Will & Rubinoff 2004).

In order to avoid biasing the results and to still conserve all sequence data, one could conceive adopting the use of RNA secondary structure for 16S, enhanced for example by weighting (as exemplified by Dixon & Hillis 1993). However, some of the rRNA secondary structure might be a 'by-product of underlying mutational processes', emerging through selectively neutral slippage events (Hancock & Vogler 2000). Besides, the tedious task of including RNA secondary structure within the alignment would make such an approach inappropriate considering the simplicity and speed demanded for molecular species identification. Along with the need for replicable procedures, this was also the reason why the hypothesis delivered by the MUSCLE alignment algorithm was hardly changed manually and alignment parameters were not modified.

If an alignment is not constructed according to the hypotheses promising the highest possible degree of positional homology, but instead deliberately takes artifacts into account, model use [and hence the use of maximum likelihood (ML) distances] has to be rejected. Assumptions of sequence evolution become invalid when some of the characters linked by them are possibly not homologous. This need not be a disadvantage, since it might not be necessary to regularly implement ML distances in taxonomy. By applying the above-described tools to different kinds of distance data for the CO1 dataset, we ascertained that ML distances did not perform any better (K2P model) or not considerably better (GTR+I+γ, empirically chosen through Modeltest ver. 3.6; Posada & Crandall 1998) than simple *p*-distances for the data (not shown). The Kimura Two-Parameter model (K2P) is often chosen arbitrarily without testing the suitability for the data (Dalebout *et al*. 1998; Moon-van der Staay *et al*. 2001; Hebert *et al*. 2003a, 2004a,b; Armstrong & Ball 2005; Barrett & Hebert 2005; Hille *et al*. 2005; Vences *et al*. 2005b; Ward *et al*. 2005; Hajibabaei *et al*. 2006; Smith *et al*. 2006). This is especially surprising considering the mitochondrial origin of many species identification datasets, since unequal base frequencies are not accounted for by K2P.

CO1 vs. 16S signal in pholcid DNA taxonomy

When time and money are constraining factors, limits are set to the application of different markers. This makes it important to determine which marker would be best suited to a taxonomic project. Differences in performance become relevant even if the markers stem from the same linkage unit (as with 16S and CO1 in this case).

As advocated recently by Hebert *et al*. (2003a,b), many studies use CO1 as single genetic marker for molecular taxonomic purposes (e.g. Baldwin *et al*. 1996; Bucklin *et al*. 2001; Agustí *et al*. 2003; Hebert *et al*. 2004a,b; Hogg & Hebert 2004; Barrett & Hebert 2005; Hille *et al*. 2005; Janzen *et al*. 2005; Lorenz *et al*. 2005; Saunders 2005; Ward *et al*. 2005; Hajibabaei *et al*. 2006; Smith *et al*. 2006). Our results indicate that in certain taxa, 16S is able to achieve better taxon separation. The 16S box plots (Fig. 1) show considerably fewer (five) outliers than the CO1 box plots (27), and a generally wider separation of categories. In 16S, the potential is enhanced to single out contaminants with less noise and observations from both subpopulations never intersected (as opposed to CO1).

Quantitatively, this is shown in Table 5: 1.2% of CO1 observations fell within the *Overlap*, whereas 16S categories were isolated from each other by a *Gap Range* of +2.8. The

5-95 Range for 16S was almost double that for CO1. Such a result is partly a sampling artifact, since the overlap for CO1 disappeared when looking at the respective 'Shared Cases' datasets, i.e. the datasets containing only individuals present in the corresponding file of the other marker. However, in the 'Shared Cases' datasets, the *Gap Range* for 16S was still more than double that of the CO1 *Gap Range*, while the 16S *5-95 Range* value was again almost twice that for CO1. Generally, the signal could not be improved by using only 1st and 2nd codon positions for CO1, nor by the translated amino acid sequence. Using 3rd positions exclusively, however, provided a slightly better *Overlap* value, a much better *5-95 Range*, but also a much longer string of intersecting outliers in the *Gap Range* (see Table 5). Although 16S was found to serve better in answering molecular taxonomic questions in pholcid spiders, dismissing the results of CO1 would clearly mean a loss of signal (see combination of both genes in Table 5).

