DNA barcoding: a practical tool for fundamental and applied biodiversity research

Edited by

Zoltán T. Nagy, Thierry Backeljau, Marc De Meyer, Kurt Jordaens

PENSOFT

Sofia–Moscow

2013

ZooKeys 365 (Special Issue)

DNA barcoding: a practical tool for fundamental and applied biodiversity research

Edited by Zoltán T. Nagy, Thierry Backeljau, Marc De Meyer, Kurt Jordaens

First published 2013
ISBN 978-954-642-710-6 (paperback)

Pensoft Publishers
12 Prof. Georgi Zlatarski Street, 1700 Sofia, Bulgaria
Fax: +359-2-870-42-82
info@pensoft.net
www.pensoft.net

Printed in Bulgaria, December 2013

Contents

ZooKeys 365: 1–3 (2013)
doi: 10.3897/zookeys.365.6681
www.zookeys.org

EDITORIAL

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Editorial

Zoltán T Nagy[1], Thierry Backeljau[1,2], Marc De Meyer[3] , Kurt Jordaens[2,3]

1 *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (JEMU), Vautierstraat 29 - B-1000 Brussels, Belgium* **2** *Evolutionary Ecology Group, University of Antwerp, Groenenborgerlaan 171 - B-2020 Antwerp, Belgium* **3** *Royal Museum for Central Africa (JEMU), Leuvensesteenweg 13 - B-3080 Tervuren, Belgium*

Corresponding author: *Zoltán T Nagy* (zoltan-tamas.nagy@naturalsciences.be)

Received 25 November 2013 | Accepted 25 November 2013 | Published 30 December 2013

Citation: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (2013) Editorial. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 1–3. doi: 10.3897/zookeys.365.6681

Since its formal introduction in 2003, DNA barcoding has become a well-accepted and popular tool for the identification of species and the detection of cryptic taxonomic diversity. Hence, it is not surprising that the past decade has witnessed a boom of DNA barcoding studies, up to the point that currently the method is becoming an integral part of taxonomic practice. This does not mean that DNA barcoding is some sort of magic technology, capable of solving all taxonomic problems. Such a view would indeed be simplistic and, in fact, was never claimed by the DNA barcoding community. Instead, DNA barcoding is a practical tool that facilitates species (or more generally, taxon) identification, without solving or considering the central taxonomic question as to what a species really is. Yet, being primarily an identification tool, DNA barcoding has a tremendous potential for a wide variety of possible applications. This point is globally well-recognized and hence, after the foundation of the overarching, worldwide International Barcode of Life Project (iBOL) and the Consortium for the Barcode of Life (CBOL), which initiated several taxon, regional or problem-oriented DNA barcoding initiatives, several countries, institutions and organizations have joined these international bodies and launched their own national or regional projects.

Also Belgium created its DNA barcoding consortium, the Belgian Network for DNA Barcoding, which embodies the Belgian Barcoding of Life (BeBoL) initiative. This network was established in January 2011 with the financial support of the Fund for Scientific Research – Flanders (FWO). It is financially administrated by the University of Antwerp, but its activities are coordinated by the Joint Experimental Molecular Unit (JEMU), a molecular systematics research facility shared by the Royal Museum for Central Africa (RMCA) and the Royal Belgian Institute of Natural Sciences (RBINS), and financed by the Belgian Science Policy Office (BELSPO). It currently involves 23 Belgian members, including not only federal research institutions such as RMCA, RBINS and the National Institute of Criminalistics and Criminology, but also universities, botanical and zoological gardens, and regional institutes dedicated to medical, agricultural, and conservation research. It aims at stimulating collaborative research by providing a discussion, training and exchange forum with respect to DNA barcoding. Therefore, BeBoL is dedicated to, amongst others, organizing meetings, workshops, symposia, and congresses.

So far, one of the most visible achievements of BeBoL was the organization of the "Third European Conference for the Barcode of Life, Brussels, 17–20 September 2012" (ECBOL3), under the thematic flag "Barcoding of organisms of policy concern". This theme was chosen in view of the increasing interest of governments, decision makers, public authorities, law enforcement entities and private companies in DNA barcoding as a practical and reliable identification tool. As such the conference also provided an overview of DNA barcoding as an example of "applied taxonomy". This formula appeared to be attractive since ECBOL3, which took place in the Royal Flemish Academy of Belgium for Science and the Arts (KVAB), was attended by about 120 researchers from Europe and beyond.

Although it was originally not planned to publish congress proceedings of ECBOL3, many participants felt that it nevertheless would be a great opportunity to produce a collection of DNA barcoding papers that emanated either from the congress or from BeBoL partners. Hence, it was decided to do so and to use this occasion to implement the unique possibilities offered by the open-access journal ZooKeys, a trend-setting taxonomic publication forum that extends papers with a whole series of extra features such as XML marking up and linking/transferring taxonomic data to ZooBank, GBIF, EOL, PLAZI and WikiSpecies. As such, ZooKeys illustrates the future of publishing freely accessible (big) biodiversity data in a global community, by what is often referred to as data hosting and the development of data publishing workflows. Therefore, ZooKeys is one of the core elements in the EU funded 7th Framework Program ViBRANT (Virtual Biodiversity Research and Access Network for Taxonomy). This network has also been instrumental for JEMU, BeBoL and ECBOL3, since these initiatives have organized their communities by means of scratchpads, one of the core products of ViBRANT.

This special ZooKeys issue on DNA barcoding is hence the fruit of all the aforementioned efforts. It deals with a wide array of animal and plant taxa, and aims at demonstrating various aspects of DNA barcoding, including fundamental biodiversity

research, applications, methodological issues, software, and limitations. Therefore, we hope that this issue may provide a modest, but lasting contribution to the already vast literature on DNA barcoding.

Brussels, December 11th, 2013

On behalf of BeBoL

This issue was realized with the support of:

ZooKeys 365: 5–24 (2013)
doi: 10.3897/zookeys.365.5873
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

The use of DNA barcoding to monitor the marine mammal biodiversity along the French Atlantic coast

Eric Alfonsi[1,2,*], Eleonore Méheust[1,2,*], Sandra Fuchs[2], François-Gilles Carpentier[1], Yann Quillivic[2], Amélia Viricel[3,4], Sami Hassani[2], Jean-Luc Jung[1]

1 *Laboratoire BioGeMME (Biologie et Génétique des Mammifères Marins dans leur Environnement), Université Européenne de Bretagne & Université de Bretagne Occidentale, UFR Sciences et Techniques, 6 Av. Victor Le Gorgeu - CS93837 - 29238 Brest Cedex 3, France* **2** *Laboratoire d'Etude des Mammifères Marins (LEMM), Océanopolis, port de plaisance, BP 91039, 29210 Brest Cedex 1, France* **3** *Observatoire PELAGIS, UMS 3462, CNRS-Université de La Rochelle, Pôle analytique, 5 allée de l'océan, 17000 La Rochelle, France* **4** *Littoral, Environnement et Sociétés, UMR 7266, CNRS-Université de La Rochelle, 2 rue Olympe de Gouges, 17000 La Rochelle, France*

Corresponding author: *Jean-Luc Jung* (jung@univ-brest.fr)

Academic editor: *T. Backeljau* | Received 28 June 2013 | Accepted 11 September 2013 | Published 30 December 2013

Citation: Alfonsi E, Méheust E, Fuchs S, Carpentier F-G, Quillivic Y, Viricel A, Hassani S, Jung J-L (2013) The use of DNA barcoding to monitor the marine mammal biodiversity along the French Atlantic coast. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 5–24. doi: 10.3897/zookeys.365.5873

Abstract

In the last ten years, 14 species of cetaceans and five species of pinnipeds stranded along the Atlantic coast of Brittany in the North West of France. All species included, an average of 150 animals strand each year in this area. Based on reports from the stranding network operating along this coast, the most common stranding events comprise six cetacean species (*Delphinus delphis*, *Tursiops truncatus*, *Stenella coeruleoalba*, *Globicephala melas*, *Grampus griseus*, *Phocoena phocoena*) and one pinniped species (*Halichoerus grypus*). Rare stranding events include deep-diving or exotic species, such as arctic seals. In this study, our aim was to determine the potential contribution of DNA barcoding to the monitoring of marine mammal biodiversity as performed by the stranding network.

We sequenced more than 500 bp of the 5' end of the mitochondrial COI gene of 89 animals of 15 different species (12 cetaceans, and three pinnipeds). Except for members of the Delphininae, all species were unambiguously discriminated on the basis of their COI sequences. We then applied DNA barcoding to identify some "undetermined" samples. With again the exception of the Delphininae, this was successful using the BOLD identification engine. For samples of the Delphininae, we sequenced a portion of the

* These authors contributed equally to this work.

mitochondrial control region (MCR), and using a non-metric multidimentional scaling plot and posterior probability calculations we were able to determine putatively each species. We then showed, in the case of the harbour porpoise, that COI polymorphisms, although being lower than MCR ones, could also be used to assess intraspecific variability. All these results show that the use of DNA barcoding in conjunction with a stranding network could clearly increase the accuracy of the monitoring of marine mammal biodiversity.

Keywords
DNA barcoding, COI, control region, marine mammals, cetaceans, pinnipeds, biodiversity monitoring, stranding network

Introduction

The aim of DNA barcoding is to concentrate the efforts of molecular taxonomists on a single part of the mitochondrial genome, chosen because it presents portions conserved across taxa that are appropriate for primer design, while including polymorphism among and within species (Hebert et al. 2003, 2004). This DNA sequence, targeted as the 5' end of the gene coding for the subunit I of the cytochrome *c* oxidase subunit I (COI), is sufficiently diverse so as to allow the specific identification of a great majority of animal species. Numerous studies have proven the success of this approach in the animal kingdom, and using various sources of tissue samples (e.g. Lambert 2005, Clare et al. 2007, Dawnay et al. 2007, Hajibabaei et al. 2007, Borisenko et al. 2008, Ward et al. 2009, Shokralla et al. 2010). Today (June 2013), a database, accessible at http://www.boldsystems.org, groups DNA barcode sequence data for more than 133,000 animal species, and offers a powerful identification tool for new specimens (Ratnasingham and Hebert 2007).

DNA barcoding also possesses some inherent limitations (Valentini et al. 2009): it is based on a single locus on the mitochondrial genome so that it is only maternally inherited (Hartl and Clark 2007), it can show heteroplasmy (Kmiec et al. 2006, Vollmer et al. 2011) or may exist as nuclear copies. Some of these limitations have been well-exposed (Ballard and Whitlock 2004, Toews and Brelsford 2012). The use of DNA barcoding for species delimitation also requires that interspecific divergence is higher than the intraspecific divergence. Although this has been shown to be true in numerous taxonomic groups, opposite examples also exist (Amaral et al. 2007, Wiemers and Fiedler 2007, Viricel and Rosel 2012).

In the present study, we assess the contributions that DNA barcoding could provide to the monitoring of the marine mammal biodiversity along the coasts of Brittany, in the northwest of France. For almost 20 years, the stranding network has been collecting data and, when possible, sampling, each time a marine mammal stranding is reported. Field correspondents are organized in a geographical area covering the entire Brittany coasts. The network is coordinated regionally by Océanopolis (Brest, France), and nationally by *Pelagis* (La Rochelle, France).

DNA barcoding could be useful for the monitoring of marine mammal strandings at different levels. First, by confirming the quality and the reproducibility of a spe-

cies identification made by the field correspondents. Beside common species, which are often encountered and easily identified, exotic or deep living species represent rare stranding events. In such cases, DNA barcoding could provide a confirmation or an additional degree of precision of taxonomic determination (Thompson et al. 2012). Second, DNA barcoding can help specifying species identifications in those cases where the taxonomic identification was made only to the genus or family levels. This is often due to incomplete or highly degraded carcasses. DNA barcoding also is a valuable and cost effective alternative to the taking of the head or skull of the animals. Third, genetic data collected for DNA barcoding generally include intraspecific variation, which allows downstream population-level analyses including the detection of genetic structure and, in some cases, monitoring population movements. A long-term use of the barcoding approach would therefore clearly increase the significance and the precision of marine mammal stranding monitoring. Migration or movement of populations or groups of a particular species can be highlighted, thus revealing e.g. environmental changes leading to these movements (Pauls et al. 2012).

We evaluated the usefulness of DNA barcoding in the monitoring of marine mammal biodiversity along the coasts of Brittany at three levels: by confirming the taxonomic identification performed by field correspondents, by identifying degraded carcasses or parts of carcasses, and by determining intraspecific variations for two species commonly found off Brittany, the harbour porpoise and the grey seal. For this last part of our study, we also compared COI and the mitochondrial control region in terms of their effectiveness in species identification.

Methods

Collection of data and samples

The CRMM (Centre de Recherche sur les Mammifères Marins, La Rochelle, France), presently the Joint Service Unit PELAGIS, UMS 3462, University of La Rochelle-CNRS has created the French marine mammal stranding recording program at the beginning of the 70s. The network comprises about 260 field correspondents, members of organizations or volunteers (Peltier et al. 2013).

Since 1995, the LEMM (Laboratoire d'Etude des Mammifères Marins, Océanopolis, Brest, France) has coordinated this network at a regional scale in Brittany, North West of France. Data are collected from the Brittany coastlines, analyzed, and then added to the central database maintained in La Rochelle. The Brittany coasts have been divided into 18 sections covering the whole coastline (Jung et al. 2009). In each of these areas, correspondents are trained in the analysis of stranded marine mammals. Taxonomic identification and characteristic measurements are performed following a standard procedure. The LEMM therefore compiles standardized data on a large proportion of cetaceans stranded on the Brittany coasts on a yearly basis. Whenever possible, skin, blubber, muscle and teeth samples are also collected in the field from each

stranded animal. Samples are then kept in absolute ethanol or dry at -20 °C until analyses. Some harbour porpoise samples, described in the Appendix 1 and in Alfonsi et al. (2012), were stranded or by-caught in the Bay of Biscay (Atlantic coast of France).

Genomic DNA extraction, amplification and sequencing of COI and MCR (mitochondrial control region)

Genomic DNA was extracted from blood samples or from muscle and skin tissues using a standardized protocol and the DNeasy Blood and Tissue kit (Qiagen), following the instructions of the manufacturer. The quality and the concentration of all the DNA extracts were estimated by agarose gel electrophoresis and by spectrophotometry using a Nanodrop 1000 (Thermo Scientific).

A 736 base-pair (bp) fragment of the 5' region of the COI fragment (position 5352 to 6087 of the complete mitochondrial genome of the harbour porpoise, GenBank acc. no. AJ554063), was amplified using two newly designed primers, LCOIea (5'-tcggccattttacctatgttcata-3') and HBCUem (5'-ggtggccgaagaatcagaata-3'). The 50 µl PCR final volume included approximately 50 ng of genomic DNA, and 25 pmole of each primer in the Hotgoldstar master mix × 1 (Eurogentec) with a final concentration of $MgCl_2$ of 2.5 mM. After an initial denaturation step of 10 min at 95 °C, the thermocycle profile consisted of 32 cycles for cetaceans or 35 cycles for pinnipeds at 95 °C for 30 s, 53 °C for 30 s and 72 °C for 60 s, with a final extension at 72 °C for 10 min.

For some animals, we also amplified and sequenced another part of the mitochondrial genome including the control region (MCR). For cetaceans, the primers and reaction conditions are described in (Alfonsi et al. 2012). For pinnipeds, two newly designed primers LMCRHgem 5'-tcatacccattgccagcattat-3' and HMCRHgem 5'-taccaaatgcatgacaccacag-3' amplified a 693 bp fragment from position 16160 to 55 of the *Halichoerus grypus* complete mitochondrial genome sequence (GenBank acc. no. X72004). PCR reaction conditions were the same as described above for pinnipeds, with the hybridization temperature set to 53 °C. PCR products were purified using the "MinElute PCR Purification Kit" and sequenced by a commercial sequence facility (Macrogen, Korea).

Electropherograms were analyzed and edited manually using the Sequence scanner software (Applied Biosystems), and alignments were produced using CLUSTAL W (Thompson et al. 1994) with default settings in Bioedit (Hall 1999). All sequences were analyzed using the Barcode of Life Data Systems (BOLD) interface (accessible at http://www.boldsystems.org), and were also compared to GenBank data using BLAST (Benson et al. 2010).

DNA sequences and specimen information have been added to two BOLD projects. The first project includes specimens for which the species had been identified without doubt using classical morphological identification, and is referred to as IMMB (Identified Marine Mammals in Brittany). The IMMB project is a part of the campaign "barcoding mammals of the world". The second project, UMMB (Unidentified Marine Mammals in Brittany), includes specimens only identified to the genus or to higher taxonomic levels. This second project is a part of the campaign "barcoding application".

Genetic distances (intraspecific, interspecific and minimal distance to the nearest neighbour) were calculated using the Kimura 2-parameter (K2P) model (Kimura 1980) and the MUSCLE alignment algorithm on the BOLD user interface or using the software MEGA5 (Tamura et al. 2011). Neighbour-Joining trees based on the K2P-model were built using the BOLD user interface. DnaSP v5.10 was used to calculate haplotype and nucleotide diversities (Librado and Rozas 2009). We used non-metric multidimensional scaling (nMDS) to represent MCR distances graphically and to discriminate closely related species within the *Stenella-Tursiops-Delphinus* complex (LeDuc et al. 1999, McGowen 2011, Perrin et al. 2013). Distance matrices were computed with the K2P-model using DNAdist (Felsenstein 1989) and were then analyzed by nMDS using Statistica (Statsoft 2005). Posterior probabilities were calculated by a LDA (linear discriminant analysis) on coordinates given by the nMDS. Phylogenetic relationships among COI sequences of harbour porpoise were depicted using a median joining network of haplotypes using Network v4.6 (www.fluxus-engineering.com).

Results

From 2003 to 2012, 1530 marine mammal strandings were recorded along the coastline of Brittany (Table 1). Fourteen species of cetaceans and five species of pinnipeds were identified. The most frequent cetaceans were six indigenous species of the Brittany waters, viz. five members of the Delphinidae (*Delphinus delphis*, *Tursiops truncatus*, *Stenella coeruleoalba*, *Globicephala melas*, *Grampus griseus*), and the harbour porpoise (*Phocoena phocoena*). Two members of the Zyphiidae (*Hyperoodon ampullatus* and *Ziphius cavirostris*), three other species of Delphinidae (*Lagenorhynchus acutus*, *Orcinus orca* and *Stenella frontalis*), one species of Physeteridae (*Physeter macrocephalus*) and two mysticete species (*Balaenoptera acutorostrata* and *Balaenoptera physalus*) were rare stranding events. *Halichoerus grypus* was by far the most commonly encountered pinniped, far before *Phoca vitulina*, and some uncommon arctic seals (*Phoca hispida*, *Cystophora cristata* and *Phoca groenlandica*). Between 9 and 12 different marine mammal species stranded each year (Figure 1).

Members of the stranding network are trained to identify the stranded animals. Nevertheless, 258 animals (16.8% of the strandings) were not characterized to the species level, generally because of an advanced state of decomposition of the animal body, sometimes in conjunction with bad field-work conditions.

COI sequencing and analysis from different marine mammal samples

DNA was extracted from 92 stranded animals, i.e. from dead cetaceans and pinnipeds, but also from 40 grey seals stranded alive, which were treated in the care center of Océanopolis (Brest, France) and from which a small blood sample was taken and kept at -20 °C. All the samples came from animals stranded at the coasts of Brittany, except for one grey seal (Hgc406), which stranded alive in Spain in 2009 and which

Table 1. Strandings of marine mammals along the coasts of Brittany, northwest of France (2003–2012)

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
Cetaceans											
Balaenoptera acutorostrata								1	1		2
Balaenoptera physalus	1	2			2	3			4	2	14
Delphinidae (undetermined)	40	30	36	22	15	9	9	6	16	8	191
Delphinus delphis	56	61	109	53	51	56	40	39	72	57	594
Globicephala melas	6	5	7	1	1	2	1	2	1	2	28
Grampus griseus	2	1	7	3	1	7	2	1	2	4	30
Hyperoodon ampullatus									1		1
Lagenorhynchus acutus				1	2				1	1	5
Orcinus orca		1									1
Phocoena phocoena	18	13	12	15	20	23	9	10	15	11	146
Physeter macrocephalus		2			1						3
Stenella coeruleoalba	1		7	9	8	4	5	9	6	3	52
Stenella frontalis			1								1
Tursiops truncatus	6	2	7	6	4	5	3	8	3	3	47
Ziphius cavirostris					1		1				2
Mysticeti (undetermined)			1	4							5
Odontoceti (undetermined)	5	1	1	3	1						11
Cetacea (undetermined)				3	3	2			1		9
Pinnipeds											
Cystophora cristata			1	3							4
Halichoerus grypus	20	29	41	37	51	41	37	13	34	24	327
Phoca groenlandica		1									1
Phoca vitulina	1	1	2	1	1	1		2	3		12
Pusa hispida			1				1				2
Phocidae (undetermined)	5		7	4	13	4	5		2	1	41
Unknown										1	1
Total	**161**	**149**	**240**	**165**	**175**	**157**	**113**	**91**	**162**	**117**	**1530**

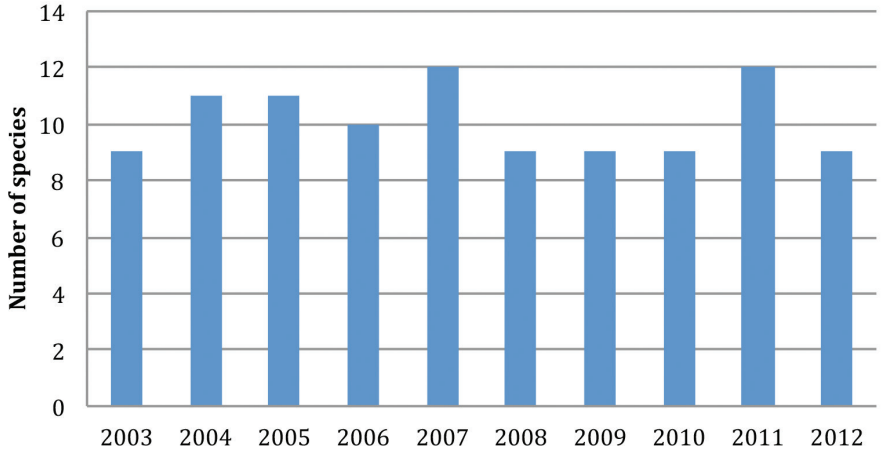


Figure 1. Numbers of different species of marine mammals stranded along the coasts of Brittany (North West of France) in the period 2003–2012.

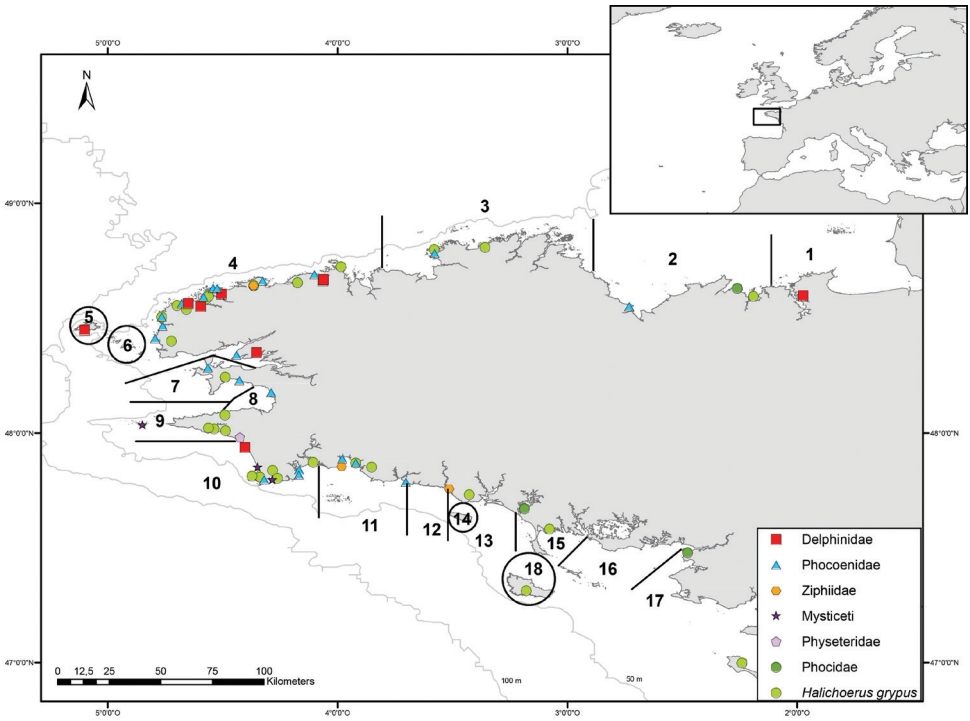


Figure 2. Organization of the stranding network in Brittany (northwest of France) and localization of the stranded specimens used in this study. Numbers indicate the 18 geographic sections of the stranding network in this area. The map was drawn using ArcGIS Desktop: Release 9.3.1 (Environmental Systems Research Institute, Redlands, CA, USA) with WGS 84 coordinates.

was transported to the care center (Figure 2). Our sampling included 12 species of cetaceans, and three species of pinnipeds (Table 2). Two species were very common, the harbour porpoise (29 samples) and the grey seal (44 samples), thus allowing intraspecific distance analyses.

A COI amplicon was recovered from 89 samples, and good quality sequences of more than 500 bp were obtained for all samples (GenBank accession numbers KF281608–KF281697). The sequence alignment used in the analyses was 507 bp long. About 32% of the positions were polymorphic in the cetaceans and 13.1% in the pinnipeds (Table 3). The maximal intraspecific distance was 0.46% for the grey seal and 0.83% for the harbour porpoise. The COI sequences of three species of the Delphininae (*Stenella frontalis*, *Stenella coeruleoalba* and *Delphinus delphis*) showed very low interspecific distances (0.84% between *D. delphis* and the nearest species *S. frontalis,* and 1.18% between the two *Stenella* species). All other interspecific distances were above 3.9% for pinnipeds and above 6% for cetaceans. The Neighbour-Joining (NJ) tree built on the BOLD interface using K2P-distances (Figure 3) confirms that, except for of the Delphininae, all the cetacean and pinniped species analyzed are distinguished unambiguously.

Table 2. Numbers of samples included in the IMMB project

Cetaceans (12 species)	
Balaenoptera acutorostrata	1
Balaenoptera physalus	1
Delphinus delphis	1
Grampus griseus	3
Hyperoodon ampullatus	2
Lagenorhynchus acutus	2
Phocoena phocoena	29
Physeter macrocephalus	1
Stenella coeruleoalba	1
Stenella frontalis	1
Tursiops truncatus	1
Ziphius cavirostris	1
Pinnipeds (3 species)	
Cystophora cristata	2
Halichoerus grypus	44
Phoca vitulina	2
Total (15 species)	92

Table 3. Polymorphism levels of COI between 12 species of marine mammals stranded in Brittany, and comparison with intra-species variation for harbour porpoise and grey seal.

	Total	**Cetaceans (12 species)**	**Pinnipeds (3 species)**	**Harbour porpoises**	**Grey seals**
Number of species	14	11	3	1	1
Number of sequences	89	41	48	45*	44
Length (bp)	507	508	656	610	658
Polymorphic sites	186	163	86	8	7
Polymorphism (%)	36.7	32.1	13.1	1.3	1.06
Minimal distance to NN	-	0.84	3.9	13.46	3.9
Maximal distance to NN	-	17.3	11.2	-	-
Maximal intraspecific distance	-	-	-	0.83	0.46

*This sampling includes 28 harbour porpoises stranded along the coasts of Brittany, and 17 more samples, stranded or by-caught in the Bay of Biscay, included to better characterize intraspecific variation. NN: nearest neighbour.

Taxonomic identification of undetermined samples

We then determined COI sequences from 10 cetacean samples whose species could not be determined accurately using morphological characters (Figure 4), either because only parts of the animal were recovered (Figure 4A) or because of the highly degraded state of the carcasses (Figure 4C). COI sequences of good qualities were obtained from all these samples, and three of them were identified unambiguously using the BOLD identification engine: Ms250511 was identified as a *Balaenoptera physalus*, Ds160111

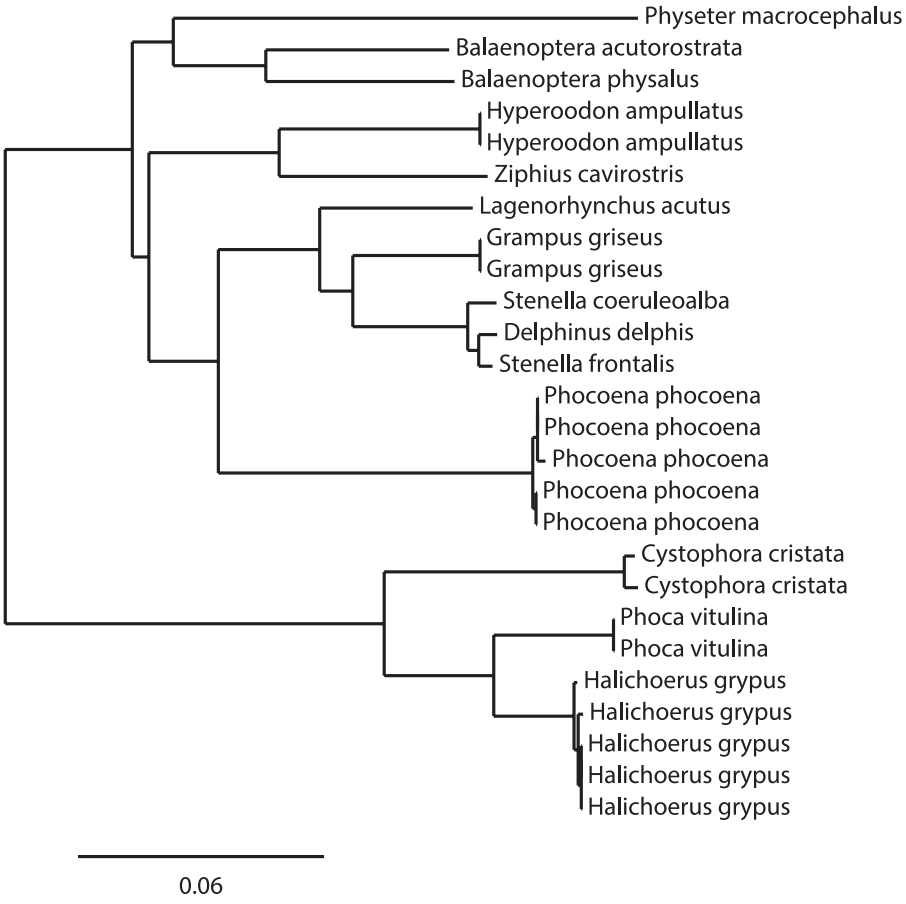


Figure 3. Neighbour-Joining tree of major species of marine mammals, based on K2P-distances calculated from 507 bp of COI. All sequences come from the IMMB project on BOLD, and only 5 harbour porpoise and 5 grey seal samples among those of the IMMB project have been included in the analysis.

as a *Grampus griseus* and Ds290811 as a *Phocoena phocoena*. The other seven samples were Delphininae, as confirmed by COI sequences. Yet, neither the BOLD identification engine, nor a BLAST search on GenBank allowed a more precise determination. We therefore sequenced MCR, which is more variable than COI, from six unidentified samples. BLAST searches on GenBank confirmed the COI results: all these samples were Delphininae, but a more precise identification could not be achieved.

We constructed a nMDS plot of the distances between MCR sequences of *S. coeruleoalba, S. frontalis* and *D. delphis* taken from GenBank: for *S. coeruleoalba*, we used sequences AM498725, AM498723, AM498721, AM498719, AM498717, AM498715, AM498713, AM498711, AM498709, AM498707 (Mace et al. unpublished), for *D. delphis* FM211560, FM211553, FM211545, FM211535, FM211527, FM211519, FM211511, FM211503, FM211495 (Mirimin et al. 2009) and DQ520121,

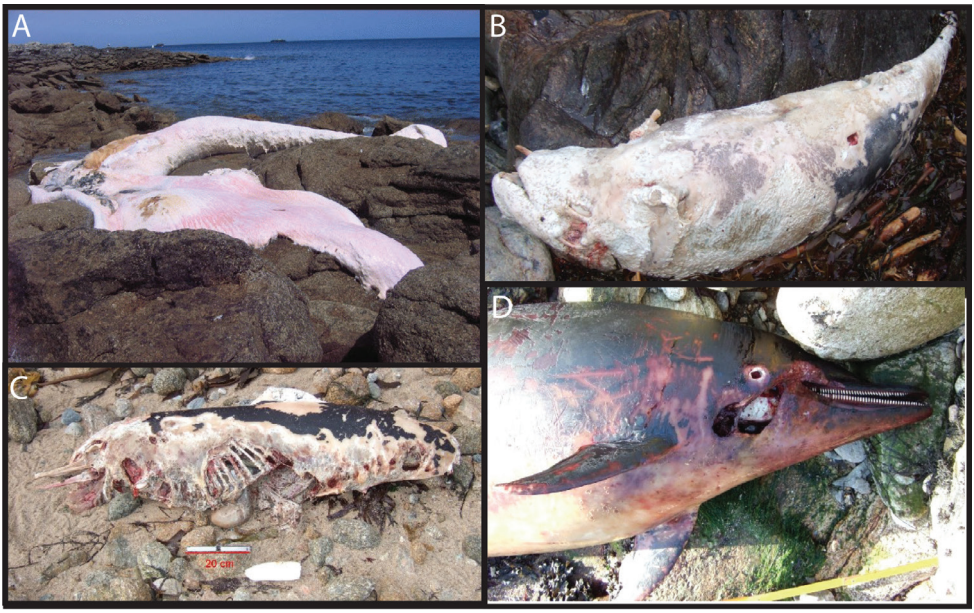


Figure 4. Examples of marine mammals stranded along the coasts of Brittany and the species-level identifications of which were determined or confirmed thanks to DNA barcoding. **A** Sample Ms250511, stranded on the "Île de Sein" during May 2011, and identified as a *Balaenoptera physalus* **B** Sample Ds160111, stranded on the Ushant Island during January 2011, and identified as a *Grampus griseus* **C** Sample Ds130211, stranded on the Ushant Island in February 2011, and identified as belonging to the Delphininae subfamily (putatively identified as a *D. delphis* on the nMDS plot in Figure 5) **D** Sample Ds080410 stranded on the Ushant Island during April 2010, and identified as belonging to the Delphininae (putatively identified as a *S. coeruleoalba* on the nMDS plot in Figure 5).

DQ520117, DQ520113, DQ520109, DQ520105 (Hildebrandt et al. unpublished) and for *S. frontalis* GQ5041986, GQ5041987, GQ5041988, GQ5041989, GQ5041990, GQ5041991, GQ5041992, GQ5041993, GQ5041994, GQ5041995 (Kingston et al. 2009).

The three species were clearly discriminated by the nMDS (Figure 5). The posterior probabilities are given in Appendix 2. This analysis suggests that five of our unidentified samples could belong to *D. delphis*, and one to *S. coeruleoalba*.

Intraspecific variation of COI and MCR in harbour porpoise and grey seal

For the intraspecific analysis of the harbour porpoise, we included 17 additional samples of animals stranded or by-caught from the Bay of Biscay (Appendix 1, Alfonsi et al. 2012). All in all, we compared 35 sequences of grey seals, and 45 of harbour porpoises. As expected, MCR sequences were more polymorphic than COI: in harbour porpoise,

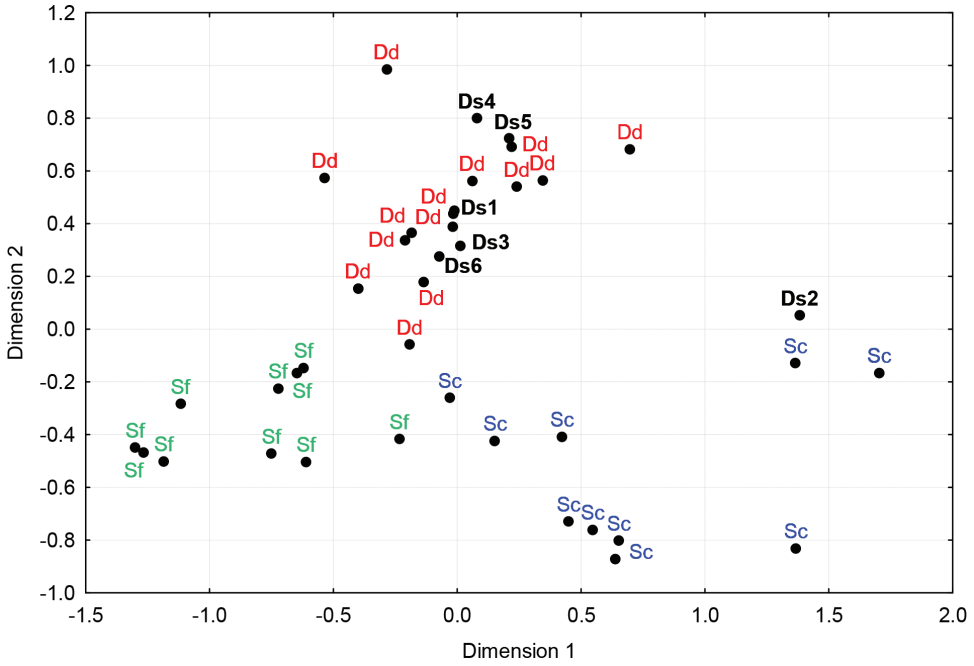


Figure 5. Non-metric Multidimensional Scaling plot of K2P-distance between MCR sequences of *S. coeruleoalba* (in blue), *D. delphis* (in red) and *S. frontalis* (in green). Individuals of each species are clearly clustered together, and unidentified samples (in black) stranded along the coasts of Brittany group with one of the three species. Dd280211A (Ds1), Ds130210 (Ds3), Ds230409 (Ds4), Ds250412 (Ds5) and Sc210910 (Ds6) are putatively identified as *D. delphis,* whereas Ds080410 (Ds2) would more likely belong to *S. coeruloalba*.

Table 4. Comparison of intraspecific COI and MCR polymorphisms for grey seal and harbour porpoise

	Harbour porpoise (*P. phocoena*)		**Grey seal (*H. grypus*)**	
Markers	**COI**	**MCR**	**COI**	**MCR**
Number of sequences	45	45	35	35
Sequence length (bp)	610	579	658	482
Haplotypes	9	14	6	14
Polymorphic sites	8	22	5	23
Polymorphism	1.30%	3.80%	0.76%	4.77%
Haplotype diversity	0.695	0.832	0.553	0.935
Nucleotide diversity	0.00242	0.00632	0.00098	0.00945

3.8% of the MCR positions were polymorphic vs. 1.30% in COI, while 4.73% of the MCR positions in the grey seal were polymorphic vs. 0.75% in COI (Table 4). Hence, MCR was 3× more polymorphic than COI in harbour porpoise and 6x in grey seals. Haplotype and nucleotide diversities were also higher for MCR than for COI.

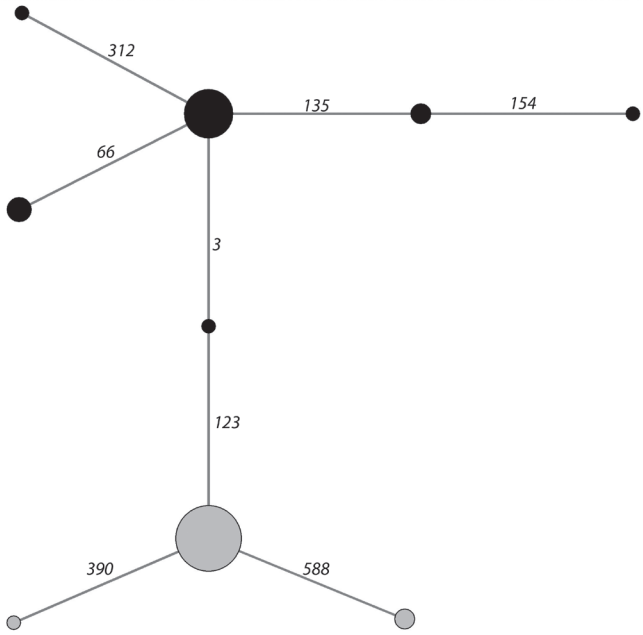


Figure 6. Haplotype network established from the COI sequences of 45 harbour porpoises stranded along the Atlantic coast of France (Appendix 1). Numbers on a line connecting two haplotypes correspond to the sequence position of the mutation differentiating these haplotypes. Two mitochondrial haplogroups appear (black circles - grey circles), that group the same individuals as the haplogroups alpha and beta determined using MCR polymorphisms and described in Alfonsi et al. (2012).

The haplotype network of the COI sequences in harbour porpoises clearly differentiated two haplogroups (Figure 6), that correspond perfectly to those described for MCR in Alfonsi et al. (2012).

Discussion

Stranding networks collect opportunistic data that are ecologically significant (Borsa 2006, Jung et al. 2009, Peltier et al. 2013), although, among other parameters, data quality control may deserve a special attention (Evans and Hammond 2004). Stranding networks can also collect skin and muscle samples that can be used for genetic analysis, therefore contributing to the construction of biological sample banks which are of high value when working with marine mammals.

The aim of this study was to evaluate the feasibility of a routine use of DNA barcoding in a stranding network; and to determine which gains this use could bring in terms of data relevance. The Brittany stranding network is a part of the French stranding network, and has to analyze an average of around 150 marine mammal strandings per year, with a high species biodiversity (19 species during 2003–2012).

Can COI be used as an appropriate species identification tool for marine mammals in the frame of a stranding network?

We obtained DNA sequences of good quality for almost all the samples studied, whatever their origin, their collectors, or even their state of degradation. This is consistent with the numerous molecular genetic studies that have used samples taken on stranded cetaceans or pinnipeds (e.g. Gaspari et al. 2006, Amaral et al. 2007, Fontaine et al. 2007, Mirimin et al. 2009, 2011, Alfonsi et al. 2012).

Viricel and Rosel (2012) previously demonstrated that COI sequences allowed identifying cetacean species, except for a few closely related Delphinidae species (see also Amaral et al. 2007). As expected, our NJ tree matched the overall classification, and the distance-based analysis identified correctly the sequences to the species levels for all cetaceans except within the Delphininae. The three species of pinnipeds analyzed were also unambiguously distinguished on the basis of their COI sequences.

The quality of the whole functioning and organization of the stranding network, from the field-work achieved by the correspondents to the preservation of the samples is therefore confirmed by our study. All the samples analyzed by DNA barcoding led to correct identification of the expected species with no exceptions.

We obtained COI good quality sequences for 10 unidentified animals, some of which were in a highly degraded body state. This showed that DNA barcoding can help to identify such specimens, which represent more than 16% of the stranded animals in the period 2003–2012. Hence, a routine use of DNA barcoding would noticeably decrease the proportion of unidentified animals.

The case of the Delphininae

Within the Delphininae, species are difficult to discriminate (Amaral et al. 2007, 2012, Viricel and Rosel 2012). In particular, *Delphinus delphis*, *Stenella coeruleoalba* and *Stenella frontalis* show very low interspecific COI distances, which do not allow distinguishing the species accurately. Other mitochondrial loci, such as MCR and cyt *b*, are neither very effective in this matter (Amaral et al. 2007, Viricel and Rosel 2012). This is attributed to recent and rapid radiation events in the subfamily, and it leads to problematic results in molecular taxonomic studies (Kingston et al. 2009, Amaral et al. 2012, Viricel and Rosel 2012, Perrin et al. 2013). In our case, these three species produced COI and MCR sequences that did not allow to associate samples with species names, neither with the identification engine on BOLD, nor with a distance tree or a BLAST search on GenBank. nMDS of genetic distances is known to uncover sample clustering (e.g. Geffen et al. 2004, Maltagliati et al. 2006, Alfonsi et al. 2012, Weckworth et al. 2012). As such, nMDS clustering of MCR sequence distances of *Delphinus delphis*, *Stenella coeruleoalba* and *Stenella frontalis*, chosen randomly on GenBank among Atlantic samples, showed that individuals of the three species formed separate groups. Moreover, each individual had a high posterior probability to belong to the

right group, except for one sample (i.e. 97.0% of the assignments were successful), so that all our unidentified samples could be putatively identified to the species level, based on the nMDS plot and its posterior probabilities.

Can DNA barcoding increase the accuracy of the data listed by the stranding network?

DNA barcoding is informative for animals that belong to species that infrequently strand along the coasts of Brittany, which can involve either species living far off the coasts or living in deep water, but also exotic species. Such species can be more difficult to identify by the field correspondents simply because of their scarcity. Along the coast of Brittany, we observed a *Stenella frontalis*, a temperate to tropical Atlantic Ocean inhabitant, and three species of arctic seals (*Phoca hispida*, *Cystophora cristata* and *Phoca groenlandica*). It is likely that other members of such rare species are listed among the "undetermined" species, just because their morphological characteristics are less well known by field correspondents. Additionally, a species that rarely strands along the French coast may be mistakenly identified as its more common sister-species. This issue can be illustrated by the case of the two pilot whale species: *Globicephala melas*, the long-finned pilot whale, commonly strands along the French Atlantic coast, while only a few stranding events of *G. macrorhynchus*, the short-finned pilot whale, have been reported (the Bay of Biscay is the northern limit of the geographical range of *G. macrorhynchus*). The two species have overlaping morphological characters, which adds to the difficulty of detecting rare stranding events of *G. macrorhynchus* based on morphological data only (Viricel and Sabatier unpublished data). A systematic use of DNA barcoding when morphological taxonomic characteristics are not straightforward, would clearly lower the percentage of exotic animals not listed. The existence of natural interspecific hybrids between the two *Globicephala* sister-species (Miralles et al. 2013), as between other cetacean species (e.g. Bérubé and Aguilar 1998, Willis et al. 2004) still reinforces the interest of such a monitoring based on molecular data.

It is important to note that a main limitation of DNA barcoding is the use of a single locus, leading to some problematic species identification such as within the Delphininae, but also to an inability to detect hybrids without complementary genetic studies. This limitation may well be removed in the near future thanks to next-generation sequencing, allowing the accumulation of large amount of DNA sequence data in a cost-effective manner. Multi-locus barcoding, including mitochondrial and nuclear polymorphic loci, will certainly represent a next step for the barcoding community.

A routine use of DNA barcoding could also allow monitoring the marine mammal biodiversity at intraspecific levels. For instance, global climate change has some effects on genetic diversity that must be studied and quantified (Pauls et al. 2012), in particular in the marine realm. Knowledge of the existence of distinct genetic groups or populations, of the history of their formation and of their movements are of a first importance to ecological understandings of natural populations, and also to the conservation efforts dedicated to them. Around the coast of Brittany, different species of marine mammals

have shown variations in abundance in the last decades (Vincent et al. 2005, Jung et al. 2009). Using samples from the French Stranding Network and MCR polymorphisms, we have recently shown that two previously separated, genetically distinct, populations of harbour porpoises are now admixing along the Atlantic coast of France (Alfonsi et al. 2012). These results were unexpected according to previous work (Tolley and Rosel 2006, Fontaine et al. 2007). In this study, we show that this genetic clustering would also have been detected using COI polymorphisms, thus reinforcing the interest of a routine use of DNA barcoding in conjunction with the stranding network.

Contributions of our study to the Barcoding of Life Database

This project is part of the collaboration between the Laboratory BioGeMME of the "Université de Bretagne Occidentale" (Brest, France), Océanopolis, a public private company (http://www.oceanopolis.com), the "Parc naturel marin d'Iroise" (http://www.parc-marin-iroise.com) and the French Stranding Network, coordinated by *Pelagis*, Université de La Rochelle, France. All the specimens and sequence data described in this manuscript are deposited in BOLD under the institution called "Oceanopolis-BioGeMME" in two projects, UMMB and IMMB. Our mixed institution became the first contributor to BOLD for the Cetacea, as well as for the Phocidae, and these two BOLD projects will be publicly available, and all the sequences published on GenBank.

Acknowledgements

The samples in this study were collected by the French Stranding Network and stored at Océanopolis (Brest) and at Pelagis (La Rochelle). We especially thank all the members of the LEMM, Christine Dumas who is in charge of the care center at Océanopolis and Willly Dabin and Olivier Van Canneyt from Pelagis. This work is part of a collaboration with "The Parc Naturel Marin d'Iroise", and especially with Phillipe Le Niliot, Cécile Lefeuvre and all the technicians working on the field.

We are indebted to Paul Marec, who collected the samples and took the photographs of the stranded animals on Ushant Island. Loriane Mendez, Anne Moulinet, Sandrine Quemener participated to some experiments during their internships in BioGeMME. Special thanks go to Chantal Hily-Mazé.

This manuscript benefited from useful comments during the reviewing and editing process.

References

Alfonsi E, Hassani S, Carpentier F-G, Le Clec'h J-Y, Dabin W, Van Canneyt O, Fontaine M C, Jung J-L (2012) A European Melting Pot of Harbour Porpoise in the French Atlan-

tic Coasts Inferred from Mitochondrial and Nuclear Data. PLoS ONE 7: e44425. doi: 10.1371/journal.pone.0044425

Amaral AR, Jackson JA, Möller LM, Beheregaray LB, Coelho MM (2012) Species tree of a recent radiation: The subfamily Delphininae (Cetacea, Mammalia). Molecular Phylogenetics and Evolution 64: 243–253. doi: 10.1016/j.ympev.2012.04.004

Amaral AR, Sequeira M, Coelho MM (2007) A first approach to the usefulness of cytochrome c oxidase I barcodes in the identification of closely related delphinid cetacean species. Marine and Freshwater Research 58: 505–510. doi: 10.1071/MF07050

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. Molecular Ecology 13: 729–744. doi: 10.1046/j.1365-294X.2003.02063.x

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2010) GenBank. Nucleic Acids Research 39 (Database), D32–D37. doi: 10.1093/nar/gkp1024

Bérubé M, Aguilar A (1998) A new hybrid between a blue whale *Balaenoptera musculus* and a fin whale *B. physalus*: frequency and implications of hybridization. Marine Mammal Science 14: 82–98. doi: 10.1111/j.1748-7692.1998.tb00692.x

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. doi: 10.1111/j.1471-8286.2007.01998.x

Borsa P (2006) Marine mammal strandings in the New Caledonia region, Southwest Pacific. Comptes Rendus Biologies 329: 277–288. doi: 10.1016/j.crvi.2006.01.004

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. doi: 10.1111/j.1471-8286.2006.01657.x

Dawnay N, Ogden R, Mcewing R, Carvalho G, Thorpe R (2007) Validation of the barcoding gene COI for use in forensic genetic species identification. Forensic Science International 173: 1–6. doi: 10.1016/j.forsciint.2006.09.013

Evans P, Hammond P (2004) Monitoring cetaceans in European waters. Mammal Review 34: 131–156. doi: 10.1046/j.0305-1838.2003.00027.x

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

Fontaine MC, Baird SJ, Piry S, Ray N, Tolley KA, Duke S, Birkun A, Ferreira M, Jauniaux T, Llavona Á, Öztürk BA, Öztürk A, Ridoux V, Rogan E, Sequeira M, Siebert U, Vikingsson GA, Bouquegneau J-M, Michaux J R (2007) Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters. BMC Biology 5: 30. doi: 10.1186/1741-7007-5-30

Gaspari S, Airoldi S, Hoelzel AR (2006) Risso's dolphins (*Grampus griseus*) in UK waters are differentiated from a population in the Mediterranean Sea and genetically less diverse. Conservation Genetics 8: 727–732. doi: 10.1007/s10592-006-9205-y

Geffen E, Anderson MJ, Wayne RK (2004) Climate and habitat barriers to dispersal in the highly mobile grey wolf. Molecular Ecology 13: 2481–2490. doi: 10.1111/j.1365-294X.2004.02244.x

Hajibabaei M, Singer GA, Clare EL, Hebert PDN (2007) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biology 5: 24. doi: 10.1186/1741-7007-5-24

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

Hartl DL, Clark AG (2007) Principles of Population Genetics. Sinauer and Associates, Sunderland, MA.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. PLoS Biology 2: e312. doi: 10.1371/journal.pbio.0020312

Jung J-L, Stéphan E, Louis M, Alfonsi E, Liret C, Carpentier F-G, Hassani S (2009) Harbour porpoises (*Phocoena phocoena*) in north-western France: aerial survey, opportunistic sightings and strandings monitoring. Journal of the Marine Biological Association of the United Kingdom 89: 1045–1050. doi: 10.1017/S0025315409000307

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120. doi: 10.1007/BF01731581

Kingston SE, Adams LD, Rosel PE (2009) Testing mitochondrial sequences and anonymous nuclear markers for phylogeny reconstruction in a rapidly radiating group: molecular systematics of the Delphininae (Cetacea: Odontoceti: Delphinidae). BMC Evolutionary Biology 9: 245. doi: 10.1186/1471-2148-9-245

Kmiec B, Woloszynska M, Janska H (2006) Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Current Genetics 50: 149–159. doi: 10.1007/s00294-006-0082-1

Lambert DM (2005) Is a Large-Scale DNA-Based Inventory of Ancient Life Possible? Journal of Heredity 96: 279–284. doi: 10.1093/jhered/esi035

LeDuc RG, Perrin WF, Dizon AE (1999) Phylogenetic relationships among the delphinid cetaceans based on full cytochrome b sequences. Marine Mammal Science 15: 619–648. doi: 10.1111/j.1748-7692.1999.tb00833.x

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452. doi: 10.1093/bioinformatics/btp187

Maltagliati F, Lai T, Casu M, Valdesalici S, Castelli A (2006) Identification of endangered Mediterranean cyprinodontiform fish by means of DNA inter-simple sequence repeats (ISSRs). Biochemical Systematics and Ecology 34: 626–634. doi: 10.1016/j.bse.2006.02.003

McGowen MR (2011) Toward the resolution of an explosive radiation—A multilocus phylogeny of oceanic dolphins (Delphinidae). Molecular Phylogenetics and Evolution 60: 345–357. doi: 10.1016/j.ympev.2011.05.003

Miralles L, Lens S, Rodriguez-Folgar A, Carrillo M, Martin V, Mikkelsen B, Garcia-Vazquez E (2013) Interspecific Introgression in Cetaceans: DNA Markers Reveal Post-F1 Status of a Pilot Whale. PLoS ONE 8: e69511. doi: 10.1371/journal.pone.0069511

Mirimin L, Westgate A, Rogan E, Rosel P, Andrew R, Coughlan J, Cross T (2009) Population structure of short-beaked common dolphins (*Delphinus delphis*) in the North Atlantic Ocean as revealed by mitochondrial and nuclear genetic markers. Marine Biology 156: 821–834. doi: 10.1007/s00227-009-1147-8

Mirimin L, Miller R, Dillane E, Berrow SD, Ingram S, Cross T, Rogan E (2011) Fine-scale population genetic structuring of bottlenose dolphins in Irish coastal waters. Animal Conservation 14: 342–353. doi: 10.1111/j.1469-1795.2010.00432.x

Pauls SU, Nowak C, Bálint M, Pfenninger M (2012) The impact of global climate change on genetic diversity within populations and species. Molecular Ecology 22: 925–946. doi: 10.1111/mec.12152

Peltier H, Baagøe HJ, Camphuysen KCJ, Czeck R, Dabin W, Daniel P, Deaville R, Haelters J, Jauniaux T, Jensen LF, Jepson PD, Keijl GO, Siebert U, Van Canneyt O, Ridoux V (2013) The Stranding Anomaly as Population Indicator: The Case of Harbour Porpoise *Phocoena phocoena* in North-Western Europe. PLoS ONE 8: e62180. doi: 10.1371/journal.pone.0062180

Perrin WF, Rosel PE, Cipriano F (2013) How to contend with paraphyly in the taxonomy of the delphinine cetaceans? Marine Mammal Science. doi: 10.1111/mms.12051

Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes 7: 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Shokralla S, Singer G, Hajibabaei M (2010) Direct PCR amplification and sequencing of specimens' DNA from preservative ethanol. BioTechniques 48: 233–234. doi: 10.2144/000113362

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680. doi: 10.1093/nar/22.22.4673

Thompson K, Baker CS, Van Helden A, Patel S, Millar C, Constantine R (2012) The world's rarest whale. Current Biology 22, R905–R906. doi: 10.1016/j.cub.2012.08.055

Toews DPL, Brelsford A (2012) The Biogeography of Mitochondrial and Nuclear Discordance in Animals. Molecular Ecology 21: 3907–3930. doi: 10.1111/j.1365-294X.2012.05664.x

Tolley KA, Rosel PE (2006) Population structure and historical demography of eastern North Atlantic harbour porpoises inferred through mtDNA sequences. Marine Ecology Progress Series 327: 297–308. doi: 10.3354/meps327297

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends in Ecology & Evolution 24: 110–117. doi: 10.1016/j.tree.2008.09.011

Vincent C, Fedak M, Mcconnell B, Meynier L, Saintjean C, Ridoux V (2005) Status and conservation of the grey seal, *Halichœrus grypus*, in France. Biological Conservation 126: 62–73. doi: 10.1016/j.biocon.2005.04.022

Viricel A, Rosel PE (2012) Evaluating the utility of COI for cetacean species identification. Marine Mammal Science 28: 37–62. doi: 10.1111/j.1748-7692.2010.00460.x

Vollmer NL, Viricel A, Wilcox L, Moore MK, Rosel P (2011) The occurrence of mtDNA heteroplasmy in multiple cetacean species. Current Genetics 57: 115-131. doi: 10.1007/s00294-010-0331-1

Ward RD, Hanner R, Hebert PDN (2009) The campaign to DNA barcode all fishes, FISH-BOL. Journal of Fish Biology 74: 329–356. doi: 10.1111/j.1095-8649.2008.02080.x

Weckworth BV, Musiani M, Mcdevitt AD, Hebblewhite M, Mariani S (2012) Reconstruction of caribou evolutionary history in Western North America and its implications for conservation. Molecular Ecology 21: 3610–3624. doi: 10.1111/j.1365-294X.2012.05621.x

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). Frontiers in Zoology 4: 8. doi: 10.1186/1742-9994-4-8

Willis PM, Crespi BJ, Dill LM, Baird RW, Bradley Hanson M (2004) Natural hybridization between Dall's porpoises (*Phocoenoides dalli*) and harbour porpoises (*Phocoena phocoena*). Canadian Journal of Zoology 82: 828–834. doi: 10.1139/z04-059

Appendix 1

List of the 46 *Phocoena phocoena* analyzed. (doi: 10.3897/zookeys.365.5873.app1) File format: Microsoft Word file (doc).

Citation: Alfonsi E, Méheust E, Fuchs S, Carpentier F-G, Quillivic Y, Viricel A, Hassani S, Jung J-L (2013) The use of DNA barcoding to monitor the marine mammal biodiversity along the French Atlantic coast. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 5–24. doi: 10.3897/zookeys.365.5873 List of the 46 *Phocoena phocoena* analyzed. doi: 10.3897/zookeys.365.5873.app1

Appendix 2

Posterior probabilities for species identification determined by the nMDS analysis. (doi: 10.3897/zookeys.365.5873.app2) File format: Microsoft Word file (doc).

Citation: Alfonsi E, Méheust E, Fuchs S, Carpentier F-G, Quillivic Y, Viricel A, Hassani S, Jung J-L (2013) The use of DNA barcoding to monitor the marine mammal biodiversity along the French Atlantic coast. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 5–24. doi: 10.3897/zookeys.365.5873 Posterior probabilities for species identification determined by the nMDS analysis. doi: 10.3897/zookeys.365.5873.app2

ZooKeys 365: 25–48 (2013)
doi: 10.3897/zookeys.365.6287
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

DNA barcoding of Dutch birds

Mansour Aliabadian[1,2], Kevin K. Beentjes[2], C.S. (Kees) Roselaar[2], Hans van Brandwijk[2], Vincent Nijman[3], Ronald Vonk[2,4]

1 *Department of Biology, Ferdowsi University of Mashhad, Mashhad, Iran* **2** *Naturalis Biodiversity Center, Leiden, the Netherlands* **3** *Department of Social Sciences, Oxford Brookes University, Oxford, UK* **4** *Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, the Netherlands*

Corresponding author: *Mansour Aliabadian* (aliabadi@ferdowsi.um.ac.ir; mansour.aliabadian@naturalis.nl)

Academic editor: *Z. T. Nagy* | Received 22 September 2013 | Accepted 6 December 2013 | Published 30 December 2013

Citation: Aliabadian M, Beentjes KK, Roselaar CS, van Brandwijk H, Nijman V, Vonk R (2013) DNA barcoding of Dutch birds. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 25–48. doi: 10.3897/zookeys.365.6287

Abstract

The mitochondrial cytochrome *c* oxidase subunit I (COI) can serve as a fast and accurate marker for the identification of animal species, and has been applied in a number of studies on birds. We here sequenced the COI gene for 387 individuals of 147 species of birds from the Netherlands, with 83 species being represented by > 2 sequences. The Netherlands occupies a small geographic area and 95% of all samples were collected within a 50 km radius from one another. The intraspecific divergences averaged 0.29% among this assemblage, but most values were lower; the interspecific divergences averaged 9.54%. In all, 95% of species were represented by a unique barcode, with 6 species of gulls and skua (*Larus* and *Stercorarius*) having at least one shared barcode. This is best explained by these species representing recent radiations with ongoing hybridization. In contrast, one species, the Lesser Whitethroat *Sylvia curruca* showed deep divergences, averaging 5.76% and up to 8.68% between individuals. These possibly represent two distinct taxa, *S. curruca* and *S. blythi,* both clearly separated in a haplotype network analysis. Our study adds to a growing body of DNA barcodes that have become available for birds, and shows that a DNA barcoding approach enables to identify known Dutch bird species with a very high resolution. In addition some species were flagged up for further detailed taxonomic investigation, illustrating that even in ornithologically well-known areas such as the Netherlands, more is to be learned about the birds that are present.

Keywords

Aves, conservation, cytochrome *c* oxidase subunit I, COI, taxonomy

Introduction

DNA barcoding is used as an effective tool for both the identification of known species and the discovery of new ones (Hebert et al. 2003, 2010, Savolainen et al. 2005). The core idea of DNA barcoding is based on the fact that just a small portion of a single gene, comprising a 650 to 700 bp fragment from the first half of the mitochondrial cytochrome *c* oxidase subunit I gene (COI), shows a lower intraspecific than interspecific variation. An attribute which characterizes a threshold of variation for each taxonomic group, above which a group of individuals does not belong to the same species but instead forms an intraspecific taxon. In other words, the recognition of patterns in sequence diversity of a small fragment from the mtDNA genome has led to an alternative approach for species identification across phyla.

Initially, DNA barcodes were proposed for the Animal Kingdom in 2003, when Hebert and colleagues tested a single gene barcode to identify species and coined the term 'DNA barcoding' (Hebert et al. 2003). Since that time COI sequences have been used as identifiers in the majority of animal phyla including vertebrates (e.g. Hebert et al. 2004, Ward et al. 2005, Kerr et al. 2007, Smith et al. 2008, Nijman and Aliabadian 2010, Luo et al. 2011) and invertebrates (Hajibabaei et al. 2006, Bucklin et al. 2011, Hausmann et al. 2011). In recent years, the practical utility of DNA barcodes proved to be an appealing tool to help resolve taxonomic ambiguity (Hebert et al. 2004, 2010), to screen biodiversity (e.g. Plaisance et al. 2009, Naro-Maciel et al. 2009, Grant et al. 2011), and to support applications in conservation biology (Neigel et al. 2007, Rubinoff 2006, Dalton and Kotze 2011).

Birds are among the best-known classes of animals and thus provide a taxonomically good model for analyzing the applicability of DNA barcoding. In the last seven years some 30 scientific papers have been published on the DNA barcoding of bird species, which combined have been cited 500 times (V. Nijman, unpubl. data April 2013). Most of the studies have shown that from this small fragment of DNA, individuals have been identified down to species level for 94% of the species in Scandinavian birds (Johnsen et al. 2010), 96% in Nearctic birds (Kerr et al. 2009a), 98% in Holarctic birds (Aliabadian et al. 2009) and 99% in Argentinean and South Korean birds (Kerr et al. 2009a, Yoo et al. 2006). Species delineation relying on the use of theshold set to differentiate between intraspecific variation and interspecific divergence has been criticized as leading to too unacceptable high error rates especially in incompletely samples groups (Meyer and Paulay 2005). However, even the critics of DNA barcoding concede that DNA barcoding holds promise for identification in taxonomically well-understood and thoroughly sampled clades. Birds are taxonomically well-known, especially those of the Western Palearctic to which the Netherlands, our study area, forms part. As noted by Taylor and Harris (2012), compared to other taxa that have been subjected to DNA barcoding, DNA barcoding studies of birds tend to represent aggregations of very large number of bird species barcodes. These often include (near) cosmopolitan species with samples from distant geographic locations potentially increasing the amount of interspecific variation in COI.

Here we explore the efficiency of identifying species using DNA barcoding from a large set of sympatric bird species in the Netherlands. Compared to previous studies on birds, our study area covers a very small geographic area, allowing to directly test the functionality of DNA barcoding 'in one's backyard'.

Methods

Sampling

The Netherlands is a small, densely populated country in northwestern Europe, with a land surface area of some 34,000 km^2, and ornithologically it is arguable one of the best-covered countries (Sovon 2002). The tissue samples used for sequencing were collected from breeding areas in the Netherlands, excluding oversees dependencies. Given the small size of the country some 95% of the samples were collected within a 50 km-radius of each other. Samples were part of the tissue collection of the Zoological Museum of Amsterdam (ZMA), which were recently relocated and deposited in the Naturalis Biodiversity Center, Leiden. Most were collected in the period 2000–2012 by a network of volunteers, ringers, airport staff, and bird asylums; no birds were specifically collected or killed to be included in the collection of the ZMA. Species and subspecies identification was based on morphology and when necessary, external measurements. These identifications were done by authors HvB and CSR, with the help of Tineke G. Prins. Individual birds were frozen upon arrival to be thawed and skinned at a later date, and indeed many birds arrived frozen. Samples were mostly taken from the bird's pectorial muscles, because of its size and easy access, and stored in 96% ethanol. Species nomenclature follows the taxonomy of Dickinson (2003). The complete list of sampled specimens including information about vouchers and trace files is available from the project 'Aves of the Netherlands' at the BOLD website (http://www.barcodinglife.com/).

PCR and sequencing

The tissue samples were subsampled and subjected to DNA extraction using DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. PCR and sequencing reactions were performed, mainly following the same protocols described in Förschler et al. (2010), but with some minor modifications. Polymerase chain reaction (PCR) amplifications were initially performed using standard primers BirdF1 (TTCTCCAACCACAAAGACATTGGCAC) and BirdR1 (ACGTGGGAGATAATTCCAAATCCTG). When amplification was unsuccessful, alternate reverse primer BirdR2 (ACTACATGTGAGATGATTCCGAATCCAG) was used in combination with BirdF1 or alternate primer pair CO1-ExtF (ACGCTTTAACACTCAGCCATCTTACC) and CO1-ExtR (AACCAGCATATGAGGGTTCGATTCCT) was used (Hebert et al. 2004, Johnsen et al. 2010).

All PCRs were run under the following thermal cycle program: 3 min at 94 °C followed by 40 cycles of 15 s at 94 °C, 30 s at 50 °C and 40 s at 72 °C, and a final elongation of 5 min at 72 °C. For each reaction the PCR mixture consisted of 2.5 µl Qiagen Coral Load 10 × PCR buffer, 1.0 µl of each 10mM primer, 0.5 µl 2.5 mM dNTPs, 0.25 µl 5U/µl QiagenTaq DNA polymerase, 18.75 µl milliQ and 1.0 µl template DNA for a total volume of 25 µl. Bi-directional sequencing was performed for all specimens at Macrogen. We checked the possible amplification of pseudogenes (Numts) by translating the protein coding genes into amino acids sequences, but we did not observe any unexpected stop codons, frameshifts or unusual amino acidic substitutions. Furthermore we amplified a longer sequence of the COI gene with primers (CO1-ExtF and CO1-ExtR) for selected samples, and also here we did not see any indication of pseudogene co-amplification. Lijtmaer et al. (2012) found that, in birds, full-length COI pseudogenes are uncommon noting that they might be more frequently encountered when working with avian blood samples as opposed to muscle tissue samples (as used in here).

Data analysis

Sequences shorter than 500 bp and containing more than 10 ambiguous nucleotides were excluded from the analyses. All sequences have been deposited in GenBank (Accession numbers KF946551 to KF946937). A full list of the museum vouchers, for all specimens applied in this study, is provided in Appendix – Table 1.

For all sequence comparisons, the Kimura 2-parameter (K2P) model was used, because it is shown to be the best metric to compare closely related taxa (Nei and Kumar 2000, but for a contrasting view see Srivathsan and Meier 2012). Average intraspecific distances were calculated for those species that were represented by at least two specimens using Mega v5.1 software (Tamura et al. 2011).

For a group of birds that expressed a larger than expected intraspecific variation, the *Sylvia* warblers, we created a phylogenetic tree and created a haplotype network. We chose GTR+G+I as the best-fitting model of nucleotide substitution based on its Akaike's information criterion as implemented in JModelTest v0.1.1 (Posada 2008). A maximum likelihood (ML) tree was constructed in PAUP* v4.0b10 (Swofford 2002) using a heuristic search with the tree-bisection-reconnection branch-swapping algorithm and random addition of taxa. Relative branch support was evaluated with 500 bootstrap replicates (Felsenstein 1985). A minimum spanning haplotype network was constructed using a statistical parsimony network construction approach as implemented in TCS software package (Clement et al. 2000). This programme calculates the number of mutational steps by which pairwise haplotypes differ and computes the probability of parsimony (Templeton et al. 1992) for pairwise differences until the probability exceeds 0.95. The number of mutational differences associated with the probability just before the 0.95 cut-off point is then the maximum number of mutational connections between pairs of sequences justified by the parsimony criterion; these justified connections are applied in the haplotype network (Clement et al. 2000).

Results

A total of 387 sequences for 141 species (representing at least 158 taxa) were retrieved, including 52% of the breeding bird species in the Netherlands (Supplementary Table 1). The average number of sequences per species was 3.36 (range 1-13), with 83 species (59%) represented by more than two sequences. The mean K2P-divergence within species bears no significant relationship with sample sizes, i.e. number of sequences per species (R^2 = 0.001, p = 0.465). The mean intraspecific K2P-distance was 0.29% (range 0-8.68%) some 30 times lower than the mean intrageneric K2P-distances (9.54%, range 0-27.71%) (Table 1, Figure 1).

In general, 95% of species (134 species) showed a unique DNA barcode (these included the 58 species for which we only sequenced single individuals), while six congeneric species shared the same barcode and the mean intraspecific distance of them fell well below the threshold of species based on distance-based criterion (Hebert et al. (2003) 10 x rule). These congeneric species mostly included circumpolar species with close morphological similarities (Table 2).

Table 1. Comparisons of K2P-pairwise distances within various taxonomic levels for 83 species of birds from the Netherlands for which two or more sequences were available. Distances are expressed in percentages.

	Individuals	Taxa	Comparisons	Distances		
				Minimum	Mean ± S.E.M.	Maximum
Within Species	340	83	805	0	0.294±0.001	8.683
Within Genera	203	23	794	0	9.544±0.004	15.849
Within Families	282	20	2519	5.809	14.467±0.001	20.473

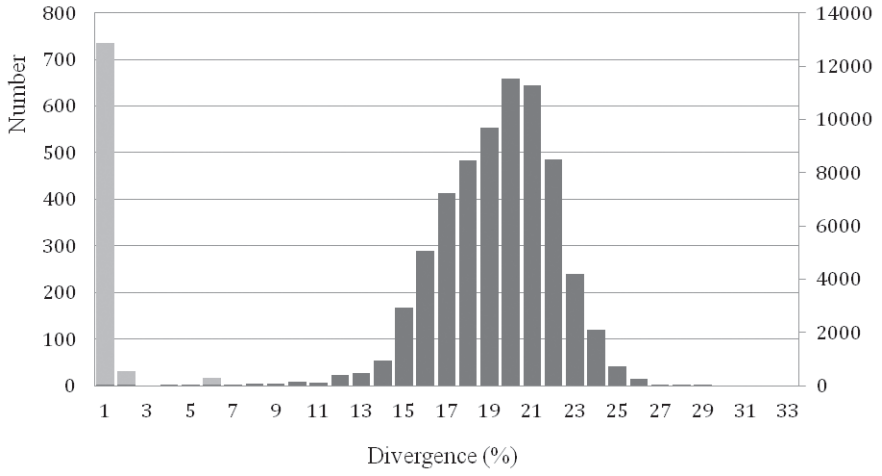


Figure 1. Comparisons of K2P-pairwise distances based on the COI gene of 141 species of birds from the Netherlands, showing a clear barcoding gap. Interspecific distances are indicated with light grey bars and intraspecific distances with dark grey bars. Left Y-axis: numbers of intraspecific comparisons; Right Y-axis: numbers of interspecific comparisons.

Table 2. Bird species (Charadriiformes) from the Netherlands with one or more shared DNA barcodes (K2P-distances of 0%). For a detailed breakdown of the individual samples involved see Appendix – Table 2.

Family	Species	Nearest species	Mean K2P-distance (%)
Laridae	Herring Gull *Larus argentatus*	Yellow-legged Gull *L. michahellis*	0.14
	Lesser Black-backed Gull *Larus fuscus*	Caspian Gull *L. cachinnans*	0
	Iceland Gull *Larus glaucoides*	Caspian Gull *L. cachinnans*	0
	Glaucous Gull *Larus hyperboreus*	Yellow-legged Gull *L. michahellis*	0.58
	Yellow-legged Gull *Larus michahellis*	Caspian Gull *L. cachinnans*	0
Stercorariidae	Great Skua *Stercorarius skua*	Pomarine Skua *S. pomarinus*	0.30

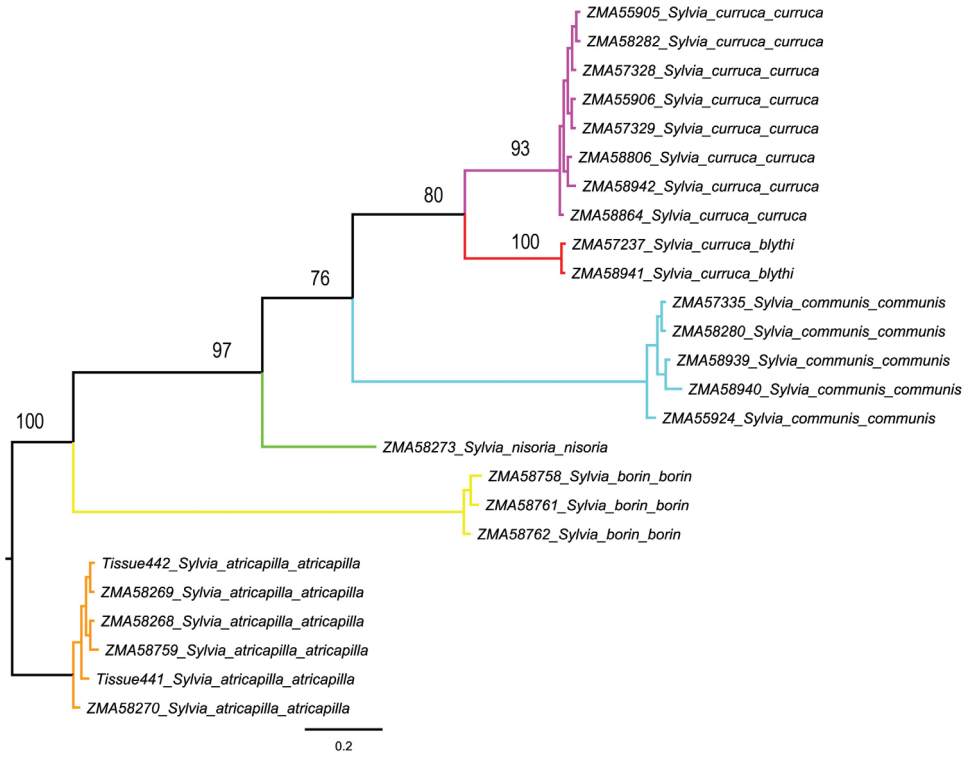


Figure 2. Phylogenetic relationships of two putative subspecies of Lesser Whitethroat, i.e. the Western Lesser Whitethroat *Sylvia curruca curruca* and the Northeastern Lesser Whitethroat *Sylvia curruca blythi* from the Netherlands, based on analysis of 694 bp of the mitochondrial cytochrome *c* oxidase subunit I gene (COI). Bootstrap values are given for the maximum likelihood (ML) analysis.

Although most species possessed low intraspecific distances, one species showed high intraspecific K2P-distances clearly above the threshold of 2 to 3 per cent sequence divergence in our data set. This is the Lesser Whitethroat *Sylvia curruca*, with a mean interspecific divergence of 5.76% and a maximum interspecific distance of 8.68%. Two subspecies occur in the Netherlands, i.e. the Western Lesser Whitethroat *S. c. curruca*

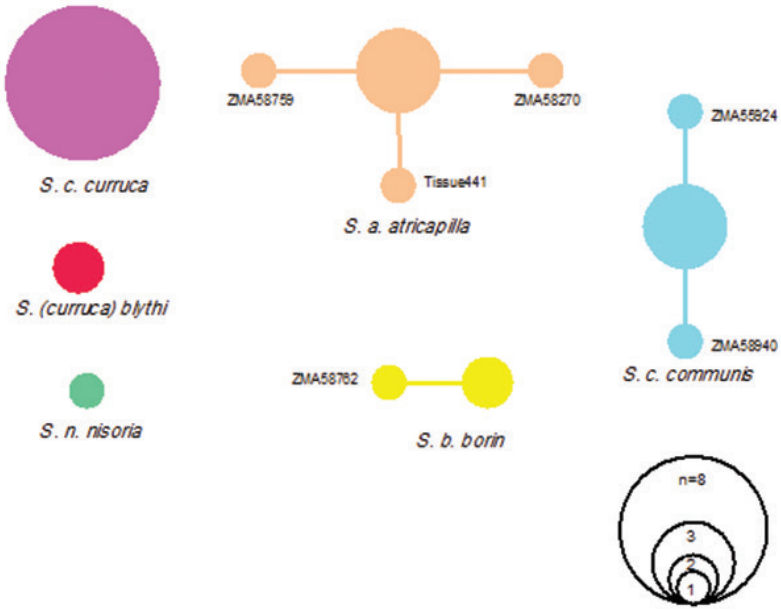


Figure 3. Haplotype networks constructed with statistical parsimony based on 694 bp of the mitochondrial cytochrome *c* oxidase subunit I gene (COI) of the *Sylvia* group (25 individuals). Each circle represents one haplotype; size of circles is proportional to haplotype frequency.

and, as a migrant, the Northeastern Lesser Whitethroat *S. c. blythi.* Both are morphologically somewhat distinct, with compared to the nominate *S. c. blythi* having a paler top of the head, separated from face by a white supercilium, and geographically the nominate occupies the western part of the species range and *S. c. blythi* the eastern part. A maximum likelihood tree for these two taxa based on Kimura 2-parameter is presented in Figure 2. Two different haplotype networks, one each for *S. c. curruca* and *S. c. blythi* were recovered by TCS (Figure 3), and given the large genetic distances between their haplotypes, the two taxa are not included in the same haplotype network.

Discussion

We here present the results of a modest effort to barcode the avifauna of the Netherlands. In terms of DNA barcoding of birds, the Netherlands form the southernmost part of one of the most densely sampled regions globally (Lijtmaer et al. 2012: figure 1). In addition, many of the species that overwinter in the country originate equally well-sampled regions to the north. As such our study adds to a growing number of studies allowing us to build up comprehensive public libraries of bird barcodes. Combined these allow us to explore new lines of scientific inquiry and practical applications (Hebert et al. 2010, Lijtmaer et al. 2012, but see Ebach and Carvalho 2010). The collection of our samples was done as part of the museum's standard collection man-

agement of newly obtained material, and as such sample collection was inexpensive and required little effort in terms of manpower. All birds were collected and processed in the Netherlands and did not require specific permits other than the ones already required to curate the collections.

Recently, Taylor and Harris (2012) expressed the opinion that proponents of DNA barcoding consistently fail to recognize its limitations (including, but not restricted to, the functioning of COI as a universal barcoding gene, whether its use is to be restricted to species identification only or whether it has a role in species discovery and delimitation and the failure to have sufficient systems in place to deal with the large amounts of data generated), do not evolve their methodologies, and do not embrace the possibilities that next-generation sequencing offers. We agree that DNA barcoding will not offer a panacea for all the issues Taylor and Harris (2012) raised, or indeed some of its earlier critics (Will et al. 2005, Moritz and Cicero 2004) but we point out that for this was probably never the intention of DNA barcoding when envisaged some ten years ago. Irrespective of the aims and goals of DNA barcoding as a 'global enterprise' (Ebach and Carvalho 2010), we found it a useful tool in our studies on birds (cf. Baker et al. 2009). The bird collection of the Zoological Museum Amsterdam, and our sample reported in this study, was well-curated by knowledgeable staff, with a very high degree of taxonomic certainty attached to each individual specimen. We see immense value to having a DNA barcoding dataset linked to this reference collection. As such this work has added to the growing library of DNA barcodes of bird species of the world and subsequent improvement in our knowledge of biodiversity.

The mean intraspecific divergences found in the birds of the Netherlands (0.29%, based on 147 species) is congruent with that of for instance Argentina (0.24%, 500 species), North America (0.23%, 643 species) and the Holarctic (0.24%, 566 species) (Kerr et al. 2009a, Aliabadian et al. 2009). More importantly, like other studies on birds, the efficiency of DNA barcode sequences to identify species is high, showing a clear barcoding gap (Figure 1), and overall it seems that for birds typically 95% or more of the species can be identified (Hebert et al. 2003, Johnsen et al. 2010, Kerr et al. 2009a, b, Yoo et al. 2006, Aliabadian et al. 2009).

Most DNA barcoding studies of birds flag a small number of deep divergences (e.g. Johnsen et al. 2010, Kerr et al. 2009b, Aliabadian et al. 2009, Nijman and Aliabadian 2013), in our study involving the two subspecies of *Sylvia curruca*, where the two lineages diverge almost 6%. Similar results were found by Olsson et al. (2013) when analyzing the cytochrome *b* gene for these two taxa, with distances in the order of 11-14%. Based on COI sequences, the two taxa appear to be sister taxa, albeit with a relatively low support (Figure 3), but no other members of the *Sylvia curruca* were included in the analysis. In contrast, having included a range of other members of this complex, Olsson et al. (2013) found *curruca* and *blythi* not to be sister taxa. Olsson et al. (2013: 81) concluded that while "due to their morphological similarity it is unclear where their ranges meet, [o]ur data suggest that *blythi* is a valid taxon, not closely related to *curruca*. It has its closest relatives to the south-east [Asia], and may have colonised the eastern taiga from this direction, ultimately coming into contact

with *curruca*". When it comes to drawing conclusion from their work with respect to taxomomy, Olsson et al. (2013) were, in our view correctly, cautious. They noted that the *Sylvia curruca* complex comprised up to 13 taxa with little consensus as to circumscription and taxonomic rank. Of these, morphologically some taxa are very similar, including *S. c. curruca* and *S. c. blythi*, and the apparent conflict between morphology and phylogeny (based in their case on cyt *b* and in our study on COI) can be explained in different ways. One would be to accept the single mitochondrial gene trees at face value in which case the morphological similarities in pelage coloration may be a result of parallel evolution possibly in response to adaptations to similar temperate forest habitats – both taxa are then best treated as different species. Alternatively, the mitochondrial gene trees do not reflect the species tree and, based on morphological similarities, *S. c. curruca* and *S. c. blythi* are best treated as sister taxa (either as one or two species). Their divergent position on the mitochondrial gene tree, and the large genetic distances between these taxa, are due to ancient mitochondrial introgression. In either case, working with single mitochondrial markers cannot not resolve this issue and a more integrative approach ideally involving the analysis of nuclear genes is paramount.

Those cases where we found species sharing the same DNA barcodes were small in number but not insignificant. Seven of the eight cases involved closely related gulls with partially overlapping ranges, or allopatric distributions, that are part of a recent Holarctic radiation (Liebers-Helbig et al. 2010). Alternatively, the the sharing of DNA barcodes may be due to hybridization or, perhaps less likely, misidentification. Likewise, skuas are part of a recent radiation with, just like gulls, frequent hybridization between species (Ritz 2009). DNA barcoding using a relative slowly evolving maternally inherited gene, with, compared to other mitochondrial genes, small amounts of rate heterogeneity (Pacheco et al. 2011), will, on its own, not be able to differentiate between these taxa.

We conclude that DNA barcoding approach makes it possible to identify known Dutch bird species with a very high resolution. Although some species were flagged for further detailed taxonomic investigation, our study reaffirms once more that a short segment of COI gene can be used to handle large number of taxa and aid in detecting overlooked taxa and hybridizing species with low deep barcode divergences.

Acknowledgments

We thank Tineke G Prins, involved in the sampling and administering of the bird specimens over the years in the Zoological Museum Amsterdam, for her commitment and hard work, and Miguel Vences, formerly of the Zoological Museum Amsterdam and currently at the Technical University Braunschweig, as the initiator of this project. Hans Breeuwer, Betsy Voetdijk, Peter Kuperus, and Lin Dong are thanked for their support and advice in the molecular laboratory of the Evolutionary Biology Department, University of Amsterdam. Finally, we thank the editors of this special issue for their patience, guidance and support, and two sets of reviewers for constructive com-

ments: combined their efforts greatly improved the quality and clarity of the work. Our molecular work is funded in part by the Fonds Economische Structuurversterking. We dedicate this paper to the memory of Jan Wattel, former curator of birds at the Zoological Museum Amsterdam, who passed away in March 2013.

References

Aliabadian M, Kaboli M, Nijman V, Vences M (2009) Molecular identification of birds: performance of distance-based DNA barcoding in three genes to delimit parapatric species. PLoS ONE 4: e4119. doi: 10.1371/journal.pone.0004119

Baker AJ, Tavares ES, Elbourne RF (2009) Countering criticisms of single mitochondrial DNA gene barcoding in birds. Molecular Ecology Resources 2009, 9: 257–267. doi: 10.1111/j.1755-0998.2009.02650.x

Bucklin A, Steinke D, Blanco-Bercial L (2011) DNA barcoding of marine metazoa. Annual Review of Marine Science 3: 471–508. doi: 10.1146/annurev-marine-120308-080950

Clement X, Posada D, Crandall K, (2000) TCS: A computer program to estimate gene genealogies. Molecular Ecology 9: 1657–1659. doi: 10.1046/j.1365-294x.2000.01020.x

Dalton DL, Kotze A (2011) DNA barcoding as a tool for species identification in three forensic wildlife cases in South Africa. Forensic Science International 207: e51–e54. doi: 10.1016/j.forsciint.2010.12.017

Dickinson EC (2003) The Howard & Moore Complete Checklist of the Birds of the World, 3rd Edition. Christopher Helm, London.

Ebach MC, Carvalho MRD (2010) Anti-intellectualism in the DNA barcoding enterprise. Zoologia (Curitiba) 27: 165–178. doi: 10.1590/S1984-46702010000200003

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783–791. doi: 10.2307/2408678

Förschler MI, Khoury F, Bairlein F, Aliabadian M (2010) Phylogenetic analyses of the Mourning Wheatear complex. Molecular Phylogenetics and Evolution 56: 758–767. doi: 10.1016/j.ympev.2010.03.022

Grant RA, Griffiths HJ, Steinke D, Wadley V, Linse K (2011) Antarctic DNA barcoding, a drop in the ocean? Polar Biology 34: 775–780. doi: 10.1007/s00300-010-0932-7

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proceedings of the National Academy of Sciences of the USA 103: 968–971. doi: 10.1073/pnas.0510466103

Hausmann A, Haszprunar G, Hebert PD (2011) DNA barcoding the geometrid fauna of Bavaria (Lepidoptera): successes, surprises, and questions. PLoS ONE 6: e17134. doi: 10.1371/journal.pone.0017134

Hebert PDN, Ratnasingham S, de Waard JR (2003) Barcoding animal life: cytochrome c oxidase subunit I divergences among closely related species. Proceedings of the Royal Society of London B (Supplement) 270: S96–S99. doi: 10.1098/rsbl.2003.0025

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. PLoS Biology 2: 1657–1663. doi: 10.1371/journal.pbio.0020312

Hebert PDN, de Waard JR, Landry JF (2010) DNA barcodes for 1/1000 of the animal kingdom. Biology Letters 6: 359–362. doi: 10.1098/rsbl.2009.0848

Johnsen A, Rindal E, Ericson PGP, Zuccon D, Kerr KCR, Stoeckle MY, Lifjeld D (2010) DNA barcoding of Scandinavian birds reveals divergent lineages in trans-Atlantic species. Journal of Ornithology 151: 565–578. doi: 10.1007/s10336-009-0490-3

Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Francis CM, Hebert PDN (2007) Comprehensive DNA barcode coverage of North American birds. Molecular Ecology Notes 7:535–543. doi: 10.1111/j.1471-8286.2007.01670.x

Kerr KCR, Lijtmaer DA, Barreira AS, Hebert PDN, Tubaro PL (2009a) Probing evolutionary patterns in Neotropical birds through DNA barcodes. PLoS ONE 4: e4379. doi: 10.1371/journal.pone.0004379

Kerr KC, Birks SM, Kalyakin MV, Red'kin YA, Koblik EA, Hebert PD (2009b) Filling the gap-COI barcode resolution in eastern Palearctic birds. Frontiers in Zoology 6(1): 29–42. doi: 10.1186/1742-9994-6-29

Liebers-Helbig D, Sternkopf V, Helbig AJ, de Knijff P (2010) The Herring Gull complex (*Larus argentatus-fuscus-cachinnans*) as a model group for recent Holarctic vertebrate radiations. In: Glaubrecht M (Ed) Evolution in Action, Springer, Berlin, 351–371. doi: 10.1007/978-3-642-12425-9_17

Lijtmaer DA, Kerr KC, Stoeckle MY, Tubaro PL (2012) DNA barcoding birds: from field collection to data analysis. In: Kress WJ, Erickson DL (Eds) DNA Barcodes: Methods and Protocols, Springer, New York, 127–152. doi: 10.1007/978-1-61779-591-6_7

Luo A, Zhang A, Ho SY, Xu W, Zhang Y, Shi W, Cameron SL, Zhu C (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. BMC Genomics 12(1): 84. doi: 10.1186/1471-2164-12-84

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS Biology 3: e422. doi: 10.1371/journal.pbio.0030422

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

Naro-Maciel E, Reid B, Fitzsimmons NN, Le M, DeSalle R, Amato G (2009) DNA barcodes for globally threatened marine turtles: a registry approach to documenting biodiversity. Molecular Ecology Resources 10: 252–263. doi: 10.1111/j.1755-0998.2009.02747.x

Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, Oxford.

Neigel J, Domingo A, Stake J (2007) DNA barcoding as a tool for coral reef conservation. Coral Reefs 26: 487–499. doi: 10.1007/s00338-007-0248-4

Nijman V, Aliabadian M (2010) Performance of distance-based DNA barcoding in the molecular identification of Primates. Comptes rendus Biologies 333: 11–16. doi: 10.1016/j.crvi.2009.10.003

Nijman V, Aliabadian M (2013) DNA barcoding as a tool for elucidating species delineation in wide-ranging species as illustrated by owls (Tytonidae and Strigidae). Zoological Science 30(11): 1005–1009. doi: 10.2108/zsj.30.1005

Olsson U, Leader PJ, Carey GJ, Khan AA, Svensson L, Alström P (2013) New insights into the intricate taxonomy and phylogeny of the *Sylvia curruca* complex. Molecular Phylogenetics and Evolution 67: 72–85. doi: 10.1016/j.ympev.2012.12.023

Pacheco MA, Battistuzzi FU, Lentino M, Aguilar RF, Kumar S, Escalante AA (2011) Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. Molecular Biology and Evolution 28: 1927–1942. doi: 10.1093/molbev/msr014

Plaisance L, Knowlton N, Paulay G, Meyer C (2009) Reef-associated crustacean fauna: biodiversity estimates using semi-quantitative sampling and DNA barcoding. Coral Reefs 28: 977–986. doi: 10.1007/s00338-009-0543-3

Posada D (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25:1253–1256. doi: 10.1093/molbev/msn083

Ritz M (2009) Speciation and hybridization in skuas (*Catharacta* spp.). PhD dissertation, Friedrich Schiller University, Jena.

Rubinoff D (2006) Utility of mitochondrial DNA barcodes in species conservation. Conservation Biology 20: 1026–1033. doi: 10.1111/j.1523-1739.2006.00372.x

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1805–1811. doi: 10.1098/rstb.2005.1730

Saunders GW (2005) Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. Philosophical Transactions of the Royal Society B 360: 1879–1888. doi: 10.1098/rstb.2005.1719

Seifert KA, Samson RA, de Waard JR, Houbraken J, Levesque CA, Moncalvo JM, Louis-Seize G, Hebert PDN (2007) Prospects for fungus identification using CO1 DNA barcodes, with *Penicillium*as a test case. Proceedings of the National Academy of Sciences of the USA 104: 3901–3906. doi: 10.1073/pnas.0611691104

Smith MA, Poyarkov NA, Hebert PDN (2008) CO1 DNA barcoding amphibians: take the chance, meet the challenge. Molecular Ecology Resources 8: 235–246. doi: 10.1111/j.1471-8286.2007.01964.x

SOVON (2002) Atlas van de Nederlandse broedvogels 1998–2000. SOVON, Nijmegen.

Srivathsan A, Meier R (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. Cladistics 28: 190–194. doi: 10.1111/j.1096-0031.2011.00370.x

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (and other methods), Version 4b10. Sinauer Associates, Sunderland, Massachusets.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12: 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132: 619–633.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B 360: 1847–1857. doi: 10.1098/rstb.2005.1716

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 5: 844–51. doi: 10.1080/10635150500354878

Yoo HS, Eah JY, Kim JS, Kim YJ, Min MS, Paek WK, Lee H, Kim CB (2006) DNA barcoding Korean birds. Molecules and Cells 22: 323–327.

Appendix

Supplementary Table 1. List of all Dutch birds that have been sequenced in this study, with voucher numbers and collection localities. Note that specimens from which only tissue samples have been taken have not been given a collection number, sine loco refers to specimens collected in the Netherlands but without a precise named collection locality. Localities in the province of Friesland are listed with their Dutch name first, followed by their Frisian name. Coordinates are given in decimal degrees.