The superiority of 16S over CO1 as a marker in DNA taxonomy has already been suggested (Vences *et al*. 2005a in frogs; Collins *et al*. 2005 in cnidarians; Steinke *et al*. 2005 in snails; on its promises as taxonomic marker in Crustacea, see Schubart *et al*. 2000).

Acknowledgements

We are grateful to Christoph Schubart and two anonymous reviewers for their helpful comments on the manuscript, to Natasha Astrina for enlightening discussions and to Claudia Etzbauer for her technical help. We are indebted to all those who contributed specimens: W.G. Eberhard, H. El-Hennawy, M.S. Harvey, H.F. Japyassú, R. Jocqué, K. van Keer, C. Kristenson, N. López Mercader, J. Miller, P. Paquin, A. Peretti, A. Pérez González, R. Pinto-da-Rocha, W. Reeves, C. Rheims, A.J. Santos, P. Schwendinger, J. Sewlal, M.C.K. Starr, A. van Harten. A considerable part of the material was collected in Venezuela and in Brazil (by BAH), and we are most thankful to Osvaldo Villarreal and Abel Pérez González for invaluable help with getting a collection permit in Venezuela (permit N° 01-11-0966, issued by the Dirección General de Fauna y Oficina Nacional de Diversidad Biológica in Caracas), and to Antonio Brescovit and Cristina Rheims for arranging permits and trips in Brazil. This study was financially supported by the Alexander Koenig Stiftung (to JJA).

References

Agustí, N., Shayler, S. P., Harwood, J. D., Vaughan, I. P., Sunderland, K. D. & Symondson, W. O. (2003). Collembola as alternative prey sustaining spiders in arable ecosystems: prey detection within predators using molecular markers. *Molecular Ecology*, *12*, 3467–3475.

Armstrong, K. F. & Ball, S. L. (2005). DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1813–1823.

Ayoub, N. A., Riechert, S. E. & Small, R. L. (2005). Speciation history of the North American funnel web spiders, *Agelenopsis* (Araneae: Agelenidae): phylogenetic inferences at the population–species interface. *Molecular Phylogenetics and Evolution*, *36*, 42–57.

Baldwin, B. S., Black, M., Sanjur, O., Gustafson, R., Lutz, R. A. & Vrijenhoek, R. C. (1996). A diagnostic molecular marker for zebra mussels (*Dreissena polymorpha*) and potentially co-occurring bivalves: mitochondrial COI. *Molecular Marine Biology and Biotechnology*, *5*, 9–14.

Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*. Chichester: John Wiley & Sons.

Barrett, R. D. H. & Hebert, P. D. N. (2005). Identifying spiders through DNA barcodes. *Canadian Journal of Zoology — Revue Canadienne de Zoologie*, *83*, 481–491.

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *359*, 669–679.

Blaxter, M., Elsworth, B. & Daub, J. (2004). DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. *Proceedings of the Royal Society of London, Series B — Biological Sciences*, *271* (Suppl. 4), S189–S192.

Bond, J. E. (2004). Systematics of the Californian euctenizine spider genus *Apomastus* (Araneae: Mygalomorphae: Cyrtaucheniidae): the relationship between molecular and morphological taxonomy. *Invertebrate Systematics*, *18*, 361–376.

Bond, J. E. & Sierwald, P. (2003). Molecular taxonomy of the *Anadenobolus excisus* (Diplopoda: Spirobolida: Rhinocricidae) species-group on the Caribbean island of Jamaica. *Invertebrate Systematics*, *17*, 515–528.

Britten, R. J., Rowen, L., Williams, J. & Cameron, R. A. (2003). Majority of divergence between closely related DNA samples is due to indels. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 4661–4665.

Bruvo-Madaric, B., Huber, B. A., Steinacher, A. & Pass, G. (2005). Phylogeny of pholcid spiders (Araneae: Pholcidae): combined analysis using morphology and molecules. *Molecular Phylogenetics and Evolution*, *37*, 661–673.

Bucklin, A., Guarnieri, M., McGillicuddy, D. J. & Hill, R. S. (2001). Spring evolution of *Pseudocalanus spp.* abundance on Georges Bank based on molecular discrimination of *P. moultoni* and *P. newmani*. *Deep-Sea Research Part II — Topical Studies in Oceanography*, *48*, 589–608.