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Accipiter gentilis gentilis	ZMA58297	skin	Zaandam	52.25N, 4.49E		KF946551
Accipiter gentilis gentilis	ZMA58724	skin	De Rips	51.32N, 5.48E		KF946552
Accipiter nisus nisus	ZMA58243	skin	Malden	51.47N, 5.52E		KF946553
Accipiter nisus nisus	ZMA58245	skin	Helden	51.21N, 5.55E		KF946554
Accipiter nisus nisus	ZMA58246	skin	Reuver	51.17N, 6.04E		KF946555
Accipiter nisus nisus	ZMA58247	skin	Culemborg	51.55N, 5.15E		KF946556
Accipiter nisus nisus	ZMA58248	skin	Amsterdam	52.21N, 4.53E		KF946557
Accipiter nisus nisus	ZMA58741	skin	Amsterdam	52.21N, 4.53E		KF946558
Accipiter nisus nisus	ZMA58742	skin	Montfort	51.07N, 5.56E		KF946559
Accipiter nisus nisus	ZMA58743	skin	Belfeld	51.18N, 6.08E		KF946560
Accipiter nisus nisus	ZMA58744	skin	Laren	52.11N, 6.22E		KF946561
Accipiter nisus nisus	ZMA58745	skin	Almere	52.22N, 5.13E		KF946562
Accipiter nisus nisus	ZMA58746	skin	Venlo	51.21N, 6.11E		KF946563
Acrocephalus palustris	ZMA56679	skin	Harderbroek reserve	52.22N, 5.35E		KF946564
Acrocephalus palustris	ZMA58811	skin	Castricum	52.32N, 4.36E		KF946565
Acrocephalus schoenobaenus	ZMA58278	skin	Almere	52.22N, 5.13E		KF946566
Acrocephalus schoenobaenus	ZMA58809	skin	Almere	52.22N, 5.13E		KF946567
Acrocephalus schoenobaenus	ZMA58810	skin	Castricum	52.32N, 4.36E		KF946568
Acrocephalus schoenobaenus	ZMA58862	skin	Wassenaar	53.08N, 5.53E		KF946569
Acrocephalus scirpaceus scirpaceus	ZMA58277	skin	Oostvaardersdijk	52.29N, 5.23E		KF946570
Acrocephalus scirpaceus scirpaceus	ZMA58725	skin	Schermerhorn	52.36N, 4.54E		KF946571
Acrocephalus scirpaceus scirpaceus	ZMA58727	skin	Lelystad	52.29N, 5.24E		KF946572
Acrocephalus scirpaceus scirpaceus	ZMA58728	skin	Lelystad	52.29N, 5.24E		KF946573
Acrocephalus scirpaceus scirpaceus	ZMA58729	skin	Castricum	52.32N, 4.36E		KF946574
Acrocephalus scirpaceus scirpaceus	ZMA58863	skin	Lauwersmeer	53.22N, 6.14E		KF946575

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Acrocephalus scirpaceus scirpaceus	ZMA58937	skin	Lelystad	52.29N, 5.24E		KF946576
Acrocephalus scirpaceus scirpaceus	ZMA58938	skin	Purmerend	52.28N, 4.58E		KF946577
Aegithalos caudatus europaeus	ZMA57353	skin	Westenschouwen	51.41N, 3.42E		KF946578
Aegithalos caudatus europaeus	ZMA57354	skin	Westenschouwen	51.41N, 3.42E		KF946579
Aegithalos caudatus europaeus	ZMA57356	skin	Hilversum	52.13N, 5.09E		KF946580
Aegithalos caudatus europaeus	ZMA58804	skin	Castricum	52.32N, 4.36E		KF946581
Alcedo atthis ispida	ZMA56216	skin	Haelen	51.13N, 5.56E		KF946582
Alcedo atthis ispida	ZMA57341	skin	Purmerland	52.28N, 4.55E		KF946583
Alcedo atthis ispida	ZMA57342	skin	Alkmaar	52.38N, 4.44E		KF946584
Alcedo atthis ispida	ZMA57343	skin	Utrecht	52.03N, 5.08E		KF946585
Alcedo atthis ispida	ZMA58869	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946586
Alle alle alle	ZMA58842	skin	Amsterdam	52.21N, 4.53E		KF946587
Alle alle alle	ZMA58917	skin	Amsterdam	52.21N, 4.53E		KF946588
Alle alle alle	ZMA58918	skin	Den Helder	52.55N, 4.46E		KF946589
Anas acuta	ZMA58228	skin	Vlieland Island	53.15N, 4.59E		KF946590
Anas strepera strepera	ZMA58913	skin	Driebond Polder	53.11N, 6.37E		KF946591
Anthus spinoletta spinoletta	ZMA58279	skin	Lelystad	52.29N, 5.24E		KF946592
Anthus spinoletta spinoletta	ZMA64552	skin	Castricum	52.32N, 4.36E		KF946593
Anthus trivialis trivialis	Tissue553	DNA sample	Castricum	52.32N, 4.36E		KF946594
Apus apus apus	ZMA58717	skin	Tegelen	51.19N, 6.09E		KF946595
Ardea cinerea cinerea	Tissue434	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946596
Ardea cinerea cinerea	Tissue435	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946597
Asio flammeus flammeus	ZMA58253	skin	Texel Island	53.04N, 4.43E		KF946598
Asio otus otus	Tissue455	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946599
Asio otus otus	ZMA58233	skin	Purmerend	52.28N, 4.58E		KF946600
Asio otus otus	ZMA58234	skin	Zutphen	52.07N, 6.12E		KF946601
Athene noctua vidalii	ZMA58493	skin	Heerhugowaard	52.4N	4.51E	KF946602
Athene noctua vidalii	ZMA58294	skin	Blerick	51.21N, 6.08E		KF946603
Bombycilla garrulus garrulus	ZMA56300	skin	Amsterdam	52.21N, 4.53E		KF946604
Bombycilla garrulus garrulus	ZMA56301	wings	Texel Island	53.04N, 4.43E		KF946605
Bombycilla garrulus garrulus	ZMA58301	wings	Hellendoorn	52.23N, 6.26E		KF946606
Bombycilla japonica	ZMA58302	skin	Amsterdam	52.21N, 4.53E		KF946607
Buteo buteo buteo	Tissue461	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946608
Buteo buteo buteo	ZMA58238	skin	Wieringermeer	52.54N, 5.01E		KF946609
Buteo buteo buteo	ZMA58239	skin	De Rips	51.32N, 5.48E		KF946610
Buteo buteo buteo	ZMA58781	wing	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946611
Buteo buteo buteo	ZMA58828	skin	Wartena	52.12N, 4.3E		KF946612
Buteo buteo buteo	ZMA58920	wings	Rolde	52.58N, 6.38E		KF946613

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Calidris alpina alpina	ZMA58700	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946614
Calonectris diomedea borealis	ZMA57255	skin	Lith	51.47N, 5.26E		KF946615
Carduelis cannabina cannabina	ZMA58911	skin	Noordijk	52.08N, 6.34E		KF946616
Carduelis carduelis	ZMA58866	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946617
Carduelis chloris chloris	ZMA57337	skin	Cadier en Keer	50.49N, 5.46E		KF946618
Carduelis chloris chloris	ZMA58947	skin	Goor	52.14N, 6.34E		KF946619
Carduelis flammea cabaret	ZMA57248	skin	Kennemerduinen	52.42N, 4.58E		KF946620
Carduelis flammea cabaret	ZMA58283	skin	Westenschouwen	51.41N, 3.42E		KF946621
Carduelis flammea flammea	ZMA57251	skin	Kennemerduinen	52.42N, 4.58E		KF946622
Carduelis flammea flammea	ZMA64564	skin	Castricum	52.32N, 4.36E		KF946623
Carduelis flavirostris	ZMA57253	skin	Castricum	52.32N, 4.36E		KF946624
Carduelis flavirostris	ZMA57254	skin	Castricum	52.32N, 4.36E		KF946625
Carduelis spinus	ZMA55904	skin	Nijverdal	52.22N, 6.28E		KF946626
Carduelis spinus	ZMA57256	skin	Westenschouwen	51.41N, 3.42E		KF946627
Carduelis spinus	ZMA58286	skin	Hellendoorn	52.23N, 6.26E		KF946628
Certhia brachydactyla megarhyncha	ZMA57322	skin	Hellendoorn	52.23N, 6.26E		KF946629
Certhia brachydactyla megarhyncha	ZMA57323	skin	Lekkerkerk	51.53N, 4.41E		KF946630
Certhia brachydactyla megarhyncha	ZMA57325	skin	Wageningen	51.58N, 5.38E		KF946631
Certhia brachydactyla megarhyncha	ZMA57326	skin	Zeist	52.05N, 5.16E		KF946632
Certhia brachydactyla megarhyncha	ZMA57327	skin	Heiloo	52.36N, 4.44E		KF946633
Certhia brachydactyla megarhyncha	ZMA58805	skin	Castricum	52.32N, 4.36E		KF946634
Certhia brachydactyla megarhyncha	ZMA58949	skin	Lekkerkerk	51.53N, 4.41E		KF946635
Certhia brachydactyla megarhyncha	ZMA64563	skin	Castricum	52.32N, 4.36E		KF946636
Charadrius hiaticula	Tissue452	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946637
Circus aeruginosus aeruginosus	ZMA58780	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946638
Circus aeruginosus aeruginosus	ZMA58826	skin	Eibergen	52.06N, 6.37E		KF946639
Circus aeruginosus aeruginosus	ZMA58874	wings	Zuid-Flevoland	52.26N, 5.16E		KF946640
Coccothraustes coccothraustes	ZMA56212	skin	Laag Keppel	51.59N, 6.13E		KF946641
Corvus corax corax	ZMA57144	skin	Appelscha/ Appelskea	52.55N, 5.2E		KF946642
Coturnix coturnix coturnix	ZMA58775	skin	Deventer	52.15N, 6.11E		KF946643
Coturnix coturnix coturnix	ZMA58776	skin	Het Bildt	53.17N, 5.4E		KF946644
Cuculus canorus canorus	ZMA56681	skin	Bergen	52.4N, 4.41E		KF946645

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Cuculus canorus canorus	ZMA64549	skin	Alkmaar	52.38N, 4.44E		KF946646
Delichon urbicum	ZMA56215	skin	Sea	,		KF946647
Delichon urbicum urbicum	ZMA55919	skin	Nieuwegein	52.01N, 5.05E		KF946648
Delichon urbicum urbicum	ZMA58300	wings	Lage Zwaluwe	51.42N, 4.42E		KF946649
Delichon urbicum urbicum	ZMA58870	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946650
Dendrocopos major pinetorum	ZMA58803	skin	Oudkerk/Aldtsjerk	53.15N, 5.53E		KF946651
Dryocopus martius martius	ZMA58766	skin	Tegelen	51.19N, 6.09E		KF946652
Emberiza citrinella citrinella	ZMA57257	skin	Westenschouwen	51.41N, 3.42E		KF946653
Emberiza melanocephala	ZMA56996	skin	Bovenkerk	52.17N, 4.49E		KF946654
Emberiza pusilla	ZMA58859	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946655
Emberiza pusilla	ZMA58860	skin	Vlieland Island	53.15N, 4.59E		KF946656
Emberiza schoeniclus schoeniclus	ZMA58857	skin	Noordpolderzijl	53.25N, 6.34E		KF946657
Emberiza schoeniclus schoeniclus	ZMA58858	skin	Oostvaardersdijk	52.29N, 5.23E		KF946658
Erithacus rubecula rubecula	Tissue436	DNA sample	Castricum	52.32N, 4.36E		KF946659
Erithacus rubecula rubecula	Tissue437	DNA sample	Castricum	52.32N, 4.36E		KF946660
Erithacus rubecula rubecula	ZMA58274	skin	Bloemendaal	52.24N, 4.33E		KF946661
Erithacus rubecula rubecula	ZMA58740	skin	Doldersum	52.52N, 6.17E		KF946662
Falco columbarius aesalon	ZMA58840	skin	Texel Island	53.04N, 4.43E		KF946663
Falco columbarius aesalon	ZMA60127	skin	Spaarndam	52.24N, 4.41E		KF946664
Falco peregrinus peregrinus	ZMA58872	skin	Haarlem	52.23N, 4.37E		KF946665
Falco subbuteo subbuteo	ZMA56231	skin	Zundert	51.28N, 4.38E		KF946666
Falco subbuteo subbuteo	ZMA56232	skin	Heerhugowaard	52.4N, 4.51E		KF946667
Falco subbuteo subbuteo	ZMA58241	skin	Hoogland	52.1N, 5.21E		KF946668
Falco subbuteo subbuteo	ZMA58242	skin	Texel Island	53.04N, 4.43E		KF946669
Falco subbuteo subbuteo	ZMA58841	skin	Amsterdam	52.21N, 4.53E		KF946670
Falco tinnunculus tinnunculus	Tissue456	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946671
Falco tinnunculus tinnunculus	ZMA58296	skin	Zaandam	52.25N, 4.49E		KF946672
Falco tinnunculus tinnunculus	ZMA58752	skin	Maasbree	51.21N, 6.03E		KF946673
Falco tinnunculus tinnunculus	ZMA58754	skin	Boekend	51.22N, 6.06E		KF946674
Falco tinnunculus tinnunculus	ZMA58774	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946675
Falco tinnunculus tinnunculus	ZMA58837	skin	Westzaan	52.26N, 4.46E		KF946676
Falco tinnunculus tinnunculus	ZMA58838	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946677
Falco tinnunculus tinnunculus	ZMA58839	wings	Reutum	52.23N, 6.5E		KF946678
Falco vespertinus	ZMA58773	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946679
Ficedula hypoleuca muscipeta	ZMA55913	skin	Otterlo	52.04N, 5.5E		KF946680
Ficedula hypoleuca muscipeta	ZMA57239	skin	Markelo	52.14N, 6.3E		KF946681
Ficedula hypoleuca muscipeta	ZMA57320	skin	Garderen	52.12N, 5.43E		KF946682
Ficedula hypoleuca	ZMA58865	skin	Eemshaven	53.26N, 6.52E		KF946683

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Fratercula arctica grabae	ZMA56226	skin	Texel Island	53.04N, 4.43E		KF946684
Fratercula arctica grabae	ZMA58226	skin	Texel Island	53.04N, 4.43E		KF946685
Fratercula arctica grabae	ZMA58227	skin	Hondsbossche Zeewering	52.44N, 4.38E		KF946686
Fringilla coelebs coelebs	ZMA58948	skin	Goor	52.14N, 6.34E		KF946687
Fringilla montifringilla	Tissue449	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946688
Fulmarus glacialis auduboni	ZMA56235	wings	Hondsbossche Zeewering	52.44N, 4.38E		KF946689
Fulmarus glacialis glacialis	ZMA60119	skin	Neeltje Jans	51.37N, 3.41E		KF946690
Fulmarus glacialis glacialis	ZMA60120	skin	Texel Island	53.04N, 4.43E		KF946691
Fulmarus glacialis glacialis	ZMA60121	skin	Hondsbossche Zeewering	52.44N, 4.38E		KF946692
Fulmarus glacialis glacialis	ZMA60123	skin	Ameland Island	53.27N, 5.39E		KF946693
Fulmarus glacialis glacialis	ZMA60124	skin	Ameland Island	53.27N, 5.39E		KF946694
Fulmarus glacialis glacialis	ZMA60125	skin	Hondsbossche Zeewering	52.44N, 4.38E		KF946695
Fulmarus glacialis glacialis	ZMA60126	skin	Petten	52.46N, 4.38E		KF946696
Fulmarus glacialis	ZMA58737	skin	Vlieland Island	53.15N, 4.59E		KF946697
Gallinula chloropus chloropus	Tissue105	DNA sample	Wijde Wormer	52.28N, 4.53E		KF946698
Gallinula chloropus chloropus	Tissue110	DNA sample	Wijde Wormer	52.28N, 4.53E		KF946699
Garrulus glandarius glandarius	ZMA58306	wings	Amsterdam	52.21N, 4.53E		KF946700
Gavia immer	Tissue214	DNA sample	Bergen aan Zee	52.39N, 4.37E		KF946701
Haematopus ostralegus ostralegus	Tissue458	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946702
Haematopus ostralegus ostralegus	Tissue459	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946703
Hirundo rustica rustica	Tissue450	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946704
Hirundo rustica rustica	Tissue451	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946705
Hirundo rustica rustica	ZMA56214	skin	Amstelveen	52.18N, 4.53E		KF946706
Hirundo rustica rustica	ZMA58289	skin	Appelscha/ Appelskea	52.55N, 5.2E		KF946707
Hirundo rustica rustica	ZMA58290	skin	Appelscha/ Appelskea	52.55N, 5.2E		KF946708
Hirundo rustica rustica	ZMA58696	skin	Rijswijk	51.57N, 5.21E		KF946709
Hirundo rustica rustica	ZMA58802	skin	Noordbergum/ Noardburgum	53.13N, 6E		KF946710
Jynx torquilla torquilla	ZMA56213	skin	Aarle-Rixtel	51.3N, 5.39E		KF946711
Jynx torquilla torquilla	ZMA57330	skin	Limmen	52.34N, 4.41E		KF946712
Jynx torquilla torquilla	ZMA58303	wings	Belfeld	51.18N, 6.08E		KF946713
Jynx torquilla torquilla	ZMA58873	skin	Wilnis	52.11N, 4.54E		KF946714
Larus argentatus argenteus	ZMA58921	wings	Eemshaven	53.26N, 6.52E		KF946715

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Larus argentatus	Tissue433	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946716
Larus cachinnans	ZMA64547	skin	Vlieland Island	53.15N, 4.59E		KF946717
Larus fuscus graelsii	Tissue432	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946718
Larus fuscus intermedius	Tissue327	DNA-sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946719
Larus fuscus intermedius	ZMA55932	skin	Neeltje Jans	51.37N, 3.41E		KF946720
Larus fuscus intermedius	ZMA56230	skin	Europoort	51.56N, 4.05E		KF946721
Larus fuscus intermedius	ZMA58834	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946722
Larus glaucoides glaucoides	ZMA58836	wings	Texel Island	53.04N, 4.43E		KF946723
Larus hyperboreus	ZMA56221	skin	Texel Island	53.04N, 4.43E		KF946724
Larus melanocephalus	ZMA57226	skin	Wijdenes	52.37N, 5.1E		KF946725
Larus michahellis michahellis	ZMA58835	skin	Afsluitdijk	52.57N, 5.04E		KF946726
Limosa lapponica lapponica	ZMA58202	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946727
Limosa lapponica lapponica	ZMA58203	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946728
Limosa lapponica taymyrensis	ZMA58204	skin	Paesens	53.24N, 6.06E		KF946729
Limosa lapponica taymyrensis	ZMA58205	skin	Paesens	53.24N, 6.06E		KF946730
Limosa lapponica taymyrensis	ZMA58206	skin	Paesens	53.24N, 6.06E		KF946731
Limosa lapponica taymyrensis	ZMA58207	skin	Paesens	53.24N, 6.06E		KF946732
Limosa lapponica taymyrensis	ZMA58208	skin	Paesens	53.24N, 6.06E		KF946733
Limosa lapponica taymyrensis	ZMA58782	wings	Castricum	52.32N, 4.36E		KF946734
Limosa lapponica taymyrensis	ZMA58783	wings	Castricum	52.32N, 4.36E		KF946735
Limosa limosa limosa	Tissue457	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946736
Limosa limosa limosa	ZMA57227	skin	Holysloot	52.24N, 5.01E		KF946737
Limosa limosa limosa	ZMA58229	skin	Waterland	52.07N, 4.19E		KF946738
Limosa limosa limosa	ZMA58230	skin	Edam	52.32N, 5.01E		KF946739
Limosa limosa limosa	ZMA58231	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946740
Limosa limosa limosa	ZMA58232	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946741
Locustella luscinioides luscinioides	ZMA64557	skin	Castricum	52.32N, 4.36E		KF946742
Locustella naevia naevia	ZMA56675	skin	Almere	52.22N, 5.13E		KF946743
Locustella naevia naevia	ZMA56678	skin	Almere	52.22N, 5.13E		KF946744
Locustella naevia naevia	ZMA57235	skin	Westenschouwen	51.41N, 3.42E		KF946745
Locustella naevia naevia	ZMA58812	skin	Castricum	52.32N, 4.36E		KF946746
Locustella naevia naevia	ZMA58936	skin	Hondsbossche Zeewering	52.44N, 4.38E		KF946747
Locustella naevia naevia	ZMA60132	skin	Kennemerduinen	52.42N, 4.58E		KF946748
Locustella naevia naevia	ZMA60133	skin	Kennemerduinen	52.42N, 4.58E		KF946749
Locustella naevia naevia	ZMA64556	skin	Castricum	52.32N, 4.36E		KF946750

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Loxia curvirostra curvirostra	ZMA57246	skin	Eesveen	52.5N, 6.06E		KF946751
Loxia curvirostra curvirostra	ZMA57247	skin	Leersum	52.01N, 5.25E		KF946752
Luscinia megarhynchos megarhynchos	ZMA58798	skin	Amsterdam	52.21N, 4.53E		KF946753
Lymnocryptes minimus	ZMA55930	skin	Heerhugowaard	52.4N, 4.51E		KF946754
Lymnocryptes minimus	ZMA58293	skin	Uitgeest	52.31N, 4.42E		KF946755
Milvus milvus milvus	ZMA58307	wings	Grolloo	52.55N, 6.39E		KF946756
Milvus milvus milvus	ZMA58824	wings	Susteren	51.03N, 5.52E		KF946757
Milvus milvus milvus	ZMA58825	skin	Heurne	51.54N, 6.34E		KF946758
Motacilla alba yarrellii	ZMA58946	skin	Haastrecht	51.59N, 4.46E		KF946759
Motacilla cinerea cinerea	ZMA57241	skin	Westenschouwen	51.41N, 3.42E		KF946760
Motacilla cinerea cinerea	ZMA58266	skin	Westenschouwen	51.41N, 3.42E		KF946761
Motacilla cinerea cinerea	ZMA58267	skin	Westenschouwen	51.41N, 3.42E		KF946762
Motacilla cinerea cinerea	ZMA58945	skin	Westenschouwen	51.41N, 3.42E		KF946763
Muscicapa striata striata	ZMA57336	skin	Ilpendam	52.27N, 4.56E		KF946764
Numenius arquata arquata	Tissue431	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946765
Numenius arquata arquata	ZMA58765	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946766
Numenius arquata arquata	ZMA58829	skin	Heemskerk	52.3N, 4.36E		KF946767
Oenanthe oenanthe leucorhoa	ZMA58868	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946768
Oenanthe oenanthe oenanthe	ZMA58275	skin	Hondsbossche Zeewering	52.44N, 4.38E		KF946769
Oenanthe oenanthe oenanthe	ZMA58800	skin	Noordbergum/ Noardburgum	53.13N, 6E		KF946770
Oriolus oriolus oriolus	ZMA58288	skin	Heteren	51.57N, 5.45E		KF946771
Oriolus oriolus oriolus	ZMA58305	wings	Zundert	51.28N, 4.38E		KF946772
Pandion haliaetus haliaetus	ZMA58823	wing	Vlieland Island	53.15N, 4.59E		KF946773
Panurus biarmicus biarmicus	ZMA57318	skin	Oostvaardersdijk	52.29N, 5.23E		KF946774
Panurus biarmicus biarmicus	ZMA58262	skin	Lelystad	52.29N, 5.24E		KF946775
Panurus biarmicus biarmicus	ZMA58263	skin	Lelystad	52.29N, 5.24E		KF946776
Panurus biarmicus biarmicus	ZMA58264	skin	Lelystad	52.29N, 5.24E		KF946777
Panurus biarmicus biarmicus	ZMA58265	skin	Lelystad	52.29N, 5.24E		KF946778
Panurus biarmicus biarmicus	ZMA58854	skin	Oostvaardersdijk	52.29N, 5.23E		KF946779
Panurus biarmicus biarmicus	ZMA58855	skin	Oostvaardersdijk	52.29N, 5.23E		KF946780
Panurus biarmicus biarmicus	ZMA58856	skin	Oostvaardersdijk	52.29N, 5.23E		KF946781
Parus ater ater	Tissue555	DNA sample	Castricum	52.32N, 4.36E		KF946782
Parus ater ater	ZMA56219	skin	Huizen	52.17N, 5.14E		KF946783
Parus ater ater	ZMA57242	skin	Arnhem	51.58N, 5.53E		KF946784
Parus ater ater	ZMA57243	skin	Amsterdam	52.21N, 4.53E		KF946785
Parus ater ater	ZMA58867	skin	Amsterdam	52.21N, 4.53E		KF946786
Parus ater ater	ZMA64562	skin	Castricum	52.32N, 4.36E		KF946787
Parus caeruleus caeruleus	Tissue438	DNA sample	Castricum	52.32N, 4.36E		KF946788

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Parus caeruleus caeruleus	Tissue439	DNA sample	Castricum	52.32N, 4.36E		KF946789
Parus caeruleus caeruleus	Tissue440	DNA sample	Castricum	52.32N, 4.36E		KF946790
Parus caeruleus caeruleus	ZMA58944	wing	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946791
Parus cristatus mitratus	ZMA56677	skin	Nijverdal	52.22N, 6.28E		KF946792
Parus cristatus mitratus	ZMA57245	skin	Hoog Buurlo	52.1N, 5.5E		KF946793
Parus major major	ZMA58796	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946794
Parus major major	ZMA58797	skin	Castricum	52.32N, 4.36E		KF946795
Parus palustris palustris	ZMA57244	skin	Castricum	52.32N, 4.36E		KF946796
Parus palustris palustris	ZMA64561	skin	Goor	52.14N, 6.34E		KF946797
Passer domesticus domesticus	ZMA58799	skin	Cadier en Keer	50.49N, 5.46E		KF946798
Passer domesticus domesticus	ZMA60138	skin	Lekkerkerk	51.53N, 4.41E		KF946799
Passer montanus montanus	ZMA58851	skin	Zuidhorn	53.14N, 6.23E		KF946800
Passer montanus montanus	ZMA58852	skin	Zuidhorn	53.14N, 6.23E		KF946801
Passer montanus montanus	ZMA58853	skin	Zuidhorn	53.14N, 6.23E		KF946802
Passer montanus montanus	ZMA58950	skin	Zuidhorn	53.14N, 6.23E		KF946803
Perdix perdix perdix	ZMA58738	skin	Texel Island	53.04N, 4.43E		KF946804
Perdix perdix perdix	ZMA58739	skin	Petten	52.46N, 4.38E		KF946805
Pernis apivorus	ZMA58827	wings	Vledder	52.53N, 6.13E		KF946806
Phalacrocorax aristotelis aristotelis	ZMA58224	skin	Wijk aan Zee	52.28N, 4.34E		KF946807
Philomachus pugnax	ZMA56680	skin	Graftermeer polder	52.33N, 4.48E		KF946808
Philomachus pugnax	ZMA58250	skin	Lelystad	52.29N, 5.24E		KF946809
Phoenicopterus chilensis	ZMA56683	skin	Ransdorp	52.23N, 4.59E		KF946810
Phoenicurus phoenicurus phoenicurus	ZMA55914	skin	Westenschouwen	51.41N, 3.42E		KF946811
Phylloscopus collybita collybita	ZMA55917	skin	Nijverdal	52.22N, 6.28E		KF946812
Phylloscopus collybita collybita	ZMA55918	wings	Leveroy	51.14N, 5.5E		KF946813
Phylloscopus collybita collybita	ZMA56217	skin	Hoogland	52.1N, 5.21E		KF946814
Phylloscopus trochilus	ZMA58284	skin	Lelystad	52.29N, 5.24E		KF946815
Phylloscopus trochilus	ZMA58710	skin	Almere	52.22N, 5.13E		KF946816
Phylloscopus trochilus	ZMA58713	skin	Egmond aan Zee	52.37N, 4.38E		KF946817
Phylloscopus trochilus	ZMA58714	skin	Lekkerkerk	51.53N, 4.41E		KF946818
Phylloscopus trochilus	ZMA58715	skin	Texel Island	53.04N, 4.43E		KF946819
Phylloscopus trochilus	ZMA58716	skin	Castricum	52.32N, 4.36E		KF946820
Phylloscopus trochilus	ZMA58861	skin	Castricum	52.32N, 4.36E		KF946821
Phylloscopus trochilus	ZMA58933	wings	Goor	52.14N, 6.34E		KF946822
Phylloscopus trochilus	ZMA58934	skin	Eemshaven	53.26N, 6.52E		KF946823
Picus viridis viridis	ZMA58718	skin	Breda	51.33N, 4.46E		KF946824
Picus viridis viridis	ZMA58719	skin	Haaksbergen	52.08N, 6.4E		KF946825
Picus viridis viridis	ZMA58720	skin	Alkmaar	52.38N, 4.44E		KF946826
Picus viridis viridis	ZMA58721	skin	Roggel	51.17N, 5.54E		KF946827
Picus viridis viridis	ZMA58722	skin	Bergen	52.4N, 4.41E		KF946828
Plectrophenax nivalis insulae	ZMA56672	skin	Castricum	52.32N, 4.36E		KF946829

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Pluvialis apricaria	ZMA58213	skin	Winsum	53.09N, 5.38E		KF946830
Pluvialis apricaria	ZMA58214	skin	Winsum	53.09N, 5.38E		KF946831
Pluvialis apricaria	ZMA58215	skin	Dronrijp/Dronryp	53.11N, 5.4E		KF946832
Pluvialis squatarola squatarola	ZMA56224	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946833
Pluvialis squatarola squatarola	ZMA56225	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946834
Puffinus gravis	ZMA64542	skin	Sexbierum/ Seisbierrum	53.14N, 5.28E		KF946835
Pyrrhula pyrrhula europoea	ZMA56673	skin	Castricum	52.32N, 4.36E		KF946836
Pyrrhula pyrrhula europoea	ZMA58793	skin	Castricum	52.32N, 4.36E		KF946837
Pyrrhula pyrrhula europoea	ZMA58794	skin	Castricum	52.32N, 4.36E		KF946838
Pyrrhula pyrrhula europoea	ZMA58795	skin	Castricum	52.32N, 4.36E		KF946839
Pyrrhula pyrrhula europoea	ZMA60137	wings	Kennemerduinen	52.42N, 4.58E		KF946840
Rallus aquaticus aquaticus	ZMA58763	skin	Lauwersmeer	53.22N, 6.14E		KF946841
Recurvirostra avosetta	ZMA58216	skin	Petten	52.46N, 4.38E		KF946842
Regulus ignicapilla ignicapilla	Tissue448	DNA sample	Castricum	52.32N, 4.36E		KF946843
Regulus ignicapilla ignicapilla	ZMA57360	skin	Zundert	51.28N, 4.38E		KF946844
Regulus ignicapilla ignicapilla	ZMA58807	skin	Castricum	52.32N, 4.36E		KF946845
Regulus ignicapilla ignicapilla	ZMA58808	skin	Castricum	52.32N, 4.36E		KF946846
Regulus regulus regulus	ZMA64560	skin	Castricum	52.32N, 4.36E		KF946847
Riparia riparia riparia	ZMA58871	skin	Zeewolde	52.21N, 5.34E		KF946848
Saxicola rubetra	ZMA60131	skin	Kennemerduinen	52.42N, 4.58E		KF946849
Saxicola rubetra	ZMA64555	skin	Castricum	52.32N, 4.36E		KF946850
Somateria mollissima mollissima	ZMA58912	skin	Lauwersoog	53.24N, 6.12E		KF946851
Stercorarius longicaudus	ZMA58779	wings	Afsluitdijk	52.57N, 5.04E		KF946852
Stercorarius longicaudus	ZMA64546	skin	Petten	52.46N, 4.38E		KF946853
Stercorarius parasiticus	ZMA56229	skin	Vlieland Island	53.15N, 4.59E		KF946854
Stercorarius parasiticus	ZMA56684	wings	Terschelling Island	53.26N, 5.29E		KF946855
Stercorarius parasiticus	ZMA58778	skin	Den Oever	52.56N, 5.02E		KF946856
Stercorarius parasiticus	ZMA58830	skin	Den Helder	52.55N, 4.46E		KF946857
Stercorarius pomarinus	Tissue211	DNA sample	Texel Island	53.04N, 4.43E		KF946858
Stercorarius pomarinus	ZMA55929	skin	Hondsbossche Zeewering	52.44N, 4.38E		KF946859
Stercorarius skua skua	ZMA64545	skin	Egmond aan Zee	52.37N, 4.38E		KF946860
Sterna albifrons albifrons	ZMA58832	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946861
Sterna hirundo hirundo	ZMA58915	skin	Eemshaven	53.26N, 6.52E		KF946862
Sterna paradisaea	ZMA58831	skin	Amsterdam	52.21N, 4.53E		KF946863
Streptopelia decaocto decaocto	ZMA58923	wing	Hoogkerk	53.12N, 6.3E		KF946864
Streptopelia turtur turtur	ZMA58757	skin	Texel Island	53.04N, 4.43E		KF946865
Sylvia atricapilla atricapilla	Tissue441	DNA sample	Castricum	52.32N, 4.36E		KF946866
Sylvia atricapilla atricapilla	Tissue442	DNA sample	Castricum	52.32N, 4.36E		KF946867
Sylvia atricapilla atricapilla	ZMA58268	skin	Bloemendaal	52.24N, 4.33E		KF946868

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Sylvia atricapilla atricapilla	ZMA58269	skin	Bloemendaal	52.24N, 4.33E		KF946869
Sylvia atricapilla atricapilla	ZMA58270	skin	Bloemendaal	52.24N, 4.33E		KF946870
Sylvia atricapilla atricapilla	ZMA58759	skin	Cadier en Keer	50.49N, 5.46E		KF946871
Sylvia borin borin	Tissue443	DNA sample	Castricum	52.32N, 4.36E		KF946872
Sylvia borin borin	ZMA58758	skin	Groningen	53.14N, 6.35E		KF946873
Sylvia borin borin	ZMA58761	skin	Almere	52.22N, 5.13E		KF946874
Sylvia borin borin	ZMA58762	skin	Purmerend	52.28N, 4.58E		KF946875
Sylvia communis communis	ZMA55924	wing	Asten	51.21N, 5.48E		KF946876
Sylvia communis communis	ZMA57335	skin	Almere	52.22N, 5.13E		KF946877
Sylvia communis communis	ZMA58280	skin	Breda	51.33N, 4.46E		KF946878
Sylvia communis communis	ZMA58939	skin	Castricum	52.32N, 4.36E		KF946879
Sylvia communis communis	ZMA58940	skin	Bloemendaal	52.24N, 4.33E		KF946880
Sylvia curruca blythi	ZMA58941	skin	Houten	52.01N, 5.1E		KF946881
Sylvia curruca blythi	ZMA57237	skin	Rotterdam	51.57N, 4.32E		KF946882
Sylvia curruca curruca	ZMA55905	skin	Westenschouwen	51.41N, 3.42E		KF946883
Sylvia curruca curruca	ZMA55906	skin	Amsterdam	52.21N, 4.53E		KF946884
Sylvia curruca curruca	ZMA57328	skin	Almere	52.22N, 5.13E		KF946885
Sylvia curruca curruca	ZMA57329	skin	Texel Island	53.04N, 4.43E		KF946886
Sylvia curruca curruca	ZMA58282	skin	Zeewolde	52.21N, 5.34E		KF946887
Sylvia curruca curruca	ZMA58806	skin	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946888
Sylvia curruca curruca	ZMA58864	skin	Eemshaven	53.26N, 6.52E		KF946889
Sylvia curruca curruca	ZMA58942	skin	Bloemendaal	52.24N, 4.33E		KF946890
Sylvia nisoria nisoria	ZMA58273	skin	Westenschouwen	51.41N, 3.42E		KF946891
Tringa ochropus	ZMA64544	skin	Castricum	52.32N, 4.36E		KF946892
Tringa totanus totanus	ZMA58212	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946893
Troglodytes troglodytes troglodytes	Tissue447	DNA sample	Castricum	52.32N, 4.36E		KF946894
Troglodytes troglodytes troglodytes	ZMA58281	skin	Bloemendaal	52.24N, 4.33E		KF946895
Turdus iliacus iliacus	ZMA58287	skin	Bloemendaal	52.24N, 4.33E		KF946896
Turdus merula merula	ZMA56669	skin	Haarlem	52.23N, 4.37E		KF946897
Turdus merula merula	ZMA56670	skin	Bergen	52.4N, 4.41E		KF946898
Turdus merula merula	ZMA57345	skin	Zwolle	52.3N, 6.06E		KF946899
Turdus merula merula	ZMA58731	skin	Alkmaar	52.38N, 4.44E		KF946900
Turdus merula merula	ZMA58732	skin	Maasbree	51.21N, 6.03E		KF946901
Turdus merula merula	ZMA58733	skin	Maasbree	51.21N, 6.03E		KF946902
Turdus merula merula	ZMA58734	skin	Steijl	51.2N, 6.07E		KF946903
Turdus merula merula	ZMA58736	skin	Schiermonnikoog Island	53.29N, 6.11E		KF946904
Turdus philomelos philomelos	Tissue453	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946905
Turdus philomelos philomelos	Tissue454	DNA sample	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946906
Turdus torquatus torquatus	ZMA56222	skin	Texel Island	53.04N, 4.43E		KF946907

Species or subspecies	ZMA number	Preparation	Locality	Coordinates		Access numbers
				N	E	
Turdus torquatus torquatus	ZMA56671	skin	Castricum	52.32N, 4.36E		KF946908
Turdus torquatus torquatus	ZMA58693	skin	Apeldoorn	52.1N, 5.58E		KF946909
Turdus torquatus torquatus	ZMA58694	skin	Vlieland Island	53.15N, 4.59E		KF946910
Turdus torquatus torquatus	ZMA58695	skin	Zuilichem	51.48N, 5.07E		KF946911
Turdus torquatus torquatus	ZMA64554	skin	Texel Island	53.04N, 4.43E		KF946912
Turdus viscivorus viscivorus	ZMA60130	skin	Kennemerduinen	52.42N, 4.58E		KF946913
Tyto alba alba	ZMA56233	skin	Burgerbrug	52.45N, 4.42E		KF946914
Tyto alba guttata	ZMA56682	skin	Wierden	52.22N, 6.34E		KF946915
Tyto alba guttata	ZMA58235	skin	Texel Island	53.04N, 4.43E		KF946916
Tyto alba guttata	ZMA58236	skin	Ouderkerk aan de Amstel	52.17N, 4.56E		KF946917
Tyto alba guttata	ZMA58843	skin	Westzaan	52.26N, 4.46E		KF946918
Tyto alba guttata	ZMA58844	skin	Zaanstreek	52.28N, 4.44E		KF946919
Tyto alba guttata	ZMA58845	skin	Roodkerk/ Readtsjerk	53.15N, 5.55E		KF946920
Tyto alba guttata	ZMA58846	skin	Garijp/Garyp	53.1N, 5.57E		KF946921
Tyto alba guttata	ZMA58847	skin	Middenmeer	52.48N, 4.59E		KF946922
Tyto alba guttata	ZMA58848	wings	Leeuwarden/ Ljouwert	53.13N, 5.45E		KF946923
Tyto alba guttata	ZMA58919	skin	Texel Island	53.04N, 4.43E		KF946924
Tyto alba guttata	ZMA64550	skin	Purmerend	52.28N, 4.58E		KF946925
Tyto alba guttata	ZMA64551	skin	Goor	52.14N, 6.34E		KF946926
Uria aalge albionis	ZMA56227	skin	Amsterdam	52.21N, 4.53E		KF946927
Uria aalge albionis	ZMA58218	skin	Vlieland Island	53.15N, 4.59E		KF946928
Uria aalge albionis	ZMA58916	skin	Petten	52.46N, 4.38E		KF946929
Vanellus vanellus	ZMA58784	wing	Valkenburg	52.09N, 4.25E		KF946930
Vanellus vanellus	ZMA58785	wing	Valkenburg	52.09N, 4.25E		KF946931
Vanellus vanellus	ZMA58786	wing	Valkenburg	52.09N, 4.25E		KF946932
Vanellus vanellus	ZMA58787	wing	Valkenburg	52.09N, 4.25E		KF946933
Vanellus vanellus	ZMA58788	wing	Valkenburg	52.09N, 4.25E		KF946934
Vanellus vanellus	ZMA58789	wing	Valkenburg	52.09N, 4.25E		KF946935
Vanellus vanellus	ZMA58790	wing	Valkenburg	52.09N, 4.25E		KF946936
Vanellus vanellus	ZMA58791	wing	Valkenburg	52.09N, 4.25E		KF946937

Supplementary Table 2. Bird species (gulls *Larus* and skuas *Stercorarius*) from the Netherlands with low (< 1.1%) K2P mean intraspecific distances.

Collection number and species		Collection number and species		Distance (%)
#ZMA58835	*Larus michahellis*	#Tissue327	*L. fuscus*	0
#ZMA58835	*Larus michahellis*	#Tissue432	*L. fuscus*	0
#ZMA58835	*Larus michahellis*	#ZMA55932	*L. fuscus*	0
#ZMA58835	*Larus michahellis*	#ZMA56230	*L. fuscus*	0
#ZMA64547	*Larus cachinnans*	#Tissue327	*L. fuscus*	0
#ZMA64547	*Larus cachinnans*	#Tissue432	*L. fuscus*	0
#ZMA64547	*Larus cachinnans*	#ZMA55932	*L. fuscus*	0
#ZMA64547	*Larus cachinnans*	#ZMA56230	*L. fuscus*	0
#ZMA64547	*Larus cachinnans*	#ZMA58835	*L. michahellis*	0
#ZMA58921	*Larus argentatus*	#ZMA55932	*L. fuscus*	0.14
#ZMA58921	*Larus argentatus*	#ZMA58835	*L. michahellis*	0.14
#ZMA58921	*Larus argentatus*	#Tissue432	*L. fuscus*	0.15
#ZMA58921	*Larus argentatus*	#ZMA56230	*L. fuscus*	0.15
#ZMA64547	*Larus cachinnans*	#ZMA58834	*L. fuscus*	0.15
#ZMA64547	*Larus cachinnans*	#ZMA58921	*L. argentatus*	0.15
#ZMA58921	*Larus argentatus*	#Tissue327	*L. fuscus*	0.16
#ZMA55932	*Larus fuscus*	#Tissue433	*L. argentatus*	0.29
#ZMA58835	*Larus michahellis*	#Tissue433	*L. argentatus*	0.29
#Tissue433	*Larus argentatus*	#Tissue432	*L. fuscus*	0.30
#ZMA56230	*Larus fusca*	#Tissue433	*L. argentatus*	0.30
#ZMA64545	*Stercorarius skua*	#ZMA55929	*S. pomarinus*	0.30
#ZMA58836	*Larus glaucoides*	#Tissue432	*L. fuscus*	0.31
#ZMA58836	*Larus glaucoides*	#ZMA55932	*L. fuscus*	0.31
#ZMA58836	*Larus glaucoides*	#ZMA56230	*L. fuscus*	0.31
#ZMA58836	*Larus glaucoides*	#ZMA58835	*L. michahellis*	0.31
#ZMA64547	*Larus cachinnans*	#Tissue433	*L. argentatus*	0.31
#ZMA64547	*Larus cachinnans*	#ZMA58836	*L. glaucoides*	0.31
#Tissue433	*Larus argentatus*	#Tissue327	*L. fuscus*	0.32
#ZMA58836	*Larus glaucoides*	#Tissue327	*L. fuscus*	0.32
#ZMA64545	*Stercorarius skua*	#Tissue211	*S. pomarinus*	0.43
#ZMA58835	*Larus michahellis*	#ZMA58834	*L. fuscus*	0.45
#ZMA58836	*Larus glaucoides*	#ZMA58834	*L. fuscus*	0.46
#ZMA58921	*Larus argentatus*	#ZMA58836	*L. glaucoides*	0.46
#ZMA56221	*Larus hyperboreus*	#ZMA55932	*L. fuscus*	0.58
#ZMA58835	*Larus michahellis*	#ZMA56221	*L. hyperboreus*	0.58
#ZMA56221	*Larus hyperboreus*	#Tissue432	*L. fuscus*	0.60
#ZMA56230	*Larus fuscus*	#ZMA56221	*L. hyperboreus*	0.60
#ZMA58921	*Larus argentatus*	#ZMA58834	*L. fuscus*	0.60
#ZMA64547	*Larus cachinnans*	#ZMA56221	*L. hyperboreus*	0.61
#ZMA58836	*Larus glaucoides*	#Tissue433	*L. argentatus*	0.62
#ZMA56221	*Larus hyperboreus*	#Tissue327	*L. fuscus*	0.64
#ZMA58921	*Larus argentatus*	#ZMA56221	*L. hyperboreus*	0.73
#ZMA58834	*Larus fuscus*	#Tissue433	*L. argentatus*	0.75
#ZMA56221	*Larus hyperboreus*	#Tissue433	*L. argentatus*	0.87
#ZMA58836	*Larus glaucoides*	#ZMA56221	*L. hyperboreus*	0.93
#ZMA58834	*Larus fuscus*	#ZMA56221	*L. hyperboreus*	1.06

ZooKeys 365: 49–65 (2013)
doi: 10.3897/zookeys.365.6409
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Applications of DNA barcoding to fish landings: authentication and diversity assessment

Alba Ardura[1], Serge Planes[2,3], Eva Garcia-Vazquez[1]

1 *University of Oviedo, Department of Functional Biology. C/ Julian Claveria s/n. 33006-Oviedo, Spain* **2** *USR 3278 CNRS – EPHE. Centre de Recherche Insulaire et Observatoire de l'Environnement (CRIOBE) BP 1013 - 98 729, Papetoai, Moorea, Polynésie française* **3** *Centre de Biologie et d'Ecologie Tropicale et Méditerranéenne, Université de Perpignan, 52 Av. Paul Alduy - 66860 Perpignan cedex, France*

Corresponding author: *Alba Ardura* (alarguti@hotmail.com)

Academic editor: *T. Backeljau* | Received 7 October 2013 | Accepted 23 October 2013 | Published 30 December 2013

Citation: Ardura A, Planes S, Garcia-Vazquez E (2013) Applications of DNA barcoding to fish landings: authentication and diversity assessmente. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 49–65. doi: 10.3897/zookeys.365.6409

Abstract

DNA barcoding methodologies are being increasingly applied not only for scientific purposes but also for diverse real-life uses. Fisheries assessment is a potential niche for DNA barcoding, which serves for species authentication and may also be used for estimating within-population genetic diversity of exploited fish. Analysis of single-sequence barcodes has been proposed as a shortcut for measuring diversity in addition to the original purpose of species identification. Here we explore the relative utility of different mitochondrial sequences (12S rDNA, COI, cyt *b*, and D-Loop) for application as barcodes in fisheries sciences, using as case studies two marine and two freshwater catches of contrasting diversity levels. Ambiguous catch identification from COI and cyt *b* was observed. In some cases this could be attributed to duplicated names in databases, but in others it could be due to mitochondrial introgression between closely related species that may obscure species assignation from mtDNA. This last problem could be solved using a combination of mitochondrial and nuclear genes. We suggest to simultaneously analyze one conserved and one more polymorphic gene to identify species and assess diversity in fish catches.

Keywords

Species identification, freshwater fisheries, marine fisheries, genetic diversity, mitochondrial DNA markers

Introduction

DNA barcoding is increasingly important in natural sciences. For ecologists it is a tool with many utilities (e.g. Valentini et al. 2009), most of which are related with biodiversity inventories. Fisheries are a field of enormous potential interest for barcoding applications. The use of genetics is increasingly required in fisheries for species authentication in fish landings (Rasmussen and Morrisey 2008, Ardura et al. 2010a). Fisheries are unsustainable if catch records are based on erroneous or inaccurate species identifications (Watson and Pauly 2001, Marko et al. 2004, Crego et al. 2012). Moreover, guaranteeing species authenticity along the commercial chain would improve consumer's security and prevent fraud, which has been proven to occur worldwide (e.g. DeSalle and Birstein 1996, Marko et al. 2004, Jacquet and Pauly 2008, Wong and Hanner 2008, Ardura et al. 2010b, Ardura et al. 2010c, Barbuto et al. 2010, Filonzi et al. 2010, Miller and Mariani 2010, Garcia-Vazquez et al. 2011). On the other hand, declines in population genetic variation diminish the ability of a population to adapt to environmental changes and decrease its chance of long-term survival (Frankham 1995, Hedrick 2001, Wang et al. 2002); thus periodical monitoring of population variation of exploited stocks is highly recommended in fisheries management.

Despite the potential importance of genetics in fisheries, the application of DNA analyses in real cases is not so easy. The economic aspect is crucial: increasing costs are making fisheries not only ecologically, but also economically unsustainable (e.g. Willmann and Kelleher 2010). The practical use of genome-wide studies in everyday management does not seem to be realistic in a near future because massive DNA analysis of catches would increase even more the costs of fish products. If the genetic tool (marker) employed for species authentication exhibits enough variation for reliable quantification of population diversity, a single analysis could solve two problems at the same time. Another practical problem for applying genetics to fisheries is the time required for DNA analysis. Catches can not be immobilized for a long time without increasing storage costs for guaranteeing the cold chain. The accelerated development of high throughput sequencing methodologies (e.g. Steemers and Gunderson 2005, Sundquist et al. 2007) can help in this issue because now it is possible to analyze thousands of samples very fast. Genomics at population level is being carried out for a few targeted marine species (Nielsen et al. 2009); the moment of applying large scale routine genetic analysis in fisheries science, including all species, seems thus to be approaching.

The potential taxonomic diversity of fish catches is enormous, since in biodiversity hotspots unknown species are landed (Worm and Branch 2012). This makes it difficult to analyze introns and SNP of the nuclear genome, whose development requires a good knowledge of each species' genome for developing primers in flanking regions. However, using universal primers is much easier. Demographic changes in fish populations can be associated with the observed amount of variation in mitochondrial DNA (e.g. Fauvelot et al. 2003, Nevado et al. 2013), and genetic erosion due to population depletion could be theoretically detected from variable mitochondrial regions. The international barcoding initiative (Hebert et al. 2003, Janzen et al. 2005) has converged

with next-generation sequencing, and ecosystem biodiversity can be better estimated through DNA information now (Hajibabei 2012). The main DNA barcode has been chosen by some authors as an initial tool for calibrating fish species diversity due to the large number of cytochrome *c* oxidase I gene (COI) sequences included in the *Barcode of Life Data Systems* (BOLD) database (April et al. 2011, Ardura et al. 2011). However, it may not be sufficient to rigorously address intraspecific variation at population level (Moritz and Cicero 2004, Rubinoff 2006). The informative value of other DNA regions with different degrees of polymorphism should therefore be evaluated. The highly conserved mitochondrial 12S rDNA has been applied for analyzing diversity in high categorical levels such as phyla (Gerber et al. 2001). In decreasing order of conservation, the protein-coding cytochrome *b* (cyt *b*) has been extensively used for diversity analysis at genera and species level (Min et al. 2004, Zhang and Jiang 2006). Finally, the D-Loop or mitochondrial control region exhibits more variation than protein-coding sequences due to reduced functional constraints and relaxed selection pressure (Onuma et al. 2006, Wu et al. 2006). Therefore, D-Loop variation would roughly inform about intraspecific diversity, whereas more conserved sequences would better reflect biodiversity (number and genetic proximity of species in a catch).