Cognato, A. I. & Vogler, A. P. (2001). Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Coleoptera: Scolytinae). *Systematic Biology*, *50*, 758–780.

Collins, A. G., Winkelmann, S., Hadrys, H. & Schierwater, B. (2005). Phylogeny of Capitata and Corynidae (Cnidaria, Hydrozoa) in light of mitochondrial 16S rDNA data. *Zoologica Scripta*, *34*, 91–99.

Crandall, K. A. & Fitzpatrick, J. E. Jr (1996). Crayfish molecular systematics: using a combination of procedures to estimate phylogeny. *Systematic Biology*, *45*, 1–26.

Dalebout, M. L., Van Helden, A., Van Waerebeek, K. & Baker, C. S. (1998). Molecular genetic identification of southern hemisphere beaked whales (Cetacea: Ziphiidae). *Molecular Ecology*, 7, 687–694.

Dalebout, M. L., Mead, J. G., Baker, C. S., Baker, A. N. & van Helden, A. L. (2002). A new species of beaked whale *Mesoplodon perrini* sp. n. (Cetacea: Ziphiidae) discovered through phylogenetic analyses of mitochondrial DNA sequences. *Marine Mammal Science*, *18*, 577–608.

Dawson, M. N. & Jacobs, D. K. (2001). Molecular evidence for cryptic species of *Aurelia aurita* (Cnidaria, Scyphozoa). *Biological Bulletin*, *200*, 92–96.

DeSalle, R., Egan, M. G. & Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1905–1916.

Dixon, M. T. & Hillis, D. M. (1993). Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Molecular Biology and Evolution*, *10*, 256–267.

Dunn, C. P. (2003). Keeping taxonomy based in morphology. *Trends in Ecology and Evolution*, *18*, 270–271.

Eberhard, W. G. (1985). *Sexual Selection and Animal Genitalia*. Cambridge, MA: Harvard University Press.

Eberhard, W. G., Huber, B. A., Rodriguez, R. L., Briceno, R. D., Salas, I. & Rodriguez, V. (1998). One size fits all? Relationships between the size and degree of variation in genitalia and other body parts in twenty species of insects and spiders. *Evolution*, *52*, 415–431.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113.

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

Floyd, R., Abebe, E., Papert, A. & Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology*, *11*, 839–850.

Funk, D. J. (1999). Molecular systematics of cytochrome oxidase I and 16S from *Neochlamisus* leaf beetles and the importance of sampling. *Molecular Biology and Evolution*, *16*, 67–82.

Funk, D. J. & Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology Evolution and Systematics*, *34*, 397–423.

Giribet, G. & Wheeler, W. C. (1999). On gaps. *Molecular Phylogenetics and Evolution*, *13*, 132–143.

Godfray, H. C. (2002). Challenges for taxonomy. *Nature*, *417*, 17–19.

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

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hancock, J. M. & Vogler, A. P. (2000). How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Molecular Phylogenetics and Evolution*, *14*, 366–374.

Hebert, P. D., Cywinska, A., Ball, S. L. & deWaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London*, *Series B — Biological Sciences*, *270*, 313–321.

Hebert, P. D., Ratnasingham, S. & deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London*, *Series B — Biological Sciences*, *270* (Suppl. 1), S96–S99.

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H. & 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, P. D., Stoeckle, M. Y., Zemlak, T. S. & Francis, C. M. (2004b). Identification of birds through DNA barcodes. *PLoS Biology*, *2*, e312.

Hedin, M. C. & Maddison, W. P. (2001). A combined molecular approach to phylogeny of the jumping spider subfamily *Dendryphantinae* (Araneae: Salticidae). *Molecular Phylogenetics and Evolution*, *18*, 386–403.

Hille, A., Miller, M. A. & Erlacher, S. (2005). DNA sequence variation at the mitochondrial cytochrome oxidase I subunit among pheromotypes of the sibling taxa *Diachrysia chrysitis* and *D. tutti* (Lepidoptera: Noctuidae). *Zoologica Scripta*, *34*, 49–56.