The objective of this study was to assess the utility of well-known public databases for identifying catches from very different fisheries, comparing genes and species for determining if there is sufficient information available for routine genetic analysis of fish catches that informs about species composition. The main areas where generating new data are necessary, if any, will be identified from the shortcomings detected in this small-scale exercise. We have employed standard primer sets for PCR amplification of four mtDNA gene fragments, then estimated standard parameters of genetic diversity and evaluated their utility for identifying landings using GenBank and BOLD. We have also estimated intrapopulation diversity in order to assess possible applications of these markers for monitoring demographic changes. Our case studies were two marine and two freshwater catches of contrasting diversity for the standard COI DNA barcode (Ardura et al. 2011).

Materials and methods

Case studies

Mediterranean Sea. It is a marine biodiversity hotspot with 713 fish species inventoried (FishBase; www.fishbase.org). Samples were obtained from fish markets in the Languedoc-Roussillon region (Gulf of Lion, France), in the north-western Mediterranean coast.

Cantabric Sea. Less diverse than the Mediterranean Sea, it contains 148 fish species inventoried. Catch from commercial fisheries was sampled from fish markets in Asturias (North of Spain).

Amazon River. It is the main freshwater biodiversity hotspot of the world (1218 inventoried fish species). We have sampled catches obtained in different fish markets

of Manaus (Brazil). This is the area where the two main Amazonian drainages (the rivers Negro and Solimões) join.

Narcea River (North of Spain). As other North Iberian rivers, it exhibits reduced biodiversity with only 17 fish species inventoried. Fisheries are strongly targeted and focused on sport angling of salmonids. Samples were obtained *in situ* from fishermen in the lower reach of the river.

The two most exploited species (those that yield more tonnes in official catch statistics) from each site were chosen for this study. They were: mackerel *Scomber scombrus* (Goode, 1884) and anchovy *Engraulis encrasicolus* (Linnaeus, 1758) from the Mediterranean Sea; mackerel and albacore tuna *Thunnus alalunga* (Bonnaterre, 1778) from the Cantabric Sea; Curimatá *Prochilodus nigricans* (Spix & Agassiz, 1829) and jaraquí *Semaprochilodus insignis* (Jardine & Schomburgk, 1841) from the Amazon River; Atlantic salmon *Salmo salar* (Linnaeus, 1758) and brown trout *S. trutta* (Linnaeus, 1758) from the Narcea River. These species do not exhibit population sub-division in the fishing areas considered. The West Mediterranean and the Eastern Atlantic Ocean populations of mackerel seem to form a panmictic unit (Zardoya et al. 2004). The highly migratory albacore tuna exhibits only inter-oceanic population differentiation or between the Atlantic and the Mediterranean, not within the same ocean (Chow and Ushiama 1995, Viñas et al. 2004). For anchovy, the whole north-western Mediterranean likely harbors a single population (Tudela et al. 1999). Curimatá and jaraquí, the main catch in the Brazilian Amazon state, have a shallow genetic structuring in the Amazon basin and can be considered homogeneous populations around Manaus (Ardura et al. 2013). Finally, Atlantic salmon and brown trout populations are not subdivided within rivers in North Spain unless there is strong habitat fragmentation (e.g. Horreo et al. 2011a, b), yet this is not the case for the lower accessible zone of River Narcea.

Ten samples were analyzed per species.

mtDNA analysis

DNA extraction was automatized with QIAxtractor robot following the manufacturer's protocol (QIAGEN DX Universal DNA Extraction Tissue Sample CorProtocol), which yields high quality DNA suitable for a wide variety of downstream applications. The procedure is divided into two sections: digestion and extraction. The digestion process favors tissue dissociation and liquid suspension, and is ready for extraction.

Briefly, a 96 well round well lysis block (Sample Block) is loaded with 420 µl DX Tissue Digest (containing 1% v/v DX Digest Enzyme) manually or using the Tissue Digest Preload run file. Once the DX Tissue Digest is loaded with the sample, the sample block is sealed and incubated at 55 °C with agitation for at least 3 h. 220 µl of supernatant is transferred from the sample block in position C1 to the lysis plate in position B1. 440 µl of DX Binding with DX Binding Additive is added to the lysis plate. The lysate is then mixed 8 × and incubated at room temperature for 5 min. 600 µl of the lysate is added into the capture plate (Pre-mixed 8 ×). A vacuum of 35 kPa is applied for 5 min. 200 µl of DX Binding with DX Binding Additive is loaded into the capture

plate. A vacuum of 35 kPa is applied for 5 min. 600 µl of DX Wash is loaded into the capture plate. A vacuum of 25 kPa applied for 1 min, repeated (2 iterations). 600 µl of DX Final Wash is loaded into the capture plate. A vacuum of 35 kPa is applied for 1 min. A vacuum of 25 kPa is applied for 5 min to dry the plate. The carriage is moved to elution chamber. 200 µl of Elution buffer is loaded into the capture plate. The sample is then incubated for 5 min. A vacuum of 35 kPa is applied for 1 min.

We employed the QIAxtractor Software application. The tube was frozen at -20 °C for long-time preservation.

Fragments of four different mitochondrial genes were amplified by polymerase chain reaction (PCR): 12S rDNA, COI, cyt *b* and D-Loop (Table 1). We employed primers commonly used for fish published by Palumbi (1996), Ward et al. (2005), Kocher et al. (1989) and Lee et al. (1995) respectively. Amplification reactions were performed in a total volume of 23 µl, including 5 PRIME Buffer 1 × (Gaithersburg, MD, USA), 1.5 mM $MgCl_2$, 0.25 mM dNTPs, 1 µM of each primer, 20 ng of template DNA, and 1.5U of DNA Taq polymerase (5 PRIME).

The PCR conditions were the following:

12S rDNA: an initial denaturation at 95 °C for 10 min, then 35 cycles of denaturation at 94 °C for 1 min, annealing at 57 °C for 1 min and extension at 72 °C for 1.5 min, followed by a final extension at 72 °C for 7 min.

COI: an initial denaturation at 94 °C for 5 min, then 10 cycles of denaturation at 94 °C for 1 min, annealing at 64–54 °C for 1 min and extension at 72 °C for 1.5 min, followed by 25 cycles of denaturation at 94 °C for 1 min, annealing at 54 °C for 1 min and extension at 72 °C for 1.5 min, finally a final extension at 72 °C for 5 min.

cyt *b*: an initial denaturation at 94 °C for 5 min, then 10 cycles of denaturation at 94 °C for 1 min, annealing at 60–50 °C for 1 min and extension at 72 °C for 1.5 min, followed by 25 cycles of denaturation at 94 °C for 1 min, annealing at 54 °C for 1 min and extension at 72 °C for 1.5 min, finally a final extension at 72 °C for 5 min.

D-Loop: an initial denaturation at 94 °C for 5 min, then 10 cycles of denaturation at 94 °C for 1 min, annealing at 57 °C for 1 min and extension at 72 °C for 1.5 min, followed by 25 cycles of denaturation at 94 °C for 1 min, annealing at 54 °C for 1 min and extension at 72 °C for 1.5 min, finally a final extension at 72 °C for 5 min.

Sequencing was carried out by the DNA sequencing service GATC Biotech (Germany).

Sequence edition

Sequences were visualized and edited employing the BioEdit Sequence Alignment Editor software (Hall 1999). Sequences were aligned with the MEGA v4.0 software (Tamura et al. 2007).

Putative proteins (amino acid sequences) from the COI and cyt *b* sequences were inferred with the software MEGA v4.0 (Tamura et al. 2007).

Table 1. Species considered within each case study; common and specific names and classification. Numbers of nucleotides obtained for each mtDNA gene fragment (length in bp) and GenBank Accession Numbers.

REGION	SPECIES		CLASSIFICATION (Order, Family)	Mitochondrial regions (length in bp)	GenBank A.N.
	Common name	Scientific name			
Amazon River	curimata	*Prochilodus nigricans*	Characiformes, Curimatidae	12S rDNA (380)	JN007487–JN007496
				COI (605)	JN007727–JN007734 HM480806–HM480807
				cyt *b* (293)	JN007647–JN007656
				D–Loop (424)	JN007567–JN007576
	jaraquí	*Semaprochilodus insignis*	Characiformes, Curimatidae	12S rDNA (380)	JN007497–JN007506
				COI (605)	JN007735–JN007744
				cyt *b* (293)	JN007657–JN007666
				D–Loop (424)	JN007577–JN007586
Cantabric Sea	mackerel tuna	*Scomber scombrus*	Perciformes, Scombridae	12S rDNA (382)	JN007507–JN007516
				COI (605)	JN007745–JN007751 HM480797 HM480799 HM480819
				cyt *b* (293)	JN007667–JN007676
				D–Loop (412)	JN007587–JN007596
		Thunnus alalunga	Perciformes, Scombridae	12S rDNA (382)	JN007517–JN007526
				COI (605)	JN007752–JN007761
				cyt *b* (293)	JN007677–JN007687
				D–Loop (412)	JN007597–JN007606
Mediterranean Sea	anchovy	*Engraulis encrasicolus*	Clupeiformes, Engraulidae	12S rDNA (384)	JN007527–JN007536
				COI (605)	JN007762–JN007768 HM480814–HM480816
				cyt *b* (293)	JN007687–JN007696
				D–Loop (462)	JN007607–JN007616
	mackerel	*Scomber scombrus*	Perciformes, Scombridae	12S rDNA (384)	JN007537–JN007546
				COI (605)	JN007769–JN007777 HM480797
				cyt *b* (293)	JN007697–JN007706
				D–Loop (462)	JN007617–JN007626
Narcea River	Atlantic salmon	*Salmo salar*	Salmoniformes, Salmonidae	12S rDNA (439)	JN007547–JN007556
				COI (635)	JN007778–JN007787
				cyt *b* (322)	JN007707–JN007716
				D–Loop (460)	JN007627–JN007636
	brown trout	*Salmo trutta*	Salmoniformes, Salmonidae	12S rDNA (439)	JN007557–JN007566
				COI (635pb)	JN007788–JN007797
				cyt *b* (322)	JN007717–JN007726
				D–Loop (460)	JN007637–JN007646

Species identification from DNA sequences

The sequences obtained were compared with those existing in the public database GenBank using the BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch). Species were

identified based on maximum BLAST scores with matching sequences, corresponding to 100% coverage and 100% identity. When the haplotype was new (i.e. not present in GenBank and BOLD), a 100% coverage with 99% identity, or in a few cases 98% identity, was found for the matching sequence. COI barcodes were also compared against the BOLD database, uploading them in the BOLD identification system in FASTA format at http://www.boldsystems.org/index.php/IDS_OpenIdEngine. The system retrieves matching sequences with the corresponding % similarity (matching nucleotides) and gives the most likely species for the query sequence. If matching sequences from more than one species are retrieved with a similar probability, then the system displays all the possible putative species the query can be assigned to.

The two databases were accessed for species identification in September 2013.

Diversity indices

Three well-known diversity indices were employed: number of haplotypes, haplotype diversity and nucleotide diversity. They were calculated with the DnaSP software (Librado and Rozas 2009). The same program was employed to generate concatenated data files with the different markers analyzed and re-estimate genetic diversity parameters.

Haplotype diversity is a measure of population variation, as the probability of two randomly chosen haplotypes in the sample being different. It is calculated with the formula described by Nei and Tajima (1981).

Nucleotide diversity indicates how different sequences are to each other. Its value is higher when sequences belong to distant taxa. It is defined as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population, and is symbolised as π (Nei and Li 1979).

We have also used the simplest diversity measure Nh/n (number of haplotypes divided by the number of samples analysed).

Statistical analysis

Comparison between genes for their polymorphic content was made based on means and variances of diversity parameters. It was performed using the software SPSS 13.0 software (SPSS Inc., Chicago, IL, USA).

Results

Species identification of the considered samples

For three study areas, the two most harvested species belonged to the same family (Table 1), viz. Curimatidae, Salmonidae and Scombridae in the Amazon River, Narcea

Table 2. Species identification based on the assayed genes in the four considered catches, measured as the number of individuals that are unambiguously assigned to a species in GenBank (all genes) and BOLD (COI). Databases accessed in September 2013.

	COI		12S rDNA	cyt *b*	D-Loop
	GenBank	**BOLD**	**GenBank**	**GenBank**	**GenBank**
Cantabric Sea					
mackerel	10	10	10	10	10
tuna	5	0	10	0	6
% catch	75%	50%	100%	50%	80%
Mediterranean Sea					
anchovy	10	0	10	10	10
mackerel	10	10	10	10	10
% catch	100%	50%	100%	100%	100%
Narcea River					
Atlantic salmon	10	10	10	10	10
brown trout	10	0	10	10	10
% catch	100%	50%	100%	100%	100%
Amazon River					
curimatá	10	0	10	10	10
jaraquí	0	0	10	0	10
% catch	50%	0%	100%	50%	100%

River and Cantabric Sea, respectively. In the Mediterranean Sea, the two most harvested species were respectively anchovy *Engraulis encrasicolus* (Engraulidae) and mackerel *Scomber scombrus* (Scombridae).

PCR yielded positive amplifications in all cases, and sequences of different length were obtained for each marker and species analyzed: 380–439, 605–635, 293–322, 412–462 base pairs (bp) for 12S rDNA, COI, cyt *b* and D-Loop respectively (Table 1). The concatenated sequences were thus 1,692–1,856 bp long. The sequences obtained were submitted to the GenBank where they are available with the accession numbers reported in Table 1.

Clear and unambiguous species identification from significant matches with the databases was not always possible (Table 2). All the 12S rDNA sequences yielded a 100% identity score with at least one GenBank reference sequence (other than those generated in the present study) belonging to only one species, and were hence considered as being unambiguously identified. However, the results were less clear for the other genes and also varied among species. All mackerel samples were well-identified by the four genes and the two databases, whereas tuna retrieved more than one species with identical scores or match probabilities (*Thunnus alalunga*, *T. thynnus* and *T. orientalis*) for all cyt *b* and many COI and D-Loop sequences (Table 3). One D-Loop sequence retrieved *Thunnus albacares* as the closest match (Table 3). Ambiguous results (more than one putative species) were obtained from BOLD also for anchovy (COI sequences assigned to any of *Engraulis encrasicolus*, *E. eurystole*, *E. australis* and *E.*

Table 3. Ambiguous or inconclusive matches between sequences in this study and reference sequences in GenBank (all sequences) and BOLD (COI). The species retrieved from each database (with maximum score for GenBank) are presented. + : Sequences for which there are > 5 entries in GenBank with a maximum score.

	GenBank	**BOLD**
Sequences of this study	**COI**	
JN007753,54,59,60,61	*Thunnus alalunga*	*Thunnus alalunga, T. orientalis, T. obesus, T. thynnus, T. atlanticus*
JN007752,55,56,57,58	*Thunnus alalunga, T. thynnus*	*Thunnus alalunga, T. orientalis, T. obesus, T. thynnus, T. atlanticus*
HM480814–15, JN007765–68	*Engraulis encrasicolus*	*Engraulis encrasicolus, E. eurystole, E. australis*
HM480816, JN007762–64	*Engraulis encrasicolus*	*Engraulis encrasicolus, E. capensis, Atherina breviceps*
JN007788 +	*Salmo trutta*	*Salmo trutta, S. ohridanus*
JN007727 +	*Prochilodus nigricans*	*Prochilodus nigricans, P. rubrotaeniatus*
JN007743 +	*Semaprochilodus insignis, S. taeniurus*	*Semaprochilodus insignis, S. taeniurus, Curimata inornata*
	cyt *b*	
JN007677 +	*Thunnus alalunga, T. orientalis*	
JN007657 +	None out of this study	
	D-Loop	
JN007604	*Thunnus albacares*	
JN007600–02	*Thunnus alalunga, T. thynnus*	

japonicus species), brown trout (assigned indistinctly to *Salmo trutta* and *S. ohridanus* by BOLD), curimatá (*Prochilodus nigricans*, *P. rubrotaeniatus*, *P. lineatus*, *P. costatus*) and jaraquí (*Semiprochilodus insignis*, *S. taeniurus*, *Curimata inornata*). In GenBank ambiguous COI species identifications occurred for five tuna haplotypes that yielded identical and maximum matching scores with *Thunnus alalunga* and *T. orientalis* sequences, and for jaraquí (*Semaprochilodus insignis* and *S. taeniurus* sequences yielded identical and maximum matching scores with our haplotypes). For cyt *b* of jaraquí (Table 3) the problem was not ambiguity but lack of external reference sequences in GenBank, viz. all the sequences yielding > 91% matching scores with ours were from the present study, and the closest identity with an external sequence (91%, unlikely the same species for a conserved coding gene) occurred with the sequence AY791437 of *Prochilodus nigricans*.

Genetic diversity in the four analyzed case studies

As expected, the four DNA regions exhibited different degrees of variability (Table 4). The non-coding D-Loop (58 haplotypes in total) was more variable than the two protein coding loci (31 and 27 haplotypes for cyt *b* and COI respectively) and the ribosomal 12S rDNA gene (15 haplotypes). The four marine species, the Amazonian jaraquí (*Semiprochilodus insignis*) and the north Spanish brown trout (*Salmo trutta*) exhibited

Table 4. Sequence diversity in each species. Nh, Hd and π are the number of haplotypes, haplotype diversity and nucleotide diversity, respectively.

		Species							
Locus	**Parameter**	**anchovy**	**mackerel (Cant.)**	**mackerel (Med.)**	**curimatá**	***A. salmon***	**brown trout**	**jaraquí**	**tuna**
12S rDNA	Nh	2	1	2	2	2	2	3	1
n = 380–439	Hd	0.2	0	0.467	0.467	0.356	0.356	0.378	0
	π	0.052	0	0.124	0.123	0.081	0.081	0.105	0
COI	Nh	2	4	5	4	1	2	3	6
n = 605–635	Hd	0.2	0.533	0.8	0.733	0	0.556	0.689	0.778
	π	0.165	0.265	1.249	0.154	0	0.088	0.136	0.191
Cyt *b*	Nh	3	4	8	1	1	5	6	3
n = 293–322	Hd	0.378	0.533	0.956	0	0	0.822	0.778	0.689
	π	0.205	0.273	1.82	0	0	0.469	0.394	0.88
D–Loop	Nh	8	10	10	6	1	5	8	10
n = 412–462	Hd	0.978	1	1	0.867	0	0.867	0.956	1
	π	1.893	2.126	3.655	0.65	0	0.358	1.268	6.362
All coding	Nh	4	6	10	5	2	8	8	7
n = 1278–1396	Hd	0.533	0.778	1	0.8	0.356	0.956	0.956	0.911
	π	0.141	0.188	1.048	0.111	0.025	0.174	0.186	0.293
All loci	Nh	10	10	10	6	2	10	10	10
n = 1682–1856	Hd	1	1	1	0.867	0.356	1	1	1
	π	0.588	0.644	1.738	0.244	0.019	0.219	0.449	1.744

ten different haplotypes in total considering the concatenated mitochondrial sequences analyzed. Fewer haplotypes were obtained for the Amazonian *Prochilodus nigricans* (6 haplotypes) and the Spanish *Salmo salar* (two haplotypes). In this latter species polymorphism occurred in the 12S rDNA gene, but not in the D-Loop, which was the most variable region in the other species. Overall nucleotide diversity was higher for marine than for freshwater settings for all markers as well as the concatenated sequence (Table 4). The highest Hd for both 12S rDNA and COI genes corresponded to the Amazonian samples, whereas marine catches were most variable at the less conserved cyt *b* and especially at the D-Loop. The least diverse Narcea River exhibited higher Hd at the highly conserved 12S rDNA than the two marine catches, due to Atlantic salmon polymorphism (likely due to a mixture of lineages remaining from past stocks transfers from North European populations; e.g. Horreo et al. 2011b).

The trade-off between using the same genetic analysis for simultaneously authenticating specimens and rapidly evaluating population diversity is that conserved species-specific sequences may not exhibit enough polymorphism. This is exemplified in Figure 1 and in the total number of variants of each marker found in this study, with 58 D-Loop versus only 15 12S rDNA haplotypes. Comparison between DNA regions for polymorphic information -measured as mean variation for each gene as in Figure 1- yielded, despite small sample sizes, highly significant differences for all parameters when the six sequences were considered at the same time ($p = 0.011$, $p = 0.006$ and $p =$

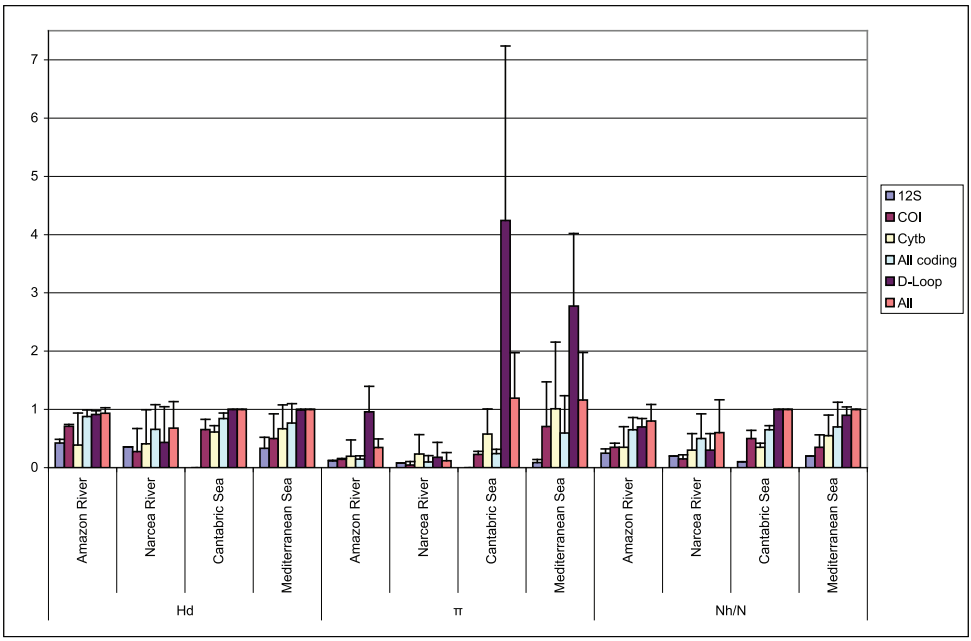


Figure 1. Summary of population genetic diversity retrieved from each mitochondrial region separately (12S rDNA, COI, cyt *b*, D-Loop), from the coding and from all regions concatenated (All), in the four case studies. Mean (standard deviation as vertical bars) is provided for Nh/n, Hd and π (mean number of different haplotypes per species, haplotype diversity and nucleotide diversity respectively).

0.000, for Hd, π and Nh/n, respectively). Most polymorphisms were provided by the non-coding D-Loop (Figure 1), and adding more nucleotides (concatenated sequence of all loci) did not increase significantly the level of polymorphism ($p = 0.639$, $p = 0.109$ and $p = 0.428$, for Hd, π and Nh/n, respectively). As expected, in relation with its length, the D-Loop was the most informative gene for quantifying diversity.

Discussion

The results presented in this study illustrate how genetic methodologies could be applied in practice for monitoring fish catches. They also suggest some caveats of the current databases that should be considered in order to improve their built-in tools for species identification, especially if massive sequencing is envisaged. We have found ambiguous catch identifications in several cases. This is due to the fact that some identical haplotypes (sequences) are labeled in the databases with different specific names. Duplicated names at species level are a problem well recognized in reference databases such as GenBank (e.g. Federhen 2012). In this sense, we encourage a thorough taxonomic revision of the existing databases. The joint work of taxonomists and molecular systematists will help in the effort of cataloguing collections and voucher specimens

(Puillandre et al. 2012). It may also happen that very closely related species share haplotypes at highly conserved genes. This could be the case of the *Thunnus* species, which are so closely related that they even give inconsistent phylogenetic signals (e.g. Chow and Kishino 1995). Mitochondrial introgression between species has been reported for this genus (Chow et al. 2006), so mitochondrial markers would not be a good choice for identifying tuna species. However, there was no ambiguity with the highly conserved 12S rDNA. Therefore, using this region may solve the problem in *Thunnus*. Although DNA barcoding through COI resolves most species, some taxa have proved intractable (Waugh 2007). We cannot explain what the reason was for all the cases found here, but it is clear that ambiguous identification would be a problem in routine large-scale fisheries barcoding. As also suggested by other authors (e.g. Savolainen et al. 2005, Austerlitz et al. 2009), incorporating nuclear genes as barcodes could help to solve these problems.

On the other hand, analyzing two DNA regions of different level of variability and recording simple polymorphism data in a database are easy actions that can be done very fast employing massive sequencing methodologies. They would hopefully allow to ascertaining the species and early detecting variation losses in catch. In a moment of stock overexploitation (Myers and Worm 2005) and urgent need of a better fisheries control in many regions (Worm and Branch 2012), these two issues are of most importance for long-term fisheries sustainability (Dahl 2000, Wessells et al. 2001, Pauly et al. 2002). For mitochondrial (haploid) sequences, simple statistical parameters for measuring sequence variation such as haplotype and nucleotide diversity could be incorporated into next-generation sequencing software, making it easier the process of diversity monitoring in fish landings. Hence, we propose to incorporate DNA barcoding as a first-instance routine surveys and periodical monitoring of catch diversity, but adding nuclear genes seems to be necessary (Markmann and Tautz 2005, Monaghan et al. 2005, Savolainen et al. 2005). If a decrease of variation is detected, further studies should follow, may be employing population genomics approaches and other biological tools. Diversity can be properly measured by using a diversity of tools and characters (Rubinoff 2006). Morphology (Wiens 2004), ecology (Crandall et al. 2000), adaptive differences (*sensu* Waples 1991) and genetic data from the mitochondrial and nuclear genomes, which can result in very different assessments of biodiversity, should be combined for having a complete perspective of the diversity of a community or ecosystem (Mouillot et al. 2011).

Conclusions

Taking into account the number of existing sequences in databases, that is essential for species identification, and the polymorphic information provided by the different mitochondrial regions examined, the use of more than one gene and preferably a combination of nuclear and mitochondrial sequences would be recommended for routine genetic monitoring of fish catches. Incorporating new sequencing technologies will speed up large-scale genetic analysis of catch.

Acknowledgements

This study has been funded by the Asturias Government, SV-PA-13-ECOEMP-41. We are grateful to three anonymous reviewers of Zookeys for useful comments on the manuscript.

References

April J, Mayden RL, Hanner RH, Bernatchez L (2011) Genetic calibration of species diversity among North America's freshwater fishes. Proceedings of the National Academy of Sciences of the USA 108: 10602–10607. doi: 10.1073/pnas.1016437108

Ardura A, Linde AR, Moreira JC, Garcia-Vazquez E (2010a) DNA barcoding for conservation and management of Amazonian commercial fish. Biological Conservation 143: 1438–1443. doi: 10.1016/j.biocon.2010.03.019

Ardura A, Pola IG, Ginuino I, Gomes V, Garcia-Vazquez E (2010b) Application of Barcoding to Amazonian commercial fish labeling. Food Research International 43: 1549–1552. doi: 10.1016/j.foodres.2010.03.016

Ardura A, Pola IG, Linde AR, Garcia-Vazquez E (2010c) DNA-based methods for species authentication of Amazonian commercial fish. Food Research International 43: 2259–2302. doi: 10.1016/j.foodres.2010.08.004

Ardura A, Planes S, Garcia-Vazquez E (2011) Beyond biodiversity: fish metagenomes. PLoS ONE 6: e22592. doi: 10.1371/journal.pone.0022592

Ardura A, Gomes V, Linde AR, Moreira JC, Horreo JL, Garcia-Vazquez E (2013) The Meeting of Waters, a possible shelter of evolutionary significant units for Amazonian fish. Conservation Genetics 14: 1185–1192. doi: 10.1007/s10592-013-0505-8

Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, Veuille M, Laredo C (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10 (Suppl 14): S10. doi: 10.1186/1471-2105-10-S14-S10

Barbuto M, Galimberti A, Ferri E, Labra M, Malandra R, Galli P, Casiraghi M (2010) DNA barcoding reveals fraudulent substitutions in shark seafood products: The Italian case of "palombo" (*Mustelus* spp.). Food Research International 43: 376–381. doi: 10.1016/j.foodres.2009.10.009

Chow S, Kishino H (1995) Phylogenetic relationships between tuna species of the genus *Thunnus* (Scombridae: Teleostei): Inconsistent implications from morphology, nuclear and mitochondrial genomes. Journal of Molecular Evolution 41: 741–748. doi: 10.1007/BF00173154

Chow S, Nakagawa T, Suzuki N, Takeyama H, Matsunaga T (2006) Phylogenetic relationships among *Thunnus* species inferred from rDNA ITS1 sequence. Journal of Fish Biology 68: 24–35. doi: 10.1111/j.0022-1112.2006.00945.x

Chow S, Ushiama H (1995) Global population structure of albacore (*Thunnus alalunga*) inferred by RFLP analysis of the mitochondrial ATPase gene. Marine Biology 123: 39–45. doi: 10.1007/BF00350321

Crandall KA, Bininda-Emonds ORP, Mace GM, Wayne RK (2000) Considering evolutionary processes in conservation biology. Trends in Ecology & Evolution 15: 290–295. doi: 10.1016/s0169-5347(00)01876-0

Crego-Prieto V, Campo D, Perez J, Martinez JL, Garcia-Vazquez E, Roca A (2012) Inaccurate labelling detected at landings and markets: The case of European megrims. Fisheries Research 129–130: 106–109. doi: 10.1016/j.fishres.2012.06.017

Dahl AL (2000) Using indicators to measure sustainability: recent methodological and conceptual developments. Marine and Freshwater Research 51: 427–433. doi: 10.1071/MF99056

DeSalle R, Birstein VJ (1996) PCR identification of black caviar. Nature 381: 197–198. doi: 10.1038/381197a0

Fauvelot C, Bernardi G, Planes S (2003) Reductions in the mitochondrial DNA diversity of coral reef fish provide evidence of population bottlenecks resulting from Holocene sea-level change. Evolution 57: 1571–1583. doi: 10.1111/j.0014-3820.2003.tb00365.x

Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Research 40(D1): D136–D143. doi: 10.1093/nar/gkr1178

Filonzi L, Chiesa S, Vaghi M, Marzano FN (2010) Molecular barcoding reveals mislabelling of commercial fish products in Italy. Food Research International 43: 1383–1388. doi: 10.1016/j.foodres.2010.04.016

Frankham R (1995) Conservation genetics. Annual Review of Genetics 29: 305–327. doi: 10.1146/annurev.ge.29.120195.001513

Garcia-Vazquez E, Perez J, Martinez JL, Pardiñas AF, Lopez B, Karaiskou N, Casa MF, Machado-Schiaffino G, Triantafyllidis A (2011) High Level of Mislabeling in Spanish and Greek Hake Markets Suggests the Fraudulent Introduction of African Species. Journal of Agricultural and Food Chemistry 59: 475–480. doi: 10.1021/jf103754r

Gerber AS, Loggins R, Kumar S, Dowling TE (2001) Does nonneutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? Annual of Review of Genetics 35: 539–566. doi: 10.1146/annurev.genet.35.102401.091106

Hajibabei M (2012) The golden age of DNA metasystematics. Trends in Genetics 28: 535–537. doi: 10.1016/j.tig.2012.08.001

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

Hebert P, Cywinska A, Ball S, deWaard J (2003) Biological identification through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Hedrick PW (2001) Conservation genetics: where are we now? Trends in Ecology & Evolution 16: 629–636. doi: 10.1016/S0169-5347%2801%2902282-0

Horreo JL, Martinez JL, Ayllon F, Pola IG, Monteoliva JA, Héland M, Garcia-Vazquez E (2011a) Impact of habitat fragmentation on the genetics of populations in dendritic landscapes. Freshwater Biology 56: 2567–2579. doi: 10.1111/j.1365-2427.2011.02682.x

Horreo JL, Machado-Schiaffino G, Ayllon F, Griffiths AM, Bright D, Stevens JR, Garcia-Vazquez E (2011b) Impact of climate change and human-mediated introgression on South European Atlantic salmon populations. Global Change Biology 17: 1778–1787. doi: 10.1111/j.1365-2486.2010.02350.x

Jacquet JL, Pauly D (2008) Trade secrets: Renaming and mislabeling of seafood. Marine Policy 32: 309–318. doi: 10.1016/j.marpol.2007.06.007

Janzen DH, Hajibabaei M, Burns JM, Hallwachs W, Remigio E, Hebert PDN (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philosophical Transactions of the Royal Society of London B 360: 1835–1845. doi: 10.1098/rstb.2005.1715

Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proceedings of the National Academy of Sciences of the USA 86: 6196–6200. doi: 10.1073/pnas.86.16.6196

Lee WJ, Conroy J, Howell WH, Kocher TD (1995) Structure and evolution of teleost mitochondrial control regions. Journal of Molecular Evolution 41: 54–66. doi: 10.1007/BF00174041

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452. doi: 10.1093/bioinformatics/btp187

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 B 360: 1917–1924. doi: 10.1098/rstb.2005.1723

Marko PB, Lee SC, Rice AM, Gramling JM, Fitzhenry TM, McAlister JS, Harper GR, Moran AL (2004) Mislabeling of a depleted reef fish. Nature 430: 309–310. doi: 10.1038/430309b

Miller DD, Mariani S (2010) Smoke, mirrors, and mislabeled cod: poor transparency in the European seafood industry. Frontiers in Ecology and the Environment 8: 517–521. doi: 10.1890/090212

Min MS, Okumura H, Jo DJ, An JH, Kim KS, Kim CB, Shin NS, Lee MH, Han CH, Voloshina IV, Lee H (2004) Molecular phylogenetic status of the Korean goral and Japanese serow based on partial sequences of the mitochondrial cytochrome b gene. Molecules and Cells 17: 365–372. doi: 10.1266/ggs.75.17

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 B 360: 1925–1933. doi: 10.1098/rstb.2005.1724

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

Mouillot D, Albouy C, Guilhaumon F, Lasram FBR, Coll M, Devictor V, Meynard CN, Pauly D, Tomasini JA, Troussellier M, Velez L, Watson R, Douzery EJP, Mouquet N (2011) Protected and Threatened Components of Fish Biodiversity in the Mediterranean Sea. Current Biology 21: 1044–1050. doi: 10.1016/j.cub.2011.05.005

Myers RA, Worm B (2005) Extinction, survival or recovery of large predatory fishes. Philosophical Transactions of the Royal Society B 360: 13–20. doi: 10.1098/rstb.2004.1573

Nei M, Wen-Hsiung L (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the USA 76: 5269–5273. doi: 10.1073/pnas.76.10.5269

Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. Genetics 97: 145.

Nevado B, Mautner S, Sturmbauer C, Verheyen E (2013) Water-level fluctuations and metapopulation dynamics as drivers of genetic diversity in populations of three Tanganyikan cichlid fish species. Molecular Ecology 22: 3933–3948. doi: 10.1111/mec.1237

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009) Population genomics of marine fishes: identifying adaptive variation in space and time. Molecular Ecology 18: 3128–3150. doi: 10.1111/j.1365-294X.2009.04272.x

Onuma M, Suzuki M, Ohtaishi N (2006) Possible conservation units of the sun bear (*Helarctos malayanus*) in Sarawak based on variation of mtDNA control region. Japanese Journal of Veterinary Research 54: 135–139.

Palumbi SR (1996) Nucleic acids II: the polymerase chain reaction. In: Hillis DM, Moritz C, Mable BK (Eds) Molecular Systematics (2nd ed.) Sinauer Associates Inc, Sunderland, Massachusetts, 205–247.

Pauly D, Christensen V, Guénette S, Pitcher TJ, Sumaila UR, Walters CJ, Watson R, Zeller D (2002) Towards sustainability in world fisheries. Nature 418: 689–695. doi: 10.1038/nature01017

Puillandre N, Bouchet P, Boisselier-Dubayle MC, Brisset J, Buge B, Castelin M, Chagnoux S, Christophe T, Corbari L, Lambourdière J, Lozouet P, Marani G, Rivasseau A, Silva N, Terryn Y, Tillier S, Utge J, Samadi S (2012) New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Molecular Ecology Resources 12: 396–402. doi: 10.1111/j.1755-0998.2011.03105.x

Rasmussen RS, Morrisey MT (2008) DNA-Based Methods for the Identification of Commercial Fish and Seafood Species. Comprehensive Reviews in Food Science and Food Safety 7: 280–295. doi: 10.1111/j.1541-4337.2008.00046.x

Rubinoff D (2006) Utility of Mitochondrial DNA Barcodes in Species Conservation. Conservation Biology 20: 1026–1033. doi: 10.1111/j.1523-1739.2006.00372.x

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1805–1811. doi: 10.1098/rstb.2005.1730

Steemers FJ, Gunderson KL (2005) Illumina, Inc. Pharmacogenomics 6: 777–782. doi: 10.2217/14622416.6.7.777

Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S (2007) Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies. PLoS ONE 2: e484. doi: 10.1371/journal.pone.0000484

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599. doi: 10.1093/molbev/msm092

Tudela S, Garcia-Marin JL, Pla C (1999) Genetic structure of the European anchovy, *Engraulis encrasicolus* L., in the north-west Mediterranean. Journal of Experimental Marine Biology & Ecology 234: 95–109. doi: 10.1016/S0022-0981(98)00142-7

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends in Ecology & Evolution 24: 110–117. doi: 10.1016/j.tree.2008.09.011

Viñas J, Alvarado-Bremer JR, Pla C (2004) Inter-oceanic genetic differentiation among albacore (*Thunnus alalunga)* populations. Marine Biology 145: 25–232. doi: 10.1007/s00227-004-1319-5

Wang SZ, Hard JJ, Utter F (2002) Genetic variation and fitness in salmonids. Conservation Genetics 3: 321–333. doi: 10.1023/A:1019925910992

Waples RS (1991) Pacific salmon, *Oncorhynchus* spp., and the definition of "species" under the Endangered Species Act. Marine Fisheries Review 53: 11–22.

Ward RD, Zemlak TS, Innes BH, Last PD, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B 360: 1847–1857. doi: 10.1098/rstb.2005.1716

Watson R, Pauly D (2001) Systematic distortions in world fisheries catch trends. Nature 414: 534–536. doi: 10.1038/35107050

Waugh J (2007) DNA barcoding in animal species: progress, potential and pitfalls. Bioessays 29: 188–197. doi: 10.1002/bies.20529

Wessells CR, Cochrane K, Deere C, Wallis P, Willmann R (2001) Product certification and ecolabelling for fisheries sustainability. FAO Fisheries Technical Paper. No. 422, FAO, Rome, 83 pp.

Wiens JJ (2004) The role of morphological data in phylogeny reconstruction. Systematic Biology 53: 653–661. doi: 10.1080/10635150490472959

Willmann R, Kelleher K (2010) Economic trends in global marine fisheries. In: Grafton RQ, Hilborn R, Squires D, Tait M, Williams M (Eds) Handbook of Marine Fisheries Conservation and Management. Oxford University Press Inc., New York, 20–42.

Wong EHK, Hanner RH (2008) DNA barcoding detects market substitution in North American seafood. Food Research International 41: 828–837. doi: 10.1016/j.foodres.2008.07.005

Worm B, Branch TA (2012) The future of fish. Trends in Ecology & Evolution 27: 594–599. doi: 10.1016/j.tree.2012.07.005

Wu HL, Wan QH, Fang SG (2006) Population structure and gene flow among wild populations of the black muntjac (*Muntiacus crinifrons*) based on mitochondrial DNA control region sequences. Zoological Science 23: 333–340. doi: 10.2108/zsj.23.333

Zardoya R, Castilho R, Grande C, Favre-Krey L, Caetano S, Marcato S, Krey G, Patarnello T (2004) Differential population structuring of two closely related fish species, the mackerel (*Scomber scombrus*) and the chub mackerel (*Scomber japonicus*), in the Mediterranean Sea. Molecular Ecology 13: 1785–1798. doi: 10.1111/j.1365-294X.2004.02198.x

Zhang F, Jiang Z (2006) Mitochondrial phylogeography and genetic diversity of Tibetan gazelle (*Procapra picticaudata*): implications for conservation. Molecular Phylogenetics and Evolution 41: 313–321. doi: 10.1016/j.ympev.2006.05.024

ZooKeys 365: 67–70 (2013)
doi: 10.3897/zookeys.365.5875
www.zookeys.org

LETTER TO THE EDITOR

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

The importance of biobanking in molecular taxonomy, with proposed definitions for vouchers in a molecular context

Jonas J. Astrin[1], Xin Zhou[2], Bernhard Misof[1]

1 *Zoological Research Museum Alexander Koenig (ZFMK), Centre for Molecular Biodiversity Research, Bonn, Germany* **2** *BGI, China National GeneBank, BGI-Shenzhen, Shenzhen, China 518083*

Corresponding author: *Jonas Astrin* (j.astrin.zfmk@uni-bonn.de)

Academic editor: *M. De Meyer* | Received 30 June 2013 | Accepted 7 September 2013 | Published 30 December 2013

Citation: Astrin JJ, Zhou X, Misof B (2013) The importance of biobanking in molecular taxonomy, with proposed definitions for vouchers in a molecular context. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 67–70. doi: 10.3897/zookeys.365.5875

DNA barcoding and molecular or integrative taxonomy projects are among the most valuable sources for biobank specimens of wild organisms, thanks to – among other aspects – the high level of specimen diversity and thanks to a thorough taxonomic coverage. Specimens used to build barcoding reference libraries tend to be accompanied by deeper and higher-quality data than samples from many other sources, as they are often contributed by taxonomists, and identifications are cross-checked through barcode analysis. Vouchering of morphological specimens in natural history collections is a prerequisite for proper barcoding, which is advantageous for biobanking as well, as biobank samples should always be linked to specimen vouchers. As a further added value, barcoding provides an inherent, molecular species ID tag to the processed biobank sample.

Banked barcoding samples can greatly catalyze taxonomy, as well as many other fields of application, such as the emerging large genome sequencing projects that are constantly increasing the demand for well-preserved samples from a multitude of different species (see Wong et al. 2012).

Considered from the opposite perspective of the synergy, barcoding can benefit greatly from biobanking as well. Biobanking enables the expansion of barcoding datasets with biobanked samples from other projects. It also offers the possibility to add new barcoding markers any time in the future, e.g. scaling up to 'next-generation barcoding' (e.g. Taylor

and Harris 2012) if feasible (manageability of data, NGS and data handling cost, performance in mixed samples, etc.), without the necessity of repeating the time-consuming and expensive steps of sample collection, data collection and identification, and vouchering.

Finally and most importantly, biobanks offer barcoding projects the possibility to adequately voucher their molecular samples and to warrant reproducibility of results.

Researchers involved in barcoding projects should make sure their samples are properly vouchered – morphologically AND molecularly. They can do this by depositing their samples at a dedicated natural history collection. Increasingly, these repositories are establishing biobanks / DNA banks / tissue banks for curated long-term, ultra cold conservation of molecular samples, are adopting standard operating procedures and making their samples available online e.g. through biobank networks like the DNA Bank Network (http://www.dnabank-network.org/) or soon also the Global Genome Biodiversity Network (http://ggbn.org/). Those museums and natural history collections that implement these features and commit themselves to provide the community with proper biobanks (although maybe called differently) offer a very efficient and elegant way to both draw on and to deposit morphological-molecular 'tandem' samples. Often underappreciated by public and policy-makers (Suarez and Tsutsui 2004), natural history collections holding and curating specimen vouchers and/or cross-referenced molecular vouchers and their data play a "major role in organizing systematic knowledge in the molecular age" (Whitfield and Cameron 1994).

Although it has been pointed out before (e.g. Hafner 1994), the importance of vouchering molecular samples is not yet fully apprehended in the scientific community (perhaps because of the way taxonomy has been traditionally carried out).

We would like to encourage authors, editors and reviewers of scientific papers to give also molecular vouchers the attention they deserve.

Vouchers – morphological and molecular alike – not only form the connection between study data and taxonomic identification. They are much more: vouchers link the data collected in individual studies with the immense wealth of data that can still be (or already have been) collected through the vouchers: repetitively or in an additive manner. Put short, vouchers link individual studies with other studies and inferences, past or future.

It becomes obvious that it is only through adequate vouchering that we can make organismic biology meaningful, warranting reproducibility and embedding our research into existing and emerging knowledge.

In a laudable approach to increasing semantic accuracy regarding the voucher concept, Pleijel et al. (2008) suggest a terminology for those specimen vouchers used to produce molecular (sub-)samples. These are coined 'genophores' (although of course molecular samples lend themselves to more than genetic analysis), and for mnemonic ease follow the taxonomic nomenclatorial codes in style:

a *hologenophore* is the specimen voucher from which the molecular sample is directly derived, an *isogenophore* is a different specimen with a clonal relationship to the study organism, while a *progenophore* represents a voucher that is linked to the specimen sampled for molecular analysis by a parent-descendant or sibling relationship. A

paragenophore is a putatively conspecific specimen voucher collected together with the 'molecular' specimen. The same applies to the *syngenophore*, except that it is collected at another place or time.

These genealogy-based distinctions made by Pleijel et al. (2008) are helpful for categorizing a specimen voucher in its relation to a molecular voucher and we endorse their use in this context. The function/purpose or the nature of vouchers was deliberately not addressed by Pleijel and colleagues. However, especially in the context of molecular samples, we perceive the necessity to do so, as varying uses of terms can be observed (e.g. "DNA voucher" used synonymously for the DNA source or for the isolated DNA). Different use of terms makes it difficult to extract data from biological collection databases or from the literature in a semantically meaningful way. Therefore, in the following we propose some voucher, sample and repository definitions, with special focus on a molecular context.

- *specimen voucher*: a specimen serving as the basis for taxonomic identification and possibly also for other queries. A specimen voucher is often, but not necessarily a whole organism, or part of it (it can be a trace or ichnofossil, scats, eggs, images, etc.).
  Narrower terms: morphological voucher, acoustic voucher, e-voucher, etc.
  - *morphological voucher*: a specimen that allows the inspection of morphological characters.
  - *e-voucher*: digital objects that serve as vouchers (morphological, acoustic, etc.), e.g. sound recordings, audiovisual material, images, etc.

- *molecular voucher*: a sample that is deliberately preserved and curated in a way that will conserve its molecular properties for analysis. A molecular voucher should always be linked to a specimen voucher (which sometimes can be the same object if sufficient characters remain, see tissue voucher).
  Narrower terms: biobank voucher, DNA voucher, tissue voucher, RNA voucher, protein voucher, genomic sample, etc.
  - *tissue voucher*: tissue subsampled from a specimen - or the entire specimen -, preserved (usu. frozen) to keep its molecular properties (either fixed tissue or viable cells) for future analysis
  - *DNA voucher*: the isolated and preserved, frozen or dried (usu. genomic) DNA. As a derived sample, a DNA voucher should not – if anyhow possible – function as specimen voucher.
  - *biobank voucher*: any molecular voucher curated in a biobank. A biobank voucher is a *biobank sample* that links to other physical objects or data (other than their metadata), i.e. most biobank samples are (biobank) vouchers, as they usually link to a separate specimen voucher
  - *genomic sample*: preserved sample containing (isolated or as a constituent) a high percentage of an organism's genome in widely unfragmented form

- *biobank*: a curated collection/repository of biological materials that warrants long-term integrity at molecular level, authenticity, availability and rights management of its samples by adhering to standard operating procedures (SOPs).
  Narrower terms: DNA bank, tissue bank, biodiversity biobank, etc.
- *biodiversity biobank*: term currently used to refer to a biobank holding non-human samples
- *genomic collection*: a molecular collection holding genomic samples

Acknowledgement

We thank Marc De Meyer for reviewing and improving this article and our Global Genome Biodiversity Network colleagues for discussions.

References

Hafner MS (1994) Molecular Extracts from Museum Specimens Can and Should Be Saved - Reply. Molecular Phylogenetics and Evolution 3: 270–271. doi: 10.1006/mpev.1994.1030

Pleijel F, Jondelius U, Norlinder E, Nygren A, Oxelman B, Schander C, Sundberg P, Thollesson M (2008) Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. Molecular Phylogenetics and Evolution 48: 369–371. doi: 10.1016/j.ympev.2008.03.024

Suarez AV, Tsutsui ND (2004) The value of museum collections for research and society. Bioscience 54: 66–74. doi: 10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12: 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Whitfield JB, Cameron SA (1994) Museum Policies Concerning Specimen Loans for Molecular Systematic Research. Molecular Phylogenetics and Evolution 3: 268–270. doi: 10.1006/mpev.1994.1029

Wong P, Wiley E, Johnson W, Ryder O, O'Brien S, Haussler D, Koepfli K-P, Houck M, Perelman P, Mastromonaco G, Bentley A, Venkatesh B, Zhang Y-p, Murphy R, G10K-COS (2012) Tissue sampling methods and standards for vertebrate genomics. GigaScience 1: 8. doi: 10.1186/2047-217X-1-8

ZooKeys 365: 71–82 (2013)
doi: 10.3897/zookeys.365.5725
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

The chloroplast DNA locus *psbZ-trnfM* as a potential barcode marker in *Phoenix* L. (Arecaceae)

Marco Ballardini[1], Antonio Mercuri[2], Claudio Littardi[1], Summar Abbas[3], Marie Couderc[4], Bertha Ludeña[4], Jean-Christophe Pintaud[4]

1 *Centro Studi e Ricerche per le Palme - Sanremo (CSRP), Corso F. Cavallotti 113, I-18038 Sanremo (IM), Italy* **2** *Consiglio per la Ricerca e la sperimentazione in Agricoltura - Unità di Ricerca per la Floricoltura e le Specie Ornamentali (CRA-FSO), Corso degli Inglesi 508, I-18038 Sanremo (IM), Italy* **3** *Institute of Horticultural Sciences, University of Agriculture, 38040 Faisalabad, Pakistan* **4** *UMR DIADE/DYNADIV, Institut de Recherche pour le Développement (IRD), 911 Av. Agropolis, F-34394 Montpellier Cedex 5, France*

Corresponding author: *Marco Ballardini* (m_ballardini@hotmail.com)

Academic editor: *K. Jordaens* | Received 1 June 2013 | Accepted 3 December 2013 | Published 30 December 2013

Citation: Ballardini M, Mercuri A, Littardi C, Abbas S, Couderc M, Ludeña B, Pintaud JC (2013) The chloroplast DNA locus *psbZ-trnfM* as a potential barcode marker in *Phoenix* L. (Arecaceae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 71–82. doi: 10.3897/zookeys.365.5725

Abstract

The genus *Phoenix* (*Arecaceae*) comprises 14 species distributed from Cape Verde Islands to SE Asia. It includes the economically important species *Phoenix dactylifera*. The paucity of differential morphological and anatomical useful characters, and interspecific hybridization, make identification of *Phoenix* species difficult. In this context, the development of reliable DNA markers for species and hybrid identification would be of great utility. Previous studies identified a 12 bp polymorphic chloroplast minisatellite in the *trnG*(GCC)-*trnfM*(CAU) spacer, and showed its potential for species identification in *Phoenix*. In this work, in order to develop an efficient DNA barcode marker for *Phoenix*, a longer cpDNA region (700 bp) comprising the mentioned minisatellite, and located between the *psbZ* and *trnfM(CAU)* genes, was sequenced. One hundred and thirty-six individuals, representing all *Phoenix* species except *P. andamanensis*, were analysed. The minisatellite showed 2-7 repetitions of the 12 bp motif, with 1-3 out of seven haplotypes per species. *Phoenix reclinata* and *P. canariensis* had species-specific haplotypes. Additional polymorphisms were found in the flanking regions of the minisatellite, including substitutions, indels and homopolymers. All this information allowed us to identify unambiguously eight out of the 13 species, and overall 80% of the individuals sampled. *Phoenix rupicola* and *P. theophrasti* had the same haplotype, and so had *P. atlantica*, *P. dactylifera*, and *P. sylvestris* (the "date palm complex" *sensu* Pintaud et al. 2013). For these species, additional molecular markers will be required for their unambiguous identification. The *psbZ-trnfM*(CAU) region therefore could be considered as a good basis for the establishment of a DNA barcoding system in *Phoenix*, and is potentially useful for the identification of the female parent in *Phoenix* hybrids.

Keywords
Chloroplast *psbZ-trnfM*(CAU) region, DNA barcode, minisatellite, palms

Introduction

Taxonomy and phylogeny of *Phoenix* L.

The genus *Phoenix* L. (*Arecaceae*) comprises 14 species (Govaerts and Dransfield 2005), distributed from the E Atlantic (Macaronesia), through Africa, the Mediterranean region, S Asia to islands in the Indian Ocean (Madagascar, Andaman) and the NW Pacific (Taïwan and N Philippines). *Phoenix* is morphologically and phylogenetically highly divergent from the other palm genera, and constitutes the monogeneric tribe *Phoeniceae* within the subfamily Coryphoideae (Asmussen et al. 2006, Dransfield et al. 2008). The position of *Phoenix* within the subfamily Coryphoideae has been confirmed by a generic-level phylogenetic analysis of the entire palm family (*Arecaceae*) that included plastid and nuclear DNA sequences, cpDNA RFLPs and morphological data (Baker et al. 2009).

The taxonomy, phylogeny and evolution of the genus itself have been assessed using morphological and molecular approaches. According to Barrow (1998), both morphological, and molecular data of the 5S intergenic spacer of the nuclear ribosomal 5S DNA unit supported the existence of two clades of closely related species. The first clade included *P. dactylifera*, *P. sylvestris*, *P. theophrasti* and *P. canariensis* -the so-called "date-palm complex"-, and *P. atlantica* (Pintaud et al. 2010). The second group comprised the sister species *P. paludosa* and *P. roebelenii*. However, Barrow's (1998) molecular analysis included only 11 out of the 13 species recognized at that time, since *P. atlantica* was left as an insufficiently known taxon. Its status as a valid species was confirmed later by Henderson et al. (2006). Using one plastid and 16 nuclear microsatellite markers, Pintaud et al. (2010) demonstrated that all members of the "date-palm complex" are distinct species. Moreover, their data suggested that *P. atlantica* and *P. dactylifera* were sister species. Unfortunately, *P. paludosa* and *P. andamanensis* were not included in their analyses. Combining sequence data of the chloroplast *psbZ-trnfM* and *rpl16-rps3* loci, Pintaud et al. (2013) depicted five distinct phylogenetic lineages within *Phoenix* (*P. loureiroi-acaulis-pusilla*, *P. roebelenii-paludosa*, *P. caespitosa*, *P. reclinata*, and *P. rupicola-theophrasti-canariensis-dactylifera-atlantica-sylvestris*), and restricted the "date palm complex" to *P. dactylifera-atlantica-sylvestris*. This complex could be distinguished by the presence of a 3-repetitions haplotype of a 20 bp minisatellite motif at the *rpl*16-*rps*3 locus, that was absent in all other species. *Phoenix andamanensis* was the only taxon not included in their study.

The cultivated date palm *P. dactylifera* L. is the most important fruit crop in the Middle East and North African countries. This species was probably domesticated around 4,000 B.C. in the Mesopotamia-Arabic Gulf area (Nesbitt 1993, Zohary and Hopf 2000, Tengberg 2012) and is nowadays distributed worldwide.

Phoenix species are largely interfertile and many interspecific hybrids have been recognized or suspected (Greuter 1967, Wrigley 1995). The spread of the domesticated *Phoenix dactylifera* resulted in situations of sympatry with wild species, promoting interspecific gene flow, in particular with the endemic *P. canariensis* in the Canary islands (González-Pérez et al. 2004), and possibly with *P. theophrasti* in Turkey (Boydak and Barrow 1995), *P. atlantica* in the Cape Verde Islands (Henderson et al. 2006), and *P. sylvestris* in NW India (Newton et al. 2013). Moreover, spontaneous and directed hybridization between species is an important aspect of *Phoenix* ornamental cultivation (Tournay 2009).

Added to the common hybridization process between *Phoenix* species, the paucity of systematically useful morphological and anatomical characters within the genus (Barrow 1998), makes it difficult to establish a comprehensive taxonomy of the genus *Phoenix*. Because of this confusing situation, a reliable DNA marker set (barcode) to discriminate among *Phoenix* species and hybrids would be extremely useful.

DNA barcoding

Hebert et al. (2003) introduced the concept of "DNA barcode" as a new approach to taxon recognition, assuming that a short standardised DNA sequence can distinguish individuals of a species because genetic differentiation between species exceeds that within species. Since then, DNA barcoding has become increasingly important as a tool in taxonomic studies and species delimitation, as well as in the discovery of new (cryptic) species (e.g. DeSalle et al. 2005, Hebert et al. 2004, Hebert and Gregory 2005, Savolainen et al. 2005, Hajibabaei et al. 2007). A consortium of scientists suggested the two-locus combination of *rbc*L + *mat*K plastid genes as the universal plant barcode (CBOL 2009), while other authors (Chen et al. 2010, Yao et al. 2010) proposed the ITS2 region as a more efficient barcode. The China Plant BOL Group (2011) highlighted the importance of both sampling multiple individuals and using markers with different modes of inheritance, and suggested to incorporate the ITS1/ITS2 region into the core barcode for seed plants.

However, despite all efforts, no locus (alone or in combination), has proven to be 100% efficient as universal DNA barcode in plants at the species level.

The first DNA barcoding analysis in palms (Jeanson et al. 2011) achieved a 92% success in species discrimination by applying a combination of three markers (the plastid *mat*K and *rbc*L, and the nuclear ITS2) to the tribe *Caryoteae* (subfamily *Coryphoideae*).

Investigating the taxonomic status of *P. atlantica*, in comparison with its close relatives *P. dactylifera*, *P. canariensis* and *P. sylvestris*, Henderson et al. (2006) identified a polymorphic cpDNA minisatellite locus, situated within the *trnG*(GCC)-*trnfM*(CAU) intergenic spacer. Its structure was based on the 12 bp motif CTAACTACTATA repeated in tandem 2-6 times. Four haplotypes were observed: one specific of *P. canariensis*, one restricted to some individuals of *P. sylvestris*, and two shared between *P. dactylifera*, *P. atlantica* and *P. sylvestris*. Pintaud et al. (2010) studied this locus in 12

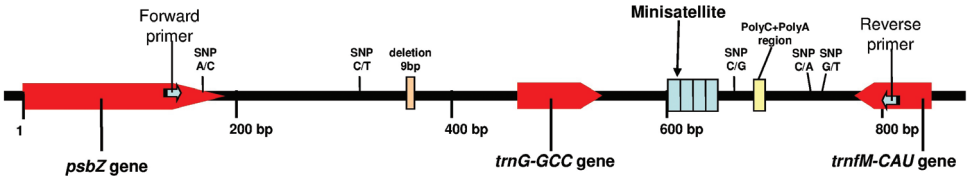


Figure 1. The sequenced cpDNA *psbZ-trnfM* region. The location of PCR primers used and polymorphisms found in this study are shown. DNA fragment length refers to the *P. dactylifera cv. Khalas* cpDNA sequence (Yang et al. 2010), characterised by a 4-repetitions minisatellite haplotype (NCBI Reference Sequence: NC_013991.2).

Phoenix species, identifying five haplotypes, whose pattern of variation was strongly associated with species. The maximum number of haplotypes per species was three (*P. roebelenii*). Yet, most of the haplotypes were shared between species, viz. the 3-repetitions haplotype was the most common haplotype within the genus, and was shared by eight out of the 12 species. *Phoenix canariensis* was the only taxon characterised by the 5-repetitions haplotype. Hence, despite the promising information obtained, the minisatellite alone did not allow to distinguish all *Phoenix* species.

Given the potential of the *trnG*(GCC)*-trnfM*(CAU) spacer for barcoding in *Phoenix*, we examined a wider cpDNA region, viz. a ~700 bp sequence *psbZ-trnfM*(CAU) (Figure 1), in search of an efficient DNA barcode locus for species delimitation and identification of female parents in hybrids in the genus *Phoenix*.

Methods

Taxon sampling

One hundred and thirty-six individuals, belonging to 13 *Phoenix* species, with emphasis on *P. dactylifera,* were analysed in this work (Appendix). *Phoenix andamanensis* was not included in the analysis due to a lack of material.

DNA sequencing

For each sample, genomic DNA was extracted from 40 mg of freeze-dried leaf tissue which was first grinded using a bead-mill homogenizer Tissuelyser (Qiagen, France). Extraction was performed using the DNeasy Plant Mini Kit protocol along with the QIAcube robotic workstation for DNA automated purification (Qiagen, France). Extracted DNA was quantified by means of a Nanodrop ND1000 spectrophotometer (Thermo Fisher Scientific Inc., USA) and visualized on 1% agarose gels stained with ethidium bromide.

The PCR amplification was carried out using the monocotyledoneous universal primers *psbZ*-IGS-F: GGTACMTCATTATGGATTGG, and *trnfM*-IGS-R: GCG-

GAGTAGAGCAGTTTGGT (Scarcelli et al. 2011). The amplified cpDNA fragment was approximately 700 bp long. PCR reactions were prepared in 25 µl of total volume, containing the following reagent concentrations: 5 ng/µl DNA template, 0.2 µM each of forward and reverse primers, 2X Failsafe PCR PreMix E (Epicentre Biotechnologies, Madison USA), 2.5 U/µl Failsafe Enzyme Mix (Epicentre Biotechnologies, USA), and DNase-free sterile water. PCR parameters were the following: an initial denaturation step at 94 °C for 3 min, then 35 cycles at 94 °C for 30 s, 60 °C for 30 s, and 72 °C for 1 min, and a final elongation step at 72 °C for 10 min. PCR products were controlled on 1% agarose gels stained with ethidium bromide and then purified using Ampure Agencourt kit (Agencourt Bioscience Corporation, USA). Their quantification was done by means of a Nanodrop ND1000 spectrophotometer (Thermo Fisher Scientific Inc., USA). Cycle sequencing was carried out using the Big Dye Terminator v3.1 kit (Applied Biosystems, USA). Cycle sequencing products were purified using the CleanSeq Agencourt Kit (Agencourt Bioscience Corporation, USA) and were then analysed on an ABI 3130 automated DNA Sequencer (Applied BioSystems, USA).

Sequence alignment and identification success

The chromatograms obtained with the forward and reverse primers were combined and edited with SeqMan II 5.00 software (DNASTAR Inc., USA), to generate consensus sequences, which were aligned in BioEdit (Hall 1999), using the Clustal W algorithm. The obtained alignment was further improved manually with MESQUITE v2.75 (Maddison and Maddison 2007). The observed polymorphisms were positioned in reference to the complete chloroplast genome sequence of *P. dactylifera* cv. Khalas, available in GenBank (accession NC_013991.2).

To assess the potential of the *psbZ-trnfM* region as a barcode for accurate species identification, we evaluated the proportion of correct identifications using TaxonDNA (Meier et al. 2006). The Best Match and Best Close Match tests were run for species with > 1 individual and with nearly complete sequences, which resulted in a reduced dataset of 11 species (excluding *P. acaulis* and *P. atlantica*) and 121 individuals. Because of this constraint, the two species represented by only one individual were analysed by direct comparison of their sequences. Moreover, direct sequence comparison included not only nucleotide substitutions as in the TaxonDNA analysis, but also indels, minisatellites and homopolymers.

Results

The amplification of the plastid target region *psbZ-trnfM*(CAU) was successful for all samples, and the sequencing with both primers was achieved for 123 individuals, while a single read (forward or reverse) was retrieved for the other 13 individuals, whose sequences were approximately 20% shorter.

Table 1. Distribution of observed polymorphisms in the region *psbZ-trnfM*(CAU)

Substitutions[a]					9 bp deletion [a]	Minisatellite[c]	Homo-polymer [a]	Species[b]
36607	36754	37099	37183	37190	36795–36803	37050–37098	37128–37139	
Haplotypes recorded in a single species[d] (80.1% total sampling)								
C	T	G	A	T	absent	**5M1+1M2**(4)	7 C + 5 A	*P. canariensis* (7)
C	T	**C**	A	T	absent	**2M1+5bp+1M2**(6)	**6 C + 5 A**	*P. reclinata* (4)
C	T	G	A	T	absent	**1M1+2M2**(7)	7 C + 5 A	*P. reclinata* (6)
C	T	G	**C**	T	absent	6M1+1M2(5)	7 C + 5 A	*P. caespitosa* (2)
C	C	G	A	**G**	absent	2M1+1M2(1)	7 C + 5 A	*P. loureiroi* (1)
A	C	G	A	T	absent	3M1+1M2(2)	6 C + 6 A	*P. loureiroi* (1)
A	C	G	A	T	absent	2M1+1M2(1)	7 C + 5 A	*P. acaulis* (1)
C	C	G	A	T	absent	2M1+1M2(1)	6 C + 6 A	*P. pusilla* (2)
C	T	G	A	T	present	2M1+1M2(1)	6 C + 6 A	*P. paludosa* (2)
C	T	G	A	T	present	4M1+1M2(3)	**5 C + 7 A**	*P. roebelenii* (3)
C	T	G	A	T	present	3M1+1M2(2)	**5 C + 7 A**	*P. roebelenii* (1)
C	T	G	A	T	absent	4M1+1M2(3)	7 C + 5 A	*P. dactylifera* (78)
C	T	G	A	T	absent	2M1+1M2(1)	**8 C + 5 A**	*P. sylvestris* (1)
Haplotypes shared by two species (5.1%)								
C	T	G	A	T	absent	6M1+1M2(5)	7 C + 5 A	*P. rupicola* (3)
								P. theophrasti (4)
Haplotypes shared by three species (14.8%)								
C	T	G	A	T	absent	3M1+1M2(2)	7 C + 5 A	*P. atlantica* (1)
								P. dactylifera (16)
								P. sylvestris (3)

[a] Position in the complete chloroplast genome of *Phoenix dactylifera* 'Khalas' accession NC_013991.2.

[b] Number of individuals analysed for each species in parentheses (total sampling of 136 specimens).

[c] Number of repetitions of the 12 bp minisatellite units, including number of units of motif 1 (M1) and motif 2 (M2) as represented in Figure 2.

[d] Species-specific mutations in bold.

(1–7) Minisatellites haplotypes as reported in Figure 2 (1 to 7).

The analysis of the intra- and interspecific variation within the sequenced region by direct observation of the sequence alignment showed four mutation types that contributed to the separation of *Phoenix* species: single nucleotide polymorphisms (SNPs), indels, length variation at the 12 bp minisatellite locus, and in homopolymers, allowing in total to identify unambiguously eight out of the 13 species (Table 1).

The minisatellite located in the *trnG*(GCC)-*trnfM*(CAU) intergenic spacer showed seven haplotypes. Most haplotypes corresponded to a Variable Number Tandem Repeat (VNTR) stepwise mutational pattern of 12 bp units. These units corresponded to two motifs: CTAACTACTATA (motif 1) and GTAGTTAGTATA (motif 2), which form between themselves a pattern of 12 bp inverted repeats shifted with respect to the boundaries of the mutational units (Figure 2). One haplotype, found in four out of ten *P. reclinata* individuals, departed from this pattern, with two complete units of motif 1 plus an incomplete third unit with a 7 bp-deletion (CTAACTA) (haplotype 6; Figure 2). These four specimens were further characterized by a SNP (C instead of G)

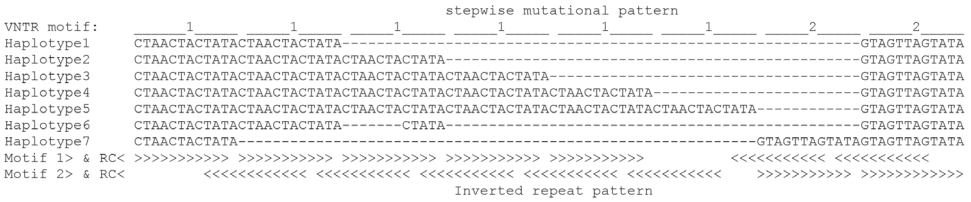


Figure 2. Structure and variation of the minisatellite in the *trnG-trnfM* intergenic spacer. The repeats of the two mutational motifs (1 and 2) are indicated above the sequence alignment of the 7 haplotypes recorded. The pattern of inverted repeats generated by the two motifs and their reverse complements (RC) is shown below the alignment. See Table 1 for haplotype distribution among species.

at position 37099. The other six *P. reclinata* samples were also unique in having two repeats of motif 2 instead of one as in the six other haplotypes (haplotype 7; Figure 2). *Phoenix canariensis* was characterised by a private haplotype with five repeats of motif 1 (haplotype 4; Figure 2). The maximum number of haplotypes per species was three; in that case, one or two of them were shared by different species (Table1).

Additional deletions and SNPs were detected in some of the analysed species, both upstream and downstream of the minisatellite (Table 1). *Phoenix roebelenii* and *P. paludosa* shared a 9 bp-deletion (GTACTTTAC, upstream to the minisatellite, in position 36795–36803), *P. acaulis, P. loureiroi*, and *P. pusilla* shared a SNP at position 36754 (C instead of T), while *P. caespitosa* had a SNP at position 37183 (C instead of A). One of the three samples of *P. loureiroi* showed a SNP in position 37190 (G instead of T), and another sample shared a SNP with the single specimen of *P. acaulis* (A instead of C in position 36607). Some more differences among species and/or individuals were found in an homopolymer region (C_n + A_n), located at position 37128–37139, downstream to the minisatellite (Table 1).

Phoenix theophrasti and *P. rupicola* shared the 6 repetitions of motif 1 minisatellite haplotype (haplotype 5; Figure 2) and could not be distinguished from each other (*P. caespitosa* had also the 6 repetitions haplotype but was further differentiated by a species-specific SNP). Within the "date palm-complex" (Pintaud et al. 2013), a 3 repetitions of motif 1 haplotype was shared by some specimens of *P. atlantica*, *P. dactylifera,* and *P. sylvestris* (haplotype 2; Figure 2), while a majority of individuals of *P. dactylifera* were uniquely identified by a 4 repetitions of motif 1 haplotype (haplotype 3; Figure 2).

The TaxonDNA pairwise comparison analysis of the 121 samples retained resulted in 115 sequences with a closest match at 0%. There were 18 allospecific matches at 0% (15.65%). At the individual level, the Best Match test, and the Best Close Match test with a threshold of 3%, resulted in 82.64% correct identifications, 14.87% ambiguous identifications and 2.47% incorrect identifications. At the species level, however, only *P. caespitosa* could be unambiguously identified, since it was the only species in the sampling with an autapomorphic SNP.

The haplotype sequences used, and the new ones obtained during this study, are deposited in GenBank under accessions: JF745571, EU043486, EU043484, EU043485, JX970915–970936.

Discussion

In this study, we tested the usefulness of the *psbZ-trnfM*(CAU) region as a barcode locus in *Phoenix*. The successful amplification and sequencing of this marker within all of the analysed species confirms its value in terms of universality. Moreover, its high performance should allow the acquisition of barcode information even with partially degraded DNA samples.

TaxonDNA unambiguously identified a single species, *P. caespitosa*, due to the scarcity of SNPs, most of them shared by two or more species, or on the contrary restricted to a subset of individuals within species. Therefore, it is important to take into account the other polymorphisms (indels, minisatellites and homopolymers) which usually represent half or more of the mutations in non-coding chloroplast DNA (Scarcelli et al. 2011). However, at the individual level, the Best Match and Best Close Match tests resulted in more than 80% correct identifications, which is indicative of the barcoding potential of the marker studied.

The 9 bp-deletion, shared by *P. roebelenii* and *P. paludosa*, supports Barrow's conclusions (1998), as well as Pintaud et al.'s (2013), regarding the close relationship between these two taxa.

Regarding the 12 bp minisatellite, our results revealed much more complexity than previously reported (Pintaud et al. 2010). This could be explained by the increased sampling of the present study, and also by differences in methodology, i.e. sequencing versus genotyping. In particular, the genotyping data of Pintaud et al. (2010) did not detect the 7 bp-deletion found within the minisatellite of some *P. reclinata* samples, and were also misled by the size homoplasy between haplotype 1 and 7 (Figure 2). We therefore recommend that sequence data should be obtained before performing any study based on genotyping, in order to have a solid basis to interpret genotyping data.

In total, considering all mutation types, our results allowed us to efficiently identify eight out of 13 species. This indicates that the locus *psbZ-trnfM*(CAU) has some potential to yield DNA barcodes that can be used for species identification within the genus *Phoenix*. This locus could also be useful to identify the female parent in many interspecific crosses, such as *P. dactylifera* × *P. canariensis*. Hybrids involving *P. canariensis* as female parents are particularly easy to track because this species is monomorphic with a private haplotype at the locus studied. Hybrids between these two species are a concern for the genetic integrity of native populations of *P. canariensis* in the Canary Islands (González-Pérez et al. 2004). Such hybrids are also very common in ornamental plantings, for which they represent a valuable horticultural resource.

Nevertheless, in order to increase resolution, other DNA regions should be examined, in search of characters allowing the identification of all taxa. Given their proven utility in palms, the *psbA-trnH* locus (Al-Qurainy et al. 2011) and/or the ribosomal ITS2 (Jeanson et al. 2011) could be investigated in combination with *psbZ-trnfM* for this purpose. Special attention should be paid to the species group sharing haplotype 2 (Figure 2): *P. atlantica*, *P. dactylifera* and *P. sylvestris*. This group is composed of very closely related species, so difficulty in DNA barcoding for these species is expected.

On the other hand, in some cases, the morphological divergence is not associated to sequence divergence in the *psbZ-trnfM* region. For example, *P. rupicola* and *P. theophrasti* share the same haplotype despite considerable morphological differentiation and geographical isolation, the former being restricted to the E Himalayan, while the latter is an Aegean endemic. These two species possibly share plesiomorphic SNP states and may show convergence in the minisatellite haplotype. In contrast, *P. dactylifera* and *P. theophrasti* are phenotypically very similar, but can easily be distinguished at the *psbZ-trnfM*(CAU) region. The relation between morphological divergence and molecular divergence at the *psbZ-trnfM*(CAU) region among the *Phoenix* species needs to be addressed with a larger sampling within species as recommended by the China Plant BOL Group (2011).

Acknowledgements

We wish to thank the following persons and institutions for their valuable help during the various phases of our study: Rita Bregliano†, Paolo Curir, Laura De Benedetti, Federica Nicoletti (CRA-FSO, Sanremo, Italy); Frédérique Aberlenc-Bertossi, Natalie Chabrillange, Nora Scarcelli (IRD Montpellier, France); Muriel Latreille, Sylvain Santoni (AMM at INRA SupAgro, Montpellier, France); Patrizia Martini (IRF, Sanremo, Italy); Mauro Roggero (Cooperativa "Il Cammino", Sanremo, Italy); Giancarlo Pignatta ("U Risveiu Burdigotu" Association, Bordighera, Italy); the Natta family (Bordighera, Italy); Robert Castellana (CRP, Nice, France), Salwa Zehdi (University of Tunis, Tunisia). We also thank the three anonymous reviewers and the scientific editor for their useful comments and suggestions.

References

Al-Qurainy F, Khan S, Al-Hemaid FM, Ajmal Ali A, Tarroum M, Ashraf M (2011) Assessing molecular signature for some potential date (*Phoenix dactylifera* L.) Cultivars from Saudi Arabia, Based on Chloroplast DNA Sequences *rpoB* and *psbA-trnH.* International Journal of Molecular Sciences 12: 6871–6880. doi: 10.3390/ijms12106871

Asmussen CB, Dransfield J, Deickmann V, Barfod AS, Pintaud J-C, Baker W (2006) A new subfamily classification of the palm family (Arecaceae): evidence form plastid DNA phylogeny. Botanical Journal of the Linnean Society 151: 15–38. http://pure.au.dk/portal/files/43885556/Asmussen_et_al_2006.pdf, doi: 10.1111/j.1095-8339.2006.00521.x

Baker WJ, Savolainen V, Asmussen-Lange CB, Chase M, Dransfield J, Forest F, Harley M, Uhl N, Wilkinson M (2009) Complete Generic-level Phylogenetic Analyses of Palms (Arecaceae) with Comparisons of Supertree and Supermatrix Approaches. Systematic Biology 58: 240–256. doi: 10.1093/sysbio/syp021

Barrow S (1998) A monograph of *Phoenix* L. (Palmae: Coryphoideae). Kew Bulletin 53: 513–575. doi: 10.2307/4110478

Boydak M, Barrow S (1995) A new locality for *Phoenix* in Turkey: Gölköy-Bödrum. Principes 39: 117–122.

CBOL Plant Working Group (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the USA 106: 12794–12797. doi: 10.1073/pnas.0905845106

Chen S, Yao H, Han J, Liu C, Song J (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5: e8613. doi: 10.1371/journal.pone.0008613

China Plant BOL Group (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proceedings of the National Academy of Sciences of the USA 108: 19641–19646. doi: 10.1073/pnas.1104551108

DeSalle R, Egan M, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1905–1916. doi: 10.1098/rstb.2005.1722

Dransfield J, Uhl NW, Asmussen CB, Baker WJ, Harley MM, Lewis CE (2008) Genera palmarum, the evolution and classification of palms. Royal Botanic Gardens, Kew, U.K.

González-Pérez MA, Caujapé-Castells J, Sosa PA (2004) Molecular evidence of hybridisation between the endemic *Phoenix canariensis* and the widespread *P. dactylifera* with Random Amplified Polymorphic DNA (RAPD) markers. Plant Systematics and Evolution 247: 165–175. doi: 10.1007/s00606-004-0166-7

Govaerts R, Dransfield J (2005) World checklist of palms. Royal Botanic Gardens, Kew, U.K.

Greuter W (1967) Beiträge zur Flora der Südägäis 8-9. Bauhinia 3: 243–254.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends in Genetics 23: 167–172. doi: 10.1016/j.tig.2007.02.001

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

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) 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 USA 101: 14812–14817. doi: 10.1073/pnas.0406166101

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Systematic Biology 54: 852–859. doi: 10.1080/10635150500354886

Henderson SA, Billotte N, Pintaud J-C (2006) Genetic isolation of Cape Verde Island *Phoenix atlantica* (Arecaceae) revealed by microsatellite markers. Conservation Genetics 7: 213–223. doi: 10.1007/s10592-006-9128-7

Jeanson ML, Labat J-N, Little DP (2011) DNA barcoding: a new tool for palm taxonomists? Annals of Botany 108: 1445–1451. doi: 10.1093/aob/mcr158

Maddison W, Maddison D (2007) Mesquite 2. http://mesquiteproject.org

Meier R, Kwong S, 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. doi: 10.1080/10635150600969864

Nesbitt M (1993) Archaeobotanical evidence for early Dilmun diet at Saar, Bahrain. Arabian Archaeology Epigraphy 4: 20–47. doi: 10.1111/j.1600-0471.1993.tb00041.x

Newton C, Gros-Balthazard M, Ivorra S, Paradis L, Pintaud J-C, Terral JF (2013) *Phoenix dactylifera* and *P. sylvestris* in Northwestern India: a glimpse of their complex relationships. Palms 57: 37–50.

Pintaud J-C, Zehdi S, Couvreur T, Barrow S, Henderson S, Aberlenc-Bertossi F, Tregear J, Billotte N (2010) Species delimitation in the genus *Phoenix* (Arecaceae) based on SSR markers, with emphasis on the identity of the Date palm (*Phoenix dactylifera*). In: Seberg O, Petersen G, Barfod A, Davis J (Eds) Taxonomy of *Phoenix*. Diversity, phylogeny, and evolution in the Monocotyledons. Aarhus University Press, Denmark, 267–286.

Pintaud J-C, Ludeña B, Aberlenc-Bertossi F, Zehdi S, Gros-Balthazard M, Ivorra S, Terral J-F, Newton C, Tengberg M, Abdoulkader S, Daher A, Nabil M, Saro Hernández I, González-Pérez MA, Sosa P, Santoni S, Moussouni S, Si-Dehbi F, Bouguedoura N (2013) Biogeography of the Date Palm (*Phoenix dactylifera* L., *Arecaceae*): Insights on the Origin and on the Structure of Modern Diversity. In: Bouguedoura N, Bennaceur M, Pintaud J-C (Eds) Proceedings of the I International Symposium on Date Palm. Acta Horticulturae 994: 19–38. http://www.actahort.org/books/994/994_1.htm

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1805–1811. doi: 10.1098/rstb.2005.1730

Scarcelli N, Barnaud A, Eiserhardt W, Treier UA, Seveno M, d'Anfray A, Vigouroux Y, Pintaud J-C (2011) A Set of 100 chloroplast primer pairs to study population genetics and phylogeny in Monocotyledons. PLoS ONE 6: e19954. doi: 10.1371/journal.pone.0019954

Tengberg M (2012) Beginnings and early history of date palm garden cultivation in the Middle East. Journal of Arid Environments 86: 139–147. doi: 10.1016/j.jaridenv.2011.11.022

Tournay F (2009) The Nabonnand family and palms. Palms 53: 119–123.

Wrigley G (1995) Date palm (*Phoenix dactylifera* L). In: Smartt J, Simmonds NW (Eds) The Evolution of Crop Plants. Longman, Essex, 399–403.

Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem I, Yu J (2010) The complete chloroplast genome sequence of Date Palm (*Phoenix dactylifera* L.). PLoS ONE 5: e12762. doi: 10.1371/journal.pone.0012762

Yao H, Song J, Liu C, Luo K, Han J, Li Y, pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE 5: e13102. doi: 10.1371/journal.pone.0013102

Zohary D, Hopf M (2000) Domestication of plants in the Old World. The origin and spread of cultivated plants in the West Asia, Europe and the Nile Valley. 3rd Ed. Oxford University Press, New York.

Appendix

List of samples. DNA bank reference: IRD = Institut de Recherche pour le Développement, 911 Av. Agropolis, F-34394 Montpellier Cedex 5, France. Tissue bank reference: CRA-FSO = Consiglio per la Ricerca e la sperimentazione in Agricoltura - Unità di Ricerca per la Floricoltura e le Specie Ornamentali, Corso degli Inglesi 508, I-18038 Sanremo (IM), Italy.

Phoenix acaulis* Roxb**: MWC5559, Kew ,UK (IRD). ***Phoenix atlantica* A. Chev.**: SH25, Cape Verde (IRD). ***Phoenix caespitosa* Chiov.**: MWC1195, MWC1802, Kew, UK (IRD). ***Phoenix canariensis* Chabaud**: 93.100, 93.101, 93.103, 93.107, Sanremo, Italy (IRD, CRA-FSO); MWC1396, Kew, UK (IRD); JCP169, JCP210, Canary Isl., Spain (IRD). ***Phoenix dactylifera* L.**: 93.003, 93.004, 93.005, 93.025, 93.027, 93.030, 93.037, 93.043, 93.045, 93.047, 93.048, 93.049, 93.052, 93.054, 93.055, 93.056, 93.059, 93.060, 93.061, 93.065, 93.066, 93.067, 93.070, 93.071, 93.072, 93.073, 93.076, 93.077, 93.080, 93.085, 90.002, 90.003, 90.004, 90.005, 90.006, 90.007, 90.008, 90.009, 90.010, 90.011, 90.012, 90.013, 90.014, 90.015, 90.025, 90.026, 90.027, 90.028, 90.029, 91.005, Sanremo, Italy (IRD, CRA-FSO); 00.01, 00.02, 00.03, 00.04, 00.05, 00.06, 00.07, 00.08, 00.09, 00.10, 00.11, 00.13, 00.14, 00.83, 00.85, 00.88, 46.02, 46.04, 46.05, 46.06, 46.08, 46.09, 46.14, 46.15, 46.16, 46.17, 46.18, 46.19, 46.20, 46.21, 46.23, JCP413, JCP414, JCP415, JCP416, JCP417, Bordighera, Italy (IRD, CRA-FSO); DAT077-365, DAT079-366, Oman (IRD); JCP260, Murcia, Spain; JCP426, Elche, Spain; SZ1, SZ2, SZ5, SZ10, Tunisia (IRD). ***Phoenix loureiroi* Kunth *var. loureiroi: JCP409, Montgomery Botanical Garden, Miami, USA (IRD); MWC1187, Kew, UK (IRD). ***Phoenix paludosa* Roxb.**: MWC1190, MWC1877, Kew, UK (IRD). ***Phoenix pusilla* Gaertn.**: JCP213_5, Sri Lanka (IRD); MWC1806, Kew, UK (IRD). ***Phoenix reclinata* Jacq.**: ECH3-A, ECH4-A, ECH5-B, 91.001, 91.007, 91.008, 91.009, 91.033, 92.003, Sanremo, Italy (IRD, CRA-FSO); MWC1397, Kew, UK (IRD). ***Phoenix roebelenii* O'Brien**: ECH1-A, ECH2-A, Sanremo, Italy (IRD, CRA-FSO); MWC1400, MWC1805, Kew, UK (IRD). ***Phoenix rupicola* T. Anderson**: ECH6-A, ECH8-A, Sanremo, Italy (IRD, CRA-FSO); MWC1399, Kew, UK (IRD). ***Phoenix sylvestris* (L.) Roxb.**: DAT057-345, Elche, Spain (IRD); JCP214, The Palm Center, UK, (IRD); JCP405-388, Thuret, France (IRD); MWC1876, Kew, UK (IRD). ***Phoenix theophrasti* Greuter**: ECH7-A, ECH9-A, Sanremo, Italy (IRD, CRA-FSO); JCP215, The Palm Center, UK (IRD); MWC1163, Kew, UK.

ZooKeys 365: 83–104 (2013)
doi: 10.3897/zookeys.365.6104
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

DNA barcodes and phylogenetic affinities of the terrestrial slugs *Arion gilvus* and *A. ponsi* (Gastropoda, Pulmonata, Arionidae)

Karin Breugelmans[1], Kurt Jordaens[2,3], Els Adriaens[4], Jean Paul Remon[4], Josep Quintana Cardona[5], Thierry Backeljau[1,3]

1 *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (JEMU), Vautierstraat 29, B-1000 Brussels, Belgium* **2** *Royal Museum for Central Africa (JEMU), Leuvensesteenweg 13, B-3080 Tervuren, Belgium* **3** *Evolutionary Ecology Group, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium* **4** *Laboratory of Pharmaceutical Technology, University of Ghent, Harelbekestraat 72, B-9000 Ghent, Belgium* **5** *Institut Catala de Paleontologia Miquel Crusafont, Universitat Autònomo de Barcelona, edifici ICP Campus de la UAB, sln 08193 Cerdanyola del Vallès, Barcelona, Spain*

Corresponding author: *Thierry Backeljau* (thierry.backeljau@naturalsciences.be)

Academic editor: *M. de Meyer* | Received 14 August 2013 | Accepted 6 December 2013 | Published 30 December 2013

Citation: Breugelmans K, Jordaens K, Adriaens E, Remon JP, Cardona JQ, Backeljau T (2013) DNA barcodes and phylogenetic affinities of the terrestrial slugs *Arion gilvus* and *A. ponsi* (Gastropoda, Pulmonata, Arionidae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 83–104. doi: 10.3897/zookeys.365.6104

Abstract

The Iberian Peninsula is a region with a high endemicity of species of the terrestrial slug subgenus *Mesarion*. Many of these species have been described mainly on subtle differences in their proximal genitalia. It therefore remains to be investigated 1) whether these locally diverged taxa also represent different species under a phylogenetic species concept as has been shown for other *Mesarion* species outside the Iberian Peninsula, and 2) how these taxa are phylogenetically related. Here, we analysed DNA sequence data of two mitochondrial (COI and 16S) genes, and of the nuclear ITS1 region, to explore the phylogenetic affinities of two of these endemic taxa, viz. *Arion gilvus* Torres Mínguez, 1925 and *A. ponsi* Quintana Cardona, 2007. We also evaluated the use of these DNA sequence data as DNA barcodes for both species. Our results showed that ITS did not allow to differentiate among most of the *Mesarion* molecular operational taxonomic units (MOTUs) / morphospecies in *Mesarion*. Yet, the overall mean p-distance among the *Mesarion* MOTUs / morphospecies for both mtDNA fragments (16.7% for COI, 13% for 16S) was comparable to that between *A. ponsi* and its closest relative *A. molinae* (COI: 14.2%; 16S: 16.2%) and to that between *A. gilvus* and its closest relative *A. urbiae* (COI: 14.4%; 16S: 13.4%). Hence, with respect to mtDNA divergence, both *A. ponsi* and *A. gilvus*, behave as other *Mesarion* species or putative species-level MOTUs and thus are confirmed as distinct 'species'.

Keywords
DNA barcoding, terrestrial slugs, Gastropoda, taxonomy, Iberian Peninsula, *Arion ponsi*, *Arion gilvus*

Introduction

The genus *Arion* Férussac, 1819 is the most species rich genus of the terrestrial slug family Arionidae (Mollusca, Pulmonata, Gastropoda). It comprizes approximately 40 species, grouped into four subgenera, viz. *Arion* s.s. Férussac, 1819, *Kobeltia* Seibert, 1873, *Carinarion* Hesse, 1926 and *Mesarion* Hesse, 1926. Species of the subgenus *Mesarion* (type species: *Limax subfuscus* Draparnaud, 1805) are characterized by 1) a medium body-size (up to 75 mm when extended), 2) an orange to dark brown dorsum, 3) two dark bands on the sides of the mantle, 4) (usually) yellow to orange body mucus, and 5) an enlarged free-oviduct with a long and V-shaped ligula (Kerney et al. 1983). Many *Mesarion* species are highly polymorphic with respect to body colour and genital anatomy. As a consequence, the species limits and phylogenetic relationships of taxa within this subgenus have been debated for decades (e.g. Garrido et al. 1995, Castillejo 1997, 1998, Pinceel et al. 2004, 2005a, b, Quinteiro et al. 2005). *Arion subfuscus* (Draparnaud, 1805) (type locality: Montagne Noire, France) is probably the most problematic "species" within *Mesarion* as it shows an overwhelming amount of variation in body pigmentation, genital anatomy, and reproductive behavior [see Garrido et al. (1995) and the references listed in their table 1]. This variation often has been interpreted as indicating reproductive isolation between geographically isolated populations, and *A. subfuscus* thus is considered a species complex (Wiktor 1973, Waldén 1976, De Winter 1986, Backeljau 1989, Altonaga et al. 1994, Backeljau et al. 1994, Garrido et al. 1995). Especially in the Pyrenees and the coastal regions of Spain there are local, morphologically diverged populations (e.g. Garrido et al. 1995, Castillejo 1998). Several of these have been described as endemic species on the basis of where the epiphallus, oviduct and pedunculus of the bursa copulatrix open into the atrium, in combination with differences in the relative lengths of the vas deferens and the epiphallus (e.g. Castillejo 1998, Garrido et al. 1995, Quintana Cardona 2007). Two of these endemic taxa occur in the eastern coastal region of Spain or the Balearic Islands, viz. *Arion gilvus* Torres Mínguez, 1925 and *A. ponsi* Quintana Cardona, 2007.

Arion ponsi (Figure 1) was described from Menorca (Balearic Islands, type locality: Barranc d'Algendar). The species has a medium body size (range: 54–66 mm), an orange to beige dorsal body colour with dark lateral bands that can be blurry in the posterior parts, a foot sole that is cream coloured with a greyish hue, and a transparent body mucus (Quintana Cardona 2007). Its genital anatomy is very similar to that of *A. gilvus*, *A. iratii* Garrido, Castillejo & Iglesias, 1995, *A. molinae* Garrido, Castillejo & Iglesias, 1995 and *A. lizarrustii* Garrido, Castillejo & Iglesias, 1995, but its epiphallus is shorter than the vas deferens (as in *A. molinae*) and opens into the genital atrium in

Figure 1. *Arion ponsi* Quintana Cardona, 2007 from Menorca (Balearic Islands, Spain).

between the oviduct and the pedunculus of the bursa copulatrix (unlike in *A. molinae*, where the pedunculus is positioned in between the epiphallus and oviduct) (figures 3–5 in Quintana Cardona 2007).

Arion gilvus (Figure 2) was described from 'Mandol' in the Spanish Province of Tarragona. However, the toponym 'Mandol' seems to be erroneous (e.g. Bech 1990) and therefore Castillejo (1990) assigned eight specimens with an *A. gilvus* morphology from Serra de Pandóls near Gandesa (Province of Tarragona) as topotypes [see also Castillejo and Rodríguez (1991)]. Subsequently, *A. gilvus* was redescribed by Garrido (1992). Afterwards, the species has also been found in the Provinces of Valencia, Teruel and Albacete [Borredà (1994), figure 15 in Castillejo (1997), figure 1 in Quinteiro et al. (2005)]. *Arion gilvus* reaches a length of up to 65 mm when extended. It has a yellowish to brown dorsum that gets lighter downwards at the sides and dark lateral bands that have a yellowish grey line on their upper side (Figure 1). The sole is white or evenly yellowish and the mucus is pale yellow (Torres Mínguez 1925, Bech 1990, Garrido 1992, Castillejo 1997). The epiphallus, the pedunculus of the bursa copulatrix, and the free oviduct join the atrium on a single line with the pedunculus of the bursa copulatrix in the middle, as in *A. molinae*, but in contrast to the latter, the epiphallus is longer than the vas deferens (Torres Mínguez 1925, Borredà 1994, Castillejo 1997, and figures 26–28 in Garrido et al. 1995).

Figure 2. *Arion gilvus* Torres Mínguez, 1925 from Serra de Pandóls (Valencia, Spain). **A** dorsal view **B** lateral view **C** ventral view.

As illustrated by *Arion ponsi* and *A. gilvus*, the alleged species-specific genital differences among the Iberian species of the *A. subfuscus* complex are very subtle and little is known about their intraspecific variation. Moreover, genital differences among arionid taxa do not necessarily imply reproductive isolation (Dreijers et al. 2013). Hence, if alleged species-specific phenotypic differences in arionids are to be interpreted under a phylogenetic species concept, then their correlation with reproductive isolation should be corroborated by molecular data. Molecular markers have been very effective in this respect (e.g. Pinceel et al. 2005a, b, Quinteiro et al. 2005, Geenen et al. 2006, Jordaens et al. 2010). As such, Quinteiro et al. (2005) investigated the taxonomic affinities of Iberian *Mesarion* species using DNA sequence data. Their analysis of the nuclear ribosomal internal transcribed spacer 1 region (ITS1) showed a polytomy of *Mesarion*

species, yet, the analysis of the mitochondrial NADH dehydrogenase I (ND1) gene suggested a strongly bootstrap supported group of Iberian *Mesarion* species with a continental-Mediterranean distribution (*A. paularensis*, *A. baeticus*, *A. urbiae*, *A. anguloi*, *A. wiktori*, and *A. gilvus*), and an unsupported group of species with an Atlantic distribution (*A. lusitanicus*, *A. nobrei*, *A. fuligineus*, *A. hispanicus* and *A. flagellus*). In addition, the positions of three Pyrenean species (*A. lizarrustii*, *A. iratii*, *A. molinae*) remained unresolved. More specifically, the ND1 data placed *A. gilvus* as sister taxon of *A. urbiae* and *A. anguloi*. Quinteiro et al. (2005) did not study individuals from the Balearic Islands and thus probably did not include *A. ponsi*.

Because DNA sequence data do not only provide phylogenetic information, but can also serve as DNA barcodes for species identification (Hebert et al. 2003, 2004), we here expand on the work of Quinteiro et al. (2005) by 1) characterizing *A. gilvus* and *A. ponsi* using mitochondrial COI and 16S rDNA gene fragments, and the larger part of the nuclear ITS1 region, 2) exploring the phylogenetic affinities of *A. gilvus* and *A. ponsi* within the subgenus *Mesarion*, and 3) providing diagnostic COI barcodes for both species.

Material and methods

Information on the species and specimens included here is provided in Table 1. In total, we screened 45 specimens (Table 1). DNA was extracted from small parts of the foot using a NucleoSpin Tissue Kit (Macherey-Nagel, Düren) following the manufacturer's instructions. PCR reactions were done in 25 µl reaction volumes that contained 1.5 mM $MgCl_2$ in 1 × PCR buffer (Qiagen), 0.2 mM of each dNTP, 0.2 µM of each primer and 0.5 units of Taq polymerase (Qiagen). A fragment of the mitochondrial COI and 16S genes was amplified using primer pairs LCO1490 and HCO2198 (Folmer et al. 1994) and 16Sar and 16Sbr (Palumbi 1996), respectively. The nuclear ITS1 region (except the ± first 30 bp) was amplified using the primer pair ITS1L and 58C (Hillis and Dixon 1991). The PCR profile was an initial denaturation step of 5 min at 95 °C, followed by 35 cycles of 45 s at 95 °C, 45 s at an annealing temperature of 40 °C (COI), 42 °C (16S) or 55 °C (ITS1) and 1.5 min at 72 °C, and ending with a final extension step of 5 min at 72 °C. PCR products were purified using the GFX PCR DNA Purification Kit (GE Healthcare) following the manufacturer's instructions. Purified DNA was diluted in 15 µl of sterile water. PCR-products were bidirectionally sequenced using the ABI PRISM BigDye® Terminator v1.1 Cycle Sequencing Kit and run on a ABI3130xl Genetic Analyzer. Sequences were assembled in SeqScape v2.5 (Life Technologies) and inconsistencies were checked by eye on the chromatogram. Sequences were submitted to GenBank under accession numbers KF305196–KF305225 for COI, KF356212–KF356245 for 16S and KF385449–KF385469 for ITS1. These datasets were supplemented with DNA sequences from GenBank [including a few species of the other *Arion* subgenera (Table 1)]. We used those of *Carinarion* as outgroup.