Hogg, I. D. & Hebert, P. D. N. (2004). Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology — Revue Canadienne de Zoologie*, *82*, 749–754.

Huber, B. A. (2000). New World pholcid spiders (Araneae: Pholcidae): a revision at generic level. *Bulletin of the American Museum of Natural History*, *254*, 1–347.

Huber, B. A. (2004). The significance of copulatory structures in spider systematics. In J. Schult (ed.) *Biosemiotik. Praktische Anwendung und Konsequenzen für die Einzelwissenschaften* (pp. 89–100). Berlin: VWB-Verlag.

Janzen, D. H., Hajibabaei, M., Burns, J. M., Hallwachs, W., Remigio, E. & Hebert, P. D. (2005). Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1835–1845.

Johnson, J., Dowling, T. & Belk, M. (2004). Neglected taxonomy of rare desert fishes: congruent evidence for two species of leatherside chub. *Systematic Biology*, *53*, 841–855.

Knapp, S., Lamas, G., Lughadha, E. N. & Novarino, G. (2004). Stability or stasis in the names of organisms: the evolving codes of nomenclature. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *359*, 611–622.

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 8369–8374.

Langor, D. W. & Sperling, F. A. H. (1995). Mitochondrial DNA variation and identification of bark weevils in the *Pissodes strobi* species group in Western Canada (Coleoptera: Curculionidae). *The Canadian Entomologist*, *127*, 895–911.

Lipscomb, D., Platnick, N. & Wheeler, Q. (2003). The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends in Ecology and Evolution*, *18*, 65–66.

López-Legentil, S. & Turon, X. (2005). How do morphotypes and chemotypes relate to genotypes? The colonial ascidian *Cystodytes* (Polycitoridae). *Zoologica Scripta*, *34*, 3–14.

Lorenz, J. G., Jackson, W. E., Beck, J. C. & Hanner, R. (2005). The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1869–1877.

Lughadha, E. N. (2004). Towards a working list of all known plant species. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *359*, 681–687.

Lutzoni, F., Wagner, P., Reeb, V. & Zoller, S. (2000). Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Systematic Biology*, *49*, 628–651.

Markmann, M. & Tautz, D. (2005). Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1917–1924.

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

Minelli, A. (2003). The status of taxonomic literature. *Trends in Ecology and Evolution*, *18*, 75–76.

Monaghan, M. T., Balke, M., Gregory, T. R. & Vogler, A. P. (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.

Monaghan, M. T., Balke, M., Pons, J. & Vogler, A. P. (2006). Beyond barcodes: complex DNA taxonomy of a South Pacific Island radiation. *Proceedings of the Royal Society Series B — Biological Sciences*, *273*, 887–893.

Moon-van der Staay, S. Y., De Wachter, R. & Vaulot, D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, *409*, 607–610.

Moritz, C. & Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biology*, *2*, e354.

Nielsen, R. & Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology*, *55*, 162–169.

Page, T. J., Choy, S. C. & Hughes, J. M. (2005). The taxonomic feedback loop: symbiosis of morphology and molecules. *Biology Letters*, *1*, 139–142.

Palumbi, A. & Cipriano, F. (1998). Species identification using genetic tools: the value of nuclear and mitochondrial gene sequences in whale conservation. *Journal of Heredity*, *89*, 459–464.

Palumbi, S. R. (1996). Nucleic acids II: the polymerase chain reaction. In D. M. Hillis, C. Moritz & B. K. Mable (eds) *Molecular Systematics* (pp. 205–246). Sunderland, MA: Sinauer.

Palumbi, S. R., Martin, A., Romano, S., McMillan, S. O., Stice, L. & Grabowski, G. (2002). *The Simple Fool's Guide to PCR*, Version 2.0. Honolulu: privately published.

Paquin, P. & Hedin, M. (2004). The power and perils of 'molecular taxonomy': a case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. *Molecular Ecology*, *13*, 3239–3255.

de Pinna, M. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics — the International Journal of the Willi Hennig Society*, 7, 367–394.

Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics*, *14*, 817–818.

Prendini, L. (2005). Comment on 'Identifying spiders through DNA barcodes'. *Canadian Journal of Zoology — Revue Canadienne de Zoologie*, *83*, 498–504.