Sequences were aligned in ClustalW (Thompson et al. 1994) with default settings and without subsequent manual adjustments. In each alignment sequences were trimmed

Table 1. List of specimens used in this study with specimen ID, sampling locality, GenBank accession numbers for the COI, 16S and ITS1 sequences, and collection number at the museums (if available). Neo-, para- and topotypes have been indicated. Specimen codes with an asterisk are data taken from Quinteiro et al. (2005); NA = not assessed. The specimen ID and GenBank accession numbers of newly sequenced specimens are given in bold.

Species/ID	Locality, country	COI	16S	ITS1	Collection number
Subgenus *Mesarion* Hesse, 1926					
Arion anguloi Martín and Gómez, 1988					
ang-SU2777	Torralba del Rio, Spain	AY987869	AY947348	AY947386	RBINS Brussels, I.G. 32471
ang-115 (topotype)	Osma, Álava, Spain	**KF305196**	**KF356212**	AJ509055	RBINS Brussels, I.G. 32471
AANG 73A*	Burgos, Spain	NA	NA	AY316291	
Arion baeticus Garrido, Castillejo and Iglesias, 1995					
bae-556 (paratype)	Malaga, Spain	AY987871	AY947350	AJ509054	MNCN Madrid 15.05/6969
Arion flagellus Colligne, 1893					
fla-130	Glasgow, UK	AY987880	AY947359	AJ509053	RBINS Brussels, I.G. 32471
fla-161	Glasgow, UK	AY987881	AY947360	AJ509052	RBINS Brussels, I.G. 32471
fla-SU672	Salamir, Spain	AY987882	AY947361	AY947388	RBINS Brussels, I.G. 32471
AFLA 44A*	Croydon, UK	NA	NA	AY316278	
Arion fuligineus Morelet, 1845					
AFUL 43A*	São Silvestre, Portugal	NA	NA	AY316277	
Arion fuscus (Müller, 1774)					
fus-SU155	Grudki, Poland	AY987885	AJ786721	AY947390	RBINS Brussels, I.G. 32471
fus-2320	Predel, Bulgaria	AY987886	AJ786722	AY947391	RBINS Brussels, I.G. 32471
fus-SU1335	Steinegg, Austria	AY987887	AJ786726	AY947392	RBINS Brussels, I.G. 32471
fus-SU2188	Kreuzen, Austria	NA	**KF356221**	NA	RBINS Brussels, I.G. 32471
Arion gilvus Torres Mínguez, 1925					
gil-46	Serra de Pandóls, Valencia, Spain	NA	NA	**KF385450**	RBINS Brussels, I.G. 32471
gil-47	Serra de Pandóls, Valencia, Spain	**KF305199**	**KF356222**	**KF385451**	RBINS Brussels, I.G. 32471
gil-73	Serra de Pandóls, Valencia, Spain	**KF305200**	**KF356223**	**KF385452**	RBINS Brussels, I.G. 32471
AGIL 49A*	Serra de Pandóls, Valencia, Spain	NA	NA	AY316282	
Arion hispanicus Simroth, 1886					
AHIS 52B*	Cáceres, Spain	NA	NA	AY316285	

Species/ID	**Locality, country**	**COI**	**16S**	**ITS1**	**Collection number**
Arion iratii Garrido, Castillejo and Iglesias, 1995					
ira-559 (paratype)	Navarra, Spain	AY987892	AY947367	AJ509042	MNCN Madrid, 15.05/18705
Arion lizarrustii Garrido, Castillejo and Iglesias, 1995					
liz-562 (paratype)	Navarra, Spain	AY987893	AY947368	AJ509046	MNCN Madrid, 15.05/18706
ALIZ 47C*	Lizarrusti, Spain	NA	NA	AY316280	
Arion lusitanicus Mabille, 1868					
lus-1613	Feitos, Portugal	**KF305203**	**KF356224**	NA	RBINS Brussels, I.G. 32471
lus-1631	Currais, Portugal	**KF305204**	**KF356225**	NA	RBINS Brussels, I.G. 32471
lus-1641	Cacia, Portugal	**KF305205**	NA	NA	RBINS Brussels, I.G. 32471
lus-1647	Cacia, Portugal	**KF305206**	NA	NA	RBINS Brussels, I.G. 32471
lus-1652	Forjães, Portugal	**KF305207**	**KF356226**	NA	RBINS Brussels, I.G. 32471
lus-1654	Currais, Portugal	NA	**KF356227**	NA	RBINS Brussels, I.G. 32471
lus-1655	Forjães, Portugal	**KF305208**	**KF356228**	NA	RBINS Brussels, I.G. 32471
lus-79	Ursel, Belgium	AY987894	AY947369	AJ509062	RBINS Brussels, I.G. 32471
lus-181	Terceira, Azores, Portugal	NA	NA	**KF385453**	RBINS Brussels, I.G. 32471
lus-186	Namur, Belgium	AY987895	AY947370	AJ509061	RBINS Brussels, I.G. 32471
lus-465	Görlitz, Germany	NA	NA	AJ509063	RBINS Brussels, I.G. 32471
lus-509	Emptinne, Belgium	**KF305209**	**KF356229**	NA	RBINS Brussels, I.G. 32471
ALUS 42A*	Serra de Arrábida, Portugal	NA	NA	AY316273	
ALUS 42B*	Serra de Arrábida, Portugal	NA	NA	AY316274	
ALUS 42C*	Serra de Arrábida, Portugal	NA	NA	AY316275	
ALUS 42G*	Alpi Carniche, Rivolato, Italy	NA	NA	AY316276	
ALUS 62E*	Montagne Noire, France	NA	NA	AY316289	
ALUS 70C*	Girona, Spain	NA	NA	AY316290	
Arion molinae Garrido, Castillejo and Iglesias, 1995					
mol-565 (paratype)	La Molina, Spain	AY987896	AY947371	AJ509043	MNCN Madrid, 15.05/18707
AMOL 48A*	Serra del Cadí, Barcelona, Spain	NA	NA	AY316281	
Arion nobrei Pollonera, 1889					
ANOB 41A*	Luso, Portugal	NA	NA	AY316271	

Species/ID	Locality, country	COI	16S	ITS1	Collection number
ANOB 41B*	Luso, Portugal	NA	NA	AY316272	
Arion paularensis Wiktor and Parejo, 1989					
pau-121	Sierra de Guadarrama, Spain	**KF305210**	NA	**KF385454**	RBINS Brussels, I.G. 32471
pau-224	Sierra de Guadarrama, Spain	AY987899	AY947374	AJ509057	RBINS Brussels, I.G. 32471
pau-226	Sierra de Guadarrama, Spain	NA	NA	**KF385455**	RBINS Brussels, I.G. 32471
APAU 51A*	Sierra de Guadarrama, Spain	NA	NA	AY316284	
Arion ponsi Quintana Cardona, 2007					
pon-1959	Ciutadella de Menorca, Spain	**KF305211**	**KF356230**	**KF385456**	RBINS Brussels, I.G. 32471
pon-1960	Ferreries, Menorca, Spain	**KF305212**	**KF356231**	**KF385457**	RBINS Brussels, I.G. 32471
pon-1962	Ciutadella de Menorca, Spain	**KF305213**	**KF356232**	**KF385458**	RBINS Brussels, I.G. 32471
pon-1965	Ciutadella de Menorca, Spain	**KF305214**	**KF356233**	**KF385459**	RBINS Brussels, I.G. 32471
Arion subfuscus (Draparnaud, 1805)					
sub1-2312	Kortrijk, Belgium	**KF305215**	**KF356238**	**KF385461**	RBINS Brussels, I.G. 32471
sub1-2318	Ingrandes, France	AY987904	AY860678	AY860729	RBINS Brussels, I.G. 32471
sub1-2317	Burnopfield, UK	AY987906	AY860672	AY860726	RBINS Brussels, I.G. 32471
sub1-SU87	Barnstable, MA, USA	NA	KF356235	NA	RBINS Brussels, I.G. 32471
sub1-1618	Wilrijk, Belgium	**KF305216**	**KF356236**	NA	RBINS Brussels, I.G. 32471
sub1-1633	Wilrijk, Belgium	**KF305217**	**KF356237**	NA	RBINS Brussels, I.G. 32471
sub2-SU2424	Heppenbach, Belgium	NA	**KF356239**	NA	RBINS Brussels, I.G. 32471
sub2-SU349	Grootrees, Belgium	NA	**KF356240**	NA	RBINS Brussels, I.G. 32471
sub2-2309	Gomzé, Belgium	**KF305218**	**KF356241**	**KF385462**	RBINS Brussels, I.G. 32471
sub2-2313	Le Landin, France	AY987908	AY860687	**KF385463**	RBINS Brussels, I.G. 32471
sub2-2314	Heppenbach, Belgium	AY987909	AY860709	**KF385464**	RBINS Brussels, I.G. 32471
sub3-2322	Bucholz, Germany	AY987910	AY860716	**KF385466**	RBINS Brussels, I.G. 32471
sub3-2310	La Salle, France	AY987911	AY860722	**KF385465**	RBINS Brussels, I.G. 32471
sub3-SU2401	La Salle, France	NA	**KF356242**	NA	RBINS Brussels, I.G. 32471
sub4-123 (topotype)	Montagne Noire, France	AY987913	AY860682	AY860733	RBINS Brussels, I.G. 32471
sub4-568 (neotype)	Montagne Noire, France	AY987914	**KF356244**	AJ509044	MNCN Madrid, 15.05/18704
sub4-2341	Oulès, France	AY987912	AY860685	AY860731	RBINS Brussels, I.G. 32471
sub4-SU1058	Col de Peyresourde, France	NA	**KF356243**	NA	RBINS Brussels, I.G. 32471

Species/ID	Locality, country	COI	16S	ITS1	Collection number
sub5-2321	Villemont-Baubiat, France	AY987915	AY860681	**KF385468**	RBINS Brussels, I.G. 32471
sub5-2311	Villemont-Baubiat, France	AY987916	AY860679	**KF385467**	RBINS Brussels, I.G. 32471
ASUB 45A*	Montagne Noire, France	NA	NA	AY316279	
Arion transsylvanus Simroth, 1885					
tra-SU1088	Covasna, Romania	AY943858	AY860798	AY947393	RBINS Brussels, I.G. 30412
tra-SU1203	Lunca Vişagului, Romania	AY943859	AY860805	AY947394	RBINS Brussels, I.G. 30412
tra-SU1296	Holda, Romania	AY943860	AY860799	AY947395	RBINS Brussels, I.G. 30412
Arion urbiae De Winter, 1986					
urb-SU2755	Saldaña, Spain	AY987919	AY947381	AY947396	RBINS Brussels, I.G. 32471
urb-99	Sierra da Urbia, Spain	NA	NA	**KF385469**	RBINS Brussels, I.G. 32471
AURB 50A*		NA	NA	AY316283	
Subgenus *Kobeltia* Seibert, 1873					
Arion distinctus Mabille, 1868					
dis-106	Mortsel, Belgium	AY987875	AY947354	AJ509040	RBINS Brussels, I.G. 32471
dis-14		AY987874	AY947353	AJ509038	RBINS Brussels, I.G. 32471
Arion hortensis Férussac, 1819					
hor-102	Mortsel, Belgium	AY987888	AJ518061	AJ509037	RBINS Brussels, I.G. 32471
hor-220	London, UK	AY987889	AY947364	AJ509036	RBINS Brussels, I.G. 32471
Arion intermedius Normand, 1852					
int-104	Rochefort, Belgium	AY987891	AY947366	AJ509031	RBINS Brussels, I.G. 32471
int-52	Flores, Azores, Portugal	AY987890	AY947365	AJ509029	RBINS Brussels, I.G. 32471
Arion obesoductus Reischütz, 1973					
alp-1610	Žďárské Vrchy, Czech Republic	DQ904249	DQ904248	NA	RBINS Brussels, I.G. 32471
alp-208	Saxony, Germany	AY987867	AY947346	AJ509041	RBINS Brussels, I.G. 32471
Arion owenii Davies, 1979					
owe-310	Devon, UK	AY987897	AY947372	AJ509033	RBINS Brussels, I.G. 32471
owe-316	Devon, UK	AY987898	AY947373	AJ509034	RBINS Brussels, I.G. 32471
Arion wiktori Parejo & Martín, 1990					
wik-SU2693	Viniegra de Abajo, Spain	AY987921	AY947383	AY947397	RBINS Brussels, I.G. 32471

Species/ID	Locality, country	COI	16S	ITS1	Collection number
wik-44	Burgos, Spain	AY987920	AY947382	AJ509060	RBINS Brussels, I.G. 32471
wik-94	Burgos, Spain	NA	**KF356245**	AJ509059	RBINS Brussels, I.G. 32471
AWIK 58A*	Demanda Sierra, Burgos, Spain	NA	NA	AY316287	
AWIK 58C*	Urbión Mountains, Soria, Spain	NA	NA	AY316288	
Subgenus *Carinarion* Hesse, 1926					
Arion circumscriptus Johnston, 1828					
cir-151	Aran Island, Kilmurvey, Ireland	AY987872	AY947351	AJ509071	RBINS Brussels, I.G. 32471
Arion fasciatus (Nilsson, 1823)					
fas-144	Görlitz, Germany	AY987877	AY947356	AJ509068	RBINS Brussels, I.G. 32471
Arion silvaticus Lohmander, 1937					
sil-142	Poulseur, Belgium	AY987917	AY947379	AJ509070	RBINS Brussels, I.G. 32471
Subgenus *Arion* s.s. Férussac, 1819					
Arion ater-rufus complex					
ate-SU517	Musland, Norway	AY987870	AY947349	AY947387	RBINS Brussels, I.G. 32471
ate/ruf-1602	Manteigas, Portugal	**KF305219**	NA	**KF385449**	RBINS Brussels, I.G. 32471
ate/ruf-1619	Santa Leocádia, Portugal	**KF305220**	**KF356213**	NA	RBINS Brussels, I.G. 32471
ate/ruf-1620	Gortmore, Ireland	**KF305221**	**KF356214**	NA	RBINS Brussels, I.G. 32471
ate/ruf-1624	Oleirinhos, Portugal	**KF305222**	**KF356215**	NA	RBINS Brussels, I.G. 32471
ate/ruf-1638	Portulezo, Portugal	**KF305223**	**KF356216**	NA	RBINS Brussels, I.G. 32471
ate/ruf-1649	Manteigas, Portugal	**KF305224**	**KF356217**	NA	RBINS Brussels, I.G. 32471
ruf-105	St.-Katelijne Waver, Belgium	**KF305225**	**KF356234**	NA	RBINS Brussels, I.G. 32471
ruf-15	Santiago de Compostela, Spain	AY987900	AY947375	AJ509066	RBINS Brussels, I.G. 32471
ruf-155	Brussels, Belgium	AY987901	AY947376	AJ509064	RBINS Brussels, I.G. 32471
ruf-180	Hoboken, Belgium	AY987902	AY947377	AJ509065	RBINS Brussels, I.G. 32471
ruf-182	Brecht, Belgium	AY987903	AY947378	AJ509067	RBINS Brussels, I.G. 32471
ruf-624	Nazareth , Belgium	NA	NA	KF385460	RBINS Brussels, I.G. 32471
AATE 39A*	Caldas de Gerês, Portugal	NA	NA	AY316268	
AATE 39E*	Valporquero Cave, Leon, Spain	NA	NA	AY316269	
ARUF 40G*	Montagne Noire, France	NA	NA	AY316270	

to equal length. The final alignments had a length of 504 bp (COI), 408 bp (16S) and 587 bp (ITS1), and of 1499 bp after concatenating the three fragments. The COI sequences were translated to amino acid sequences to check for stop codons (but none were found). The ITS1 sequences were also analysed together with those of Quinteiro et al. (2005). In this way we could extend our taxon coverage to *A. hispanicus* Simroth, 1886, *A. fuligineus* Morelet, 1845 and *A. nobrei* Pollonera, 1889 (Table 1). Because Quinteiro et al. (2005) used other ITS1 primers, we had to trim this dataset to a length of 378 bp. For each gene fragment, and for the concatenated dataset, we constructed Neighbour-Joining (NJ) trees (Saitou and Nei 1987) using the Kimura 2-parameter (K2P) model in MEGA v5 (Tamura et al. 2011) with complete deletion of insertions and deletions (indels). Branch support was evaluated with 1000 bootstrap replicates (Felsenstein 1985). Only bootstrap values ≥ 70% were considered as indicating strong support (Hillis and Bull 1993). Uncorrected p-distances (hereafter simply referred to as p-distance) were calculated in MEGA v5 (Tamura et al. 2011). For these calculations we considered the following Molecular Operational Taxonomic Units (MOTUs): 1) the five 16S rDNA clades of *A. subfuscus* (S1 to S5) defined by Pinceel et al. (2005a), 2) *A. anguloi* and *A. urbiae* jointly as a single MOTU (Backeljau et al. 1994, Quinteiro et al. 2005), 3) *A. wiktori* and *A. paularensis* jointly as a single MOTU (Backeljau et al. 1996, Quinteiro et al. (2005), and 4) *A. lusitanicus* from Portugal vs. *A. lusitanicus* from elsewhere as two different MOTUs (Davies 1987, Castillejo 1998, Quinteiro et al. 2005). Standard errors of mean p-distances among taxa and MOTUs were calculated on 1000 bootstrap replicates.

Results

Overall

The alignments comprized 504 bp for COI (196 variable sites), 408 bp for 16S (121 sites with alignment gaps, 122 variable sites) and 587 bp for ITS1 (277 sites with alignment gaps, 64 variable sites). For the concatenated dataset, there was strong support for the subgenera *Carinarion, Kobeltia* (excluding *A. wiktori*) and *Arion* s.s., and for a clade of *Arion* s.s. + *Mesarion* (including *A. wiktori*) (Figure 3). The subgenus *Mesarion* was not monophyletic but consisted of (1) a clade of *A. flagellus*, *A. wiktori*, *A. paularensis*, *A. baeticus*, *A. urbiae*, *A. anguloi*, and *A. gilvus*, (2) two haplotypes of *A. lusitanicus* (lus-79 and lus-186) that formed a sister group of *Arion* s.s. [insofar *A. lusitanicus* is, of course, considered as a member of *Mesarion*; see e.g. Backeljau (1989)], and (3) a number of species/clades among which the relationships were mostly unresolved. Within *A. subfuscus* (for which the monophyly was not supported) there were five clades (S1 to S5), with strong support for (S1,S5),S4) and (S2,S3). The mean p-distance (± SE) among the *Mesarion* OTUs (including *A. ponsi* and *A. gilvus*) was 0.168 ± 0.011 (range: 0.11–0.22) for COI, 0.134 ± 0.012 (range: 0.058–0.195) for 16S, and 0.022 ± 0.004 (range: 0.000–0.048) for ITS1 (a minimum distance of zero means that the two sequences only differed in a number of indels). The mean p-distances (± SE) excluding *A. ponsi* and *A.*

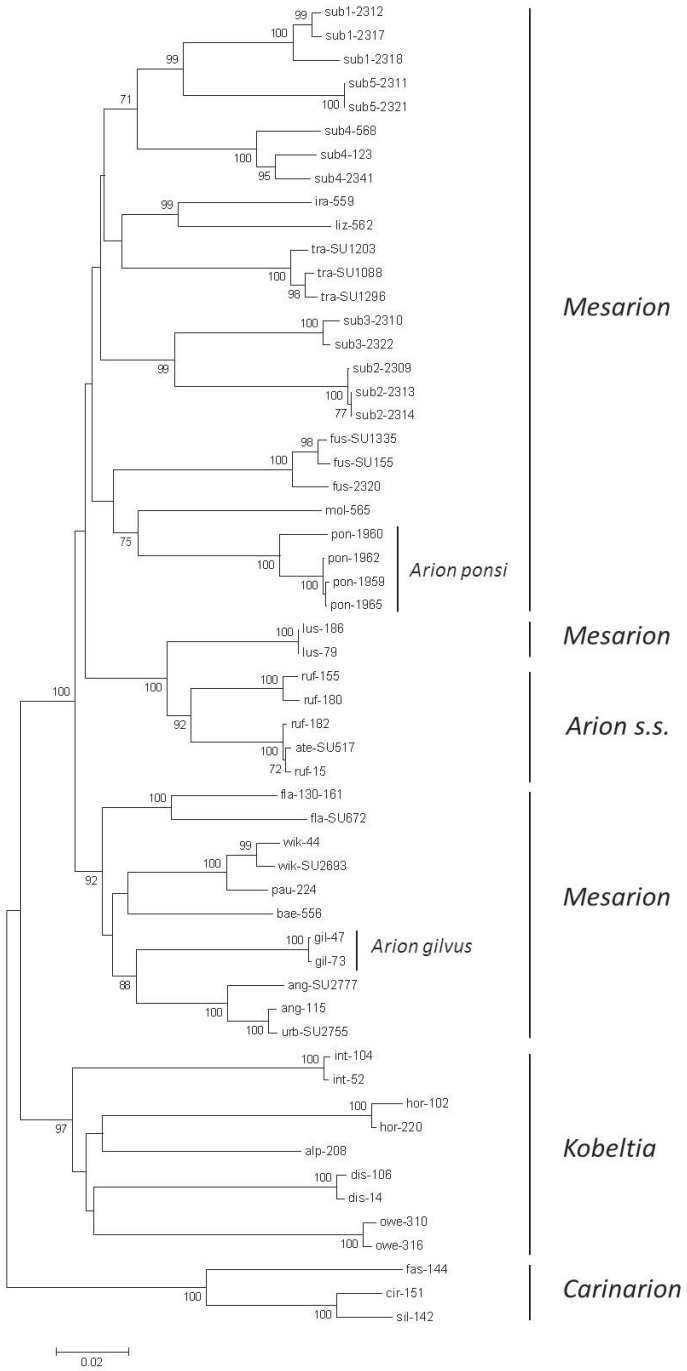


Figure 3. Neighbour-Joining tree (Kimura 2-parameter model) of a 1499 bp concatenated fragment (504 bp of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene, 408 bp of the mitochondrial 16S rDNA gene, 587 bp fragment of the nuclear internal transcribed spacer 1 (ITS1) region) for the land slug subgenus *Mesarion*. Bootstrap values ≥ 70% are shown at the nodes. For sample codes see Table 1.

gilvus were 0.167 ± 0.011 (range: 0.11–0.22) for COI, 0.130 ± 0.012 (range: 0.058–0.195) for 16S, and 0.023 ± 0.004 (range: 0.000–0.048) for ITS1. For the concatenated dataset these values were 0.108 ± 0.006 (range: 0.071–0.137) (including *A. ponsi* and *A. gilvus*) and 0.107 ± 0.006 (range: 0.071–0.137) (excluding *A. ponsi* and *A. gilvus*). The phylogenetic trees inferred from the three gene fragments and from the concatenated dataset are shown in Appendix, Supplementary Figures 1–4 and Figure 3, respectively.

Arion ponsi

The four individuals of *A. ponsi* yielded four COI and three 16S haplotypes (Appendix, Supplementary Figures 1–2), yet two 16S haplotypes only differed by an indel of two base pairs at positions 291–292. For both genes *A. molinae* showed the smallest p-distance with *A. ponsi* (COI: mean p-distance 0.142 ± 0.014; 16S: mean p-distance 0.162 ± 0.019), but a sister species relationship with *A. molinae* was only well-supported by 16S. There were three ITS1 haplotypes for *A. ponsi*; one of these had a deletion of a poly-T stretch of six base pairs at positions 556–561; the other two differed by a deletion of a G at position 554. These three ITS1 haplotypes of *A. ponsi* clustered within a clade of *A. subfuscus* S1–5, *A. lizarrustii*, *A. molinae*, *A. iratii* and *A. transsylvanus* (Appendix, Supplementary Figure 3). The ITS1 analysis with the sequences of Quinteiro et al. (2005), placed the single remaining *A. ponsi* haplotype in the same clade (mean p-distance with the other taxa of this clade = 0.046 ± 0.004), but without bootstrap support (Appendix, Supplementary Figure 4).

As for 16S, the concatenated tree of the three gene fragments showed a sister species relationship between *A. ponsi* and *A. molinae* (Figure 3).

Arion gilvus

The three *A. gilvus* specimens yielded two COI (one synonymous A-G substitution at position 366) and one 16S haplotypes. For both genes the smallest mean p-distances were observed relative to *A. urbiae* and *A. anguloi* (COI: mean p-distance = 0.145 ± 0.013; 16S: mean p-distance = 0.134 ± 0.016). The two *A. gilvus* ITS1 haplotypes reduced to one when considering the stretch that overlapped with the Quinteiro et al. (2005) sequences. In this stretch it differed from that of Quinteiro et al. (2005) by a deletion of a T at position 349. Separately, none of the three genes provided reliable evidence about the sister group relationships of *A. gilvus* (Appendix, Supplementary Figures 1–4). Yet, the concatenated tree showed a well-supported sister species relationship between *A. gilvus* and the *A. urbiae* / *A. anguloi* clade (mean p-distance = 0.021 ± 0.003) (Figure 3). Mean p-distances within this *A. urbiae* / *A. anguloi* clade (in which *A. anguloi* was paraphyletic) were P = 0.041 ± 0.006 for COI, P = 0.023 ± 0.006 for 16S, P = 0.004 ± 0.002 for ITS1 and P = 0.020 ± 0.003 for the concatenated dataset.

Discussion

The NJ-tree of the concatenated dataset confirms the major outcomes of previous phylogenetic studies, viz. 1) a strong support for the monophyly of the subgenus *Carinarion* (Geenen et al. 2006), 2) a clade of *Arion* s.s. and non-Portuguese *A. lusitanicus* (Quinteiro et al. 2005), 3) *A. wiktori* clustering with *Mesarion* species, in particular with *A. paularensis* (Quinteiro et al. 2005) instead of with *Kobeltia* species (Castillejo 1998), and 4) the strong differentiation within *A. subfuscus* s.s. that consists of, at least, five phylogenetic species (Pinceel et al. 2005a). It therefore seems that the analysis of COI, 16S and ITS1 DNA sequences yields relevant taxonomic information with respect to the characterisation of arionid species that have been described under the morphospecies concept.

Because *Arion gilvus* and *Arion ponsi* were originally described on morphological grounds they are to be interpreted as morphospecies. This phenetic morphological distinction, however, correlates well with a phenetic separation based on mtDNA distances. Indeed, the overall mean p-distance among the *Mesarion* MOTUs (excluding *A. ponsi* and *A. gilvus*) dealt with in this study is 16.7% for COI and 13% for 16S. As such, the mean p-distances between *A. ponsi* and *A. molinae* (COI: 14.2%; 16S: 16.2%) or between *A. gilvus* and *A. urbiae* (COI: 14.5%; 16S: 13.4%) are perfectly comparable with the mean p-distances among the other MOTUs and morphospecies in *Mesarion.* Hence, with respect to mtDNA divergence, both *A. ponsi* and *A. gilvus*, behave as other *Mesarion* species or putative species-level MOTUs.

Obviously, the strong COI differentiation among *Mesarion* taxa, and of *A. ponsi* and *A. gilvus* in particular, suggests that DNA barcoding may be a suitable identification tool for these animals. Yet, this may be a too simplistic conclusion, since stylommatophorans may show extremely high intraspecific mtDNA divergences of sometimes up to 27% (K2P-distances, but note the uncorrected p-distances are almost similar) (Thomaz et al. 1996, Chiba 1999). In addition, Davison et al. (2009) showed that in the Stylommatophora the mean interspecific K2P-distances (± 3%) can be substantially lower than the mean intraspecific K2P-distances (± 12%). Under these conditions, it becomes very difficult to define generally applicable thresholds that distinguish between intra- and interspecific sequence divergences. Such thresholds are normally associated with DNA barcoding gaps (Hebert et al. 2003), but Davison et al. (2009) were unable to detect DNA barcoding gaps in the taxa they studied. Nevertheless, Davison et al. (2009) suggested a pragmatic 4% threshold to separate intra- and interspecific values, but at the same time they also concluded that DNA barcoding in itself is insufficient to identify and/or detect stylommatophoran species. Unfortunately, our sample sizes were too small to explore eventual DNA barcoding gaps in *Mesarion*.

Because DNA barcoding on its own may be unreliable for identifying and detecting species-level taxa in stylommatophorans, it its necessary to backup this sort of data with, amongst others, phylogenetic analyses. As such, our phylogenetic trees of the DNA sequence data show that the morphospecies *A. ponsi* and *A. gilvus*, also represent phylo-

genetic species, since both form well-supported clades that are "significantly" associated with well-defined, but morphologically different sister species. For *A. ponsi*, the sister species appears to be *A. molinae*, the distribution range of which is located in NE continental Spain (Castillejo 1997), i.e. north of, and facing, the Balearic Islands. Conversely, the sister taxon of *A. gilvus* is the "tandem" of *A. urbiae* and *A. anguloi*, two species that have been synonymized by Backeljau et al. (1994) and that jointly should be referred to as *A. urbiae*. Our DNA sequence data on COI, 16S and ITS1 (e.g. Figure 3), as well as those on ND1 and ITS1 of Quinteiro et al. (2005) are in line with this. As such, the distribution range of *A. urbiae* is situated northwest of, and probably adjacent to, that of *A. gilvus*. Thus, for both the species pairs *A. ponsi* / *A. molinae* and *A. gilvus* / *A. urbiae*, the distribution ranges appear at least consistent with the suggested sister group relationships.

In conclusion, the present work shows that *A. ponsi* and *A. gilvus* clearly differ from *A. subfuscus* or any other currently recognized arionid species. As such, former records of *A. subfuscus* from Menorca (e.g. Gasull and van Regteren Altena 1970, Mateo 1993, Beckmann 2007) almost certainly refer to *A. ponsi*. Similarly, probably all reports of *A. subfuscus* in the regions of Valencia and Albacete involve *A. gilvus* (e.g. Borredà 1994, Borredà and Collado 1996). Finally, Borredà (1994) wondered about the eventual relationship between *A. subfuscus* from Menorca and *A. gilvus*. The current data confirm unambiguously that these are two different species, with the former being *A. ponsi*. Yet, the overall phylogenetic relationships within *Mesarion* and many other *A. subfuscus*-like taxa remain to be resolved. In this context, one of the main questions is whether *Mesarion* in its present use is a monophyletic taxon. At the same time one may wonder about the relationships with the subgenus *Arion* s.s., with which *Mesarion* seems to form a well-supported clade (Figure 3).

Acknowledgements

This work was supported by BELSPO Action 1 project MO/36/017 and FWO Research Network W0.009.11N "Belgian Network for DNA Barcoding". We are indebted to Dr. Ramón Martín (Bilbao) for providing us with specimens of *Arion gilvus*, and to the editor of Spira to give us permission to use the pictures of *A. ponsi*. We wish to thank Ben Rowson and one anonymous referee for their helpful comments.

References

Altonaga K, Gómez B, Martín R, Prieto CE, Puente AI, Rallo A (1994) Estudio faunístico y biogeográfico de los moluscos terrestres del norte de la Península Ibérica. Eusko Legebiltzarra-Parlemento Vasco. Vitoria-Gazteiz, 503 pp.

Backeljau T (1989) A review of the genus *Arion* in Belgium (Mollusca, Pulmonata). Verhandelingen van het Symposium "Invertebraten van België", Brussels, 95–99.

Backeljau T, De Winter AJ, Martín R, Rodriguez T, De Bruyn L (1994) Genital and allozyme similarity between *Arion urbiae* and *A. anguloi* (Mollusca: Pulmonata). Zoological Journal of the Linnean Society 110: 1–18. doi: 10.1006/zjls.1994.1001

Backeljau T, Winnepenninckx B, Jordaens K, De Wolf H, Breugelmans K, Parejo C, Rodríguez T (1996) Protein electrophoresis in arionid taxonomy. British Crop Protection Council Symposium Proceedings No 66. Slug & Snail Pests in Agriculture, 21–28.

Bech M (1990) Fauna malacològica de Catalunya. Molluscs terrestres i d'aigua dolça. Treballs de la Institució Catalana d'Història Natural 12: 1–229.

Beckmann KH (2007) Die Land- und Süsswassermollusken der Balearischen Inseln. Conch-Books, Hackenheim, 255 pp.

Borredà V (1994) Datos sobre la distribución geográfica de *Arion gilvus* Torres Mínguez, 1925 (Gastropoda, Pulmonata, Arionidae). Abstractbook of the X Congreso Nacional de Malacología, Barcelona, 143–144.

Borredà V, Collado MA (1996) Pulmonados desnudos (Gastropoda, Pulmonata) de la provincia de Castellón (E España). Iberus 14: 9–24.

Castillejo J (1990) Babosas de la Peninsula Ibérica. I. Los arionidos. Catalogo crítico y Mapas de Distribución. Iberus 9: 331–345. http://www.usc.es/export/sites/default/gl/investigacion/grupos/malaterra/publicaciones/articulos/087_Babosas_Penxnsula_Ibxrica_I.pdf.

Castillejo J (1997) Las babosas de la familia Arionidae Gray, 1840 en la peninsula Ibérica e islas Baleares. Morfología y Distribución (Gastropoda, Pulmonata, Terrestria nuda). Revista Real Academia Galega de Ciéncias, Vol. XVI, 118 pp.

Castillejo J (1998) Guía de las babosas ibéricas. Real Academia Gallega de Ciéncias, Santiago de Compostela.

Castillejo J, Rodríguez T (1991) Babosas de la Península Ibérica y Baleares. Inventario crítico, Citas y Mapas de Distribución (Gastopoda, Pulmonata, Terrestria nuda). Monografías da Universidade de Santiago de Compostela, No 162. Santiago de Compostela, 211 pp.

Chiba S (1999) Accelerated evolution of land snails *Mandarina* in the oceanic Bonin Islands: evidence from mitochondrial DNA sequences. Evolution 53: 460–471. doi: 10.2307/2640782

Davies SM (1987) *Arion flagellus* Collinge and *A. lusitanicus* Mabille in the British Isles: A morphological, biological and taxonomic investigation. Journal of Conchology 32: 339–354.

Davison A, Blackie RLE, Scothern GP (2009) DNA barcoding of stylommatophoran land snails: a test of existing sequences. Molecular Ecology Resources 9: 1092–1101. doi: 10.1111/j.1755-0998.2009.02559.x

De Winter A (1986) Little known and new south-west European slugs (Pulmonata: Agriolimacidae, Arionidae). Zoologische Mededelingen Leiden 60: 135–158. http://www.repository.naturalis.nl/document/149699

Dreijers E, Reise H, Hutchinson JMC (2013) Mating of the slugs *Arion lusitanicus* auct. non Mabille and *A. rufus* (L.): Different genitalia and mating behaviours are incomplete barriers to interspecific sperm exchange. Journal of Molluscan Studies 79: 51–63. doi: 10.1093/mollus/eys033

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783–791. http://statweb.stanford.edu/~nzhang/Stat366/Felsenstein85.pdf, doi: 10.2307/2408678

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. http://www.mbari.org/staff/vrijen/PDFS/Folmer_94MMBB.pdf

Garrido C (1992) A Fauna de Ariónidas da Parte Nor-Oriental da Península Ibérica (Gastropoda: Pulmonata: Arionidae). MSc Thesis, Universidade de Santiago de Compostela, Santiago de Compostela, 236 pp.

Garrido C, Castillejo J, Iglesias J (1995) The *Arion subfuscus* complex in the eastern part of the Iberian Peninsula, with redescription of *Arion subfuscus* (Draparnaud, 1805) (Gastropoda: Pulmonata: Arionidae). Archiv für Molluskenkunde 124: 103–118. http://www.usc.es/export/sites/default/gl/investigacion/grupos/malaterra/publicaciones/articulos/061_The_Arion_subfuscus_complex_Iberian_Peninsula_redescription_Arion_subfuscus.pdf

Gasull L, van Regteren Altena CO (1970) Pulmonados desnudos de las Baleares. Bolleti de la Societat d'Historia Natural de les Balears 15: 121–134. http://www.raco.cat/index.php/BolletiSHNBalears/article/view/171385

Geenen S, Jordaens K, Backeljau T (2006) Molecular systematics of the *Carinarion* complex (Mollusca: Gastropoda: Pulmonata): a taxonomic riddle caused by a mixed breeding system. Biological Journal of the Linnean Society 89: 589–604. doi: 10.1111/j.1095-8312.2006.00693.x

Hebert PDN, Ratnasingham S, De Waard JR (2003) Barcoding animal life: cytochrome c oxidase subunit I divergences among closely related species. Proceedings of the Royal Society of London B (Supplement) 270: S96-S99. doi: 10.1098/rsbl.2003.0025

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) 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 USA 101: 14812–14817. doi: 10.1073/pnas.0406166101

Hillis DM, Dixon MT (1991) Ribosomal DNA–molecular evolution and phylogenetic inference. Quarterly Review of Biology 66: 410–453. doi: 10.1086/417338

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42: 182–192. doi: 10.1093/sysbio/42.2.182

Jordaens K, Geenen S, Reise H, Van Riel P, Verhagen R, Backeljau T (2000) Is there a geographical pattern in the breeding system of a complex of hermaphroditic slugs (Mollusca: Gastropoda: *Carinarion*)? Heredity 85: 571–579. doi: 10.1046/j.1365-2540.2000.00793.x

Jordaens K, Pinceel J, Van Houtte N, Breugelmans K, Backeljau T (2010) *Arion transsylvanus* (Mollusca, Pulmonata, Arionidae): rediscovery of a cryptic species. Zoologica Scripta 39: 343–362. doi: 10.1111/j.1463-6409.2010.00425.x

Kerney MP, Cameron RAD, Jungbluth JH (1983) Die Landschnecken Nord- und Mitteleuropas. Hamburg, Parey, 384 pp.

Mateo B (1993) Invertebrats no artròpodes. In: Enciclopèdia de Menorca, vol. 3. Obra Cultural de Menorca, Menorca.

Palumbi SR (1996) Nucleic acids II: The polymerase chain reaction. In: Hillis DM, Moritz C, Mable BK (Eds) Molecular systematics. Sinauer Associates Inc, 205–247.

Pinceel J, Jordaens K, Van Houtte N, De Winter AJ, Backeljau T (2004) Molecular and morphological data reveal cryptic taxonomic diversity in the terrestrial slug complex *Arion subfuscus/fuscus* (Mollusca, Pulmonata, Arionidae) in continental north-west Europe. Biological Journal of the Linnean Society 83: 23–38.

Pinceel J, Jordaens K, Backeljau T (2005a) Extreme mtDNA divergences in a terrestrial slug (Arionidae, Pulmonata, Gastropoda): accelerated evolution, allopatric divergence and secondary contacts. Journal of Evolutionary Biology 18: 1264–1280. doi: 10.1111/j.1420-9101.2005.00932.x

Pinceel J, Jordaens K, Pfenninger M, Backeljau T (2005b) Rangewide phylogeography of a terrestrial slug in Europe: evidence for Alpine refugia and rapid colonization after the Pleistocene glaciations. Molecular Ecology 14: 1133–1150. doi: 10.1111/j.1365-294X.2005.02479.x

Quintana Cardona J (2007) Un nuevo molusco terrestre para la fauna balear: *Arion* (*Mesarion*) *ponsi* sp. nov. (Gastropoda: Pulmonata: Arionidae). Spira 2: 139–146. http://www.molluscat.com/SPIRA/PDF/Spira_2_3_2.pdf

Quinteiro J, Rodríguez-Castro J, Castillejo J, Iglesias-Piñeiro J, Rey-Méndez M (2005) Phylogeny of slug species of the genus *Arion*: evidence of monophyly of Iberian endemics and of the existence of relict species in Pyrenean refuges. Journal of Zoological Systematics and Evolutionary Research 43: 139–148.

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406–425. http://mbe.oxfordjournals.org/content/4/4/406.full.pdf+html

Selander RK, Hudson RO (1976) Animal population structure under close inbreeding: the land snail *Rumina* in southern France. American Naturalist 110: 695–718. doi: 10.1086/283098

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2737. doi: 10.1093/molbev/msr121

Thomaz D, Guiller A, Clarke B (1996) Extreme divergence of mitochondrial DNA within species of pulmonate land snails. Proceedings of the Royal Society of London B 263: 363–368. doi: 10.1098/rspb.1996.0056

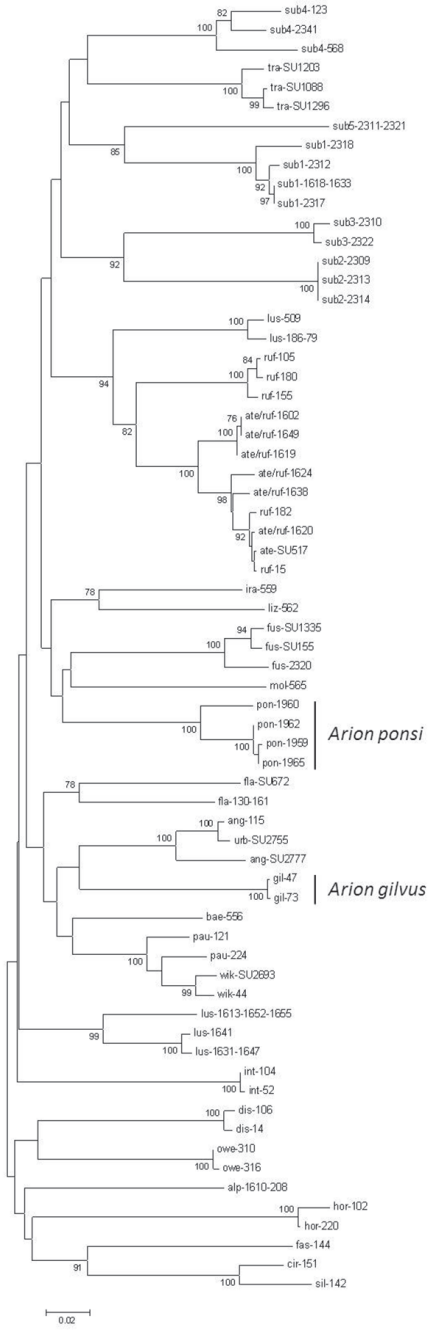
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680. doi: 10.1093/nar/22.22.4673

Torres Mínguez A (1925) Notas malacológicas VII: Cuatro nuevos *Arion* ibéricos y dos nuevos Limácidos de Guinea. Butlletí de la Institución Catalana de Historia Natural 8: 228–243.

Waldén HW (1976) A nomenclatural list of the land Mollusca of the British Isles. Journal of Conchology 29: 21–25.

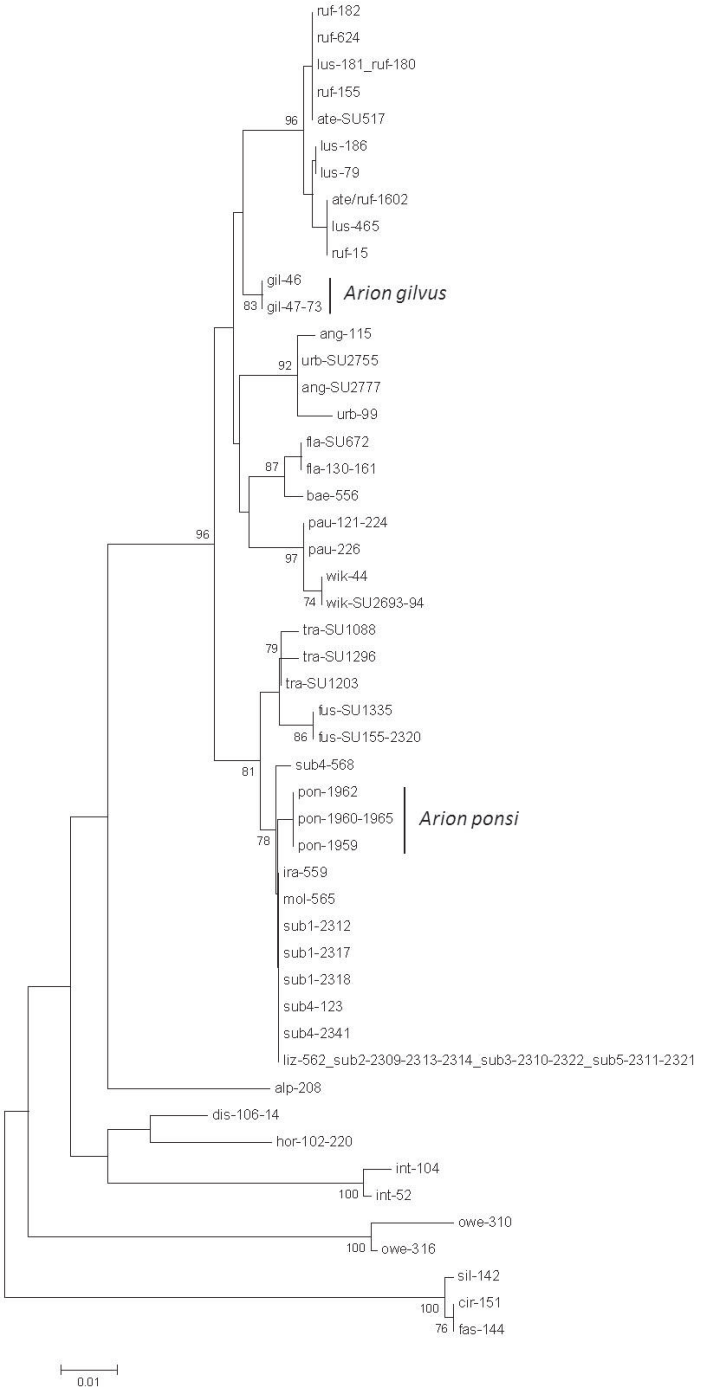
Wiktor A (1973) Die Nacktschnecken Polens (Arionidae, Milacidae, Limacidae) (Gastropoda, Stylommatophora). Monografie Fauny Polski 1: 1–180.

Appendix

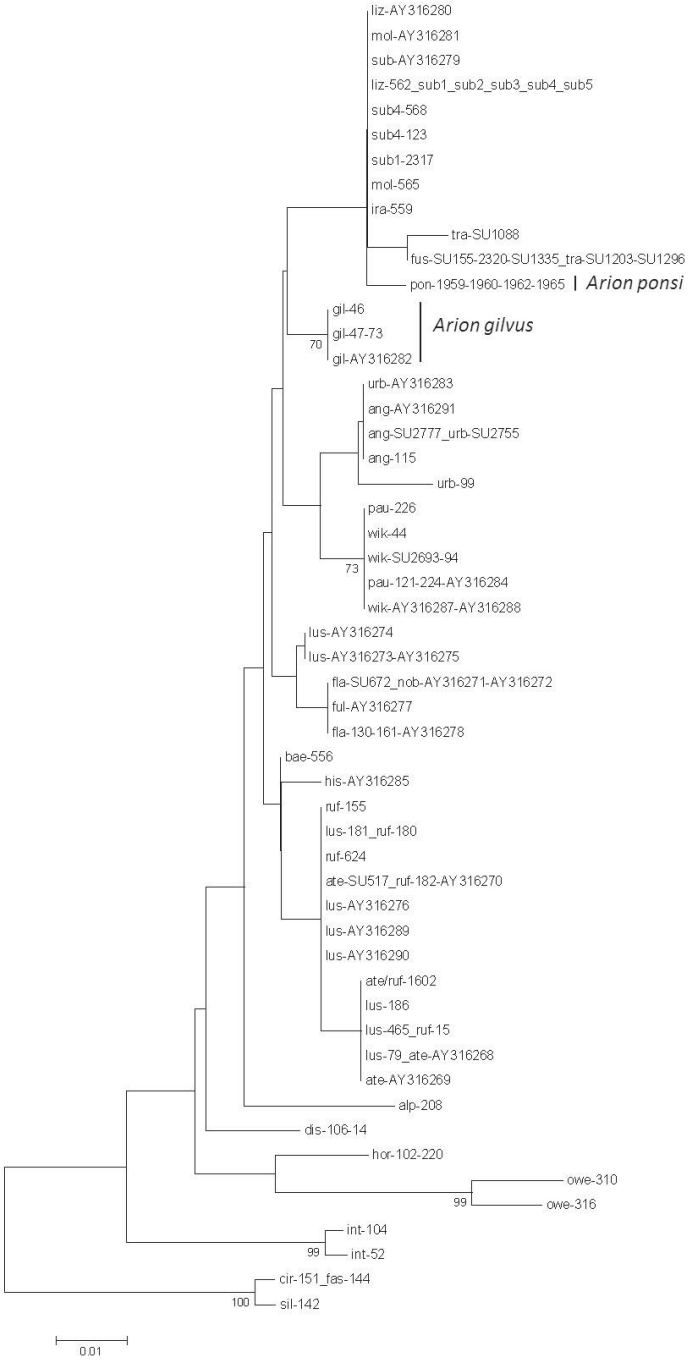


Supplementary Figure 1. Neighbour-Joining tree (Kimura 2-parameter model) of a 504 bp fragment of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene for the land slug subgenus *Mesarion*. Bootstrap values ≥ 70% are shown at the nodes. For sample codes see Table 1.

Supplementary Figure 2. Neighbour-Joining tree (Kimura 2-parameter model) of a 408 bp fragment of the mitochondrial 16S rDNA gene for the land slug subgenus *Mesarion*. Bootstrap values ≥ 70% are shown at the nodes. For sample codes see Table 1.



Supplementary Figure 3. Neighbour-Joining tree (Kimura 2-parameter model) of a 587 bp fragment of the nuclear internal transcribed spacer 1 (ITS1) region for the land slug subgenus *Mesarion*. Bootstrap values ≥ 70% are shown at the nodes. For sample codes see Table 1.



Supplementary Figure 4. Neighbour-Joining tree (Kimura 2-parameter model) of a 378 bp fragment of the nuclear internal transcribed spacer 1 (ITS1) region for the land slug subgenus *Mesarion*. This figure also includes the Iberian *Mesarion* ITS1 sequences of Quinteiro et al. (2005) Bootstrap values ≥ 70% are shown at the nodes. For sample codes see Table 1.

ZooKeys 365: 105–126 (2013)
doi: 10.3897/zookeys.365.5526
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Testing the performance of a fragment of the COI gene to identify western Palaearctic stag beetle species (Coleoptera, Lucanidae)

Karen Cox[1], Arno Thomaes[1], Gloria Antonini[2], Michele Zilioli[3], Koen De Gelas[1,4], Deborah Harvey[5], Emanuela Solano[2], Paolo Audisio[2], Niall McKeown[6], Paul Shaw[6], Robert Minetti[7], Luca Bartolozzi[8], Joachim Mergeay[1]

1 *Research Institute for Nature and Forest, Gaverstraat 4, B-9500 Geraardsbergen, Belgium* **2** *Department of Biology and Biotechnology "Charles Darwin", Sapienza - University of Rome, via A. Borelli 50, I-00161 Rome, Italy* **3** *Natural History Museum, Entomological section, Corso Venezia 55, I-20121 Milano, Italy* **4** *Royal Belgian Institute of Natural Sciences, Vautierstraat 29, B-1000 Brussels, Belgium* **5** *School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey, UK* **6** *Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Penglais, Aberystwyth, UK* **7** *7 Avenue Marc Sangnier, 13600 La Ciotat, France* **8** *Natural History Museum, Zoological Section "La Specola", via Romana 17, 50125 Firenze, Italy*

Corresponding author: *Karen Cox* (karen.cox@inbo.be)

Academic editor: *M. de Meyer* | Received 15 May 2013 | Accepted 16 October 2013 | Published 30 December 2013

Citation: Cox K, Thomaes A, Antonini G, Zilioli M, De Gelas K, Harvey D, Solano E, Audisio P, McKeown N, Shaw P, Minetti R, Bartolozzi L, Mergeay J (2013) Testing the performance of a fragment of the COI gene to identify western Palaearctic stag beetle species (Coleoptera, Lucanidae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 105–126. doi: 10.3897/zookeys.365.5526

Abstract

The taxonomy of stag beetles (Coleoptera: Lucanidae) remains challenging, mainly due to the sexual dimorphism and the strong allometry in males. Such conjecture confounds taxonomic based conservation efforts that are urgently needed due to numerous threats to stag beetle biodiversity. Molecular tools could help solve the problem of identification of the different recognized taxa in the "*Lucanus cervus* complex" and in some related Palaearctic species. We investigated the potential use of a 670 bp region at the 3' end of the mitochondrial cytochrome *c* oxidase subunit I gene (COI) for barcoding purposes (different from the standard COI barcoding region). Well resolved species and subspecies were *L. tetraodon*, *L. cervus akbesianus*, *L. c. laticornis*, as well as the two eastern Asian outgroup taxa *L. formosanus* and *L. hermani*. Conversely, certain taxa could not be distinguished from each other based on K2P-distances and tree

topologies: *L. c. fabiani* / *L.* (*P.*) *barbarossa*, *L. c. judaicus* / an unknown *Lucanus* species, *L. c. cervus* / *L. c. turcicus* / *L. c. pentaphyllus* / *L.* (*P.*) *macrophyllus* / *L. ibericus*. The relative roles of phenotypic plasticity, recurrent hybridisation and incomplete lineage sorting underlying taxonomic and phylogenetic discordances are discussed.

Keywords
Lucanus spp., Stag beetle, Western Palaearctic, DNA barcoding, COI

Introduction

Lucanidae Latreille, 1804 is a family of Coleoptera showing in most species pronounced sexual dimorphism and strong external morphological allometry in males. The species of the Holarctic and Oriental distributed genus *Lucanus* Scopoli, 1763 are renowned for the striking appearance of the males. With their large body size and prominent mandibles, the male stag beetles are very popular among amateur entomologists and as terrarium pets, mainly in Japan. Currently, there are more than 90 *Lucanus* species described, however, validity of these designations is considered questionable in many cases. Sexual dimorphism and size variation complicate the taxonomy (Didier and Séguy 1953, Clark 1977, Harvey and Gange 2006), as does the lack of informative phenotypic characters among larvae. Consequently, their classification has changed over time and is still under discussion. In this study we focus on taxa of the *Lucanus* species in the western Palaearctic.

The genus *Lucanus* is subdivided into the subgenera *Lucanus* sensu stricto and *Pseudolucanus* Hope & Westwood, 1845. Members of the latter have a peculiar stout body and substantial analogy of morphology that makes it quite easy to distinguish them from members of the subgenus *Lucanus* (Planet 1899). The male mandibles of *Pseudolucanus* are sickle shaped, their internal edge has a single denticle in most species (*Lucanus* has small denticles and one large denticle) and the apex is usually simple (*Lucanus* is mostly bifid) (Planet 1899, Baraud 1993). Furthermore, the integument of *Pseudolucanus* is relatively smooth with scattered and superficial punctuation whereas it is more stippled in *Lucanus*. Also, the sides of the pronotum of *Pseudolucanus* are strongly sinuate before the posterior angles (Baraud 1993). Previous studies (Didier and Séguy 1953, Benesh 1960, Krajcik 2001, Bartolozzi and Sprecher-Uebersax 2006, Hallan 2008, Fujita 2010) describe between four and seven species of *Lucanus* in western Palaearctic: i.e. *L.* (*Lucanus*) *cervus* (Linnaeus, 1758), *L.* (*L.*) *ibericus* Motschulsky, 1845, *L.* (*L.*) *orientalis* Kraatz, 1860, *L.* (*L.*) *tetraodon* Thunberg, 1806, *Lucanus* (*Pseudolucanus*) *barbarossa* Fabricius, 1801, *L.* (*P.*) *busignyi* Planet, 1909 and *L.* (*P.*) *macrophyllus* Kraatz, 1860.

The distribution of many of these taxa remains poorly resolved, however, we can consider some of them as endangered. The practice of removing old trees and dead wood in past and current forest management, has had detrimental effects on this group of saproxylic beetles (Jansson and Coskun 2008, Nieto and Alexander 2010). Con-

sequently, the loss of habitat might have reduced the range of some taxa, especially the Mediterranean taxa where deforestation started a few millennia ago (Jansson and Coskun 2008, Buse et al. 2010). At least *L. c. cervus* seems to be able to cope with urbanisation (Thomaes et al. 2008) as long as the habitat turnover allows recolonisation (Thomaes 2009). In addition, beetle collecting can be considered as a threat when it goes hand in hand with large scale habitat destruction or when species rarity causes overexploitation (Holden 2007, Tournant et al. 2012). Another possible consequence of the international stag beetle trade is the introduction of non-native specimens which may cause genetic introgression (Goka et al. 2004) and transmission of parasites potentially pathogenic to native stag beetles (cf. Goka et al. 2004, Kanzaki et al. 2011). Unfortunately, legal protection is often missing or inadequate. The widely distributed *L. c. cervus* is protected by the Habitats Directive of the European Union from 1992 (Luce 1996) and is listed as "near threatened" in the Red Data list of Europe (Nieto and Alexander 2010). *Lucanus* (*P.*) *barbarossa* and *L. tetraodon* are mentioned in the IUCN list, but are rated "of least concern" (IUCN 2012), while *L. ibericus* is considered to be "vulnerable" within the EU 27 (Nieto and Alexander 2010).

More detailed information on the distribution and ecology of this species group is needed to get a clear view on their conservation status. But unless the problem of identification of European and West Asian *Lucanus* is solved, it becomes difficult to set specific conservation priorities, without which rare, neglected and endangered species or Evolutionarily Significant Units (ESUs) may be unrecognised and thus, not given adequate conservation prioritisation (Ryder 1986, Waples 1991, Moritz 1994a, Moritz 1994b, Fraser and Bernatchez 2001). Molecular tools could help identification of stag beetles. The mitochondrial cytochrome *c* oxidase subunit I (COI) is the most widely used gene in barcoding animals (Hebert et al. 2003). The barcoding practice entails the analysis of the DNA sequence of a part of this mitochondrial gene, typically between 600 and 900 bp. In this study, we investigated the use of the 3' end of the COI gene, different from the standard barcoding region, for the identification of western Palaearctic *Lucanus* species and subspecies.

Material and methods

Taxonomy and morphology

Lucanus cervus has the widest geographical distribution in the genus and is very variable in form, size and colour (Harvey et al. 2011). Many subdivisions (i.e. subspecies or morphotypes) have been proposed and discussed. *Lucanus cervus cervus* (Linnaeus, 1758), the main subspecies found throughout Europe, has, in general, four lamellae on the antennal clubs and is typically bicoloured (black head and thorax, and reddish brown elytra and mandibles). *Lucanus cervus akbesianus* Planet, 1896 with generally six lamellae and large mandibles with a very open apical fork, inhabits southern Turkey and Syria. *Lucanus cervus turcicus* Sturm, 1843 also has a six lamellate club, but its

mandibles are comparable to *L. c. cervus*. It is reported in Greece, Bulgaria and Trakya (European part of Turkey). Furthermore, *L. c. judaicus* Planet, 1902 with a four lamellate club and reddish brown colour, is found in the more eastern parts of Turkey and in northern Syria. *Lucanus cervus laticornis* Deyrolle, 1864, found in central and southern Turkey, has six long lamellae and the inner denticle of the mandibles is followed by two or three denticles. *Lucanus cervus fabiani* Mulsant & Godart, 1855 is an endemic taxon inhabiting southern France and shows a five lamellate club and slender, slightly curved mandibles with a simple apex and post-median denticle along with a few other denticles. The taxa *fabiani* and *pentaphyllus* Reiche, 1853 are listed as synonyms of *L. c. cervus* by Bartolozzi and Sprecher-Uebersax (2006), but *fabiani* could well be considered as a valid species according to Boucher (unpublished data) while *pentaphyllus* may represent a small form of *L. cervus* with five lamellate clubs, a character that can also be found in *L. c. cervus*. Other taxa [*tauricus* Motschulsky, 1845 (described from Crimea), *poujadei* Planet, 1897 (Kurdistan), *mediadonta* Lacroix, 1978 (Georgia) and *pontbrianti* Mulsant, 1839 (France)], recognised by some authors as valid subspecies or simple synonyms, were not included in this study. Bartolozzi and Sprecher-Uebersax (2006) only list *cervus* and *judaicus* as separate subspecies. Hallan (2008) adds *akbesianus, fabiani, mediadonta, tauricus* and *turcicus*, while Krajcik (2001) further includes *pontbrianti* and *laticornis*, although Schenk and Fiedler (2011) perceived *laticornis* as a separate species. On the other hand, Didier and Séguy (1953) also list *capreolus* Fuessly, 1775 (considered a small form of *L. cervus*) and *poujadei* while Fujita (2010) only recognises *poujadei* but does not list *tauricus* and *mediadonta* or the [*pentaphyllus* + *fabiani* + *pontbrianti*] complex.

Lucanus ibericus can be found from Albania to Iran and is sometimes considered a synonym of *L. orientalis*. Unlike *L. cervus, L. ibericus* is entirely reddish brown, has a pronotum without a smooth discal line, but with a sinuate posterior and distinct toothed posterior angles (non-sinuate pronotum and blunt angles in *L. cervus*). The mandibles of the males, which are shorter than those of a typical male *L. cervus* of equal size, can have an apex with two equal teeth or with the inner tooth fainted and a large internal denticle in the middle. In addition, *L. ibericus* has six, rarely five, long lamellae on the antennal club.

Lucanus tetraodon described from France, Italy, North Africa, Albania and Greece, can be perceived as a central Mediterranean species. In contrast to *L. cervus* and *L. ibericus*, the basal denticle of the mandibles of *L. tetraodon* is placed in the lower half. Like *L. ibericus*, the pronotal sides have sharp posterior angles, but the pronotal disc misses the central smooth line. *Lucanus tetraodon* has six, occasionally five, lamellae on the antennal club. *Lucanus tetraodon* is by some authors subdivided in subspecies *L. t. argeliensis* Maes, 1995 in North Africa, *L. t. provincialis* Colas, 1949 in South France, *L. t. corsicus* Gautier des Cottes, 1860 in Corsica, *L. t. sicilianus* Planet, 1899 in Sicily and finally *L. t. tetraodon* Thunberg, 1806 elsewhere. In addition, specimens of problematic populations of *L. cervus* from a series of localities in central Italy (northern Latium and Umbria), are known to exhibit apparently intermediate morphological characters between *L. cervus* and *L. tetraodon*, which are sympatric in these areas (Santoro et al. 2009).

The *Pseudolucanus* species all have six long lamellae forming the antennal club, their body is stout and entirely reddish or blackish brown. Included in this study are *L.* (*P.*) *barbarossa* from the Iberian peninsula and the Maghreb, and *L.* (*P.*) *macrophyllus* reported in south-west Turkey. Krajcik (2001) and Hallan (2008) list the latter as a subspecies of *L. ibericus*. Schenk and Fiedler (2011) recently quoted populations of *L.* (*P.*) *busignyi* in western Turkey, but this taxon is not included in this study.

Taxon sampling and DNA extraction

A large number of entomologists was contacted to obtain material from the different taxa and from different regions. The samples included whole beetles, especially in regions where identification is problematic, as well as parts of a beetle, sometimes found as road kill or as prey leftovers from birds. Samples were dried and kept at room temperature or preserved in absolute ethanol. In total 76 samples were collected. The species identification was performed, using comparative material and available identification keys. Six samples from Israel and Lebanon could not be identified to species. These unidentified *Lucanus* specimens have a shape resembling in general the medium to small males of *L. c. akbesianus* but with a mandibular structure similar to that of *L. c. turcicus* (Zilioli et al. unpublished data). The tissue samples used for DNA extraction depended on what was available, but were mostly legs, which contain large muscles and are therefore rich in mitochondrial DNA (mtDNA). DNA was extracted from ground samples with the E.Z.N.A.® Forensic DNA Kit (Omega Bio-Tek), except for samples K1 and U6 (Table 1) from which DNA was extracted following the salting out procedure described by Aljanabi and Martinez (1997). The integrity of the extracted DNA was checked spectrophotometrically on a ND-1000 Nano-Drop (NanoDrop Technologies) and its quality on 1% agarose gels.

Sequencing

We first attempted to sequence the COI barcoding region with the primers developed by Folmer et al. (1994) on a subset of samples. Despite PCR optimization trials, amplification of this fragment largely failed. Instead, a 800 bp fragment of the 3' end of the COI gene was amplified using the primer set C1-J-2183 (5' CAACATTTATTTTGATTTTTTGG 3') and TL2-N-3014 (5' TCCAATGCACTAATCTGCCATATTA 3') (Simon et al. 1994). This fragment does not overlap with the standard barcoding region. For samples O9 and V44 (Table 1) we used species-specific primers (F - 5' GGGGCATCAGTAGACCTAGC 3' and R – 5' TTCAGCAGGTGGTATTAGTTGG 3'), designed from sequences on GenBank and used to PCR amplify a 1089 bp stretch of the COI gene. Reactions were performed in total volumes of 40 µl containing 5.2 µl of 10 × Taq buffer with 500 mM KCl (Fermentas, Thermo Scientific), 3.12 µl of $MgCl_2$ (25 mM), 0.78 µl dNTP (10 mM), 2.08 µl of each

Table 1. List of samples included in the analysis. Primers used are denoted with '1': C1-J-2183 and TL2-N-3014; '2': LCint1F, LCint2F, LCint3F and LCint4F (for sample SB6 also the reverse primers were used); '3': F - 5' GGGGCATCAGTAGACCTAGC 3' and R – 5' TTCAGCAGGTGGTATTAGTTGG 3'.

Species / subspecies	Code	Primers	Haplotype	GenBank acc. no.	Country	Longitude	Latitude	Date of sampling	Type of conservation	Gender
Lucanus cervus akbesianus	UA1	1	UA1	KF737127	Turkey	30.828278	37.721833	Jun 2010	ethanol	Female
	UA2	1	UA2	KF737128	Turkey	30.828278	37.721833	Jun 2010	ethanol	Male
	UA3	1	UA3	KF737129	Turkey	30.828278	37.721833	Jun 2010	ethanol	Male
	UA4	1	UA4	KF737130	Turkey	35.862100	37.676200	2010	ethanol	Male
	UA5	1	UA5	KF737131	Turkey	35.862100	37.676200	2010	ethanol	Male
	UX1	2	UX1	KF737132	Turkey	31.000000	36.900000	Jun 2010	ethanol	Male
	U10	1	U10	KF737125	Turkey	30.828278	37.721833	Jun 2010	ethanol	Male
	U11	1	U10	KF737126	Turkey	30.828278	37.721833	Jun 2010	ethanol	Male
Lucanus cervus cervus	A1	1	A1	KF737071	Belgium	4.537656	50.772652	Jul 2008	ethanol	Male
	A3	1	A3	KF737072	Belgium	4.331784	50.736622	Jun 2009	ethanol	Female
	C1	2	C1	KF737093	Czech rep.	16.803576	48.797935	May 2009	ethanol	Male
	D13	2	A3	KF737078	France	1.139310	45.391800	Jul 2010	ethanol	Male
	D4	1	D4	KF737088	France	1.431787	43.458090	Aug 2010	ethanol	Male
	D22	1	D22	KF737092	France	2.820327	47.861145	2009	ethanol	Female
	F12	1	A3	KF737079	Greece	22.653889	39.808333	Jun 2009	ethanol	Female
	F16	1	F16	KF737083	Greece	22.653889	39.808333	Jun 2009	ethanol	Female
	F23	1	F23	KF737082	Greece	21.663281	39.762333	Jun 2009	ethanol	Male
	G3	2	G3	KF737081	Hungary	18.834592	47.701586	Jul 2009	ethanol	Female
	I2	1	I2	KF737084	Italy	8.732981	45.779241	Jun 2009	ethanol	Male
	I3	1	A3	KF737080	Italy	8.732981	45.779241	Jun 2009	ethanol	Male
	I4	1	I4	KF737085	Italy	8.732981	45.779241	Jun 2009	ethanol	Male
	N3	1	N3	KF737086	Portugal	-9.397390	38.795900	Jul 2010	ethanol	Male
	O9	3	O9	KF737087	Romania	24.450700	47.102400			
	S15	1	S15	KF737094	Spain	-6.608460	40.385100	Aug 2009	ethanol	Male
	S19	1	A3	KF737076	Spain	-4.814970	43.304009	Jul 2009	ethanol	Female

Species / subspecies	Code	Primers	Haplotype	GenBank acc. no.	Country	Longitude	Latitude	Date of sampling	Type of conservation	Gender
	V2	1	A3	KF737077	UK	1.067369	52.028936	Aug 2009	dried	Female
	V26	3	V26	KF737091	UK	-0.209294	50.966300			
	V44	3	V44	KF737089	UK	0.844280	51.260100			
	W9	2	W9	KF737090	Ukraine	36.325800	49.826900	Jun 2007	dried	Male
	X1		X1	FJ606555	France	(Lin et al. 2011)				
Lucanus cervus fabiani	D11	1	D11	KF737121	France	5.753740	43.195300	Jun 2010	ethanol	Male
Lucanus cervus judaicus	UJ1	1	UJ1	KF737112	Turkey	36.261600	37.068100	Jul 2010	dried	Male
Lucanus cervus laticornis	UL2	1	UL2	KF737119	Turkey	30.457431	36.875669	Jun 2007	ethanol	Male
	UL3	1	UL3	KF737120	Turkey	30.558900	37.763600	1995	dried	Male
Lucanus cervus pentaphyllus	C2	1	A3	KF737075	Czech rep.	16.803576	48.797935	May 2009	ethanol	Male
	F13	1	F13	KF737104	Greece	22.653889	39.808333	Jun 2009	ethanol	Female
	I1	1	A3	KF737073	Italy	8.732981	45.779241	Jun 2009	ethanol	Male
	W7	2	A3	KF737074	Ukraine	38.497600	48.950200	Jul 2002	dried	Male
Lucanus cervus turcicus	B1	1	B1	KF737096	Bulgaria	27.737650	42.162733	Jul 2009	ethanol	Male
	B2	1	B2	KF737098	Bulgaria	25.578583	41.407800	Jul 2009	ethanol	Male
	B7	1	B7	KF737099	Bulgaria	27.977000	42.060792	Jul 2009	ethanol	Male
	B9	1	B1	KF737097	Bulgaria	27.900405	42.120183			
	F15	2	F15	KF737105	Greece	22.653889	39.808333	Jun 2009	ethanol	Male
	F7	1	F7	KF737107	Greece	22.733333	39.866667	Jun 2009	ethanol	
	F8	2	F7	KF737108	Greece	22.733333	39.866667	Jun 2009	ethanol	
	F9	1	F9	KF737106	Greece	22.653889	39.808333	Jun 2009	ethanol	Female
	F11	1	F11	KF737100	Greece	22.653889	39.808333	Jun 2009	ethanol	Male
	F17	2	F17	KF737101	Greece	22.653889	39.808333	Jun 2009	ethanol	
	F20	1	F20	KF737102	Greece	22.653889	39.808333	Jun 2009	ethanol	Male
	F21	1	F21	KF737103	Greece	22.653889	39.808333	Jun 2009	ethanol	Male
	U3	2	U3	KF737109	Turkey	27.950000	41.800000	Jul 2009	ethanol	Male

Species / subspecies	Code	Primers	Haplotype	GenBank acc. no.	Country	Longitude	Latitude	Date of sampling	Type of conservation	Gender
Unknown species of *Lucanus*	H1	2	H1	KF737116	Israel	35.753500	33.217100	Aug 2009	ethanol (after freezing)	Male
	H2	1	H2	KF737113	Israel	35.753500	33.217100	Aug 2009	dried	Female
	H3	2	H3	KF737117	Israel	35.753500	33.217100	Jul 2009	dried	Male
	H4	1	H4	KF737114	Israel	35.753500	33.217100	Jul 2009	dried	Male
	H5	2	H5	KF737115	Israel	35.864500	32.959600	1998	dried	Male
	J2†	2	J2	KF737118	Lebanon			Jul 2009	dried	Male
Lucanus ibericus	U6	1	U6	KF737110	Turkey	38.424200	40.290300			
Lucanus tetraodon provincialis	D6	1	D6	KF737111	France	5.850000	43.066700	Jun 2010	ethanol	Male
Lucanus tetraodon	X2		X2	EF487727		(Hunt et al. 2007)				
Lucanus (Pseudolucanus) barbarossa	SB1	1	SB1	KF737122	Spain	-3.831811	40.828139	Jul 2004	dried, later on ethanol	Male
	SB6†	2	SB6	KF737124	Spain	-3.585322	41.067361	Sep 2010	ethanol	Female
	SB7	1	SB7	KF737123	Spain	-3.982000	36.885000	May 2010	ethanol	Male
Lucanus (Pseudolucanus) macrophyllus	UB1†	2	UB1	KF737095	Turkey	33.089167	36.501944	Aug 2006	dried	Male
Dorcus parallelipipedus		1	K1	KF737133	Montenegro					
			X3	DQ156023		(Hunt et al. 2007)				
Lucanus formosanus			X4	FJ606632		(Huang and Lin 2010)				
			X5	FJ606630						
			X6	FJ606628						
			X5	FJ606626						
			X5	FJ606624						
			X5	FJ606622						
			X8	FJ606583						
Lucanus hermani			X9	FJ606552		(Lin et al. 2011)				

† sequences with a maximum of seven double peaks.

primer (10 µM), 0.8 U Taq DNA polymerase (Fermentas, Thermo Scientific), 26.42 µl sterile distilled water. 12 µl of diluted DNA (3.5–5 ng/ µl) was added. The temperature cycle was 94 °C for 1 min, then 5 cycles of 94 °C for 1 min, 45 °C for 1 min 30 s and 72 °C for 1 min and 30 s. This was followed by 40 cycles of 94 °C for 1 min, 50 °C for 1 min 30 s and 72 °C for 1 min, and finally a single cycle at 72 °C for 5 min. PCR products were cleaned enzymatically with DNA Clean & Concentrator™-5 (Zymo Research). When samples failed to amplify, mostly dried or bad quality samples, internal primers were used to allow amplification of four overlapping fragments of about 250 bp within the same 3' end of the COI gene: LCint1 (F – 5' CTTCGGCCACCCAGAAGT 3' and R – 5' TCCAGTAGGAACAGCAATRAT 3'), LCint2 (F – 5' CGAGCCTACTTCACATCAGC 3' and R – 5' GCAAAAACTGCACCTATTGAAA 3'), LCint3 (F – 5' GCTCACTTCCATTATGTACTTTCAA 3' and R – 5' GAGAGCCAAATGATGAAATAATGTT 3') and LCint4 (F – 5' CCCTGATGCCTACACCACAT 3' and R – 5' CCAATGCACTAATCTGCCATA 3'). PCR amplification was performed in 2.6 µl of 10 × Taq buffer with 500 mM KCl, 2.08 µl of MgCl2 (25 mM), 0.39 µl dNTP (10 mM), 2.6 µl of each primer (10 µM), 0.8 U Taq DNA polymerase (Fermentas, Thermo Scientific), 9.57 µl sterile distilled water, resulting in a total volume of 20 µl to which 6 µl of diluted DNA (3.5–5 ng/ µl) was added. The PCR reaction was then conducted with the following cycle: 94 °C for 3 min, then 45 cycles of 94 °C for 45 s, 59 °C for 45 s and 72 °C for 1 min 30 s, and finally a single cycle at 72 °C for 6 min. PCR products were checked on 2% agarose horizontal gels and purified using USB® ExoSAP-IT® (Isogen Life Science). DNA sequencing was performed by a commercial company (BaseClear, Leiden, the Netherlands) or on an automatic ABI 3500 Genetic Analyzer (Applied Biosystems). Both forward and reverse primers were used except when internal primers were used for PCR, in which case sequencing was performed using the respective forward primers (except for five samples of *L.* (*P.*) *barbarossa*, where both forward and reverse primers were used).

COI sequences available on GenBank were added. The COI sequence of *L. c. cervus* obtained by Lin et al. (2011; GenBank acc. no. FJ606555) was used as a reference for the subspecies with the highest number of specimens in this study. We selected two Asian stag beetle species, *L. formosanus* Planet, 1899 and *L. hermani* DeLisle, 1973, and *Dorcus parallelipipedus* (Linnaeus, 1758) (lesser stag beetle; Lucanidae) as outgroup species. Except for one available sample of the latter, the COI gene sequences of the taxa were obtained from GenBank (*D. parallelipipedus*: Hunt et al. 2007; GenBank acc. no. DQ156023; *L. formosanus*: Huang and Lin 2010; GenBank acc. no. FJ606632, FJ606630, FJ606628, FJ606626, FJ606624, FJ606622, FJ606583; *L. hermani*: Lin et al. 2011; GenBank acc. no.: FJ606552). In the study of Hunt et al. (2007) the Dorcinae formed a sisterclade of the Lucaninae. Finally, part of the COI sequence of *L. tetraodon* obtained by Hunt et al. (2007; GenBank acc. no. EF487727) was used in addition to the sequence of *L. t. provincialis*.

DNA sequences have been deposited in GenBank under accession numbers KF737071 to KF737133 (Table 1).

Alignment and sequence quality control

Overall quality of the sequences was evaluated manually. Only samples with high quality chromatograms for at least 300 bp were retained for further analyses. Sequences were aligned by hand and using CLUSTALW v1.4 (Thompson et al. 1994) in BIOEDIT v7.0.0 (Hall 1999). Sequences were trimmed to 670 bases. Duplicate haplotypes were removed using DUPLICATESFINDER v1.1 (http://bioinfotutlets.blogspot.be/2009/09/duplicates-finder-java-standalone.html). We searched for potential NUMTs (nuclear mitochondrial pseudogene sequences) or heteroplasmy by manually checking for the presence of double peaks and indels, and by looking for stop codons (Song et al. 2008, Calvignac et al. 2011) using MEGA v5.01 with the implemented invertebrate mtDNA genetic code to translate the sequences (Tamura et al. 2011). We only retained sequences with a maximum of 7 polymorphic positions, which were treated as ambiguities. Finally, we constructed a Neighbour-Joining (NJ) tree with MEGA v5.01 using 10,000 bootstraps, based on Kimura 2-parameter distances (K2P) (Kimura 1980). For comparison, a Bayesian inference approach (BI) was used as well. The Bayesian analysis was conducted with MRBAYES v3.1.2 (Huelsenbeck and Ronquist 2001, Ronquist and Huelsenbeck 2003) under the GTR+I+G model, simulating 4 Monte Carlo Markov Chains (MCMC) for 2,000,000 generations each. Trees were sampled every 100 generations and the first 300,000 generations were excluded as burn-in. A consensus tree was constructed with posterior probabilities. The MRBAYES analyses were carried out on the Bioportal at Oslo University (http://www.bioportal.uio.no). The GTR+I+G model used in MRBAYES is closely related to the TIM3+I+G model, which was selected by JMODELTEST v0.1.1 (Guindon and Gascuel 2003, Posada 2008) as the best-fit model under the Akaike information criterion (AIC).

Genetic distances and nucleotide diagnostics

As K2P-distance is the most commonly used distance metric in DNA barcoding (Hebert et al. 2003), it was employed here for comparison. It allows to compare the behavior of the DNA fragment we used to the standard barcode region which is situated in the same gene. When possible, simple nucleotide diagnostics were identified for each (sub)species. If less than two simple nucleotide diagnostics were present (Sarkar et al. 2002), a compound diagnostic was detected using the algorithm of Wong et al. (2009).

Results

Alignment and sequence quality

Of a total of 76 samples, thirteen samples with low quality sequences were removed: five *L. c. cervus*, one *L. c. pentaphyllus*, three *L. c. turcicus* and four *L.* (*P.*) *barbarossa*.

Three sequences showed a few double peaks: one *L.* (*P.*) *barbarossa* (SB6: 5 ambiguous sites), one *L.* (*P.*) *macrophyllus* (UB1: 7 ambiguous sites) and one unidentified species of *Lucanus* (J2: 2 ambiguous sites) (Table 1). None exhibited indels or stop codons which are indicative of the presence of NUMTs (Buhay 2009). The remaining 63 samples and 11 sequences obtained from GenBank are listed in Table 1. The final alignment entailed 74 sequences, representing 60 haplotypes. Incomplete sequences were obtained for the following taxa: taxon H4 with 500 bp of which the reverse sequence failed and taxon J2 of which forward sequences of only the first and third smaller fragments could be produced, resulting in a total of 383 bp. Both taxa were specimens of the unidentified *Lucanus* specimens (Table 1). Likewise, the sequence of *L. tetraodon* found in GenBank (named X2), was 122 bp short at the 3' end. One other taxon, H3 (*Lucanus* sp.) missed a mere 5 bp at the 5' end.

Both the NJ tree and the BI tree showed the same overall configuration (Figure 1 and Appendix 1, respectively) except for the position of the unidentified *Lucanus* specimens. In the NJ tree these specimens fall into two clusters with unresolved affinities (Figure 1). In the BI tree they form a single well-supported clade together with specimens identified as *L. c. judaicus* and *L. c. laticornis* (Appendix 1). The unidentified specimens fail to form a single monophyletic cluster as one subclade also includes *L. c. judaicus*. The BI tree showed *L. c. laticornis* to be monophyletic with probability 1, instead of paraphyletic as was shown in the NJ tree with bootstrap support below 70%. In both trees, several species as well as subspecies fall into distinct clades, whereas *L. c. cervus*, *L. c. turcicus*, *L. c. pentaphyllus*, *L.* (*P.*) *macrophyllus* and *L. ibericus* cluster in the same shallow clade (called the '*L. c. cervus* clade' hereafter). In addition, three out of four samples of *L. c. pentaphyllus* share a haplotype with *L. c. cervus* (haplotype A3) which is the most common haplotype among *L. cervus* sequences (Table 1). Within this clade *L. c. cervus*, *L. c. turcicus* and *L. c. pentaphyllus* are polyphyletic. Unexpectedly, one sample of *L.* (*P.*) *barbarossa* and the sample of *L.* (*P.*) *macrophyllus* are also embedded in this clade. Looking at the sequences, they only differ from haplotype A3 at their five and seven ambiguous sites, respectively. Because the two other specimens of *L.* (*P.*) *barbarossa* form a separate clade with *L. c. fabiani*, sample SB6 is excluded from further calculations but will be discussed below.

Genetic distances

The nucleotide composition of all the sequences was AT-rich, with 29.5% A, 35.2% T, 15.5% G and 19.7% C. There were 36.4% nucleotide sites variable and 12.1% variable amino acid sites, of which 94.3% and 77.8% were parsimony informative, respectively. When *D. parallelipipedus* was excluded from the dataset, variable sites decreased to 33.3% for nucleotides and 7.2% for amino acids (94.2% and 56.2% parsimony informative, respectively). Nucleotide composition and K2P-distances calculated for each codon position are shown in Table 3.

Although specimen J2 of the unidentified specimens of *Lucanus* clustered with the other specimens of the same taxon in the NJ and BI trees, the pairwise interspecific

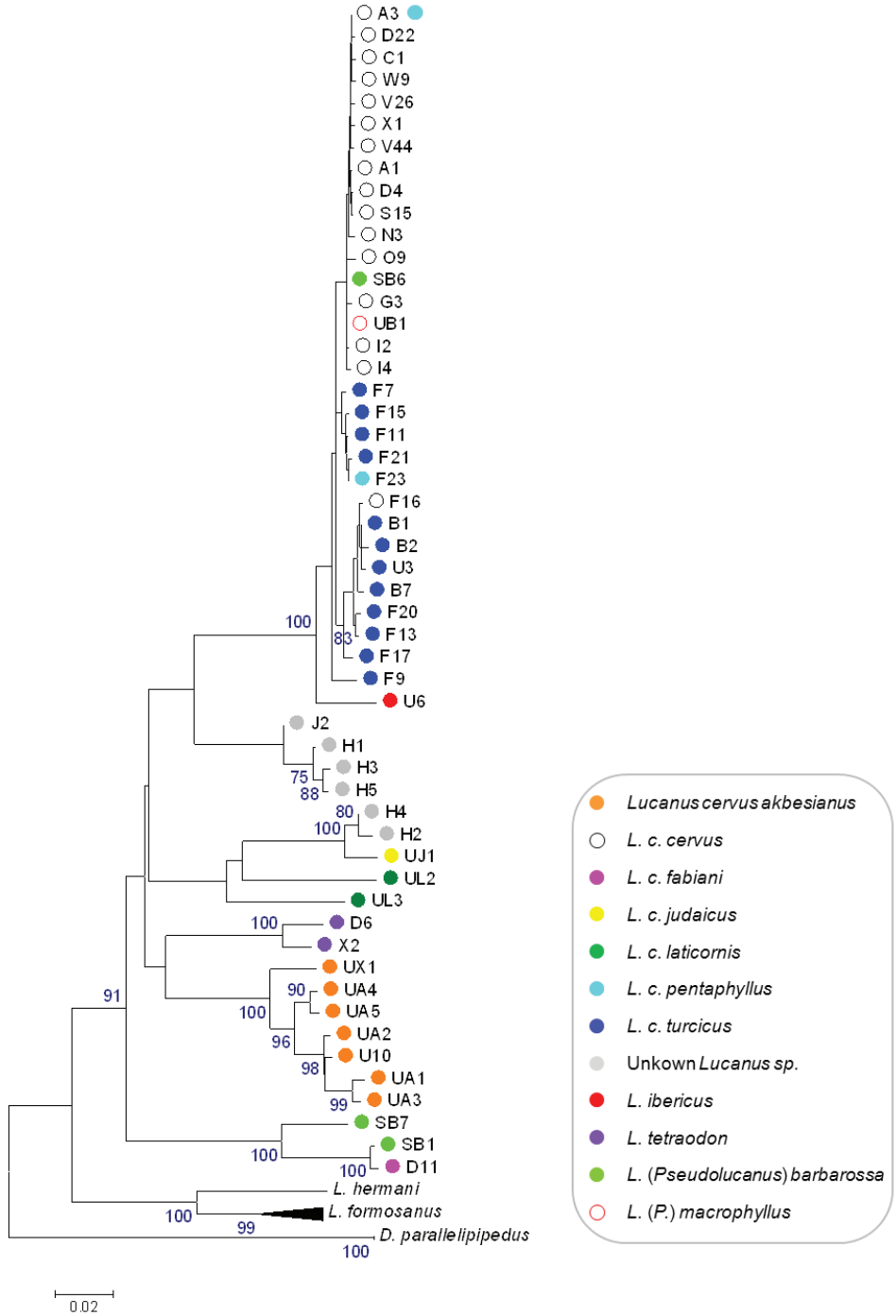


Figure 1. Bootstrap consensus NJ tree inferred from 10,000 replicates, with a cut off value of 70%, based on K2P-distances between 60 haplotypes of the 3' end of the COI gene.

K2P-distances with J2 differed substantially from those with H1 to H5 (comparisons with *L. c. judaicus* not included). More specifically, the minimum pairwise interspecific K2P-distance between J2 and the other unidentified taxa was 0.064 opposed to 0.087–0.095 when taking H1 to H5 into account. J2 is one of three incomplete sequences and missing information from position 179 to 399 in the sequence of J2 where several simple nucleotide diagnostics are present (Appendix 2). Therefore, this sample was removed from the dataset for subsequent analysis.

The congeneric interspecific K2P-distances between the western Palaearctic taxa and the eastern Asian species *L. formosanus* and *L. hermani* range from 0.156 to 0.198. Distances between taxa of *Lucanus* and *Dorcus* went from 0.211 until 0.259. K2P-distances within and between the investigated western Palaearctic taxa of *Lucanus* are shown in Table 2. As indicated by the NJ and BI trees, the taxa *L. c. cervus*, *L. c. pentaphyllus*, *L. c. turcicus* and *L.* (*P.*) *macrophyllus* cannot be distinguished based on the COI fragment; K2P-distances range from 0 to 0.021, and all taxa are reciprocally polyphyletic. Whereas the first three subspecies of *L. cervus* are distinguished solely on the basis of the number of lamellae on the antennal club, *L.* (*P.*) *macrophyllus* is morphologically much more distinctive, resembling *L. ibericus*. Although *L. ibericus* is part of the *L. c. cervus* clade, it shows slightly higher K2P-distances with the other members of this clade (0.028–0.032). Note that we only had a single specimen. Moderate to relatively high within (sub)species distances were found for *L. c. laticornis* (0.085), certain specimens of the unidentified *Lucanus* sp. (max. 0.054) and *L.* (*P.*) *barbarossa* (0.53). On the other hand, between the latter and *L. c. fabiani* a small to moderate distance exists (0.004 and 0.058). This is also the case between taxa H2 and H4 of the unknown *Lucanus* sp. and *L. c. judaicus* (K2P-distance of 0.018 and 0.016, respectively). The remaining distances between (sub)species ranged from 0.087 and 0.179.

These results do not show a distinct barcoding gap or other threshold to distinguish putative species, which is chiefly due to a lack of phylogenetic resolution to differentiate the said species and subspecies. If we consider the taxa of the *L. c. cervus* clade to be members of the same species, 99.4% of all intra(sub)specific comparisons showed K2P-distances below 5% and 99.8% of the pairwise inter(sub)specific distances were above 5%. Nucleotide diagnostics are listed in Appendix 2. No diagnostic combination of nucleotide positions and characters were found for the taxa of the *L. c. cervus* clade, *L. ibericus* not included. As the number of species and the sample size per species are rather limited, the nucleotide diagnostics should be considered with caution.

Discussion

The present study shows that the sequenced COI fragment could discriminate several of the investigated western Palaearctic *Lucanus* species and alleged subspecies of *L. cervus*. Well differentiated species and subspecies were *L. c. akbesianus*, *L. c. laticornis* and *L. tetraodon*, as well as the two eastern Asian species *L. formosanus* and *L. hermani*. Difficulties in molecular identification remained between *L. c. fabiani* and *L.* (*P.*) *barba-*

Table 2. Intra- and interspecific K2P-distances for the 670 bp COI gene of western Palaearctic *Lucanus* (sub)species. NA: intraspecific K2P-distance cannot be presented because only one sample is available.

	Lucanus cervus cervus	*L. c. pentaphyllus*	*L. c. turcicus*	*L. c. fabiani*	*L. c. akbesianus*	*L. c. judaicus*	*L. c. laticornis*	*L. ibericus*	*L. tetraodon*	*L. (P.) macrophyllus*	*L. (P.) barbarossa*	unknown *Lucanus* sp.
L. c. cervus	0–0.018											
L. c. pentaphyllus	0–0.018	0–0.014										
L. c. turcicus	0.001–0.021	0.003–0.017	0–0.017									
L. c. fabiani	0.161–0.167	0.160–0.163	0.159–0.169	NA								
L. c. akbesianus	0.118–0.161	0.121–0.155	0.121–0.165	0.159–0.174	0–0.045							
L. c. judaicus	0.151–0.164	0.153–0.160	0.155–0.170	0.167	0.144–0.154	NA						
L. c. laticornis	0.134–0.160	0.134–0.155	0.132–0.164	0.162–0.165	0.135–0.150	0.089–0.094	0.085					
L. ibericus	0.029–0.039	0.034–0.035	0.028–0.037	0.174	0.132–0.151	0.174	0.141–0.168	NA				
L. tetraodon	0.125–0.129	0.124–0.128	0.122–0.130	0.168–0.179	0.098–0.123	0.151–0.156	0.132–0.151	0.131–0.136	0.024			
L. (P.) macrophyllus	0–0.012	0–0.014	0.006–0.015	0.159	0.116–0.141	0.147	0.130–0.145	0.028	0.120–0.124	NA		
L. (P.) barbarossa	0.153–0.163	0.155–0.161	0.155–0.167	0.004–0.058	0.127–0.171	0.153–0.165	0.146–0.167	0.166–0.172	0.159–0.177	0.149–0.157	0.053	
unknown *Lucanus* sp.	0.091–0.162	0.093–0.159	0.95–0.168	0.143–0.172	0.119–0.150	0.016–0.066	0.088–0.113	0.109–0.169	0.120–0.152	0.087–0.147	0.136–0.170	0.002–0.054

Table 3. Nucleotide composition and K2P-distances at each codon position of the 670 bp COI region.

	Codon position		
	1st	**2nd**	**3rd**
% A	31.4	18.9	38.2
% T	26.6	42.5	36.6
% G	25.6	16.2	4.9
% C	16.4	22.4	20.4
K2P-distance	0–0.107	0–0.032	0–0.999

rossa, *L. c. judaicus* and the unidentified *Lucanus* species, and between taxa of the *L. c. cervus* clade. Although thoroughly sampled within their distribution range, *L. c. cervus* and *L. c. turcicus* could not be discriminated with a barcoding approach. Likewise, three out of four samples of *L. c. pentaphyllus* possessed the most common haplotype of *L. c. cervus*. Next to introgression following recent or past hybridisation events, incomplete sorting of ancestral variation may be the reason for the polyphyletic pattern. It is not known if *Lucanus* can be infected with the endosymbiotic bacteria *Wolbachia*, which can cause mitochondrial introgression between closely related species (e.g. Whitworth et al. 2007). Nonetheless, infections with *Wolbachia* are quite common among insects, and should be taken into account (Hilgenboecker et al. 2008). However, the shift from four to five or even six lamellar segments on the antennal club is, at least in this tree of maternal inheritance, not synapomorphic among all individuals, and the number of lamellae may represent a case of parallel evolution or a phenotypically plastic trait within *L. cervus*, such that *pentaphyllus* and *turcicus* may merely represent morphotypes of *L. cervus*. This hypothesis seems less likely for *L.* (*P.*) *macrophyllus*. Although this taxon's haplotype only differed from the main *L. c. cervus* haplotype, A3, by its seven ambiguous sites, it has a very distinct morphology. The same can be said about *L. ibericus*, which was part of the same clade, but showed higher pairwise K2P-distances (0.028–0.032) when comparing it to the other taxa of the clade. Lumping *L. ibericus* and *L.* (*P.*) *macrophyllus* together with the *L. cervus* subspecies *cervus*, *turcicus* and *pentaphyllus* seems therefore ill advice.

Like *L.* (*P.*) *macrophyllus*, one sample of *L.* (*P.*) *barbarossa*, SB6, was embedded in the *L. c. cervus* clade, opposed to the other two samples that clustered with *L. c. fabiani*. The taxa of the latter group showed K2P-distances between 0.004 and 0.058, which indicates a close relationship between *L. c. fabiani* and *L.* (*P.*) *barbarossa*, as well as *L.* (*P.*) *barbarossa* being very variable. High intraspecific variability could be indicative of cryptic diversity or population structure (Diptera: Meier et al. 2006; Lepidoptera, Lycaenidae: Wiemers and Fiedler 2007; Coleoptera, Nitidulidae: De Biase et al. 2012; Hemiptera, Cicadidae: Nunes et al. 2013). Despite the moderate to low genetic distance between *L.* (*P.*) *barbarossa* and *L. c. fabiani*, these taxa are morphologically very distinct. This leaves us with either incomplete lineage sorting or introgression. Considering that both taxa have very proximate distribution ranges, introgressive hybridisation is likely. Even complete loss of the original mitochondrial genome of a species,

resulting in a species with only mitochondrial genomes of the introgressed species is not unheard of (Hailer et al. 2012). Likewise, as *L. c. cervus* and *L.* (*P.*) *barbarossa* occur sympatrically in Spain and Portugal (Méndez 2003), recent hybridisation and introgression cannot be ruled out as another or supplementary cause of the polyphyletic status of *L.* (*P.*) *barbarossa* (Avise 2000). Because SB6 merely differed from A3 at its five ambiguous sites, it could be perceived as a shared haplotype, which would corroborate this hypothesis (e.g. Nicholls et al. 2012). *Lucanus cervus akbesianus*, *L. c. laticornis* and *L. c. judaicus* also have overlapping distributions. The former two were even sampled on the same tree in a Turkish forest (M. A. Cimaz, personal communication). In captivity, they do not seem to interbreed, which is concordant with our reporting of no shared haplotypes.

Finally, the *Lucanus* samples from Israel and Lebanon that were unidentified at the species level, seemed closely related and formed a paraphyletic clade with *L. c. judaicus*. Nevertheless, some of these samples could well be of a different species, indicated by the higher pairwise genetic distances (0.042–0.066). A detailed morphological and phylogenetic study is required here to investigate the number of species and relationship with *L. c. judaicus*.

A distinct barcoding gap was absent for several species and subspecies of *Lucanus*. This may either represent a low phylogenetic signal from the COI fragment for some relationships, a problem of basing a taxonomy on just one or a few morphological traits, or both. The use of the COI gene for barcoding purposes has had mixed results. High intraspecific variability (DeSalle et al. 2005) and closely related species (e.g. Funk and Omland 2003, Hajibabaei et al. 2006) can lead to an overlap in genetic distances, making the technique ineffective, as was shown here. In addition, NUMTs may complicate results and could cause the number of species to be overestimated (Song et al. 2008). Besides, the evolutionary history of the gene in question could be different from that of the studied species (Maddison 1997, Edwards 2009). Consequently, other or additional genes, ribosomal or nuclear, are recommended for barcoding purposes (Dupuis et al. 2012).