Puorto, G., Salomao, M. D., Theakston, R. D. G., Thorpe, R. S., Warrell, D. A. & Wuster, W. (2001). Combining mitochondrial DNA sequences and morphological data to infer species boundaries: phylogeography of lanceheaded pitvipers in the Brazilian Atlantic forest, and the status of *Bothrops pradoi* (Squamata: Serpentes: Viperidae). *Journal of Evolutionary Biology*, *14*, 527–538.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*, 406–425.

Saunders, G. W. (2005). Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1879–1888.

Schander, C. & Willassen, E. (2005). What can biological barcoding do for marine biology? *Marine Biology Research*, *1*, 79–83.

Schubart, C. D., Neigel, J. E. & Felder, D. L. (2000). Use of the mitochondrial 16S rRNA gene for phylogenetic and population studies in Crustacea. *Crustacean Issues*, *12*, 817–830.

Scoble, M. J. (2004). Unitary or unified taxonomy? *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *359*, 699–710.

Scotland, R., Hughes, C., Bailey, D. & Wortley, A. (2003). The Big Machine and the much-maligned taxonomist. *Systematics and Biodiversity*, *1*, 139–143.

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 in Ecology and Evolution*, *18*, 63–65.

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. *Annals of the Entomological Society of America*, *87*, 651–701.

Smith, M. A., Woodley, N. E., Janzen, D. H., Hallwachs, W. & Hebert, P. D. (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.

Sperling, F. A. (2003). DNA barcoding: Deus ex machina. *Newsletter of the Biology Survey of Canada (Terrestrial Arthropods)*, *22*, 50–53.

Steinke, D., Vences, M., Salzburger, W. & Meyer, A. (2005). TaxI: a software tool for DNA barcoding using distance methods. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1975–1980.

Swofford, D. L. (1998). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)* [Computer software and manual]. Sunderland, MA: Sinauer.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. (2002). DNA points the way ahead in taxonomy. *Nature*, *418*, 479.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology and Evolution*, *18*, 70–74.

Therriault, T. W., Docker, M. F., Orlova, M. I., Heath, D. D. & MacIsaac, H. J. (2004). Molecular resolution of the family Dreissenidae (Mollusca: Bivalvia) with emphasis on Ponto-Caspian species, including first report of *Mytilopsis leucophaeata* in the Black Sea basin. *Molecular Phylogenetics and Evolution*, *30*, 479–489.

Thiele, K. & Yeates, D. (2002). Tension arises from duality at the heart of taxonomy. *Nature*, *419*, 337.

Townson, H., Harbach, R. E. & Callan, T. A. (1999). DNA identification of museum specimens of the *Anopheles gambiae* complex: an evaluation of PCR as a tool for resolving the formal taxonomy of sibling species complexes. *Systematic Entomology*, *24*, 95–100.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Vences, M., Thomas, M., van der Meijden, A., Chiari, Y. & Vieites, D. R. (2005a). Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, *2*, 5.

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

Wägele, J.-W. (2005). *Foundations of Phylogenetic Systematics*. München: Pfeil.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *360*, 1847–1857.

Westheide, W. & Hass-Cordes, E. (2001). Molecular taxonomy: description of a cryptic *Petitia* species (Polychaeta: Syllidae) from the island of Mahe (Seychelles, Indian Ocean) using RAPD markers and ITS2 sequences. *Journal of Zoological Systematics and Evolutionary Research*, *39*, 103–111.

Wiens, J. J. & Penkrot, T. A. (2002). Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology*, *51*, 69–91.

Will, K. W. & 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.

Wilson, E. O. (2003). The encyclopedia of life. *Trends in Ecology and Evolution*, *18*, 77–80.

Wilson, E. O. (2004). Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London Series B — Biological Sciences*, *359*, 739–739.

Winker, K. (1999). How to bring collections data into the net. *Nature*, *401*, 524–524.

Supplementary material

The following material is available for this article online:

Supplementary material DNA sequence alignment files for the analyzed genes (fasta format). Organism names are followed by the corresponding voucher number (cf. Table 1) or in case of GenBank sequences, by the accession number. The .fas files can be read and edited using the program BioEdit (available to download free from http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

This material is available as part of the online article from http://www.blackwell-synergy.com