Conclusions

This study revealed that while the 3' terminus of COI contained sufficient information to resolve relationships among a number of closely related taxa, many others could not be robustly discriminated. Genotyping of additional specimens, especially of *L.* (*P.*) *macrophyllus*, *L. ibericus*, *L. c. judaicus*, *L. c. fabiani* and *L. c. laticornis*, as well as all western Palaearctic taxa is needed to fully explore COI genetic diversity and to investigate the roles of phenotypic plasticity, hybridisation and incomplete lineage sorting underlying stag beetle biodiversity and inform taxonomic investigations. We therefore see this study as a starting point for future research which should also endeavour to combine analysis of nuclear markers, such as the internal transcribed spacer (ITS) and 28S rRNA gene (e.g. Smith et al. 2007), in combination with a detailed morphological investigation, to find a useful molecular identification tool for all western Palaearctic *Lucanus* sp.

Authors' contributions

The work presented here was carried out in collaboration between all authors. AT, KDG, GA, PA and LB defined the subject and the design of the study. KDG designed methods and experiments in the laboratory and supervised laboratory work. KC analysed the data, interpreted results and wrote the paper. AT was responsible for collecting the samples and co-wrote the taxonomical part of the paper. JM discussed analyses. GA, ES, NMcK and PS provided five sequences and revised primarily the material and methods section and the interpretation of the results. MZ, LB and PA provided samples and co-wrote the paper, particularly the taxonomical section. DH and RM provided samples. All authors have contributed to, revised and approved the manuscript.

Acknowledgements

We want to thank the following people for generously providing tissue samples of stag beetles or information: E. Atay, M. Avci, L. Barbiero, R. Bekchiev, G. Bonamie, S. Boucher, C. Bouget, H. Brustel, D. G. Carrilero, L.R. Castro, G. De Coninck, J. Ibero Caballero, A.M. Cimaz, I. de las Monjas, M. Fremlin, N. Gouix, C. Hawes, J-P. Huang, N. Jansson, A. Kairouz, S. Korneyev, V.A. Korneyev, D. Kovalchuk, I. López Pérez, Á. Martínez García, M. Méndez, M. Murat, L. Nádai, I. Nel, E. Orbach, H. Podskalská, S. Rastrero Sánchez, S. Reicher, O. Rittner, F. Roviralta Peña, P. Šípek, L. Valladares, J.T. Smit, Á. R. Quirós Menéndez. Also many thanks to Leen Verschaeve, Nancy Van Liefferinge, An Van Breusegem, David Halfmaerten and Sabrina Neyrinck (INBO) for laboratory assistance. We appreciate the constructive comments of three anonymous reviewers.

References

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Research 25: 4692–4693. doi: 10.1093/nar/25.22.4692

Avise JC (2000) Phylogeography: the history and formation of species. Harvard University Press, Cambridge, MA, 447 pp.

Baraud J (1993) Les Coléoptères Lucanidae de l'Europe et du Nord de l'Afrique. Bulletin mensuel de la Société Linnéenne de Lyon 62: 42–64

Bartolozzi L, Sprecher-Uebersax E (2006) Lucanidae. In: Löbl I, Smetana A (Eds) Catalogue of Palaearctic Coleoptera. Apollo Books, Stenstrup, 63–76.

Benesh B (1960) Lucanidae. W. Junk, The Hague, 178 pp.

Buhay JE (2009) "COI-like" Sequences are becoming problematic in molecular systematic and DNA barcoding studies. Journal of Crustacean Biology 29: 96–110. doi: 10.1651/08-3020.1

Buse J, Levanony T, Timm A, Dayan T, Assmann T (2010) Saproxylic beetle assemblages in the Mediterranean region: Impact of forest management on richness and structure. Forest Ecology and Management 259: 1376–1384. doi: 10.1016/j.foreco.2010.01.004

Calvignac S, Konecny L, Malard F, Douady CJ (2011) Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts). Mitochondrion 11: 246–254. doi: 10.1016/j.mito.2010.10.004

Clark JT (1977) Aspects of variation in the stag beetle *Lucanus cervus* (L.) (Coleoptera: Lucanidae). Systematic Entomology 2: 9–16. doi: 10.1111/j.1365-3113.1977.tb00350.x

De Biase A, Antonini G, Mancini E, Trizzino M, Cline A, Audisio P (2012) Discordant patterns in the genetic, ecological, and morphological diversification of a recently radiated phytophagous beetle clade (Coleoptera: Nitidulidae: Meligethinae). Rendiconti Lincei 23: 207–215. doi: 10.1007/s12210-012-0174-4

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1905–1916. doi: 10.1098/rstb.2005.1722

Didier DR, Séguy E (1953) Catalogue illustré des Lucanides du globe. Encyclopédie Entomologique Serie A. Paul Lechevalier, Paris, 223 pp.

Dupuis JR, Roe AD, Sperling FAH (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. Molecular Ecology 21: 4422–4436. doi: 10.1111/j.1365-294X.2012.05642.x

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63: 1–19. doi: 10.1111/j.1558-5646.2008.00549.x

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. Molecular Ecology 10: 2741–2752. doi: 10.1046/j.0962-1083.2001.01411.x

Fujita H (2010) The Lucanid beetles of the world. Mushi-Sha, Tokyo, 736 pp.

Funk DJ, Omland KE (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. doi: 10.1146/annurev.ecolsys.34.011802.132421

Goka K, Kojima H, Okabe K (2004) Biological invasion caused by commercialization of stag beetles in Japan. Global Environmental Research 8: 67–74.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704. doi: 10.1080/10635150390235520

Hailer F, Kutschera VE, Hallström BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A (2012) Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336: 344–347. doi: 10.1126/science.1216424

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proceedings of the National Academy of Sciences of the USA 103: 968–971. doi: 10.1073/pnas.0510466103

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

Hallan J (2008) Synopsis of the described Coleoptera of the World. Biology Catalogue of the Texas A&M University [WWW document]. URL http://insects.tamu.edu/research/collection/hallan/test/Arthropoda/Insects/Coleoptera/Family/Lucanidae.txt [accessed on 7 January 2013]

Harvey DJ, Gange AC (2006) Size variation and mating success in the stag beetle, *Lucanus cervus*. Physiological Entomology 31: 218–226. doi: 10.1111/j.1365-3032.2006.00509.x

Harvey DJ, Gange AC, Hawes CJ, Rink M (2011) Bionomics and distribution of the stag beetle, *Lucanus cervus* (L.) across Europe. Insect Conservation and Diversity 4: 23–38. doi: 10.1111/j.1752-4598.2010.00107.x

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society Series B 270: 313–321. doi: 10.1098/rspb.2002.2218

Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH (2008) How many species are infected with *Wolbachia*? – a statistical analysis of current data. FEMS Microbiology Letters 281: 215–220. doi: 10.1111/j.1574-6968.2008.01110.x

Holden C (2007) Beetle battles. Science 318: 25. doi: 10.1126/science.318.5847.25a

Huang J-P, Lin C-P (2010) Diversification in subtropical mountains: Phylogeography, Pleistocene demographic expansion, and evolution of polyphenic mandibles in Taiwanese stag beetle, *Lucanus formosanus*. Molecular Phylogenetics and Evolution 57: 1149–1161. doi: 10.1016/j.ympev.2010.10.012

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755. doi: 10.1093/bioinformatics/17.8.754

Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, John OS, Wild R, Hammond PM, Ahrens D, Balke M, Caterino MS, Gómez-Zurita J, Ribera I, Barraclough TG, Bocakova M, Bocak L, Vogler AP (2007) A Comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science 318: 1913–1916. doi: 10.1126/science.1146954

IUCN Red List of Threatened Species. http://www.iucnredlist.org [accessed 9 October 2012]

Jansson N, Coskun M (2008) How similar is the saproxylic beetle fauna on old oaks (*Quercus* spp.) in Turkey and Sweden? Revue de Ecologie-La Terre et la Vie 10: 91–99. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-81857

Kanzaki N, Taki H, Masuya H, Okabe K, Tanaka R, Abe F (2011) Diversity of stag beetle-associated nematodes in Japan. Environmental Entomology 40: 281–288. doi: 10.1603/EN10182

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120. doi: 10.1007/BF01731581

Krajcik M (2001) Lucanidae of the world. Catalogue - Part I. Checklist of the stag beetles of the world (Coleoptera: Lucanidae). Milan Krajcik, Most, 108 pp.

Lin C-P, Huang J-P, Lee Y-H, Chen M-Y (2011) Phylogenetic position of a threatened stag beetle, *Lucanus datunensis* (Coleoptera: Lucanidae) in Taiwan and implications for conservation. Conservation Genetics 12: 337–341. doi: 10.1007/s10592-009-9996-8

Luce JM (1996) *Lucanus cervus* (Linnaeus, 1758). In: van Helsdingen PJ, Willemse L, Speight MCD (Eds) Background information on invertebrates of the Habitat Directive and the Bern Convention. Council of Europe Publishing, Strasbourg, 53–58.

Maddison WP (1997) Gene trees in species trees. Systematic Biology 46: 523–536. doi: 10.1093/sysbio/46.3.523

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Méndez M (2003) Conservation of *Lucanus cervus* in Spain: an amateur's perspective. In: Proceedings of the second pan-European conference on saproxilic beetles. People's Trust for Endangered Species, 1–3.

Moritz C (1994a) Applications of mitochondrial DNA analysis in conservation: a critical review. Molecular Ecology 3: 401–411. doi: 10.1111/j.1365-294X.1994.tb00080.x

Moritz C (1994b) Defining 'evolutionarily significant units' for conservation. Trends in Ecology & Evolution 9: 373–374. doi: 10.1016/0169-5347(94)90057-4

Nicholls JA, Challis RJ, Mutun S, Stone GN (2012) Mitochondrial barcodes are diagnostic of shared refugia but not species in hybridizing oak gallwasps. Molecular Ecology 21: 4051–4062. doi: 10.1111/j.1365-294X.2012.05683.x

Nieto A, Alexander KNA (2010) European red list of saproxylic beetles. Publications Office of the European Union, Luxembourg, 45 pp.

Nunes VL, Mendes R, Marabuto E, Novais BM, Hertach T, Quartau JA, Seabra SG, Paulo OS, Simões PC (2013) Conflicting patterns of DNA barcoding and taxonomy in the cicada genus *Tettigettalna* from southern Europe (Hemiptera: Cicadidae). Molecular Ecology Resources. doi: 10.1111/1755-0998.12158

Planet LM (1899) Essai monographique sur les Coléoptères des genres Pseudolucane & Lucane. E. Deyrolle, Paris, 144 pp.

Posada D (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25: 1253–1256. doi: 10.1093/molbev/msn083

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3 : Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574. doi: 10.1093/bioinformatics/btg180

Ryder OA (1986) Species conservation and systematics: the dilemma of subspecies. Trends in Ecology & Evolution 1: 9–10. doi: 10.1016/0169-5347(86)90059-5

Santoro F, Pivotti I, D' Allestro V, Fabrizi A, Di Veroli A, Di Giovanni MV, Corallini C, Mandrioli M, Goretti E (2009) *Lucanus cervus* e *Lucanus tetraodon* (Coleoptera – Lucanidae) in Umbria. Bollettino del Museo e Istituto di Biologia dell'Università di Genova 71: 202

Sarkar IN, Thornton JW, Planet PJ, Figurski DH, Schierwater B, DeSalle R (2002) An automated phylogenetic key for classifying homeoboxes. Molecular Phylogenetics and Evolution 24: 388–399. doi: 10.1016/S1055-7903(02)00259-2

Schenk K-D, Fiedler F (2011) A new discovery of *Lucanus* (*Pseudolucanus*) *busignyi* in Turkey. Beetles World 5: 11–16.

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Floors 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 MA, Wood DM, Janzen DH, Hallwachs W, Hebert PDN (2007) DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae)

are not all generalists. Proceedings of the National Academy of Sciences of the USA 104: 4967–4972. doi: 10.1073/pnas.0700050104

Song H, Buhay JE, Whiting MF, Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences of the USA 105: 13486–13491. doi: 10.1073/pnas.0803076105

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Thomaes A (2009) A protection strategy for the stag beetle (*Lucanus cervus*, (L., 1758), Lucanidae) based on habitat requirements and colonisation capacity. In: Buse J, Alexander KNA, Ranius T, Assmann T (Eds) Proceedings of the 5th Symposium and Workshop on the Conservation of Saproxylic Beetles. Pensoft Publishers, 149–160.

Thomaes A, Kervyn T, Maes D (2008) Applying species distribution modelling for the conservation of the threatened saproxylic Stag Beetle (*Lucanus cervus*). Biological Conservation 141: 1400–1410. doi: 10.1016/j.biocon.2008.03.018

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680. doi: 10.1093/nar/22.22.4673

Tournant P, Joseph L, Goka K, Courchamp F (2012) The rarity and overexploitation paradox: stag beetle collections in Japan. Biodiversity and Conservation 21: 1425–1440. doi: 10.1007/s10531-012-0253-y

Waples RS (1991) Pacific salmon, *Oncorhynchus* spp., and the definition of "species" under the Endangered Species Act. Marine Fisheries Review 53: 11–22.

Whitworth TL, Dawson RD, Magalon H, Baudry E (2007) DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). Proceedings of the Royal Society B 274: 1731–1739. doi: 10.1098/rspb.2007.0062

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). Frontiers in Zoology 4: 8. doi: 10.1186/1742-9994-4-8

Wong EHK, Shivji MS, Hanner RH (2009) Identifying sharks with DNA barcodes: assessing the utility of a nucleotide diagnostic approach. Molecular Ecology Resources 9: 243–256. doi: 10.1111/j.1755-0998.2009.02653.x

Appendix 1

Consensus Bayesian tree of 60 haplotypes of the 3' end of the COI gene. Values given by the nodes are posterior probabilities above 0.70. (doi: 10.3897/zookeys.365.5526.app1) File format: Adobe PDF file (pdf).

Citation: Cox K, Thomaes A, Antonini G, Zilioli M, De Gelas K, Harvey D, Solano E, Audisio P, McKeown N, Shaw P, Minetti R, Bartolozzi L, Mergeay J (2013) Testing the performance of a fragment of the COI gene to identify western Palaearctic stag beetle species (Coleoptera, Lucanidae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 105–126. doi: 10.3897/zookeys.365.5526 Consensus Bayesian tree. doi: 10.3897/zookeys.365.5526.app1

Appendix 2

Nucleotide diagnostics for (sub)species or species groups according to the Neighbour-Joining and Bayesian Inference tree topology. (doi: 10.3897/zookeys.365.5526.app2) File format: Adobe PDF file (pdf).

Citation: Cox K, Thomaes A, Antonini G, Zilioli M, De Gelas K, Harvey D, Solano E, Audisio P, McKeown N, Shaw P, Minetti R, Bartolozzi L, Mergeay J (2013) Testing the performance of a fragment of the COI gene to identify western Palaearctic stag beetle species (Coleoptera, Lucanidae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 105–126. doi: 10.3897/zookeys.365.5526 Nucleotide diagnostics for (sub)species. doi: 10.3897/zookeys.365.5526.app2

ZooKeys 365: 127–147 (2013)
doi: 10.3897/zookeys.365.5728
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Incorporating *trnH-psbA* to the core DNA barcodes improves significantly species discrimination within southern African Combretaceae

Jephris Gere[1], Kowiyou Yessoufou[1,2], Barnabas H. Daru[1], Ledile T. Mankga[1], Olivier Maurin[1], Michelle van der Bank[1]

1 *African Centre for DNA Barcoding, Department of Botany & Plant Biotechnology, University of Johannesburg, PO Box 524, South Africa* **2** *C4 EcoSolutions, 9 Mohr Road Tokai, Cape Town, South Africa 7945*

Corresponding author: *Jephris Gere* (gerejephris@gmail.com)

Academic editor: *K. Jordaens* | Received 1 June 2013 | Accepted 13 September 2013 | Published 30 December 2013

Citation: Gere J, Yessoufou K, Daru BH, Mankga LT, Maurin O, van der Bank M (2013) Incorporating *trnH-psbA* to the core DNA barcodes improves significantly species discrimination within southern African Combretaceae. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 127–147. doi: 10.3897/zookeys.365.5728

Abstract

Recent studies indicate that the discriminatory power of the core DNA barcodes (*rbc*La + *mat*K) for land plants may have been overestimated since their performance have been tested only on few closely related species. In this study we focused mainly on how the addition of complementary barcodes (nrITS and *trnH-psbA*) to the core barcodes will affect the performance of the core barcodes in discriminating closely related species from family to section levels. In general, we found that the core barcodes performed poorly compared to the various combinations tested. Using multiple criteria, we finally advocated for the use of the core + *trnH-psbA* as potential DNA barcode for the family Combretaceae at least in southern Africa. Our results also indicate that the success of DNA barcoding in discriminating closely related species may be related to evolutionary and possibly the biogeographic histories of the taxonomic group tested.

Keywords

DNA barcoding, closely related species, Combretaceae, southern Africa

Introduction

Combretaceae is a medium-sized family within Myrtales, comprising about 500 species in 17 to 23 genera. It has long been referred to as a complex phylogenetic and taxonomic group (Tan et al. 2002, Maurin et al. 2010, Stace 2010, Jordaan et al. 2011). Based on morphological characters and phylogenetic analysis, the family Combretaceae has been recovered as monophyletic and sister to the rest of Myrtales (Brown 1810, Dahlgren and Thorne 1984, Tan et al. 2002, Sytsma et al. 2004, Maurin et al. 2010, Stace 2010). Members of Combretaceae are mainly trees, shrubs or lianas, occupying a wide range of habitats from savannas, forests, to woodlands (Maurin et al. 2010) and are distributed in tropical and subtropical regions across the globe. With ca. 350 species, *Combretum* Loefl., the largest genus in the family has its centre of diversity in Africa, with approximately 63 species described in southern Africa – south of the Zambezi river and includes South Africa, Zimbabwe, Namibia, Botswana, Lesotho, Swaziland, and Mozambique (Maurin et al. 2010, Jordaan et al. 2011).

The major distinguishing feature of the family is the presence of unicellular combretaceous hairs on the abaxial leaf surfaces, a diagnostic trait in many other species of Myrtales and even beyond the group e.g. the Cistaceae Juss. family, tribe Cisteae (Maurin et al. 2010, Stace 2010). However, other morphological features such as presence of trichomes, stalked glands, domatia, inflorescence, fruit shape, leaf and pollen morphology are also important for species delimitation in Combretaceae (Exell and Stace 1966, Stace 2007, 2010, Maurin et al. 2010, Jordaan et al. 2011). Nonetheless, all these characters are not adequate enough to delimit species within the family because none is unique to a specific clade. As a result, the family has experienced several splitting and lumping in the past (El Ghazlai et al. 1998, Tan et al. 2002, Maurin et al. 2010, Stace 2010, Jordaan et al. 2011). Also, the taxonomy is further confounded by the high morphological similarity between members of different sections. For instance, inflorescence and fruit shapes are very similar between species and across clades (Figures 1 and 2). Such homoplasious morphological similarities have also been identified as the root of difficulties in delimiting the genera; for example in the *Combretum–Quisqualis* clade (Jordaan et al. 2011). Consequently, it becomes necessary to search for an alternative method to augment traditional morphology-based taxonomy of Combretaceae.

Here, we propose that DNA barcoding may provide such a complementary tool to ease species delimitation within the group. DNA barcoding involves the use of a short and standardised DNA sequence that can help assign, even biological specimens devoid of diagnostic features, to species (Hebert et al. 2004, 2010, Hajibabaei et al. 2006, Roy et al. 2010, Van der Bank et al. 2012, Franzini et al. 2013). Two DNA regions defined as 'core barcodes', i.e. *rbc*La and *mat*K have been standardised as DNA barcodes for land plants (CBOL Plant Working Group 2009). In addition to the core barcodes, two other regions, *trnH-psbA* and nrITS were suggested as supplementary DNA barcodes for plants (Hollingsworth et al. 2011, Li et al. 2011). The rationale for adopting these two regions (*rbc*La and *mat*K) is high levels of recoverability of high-

Figure 1. Selected inflorescences of seven *Combretum* species indicating closely related species evaluated based upon floral characters. **a** *C. paniculatum* **b** C. *microphyllum* **c** *C. platypetalum* **d** *C. hereroense* **e** *C. apiculatum* **f** *C. molle* **g** *C. kraussii*. All photographs by O. Maurin.

quality sequences and acceptable levels of species discrimination (Burgess et al. 2011). The discriminatory power of the core DNA barcodes for land plants was estimated at 70–80% (CBOL Plant Working Group 2009, Fazekas et al. 2009, Kress and Erickson 2007). However, a recent study suggests that efficacy of core barcodes may have been overestimated, arguing that taxon sampling has been biased towards less-related species (Clement and Donoghue 2012). Furthermore, barcoding efficacy is rarely evaluated in a phylogenetic context (but see Clement and Donoghue 2012), resulting in potentially biased estimates of discriminatory power.

In this study, we evaluated the efficacy of DNA barcoding as a tool to augment morphological species discrimination within Combretaceae. Specifically, we (1) assessed the potential of four markers to discriminate southern African species of the family, and (2) assessed the efficacy of barcodes across major clades including subgenera and sections within the largest genus *Combretum*.

Figure 2. Selected mature dry four-winged fruits of closely related species of genus *Combretum*. **a** *C. mkuzense* **b** *C. microphyllum* **c** *C. englerii* **d** *C. apiculatum* **e** *C. moggii* **f** *C. albopunctatum* **g** *C. collinum*. All photographs by O. Maurin.

Methods

Sampling includes one to six accessions of 58 species out of the 63 species representing the six genera of Combretaceae in southern Africa. These genera include *Combretum* (43 species included in this study), *Lumnitzeria* Wild. (one species included), *Meiostemon* Exell and Stace (one species included), and *Quisqualis* L. (one species included), *Pteleopsis* Engl. (two species included), and *Terminalia* (nine species included).

Collection details, taxonomy, voucher numbers, GPS coordinates, field pictures, and sequence data (only *mat*K and *rbc*La) are archived online on the BOLD system (www.boldsystems.org). Voucher information, name of herbarium, GenBank and BOLD accession numbers are listed in Appendix 1.

DNA extraction, amplification and alignment

Genomic DNA was extracted from silica gel-dried and herbarium leaf material following a modified cetyltrimethyl ammonium bromide (CTAB) method of Doyle and Doyle (1987). To ease the effects of high polysaccharide concentrations in the DNA samples, we added polyvinyl pyrolidone (2% PVP). Purification of samples was done using QIAquick purification columns (Qiagen, Inc, Hilden, Germany) following the manufacturer's protocol.

All PCR reactions were carried out using Ready Master Mix (Advanced Biotechnologies, Epsom, Surrey, UK). We added 4.5% of dimethyl sulfoxide (DMSO) to the PCR reactions of nrITS to improve PCR efficiency. Amplification of *rbc*La was done using the primer combination: 1F: 724R (Olmstead et al. 1992, Fay et al. 1998). For *mat*K, the following primer combination was used 390F: 1326R (Cuénoud et al. 2002). Intergenic spacers *trnH-psbA* and *psaA-ycf3* were amplified using the primers *trnH*: *psbA* (Sang et al. 1997) and PG1F: PG2R (Huang and Shi 2002), respectively. Intergenic spacer *psaA-ycf3* was included in this study for the purpose of reconstructing phylogeny of Combretaceae. The nrITS region was amplified into two overlapping fragments using the following two pairs of internal primer combinations: 101F: 2R and 3F: 102R (White et al. 1990, Sun et al. 1994).

The following programme was used to amplify *rbc*La and *trnH-psbA*: pre-melt at 94 °C for 60 s, denaturation at 94 °C for 60 s, annealing at 48 °C for 60 s, extension at 72 °C for 60 s (for 28 cycles), followed by a final extension at 72 °C for 7 min; for *mat*K, the protocol consisted of pre-melt at 94 °C for 3 min, denaturation at 94 °C for 60 s, annealing at 52 °C for 60 s, extension at 72 °C for 2 min (for 30 cycles), final extension at 72 °C for 7 min. For nrITS and spacer *psaA-ycf3* the protocol consisted of pre-melt at 94 °C for 1 min, denaturation at 94 °C for 60 s, annealing at 48 °C for 60 s, extension at 72°C for 3 min (for 26 cycles), final extension at 72 °C for 7 min.

Purification of the amplified products was done using QIAquick columns (QIAgen, Germany) following the manufacturer's manual. The purified products were then cycle-sequenced with the same primers used for amplification using BigDye™ v3.1 Terminator Mix (Applied Biosystems, Inc, ABI, Warrington, Cheshire, UK). Cleaning of cycle-sequenced products was done using EtOH-NaCl, followed by sequencing on an ABI 3130xl genetic analyser.

Sequences were assembled, trimmed and edited using Sequencher v4.6 (Gene Codes Corp, Ann Arbor, Michigan, USA). Alignment was done using Multiple Sequence Comparison by Log-Expectation v3.8.31 (Edgar 2004) followed by subsequent manual adjustments to refine alignments.

Data analysis

Performance of DNA markers in species delimitation was tested at three taxonomic levels (family, subgenus, and section). At family level, we evaluated four single markers:

*rbc*La, *mat*K, *trnH-psbA*, and nrITS. We also tested the core barcodes, i.e. *rbc*La + *mat*K (CBOL Plant Working Group 2009) and the following combinations: core + nrITS, core + *trnH-psbA,* and core + *trnH-psbA* +nrITS. Four criteria were used to assess their barcoding potential: presence of 'barcode gap' (Meyer and Paulay 2005), discriminatory power, species monophyly, and PCR success rate.

Barcode gap was evaluated in two ways: (1) we compared genetic variation within species (intraspecific genetic distance) versus between species (interspecific genetic distance). This comparison was based on the mean, median, and range of both distances; (2) in addition, we also used Meier et al.'s (2008) approach of evaluating the gap comparing the smallest interspecific distance with the greatest intraspecific distance. The genetic distances were calculated using the Kimura 2-parameter (K2P) model. We also assessed the index of sequence divergence, K, for each region, measured as the mean number of substitutions between any two sequences.

The discriminatory power of DNA regions was conducted using three distance-based methods including Near Neighbour, Best Close Match (Meier et al. 2006) and the BOLD identification criteria. A good barcode should exhibit the highest rate of correct species identification by assigning the highest proportion of DNA sequences to the corresponding species names. All the sequences were labelled according to species names prior to testing. For the Best Close Match test, we determined, for each dataset (family, subgenera and sections), the optimised genetic distance suitable as threshold for species delimitation. Optimised thresholds were determined using the function "localMinima" implemented in the R package Spider 1.1-1 (Brown et al. 2012).

We also used the PCR success rate to evaluate the DNA regions. This evaluation was conducted based on the percentage of successful amplification.

The test for species monophyly was conducted on a Neighbour-Joining (NJ) tree. We considered that a species is monophyletic when all individuals of the same species cluster on the NJ phylogram that we reconstructed. As such, the best barcode should provide the highest proportion of monophyletic species. We then evaluated for each DNA region and concatenated regions, the proportion of monophyletic (i.e. correct identification) and non-monophyletic species (incorrect identification). All our analyses were conducted in the R package Spider 1.1-1 (Brown et al. 2012).

Finally, we evaluated the barcoding potential in discriminating phylogenetically deliminated clades in the phylogeny of the genus that was reconstructed based on the combination of five DNA regions (*rbc*L, *mat*K, *trnH-psbA*, *psaA-ycf3* and nrITS). The phylogeny was reconstructed based on maximum parsimony (MP) implemented in PAUP* v4.0b10 (Swofford 2002). Tree searches were conducted using heuristic searches with 1000 random sequence additions, retaining 10 trees per replicate, with tree-bisection-reconnection (TBR) branch swapping and MulTrees in effect (saving multiple equally parsimonious trees). Based on Maurin et al. (2010) we used *Strephonema mannii* Hook. f. and *S. pseudocola* A. Chev. as outgroups. Node support was assessed using bootstrap (BP) values: BP > 70% for strong support (Hillis and Bull 1993, Wilcox et al. 2002).

At subgeneric and sectional levels, we only tested the performance of core barcodes and best gene combination identified using the three criteria mentioned above (barcode gap, discriminatory power and species monophyly).

Results

The overall characteristics of single and combined DNA regions are presented in Table 1. In general, our results indicate that the ranges and mean intraspecific distances were both lower than those of interspecific distances. Among single regions, *rbc*La showed the lowest interspecific distance (mean = 0.009) with nrITS exhibiting the highest genetic variation between species (mean = 0.110). For all marker combinations, the mean interspecific distances varied between 0.011 and 0.014. Assessing the index of sequence divergence K for each region, we found that nrITS showed the highest divergence (K = 21) whereas *trnH-psbA* exhibited the lowest divergence (K = 3). For the combined regions, K varied between 10 and 13, with an average of 10 substitutions between sequence-pairs (Table 1).

The distribution ranges of inter- versus intraspecific distances for all regions, showed a clear overlap between both distances (Figures 3a,b and 4), indicating the existence of a barcode gap. Comparing the smaller inter- versus the largest intraspecific distances for each region, our results further support the existence of barcode gap in all regions, but the proportion of sequences with barcode gap varied significantly with the regions tested (Table 2). Notably, the combination of all four regions exhibited the highest proportion of sequences with barcode gap (84%) followed by nrITS (73%), then core + nrITS (64%), and core + *trnH-psbA* (57%), with the lowest proportion found in *rbc*La (13%) (Table 2).

Optimised genetic distances used as threshold for species delimitation in Best Close Match method are shown in Table 1. Apart from *rbc*La (threshold = 0.04%), core + *trnH-psbA* (threshold = 0.5%) and core + nrITS (threshold = 0.7%), the thresholds for the remaining single and gene combinations were greater than 1%.

Table 1. Statistics of all gene regions for the southern African Combretaceae included in the study.

DNA regions	No. of seq	Seq length	K	Range inter	Mean inter (±SD)	Range intra	Mean intra (±SD)	Threshold (%)
*rbc*La	152	552	4	0-0.09	0.009±0.012	0-0.08	0.002±0.009	0.04
*mat*K	133	771	6	0-0.07	0.014±0.011	0-0.02	0.002±0.004	1.10
trnH-psbA	116	1034	3	0-0.15	0.047±0.035	0-0.03	0.003±0.007	1.80
nrITS	91	821	21	0-0.21	0.110±0.045	0-0.05	0.004±0.010	1.70
*rbc*La+*mat*K	129	1323	10	0-0.78	0.012±0.009	0-0.05	0.002±0.006	1.31
*rbc*La+*mat*K+*trnH-psbA*	87	2358	11	0-0.04	0.012±0.007	0-0.02	0.002±0.004	0.5
*rbc*La+*mat*K+nrITS	74	2144	9	0-0.04	0.011±0.006	0-0.02	0.002±0.004	0.70
*rbc*La+*mat*K+nrITS+*trnH-psbA*	70	3178	13	0-0.04	0.014±0.007	0-0.02	0.002±0.004	1.17

Table 2. Percentage barcode gap in all sequences for each region using the Meier et al. (2008) approach.

DNA region	Number of sequences without gap	Proportion of sequences with gap (%)
*rbc*La	132	13
*mat*K	86	35
trnH-psbA	54	53
nrITS	25	73
*rbc*La+*mat*K	82	36
*rbc*La+*mat*K+*trnH-psbA*	37	57
*rbc*La+*mat*K+nrITS	27	64
*rbc*La+*mat*K+nrITS+*trnH-psbA*	11	84

Table 3. Identification efficacy of DNA barcodes using distance based methods. F = False and T = True.

DNA region	Near Neighbour		BOLD (1%)				Best Close Match			
	F	T	Ambiguous	Correct	Incorrect	No ID	Ambiguous	Correct	Incorrect	No ID
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
*rbc*La	59	41	61	18	14	7	61	18	14	7
*mat*K	46	54	81	11	7	1	47	38	14	1
trnH-psbA	72	28	65	22	10	3	18	60	18	4
nrITS	35	65	29	47	10	14	10	63	19	8
*rbc*La+*mat*K	39	61	86	10	2	2	35	51	12	2
*rbc*La+*mat*K+*trnH-psbA*	38	62	79	16	2	3	6	80	8	6
*rbc*La+*mat*K+nrITS	43	57	62	30	7	1	3	70	19	8
*rbc*La+*mat*K+nrITS+ *trnH-psbA*	36	64	52	41	3	4	0	87	9	4

Our results for the discriminatory power analysis varied with the methods applied (Table 3) at family level. Based on the Near Neighbour method, nrITS provided the highest discriminatory power (65%) followed by *rbc*La + *mat*K + *trnH-psbA* + nrITS (64%), *rbc*La + *mat*K + *trnH-psbA* (62%), and *rbc*La + *mat*K (61%). The lowest discriminatory power was found for *trnH-psbA* (28%).

BOLD species delimitation criteria of 1% threshold provided the lowest rate of correct identification among all three methods used. However, we found that nrITS remains the most efficient region with 47% discriminatory power. The second most successful combination of regions were core + *trnH-psbA* + nrITS (41%) followed by core + nrITS (30%) and *trnH-psbA* (22%); the core barcodes were identified as the least performing regions (10%) with the highest proportion of ambiguity (86%).

In contrast to the two previous methods, the Best Close Match provided the highest rate of species discrimination for the combined dataset (core + *trnH-psbA* + nrITS) yielding the best discriminatory power (87%) with no ambiguity. This was followed by core+ *trnH-psbA* (80%), core + nrITS (70%) and nrITS (63%), with the poorest performance for *rbc*La (18%) at family level.

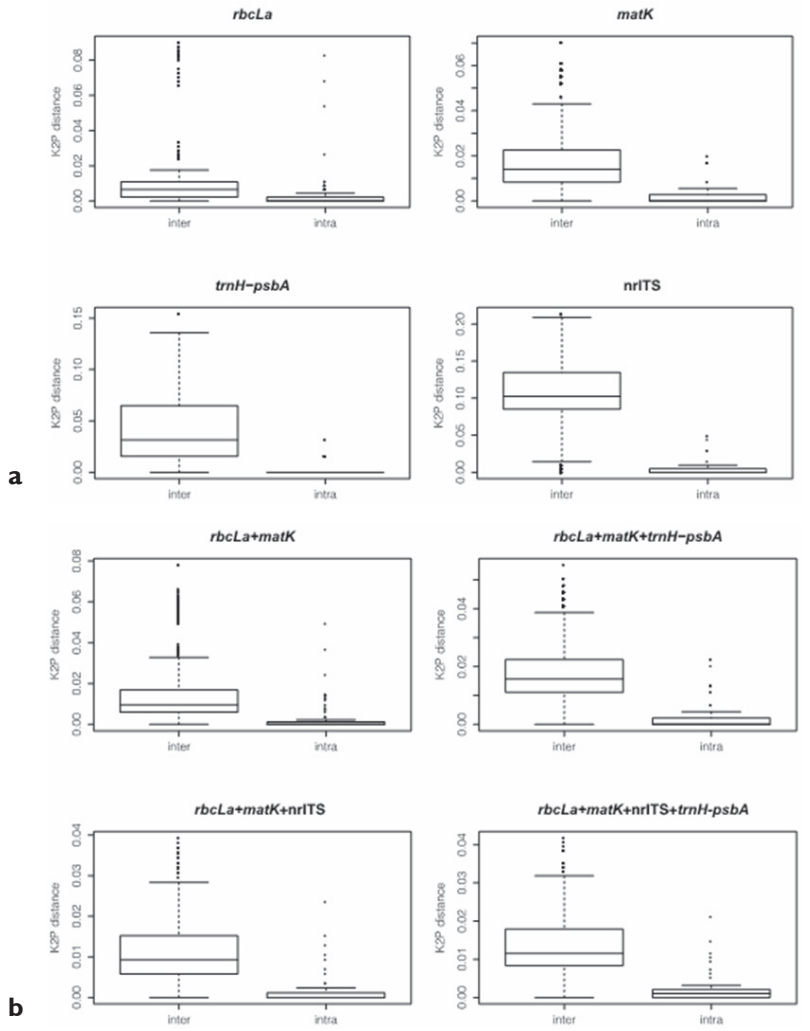


Figure 3. Comparisons of the distribution range of inter- versus intraspecific distances using boxplot **a** indicates comparison of single barcode gene regions **b** indicates the results of gene combinations.

The last criterion used to evaluate the potential of DNA region was PCR efficiency. We found that *rbc*La (87%) followed by *trnH-psbA* (85%) and *mat*K (68%) were easy to amplify, with nrITS being the most difficult (47%; Figure 5).

We complemented previous analyses using species monophyly criteria after verifying the monophyly of Combretaceae. Among all regions, core + *trnH-psbA* isolated the highest proportion of monophyletic species (83%), followed by *trnH-psbA* (78%), nrITS (76%), and combination of all four regions (65%). Again, *rbc*La provided the lowest performance in identifying species as monophyletic (37%; Figure 6).

In summary, all regions provided evidence for barcode gaps (Figure 3a, b and 4), but the strength of evidence varied with approaches used. Furthermore, the Best Close

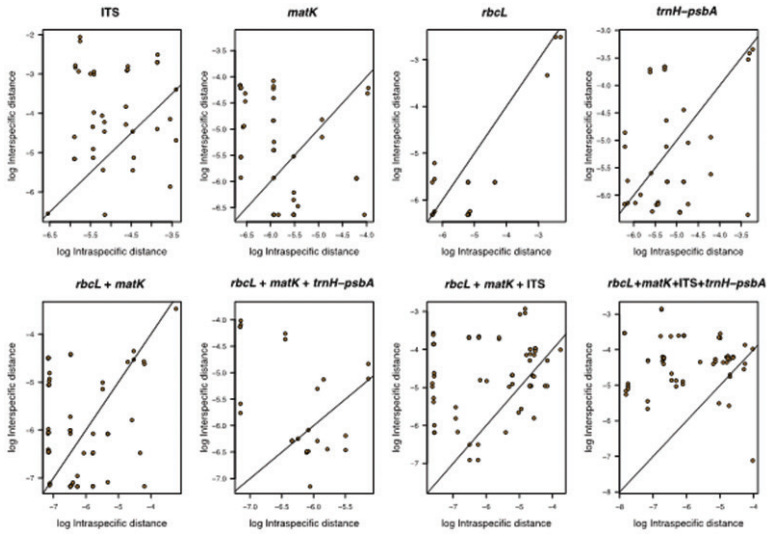


Figure 4. Relationships between inter- and intraspecific distances indicating barcoding gap for all regions tested.

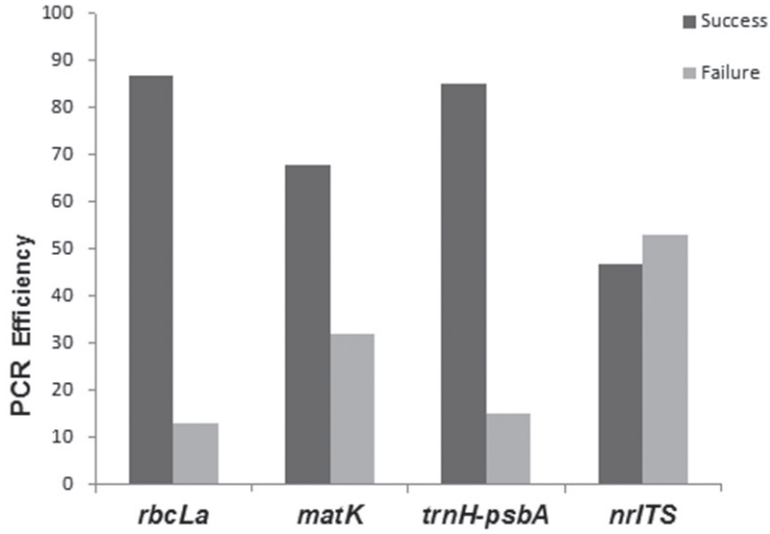


Figure 5. PCR efficiency for the four candidate barcodes (*rbc*La, *mat*K, *trnH-psbA*, nrITS).

Match method provided the highest identification accuracy among the three distance-based methods used irrespective of genes or combinations tested. Under this method, the two best potential barcodes for southern African Combretaceae were first, core + *trnH-psbA* and second, core + *trnH-psbA* + nrITS. However, based on species monophyly criteria, the single region *trnH-psbA* and the combination core + *trnH-psbA*

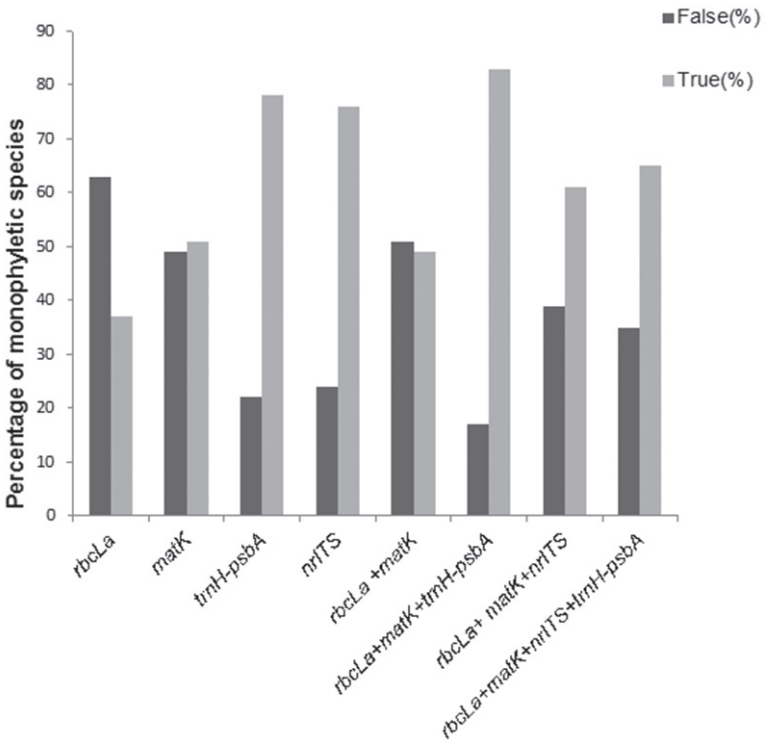


Figure 6. Gene performance based on monophyly criteria. False = proportion of non-monophyletic species; True = proportion of monophyletic species.

showed high barcode potential, with *trnH-psbA* being the second best easy-to-amplify region after *rbc*La.

We further evaluated the potential of each region as candidate barcode using a phylogeny of southern African Combretaceae (Appendix 2). Our results are congruent to the corresponding subset in the most recent and largest phylogeny assembled for the family (Appendix 3). Our evaluation for the discriminatory power at subgeneric level using the thresholds determined for the family (1.31% for the core and 0.5% for the core + *trnH-psbA)* revealed that the core barcodes alone were able to correctly identify 78% of species within the subgenus *Cacoucia*. However, the core barcodes could discriminate only 50% of species within the subgenus *Combretum*. In particular, the discriminatory power of the core barcodes within both subgenera increased markedly to 100% when we added the *trnH-psbA* region (Table 4). This trend was consistent even when we applied the thresholds that have been optimised for the subgenera.

At sectional level, we observed similar trends – the addition of *trnH-psbA* increased the performances of the core barcodes drastically except for *Macrostigmatea* (Table 5): *Angustimarginata* (core: 11%; core + *trnH-psbA*: 86%); *Ciliatipetala* (core: 55%; core + *trnH-psbA*: 73%); *Conniventia* (core: 38%; core + *trnH-psbA*: 88%); *Hypocrateropsis* (core: 63%; core + *trnH-psbA*: 80%). However, *Macrostigmatea* (core 34%, core +

Table 4. Comparisons of efficacy of core barcodes and best barcode within subgenera *Combretum* and *Cacoucia*.

Subgenus	DNA region	No. of seq	Mean Inter (±SD)	Threshold (%)	Best Close Match			
					Ambiguous (%)	Correct (%)	Incorrect (%)	No ID (%)
Cacoucia	*rbc*La+*mat*K	23	0.004±0.002	1.31	13	78	9	0
	*rbc*La+*mat*K+*trnH-psbA*	16	0.006±0.002	0.5	0	100	0	0
Combretum	*rbc*La+*mat*K	84	0.009±0.009	1.31	36	50	12	2
	*rbc*La+*mat*K+*trnH-psbA*	16	0.006±0.002	0.5	0	100	0	0

Table 5. Comparisons of core barcodes and the best barcode within five sections of the subgenera *Combretum* and *Cacoucia*.

Sections	DNA regions	No. of seq	Mean inter (±SD)	Threshold (%)	Best Close Match			
					Ambiguous (%)	Correct (%)	Incorrect (%)	No ID (%)
Angustimarginata	*rbc*La+*mat*K	19	0.007±0.014	2.6	58	11	26	5
	*rbc*La+*mat*K+*trnH-psbA*	15	0.006±0.006	*0.7*	0	86	7	7
Ciliatipetala	*rbc*La+*mat*K	20	0.004±0.002	0.3	45	55	0	0
	*rbc*La+*mat*K+*trnH-psbA*	15	0.006±0.003	*0.5*	0	73	27	0
Conniventia	*rbc*La+*mat*K	8	0.005±0.004	0.8	37	38	12	13
	*rbc*La+*mat*K+*trnH-psbA*	8	0.010±0.006	*2.4*	*0*	88	12	0
Hypocrateropsis	*rbc*La+*mat*K	8	0.012±0.005	1.31	25	63	12	0
	*rbc*La+*mat*K+*trnH-psbA*	5	0.020±0.004	*0.8*	*0*	80	20	0
Macrostigmatea	*rbc*La+*mat*K	15	0.002±0.001	0.1	53	34	13	0
	*rbc*La+*mat*K+*trnH-psbA*	9	0.003±0.002	*0.2*	*0*	44	56	0

(Only sections with at least three different species are included).

trnH-psbA 44%) showed the least performance, even with the addition of *trnH-psbA* to the core barcode, with just 10% increment being observed. This trend is not sensitive to the thresholds applied for the family or the sections.

Finally, we compared the mean number of substitutions between any two species within each section. We found that the mean number of substitutions between representatives of *Macrostigmatea* is lowest (mean = 4) whereas it ranges between 5 and 19 substitutions in other sections of subgenus *Combretum*.

Discussion

We evaluated genetic variation for both single and various combinations of *rbc*La, *mat*K, *trnH-psbA* and nrITS. Comparing ranges of intra- versus interspecific distances, our results indicate that all markers show a barcode gap (Meyer and Paulay 2005); and

this is also true for the stringent Meier et al.'s (2008) approach, although the proportion of sequences with gap varies greatly with the marker used.

The discriminatory power of the DNA regions in species identification also varies with the distance-based methods applied. From the methods tested, Near Neighbour and Best Close Match yielded high performance, with the latter giving the best results for the possible three and four different gene combinations. The core barcodes were not recognised among the three best options, and its discriminatory power has been questioned in a number of studies (Hollingsworth et al. 2009, Pettengill and Neel 2010, Roy et al. 2010, Wang et al. 2010, Clement and Donoghue 2012). Based on all three distance methods, nrITS emerges as the most suitable single region (as indicated under both Near Neighbour and BOLD; see also Kress et al. 2005, Kress and Erickson 2007, Chen et al. 2010, Gao et al. 2010, Ren et al. 2010, China Plant BOL Group et al. 2011, Muellner et al. 2011, Pang et al. 2011, Wang et al. 2011, Liu et al. 2012, Yang et al. 2012). Among combined regions, core + nrITS + *trnH-psbA* (under Best Close Match) emerges as most suitable for barcoding Combretaceae.

However, our study indicates some important drawbacks that discount the inclusion of nrITS as a good barcode. For example, based on amplification success criteria, nrITS was the most difficult of all regions tested with *rbc*La and *trnH-psbA* being the easiest regions to amplify. The technical hurdles in PCR amplification and sequencing of nrITS may be linked to the presence of retro-transposons and other repetitive elements within plant nuclear genomes, resulting in paralogous gene copies (Gao et al. 2010, Hollingsworth 2011, Hollingsworth et al. 2011, Li et al. 2011). This is likely the case for nrITS in Combretaceae as we found evidence of multiple copies that may not be identical to each other (see CBOL Plant Working Group 2009, Hollingsworth 2011, Hollingsworth et al. 2011, Yang and Berry 2011). As such, the addition of *trnH-psbA* to the core barcodes (*rbc*La + *mat*K + *trnH-psbA*) emerge as the best gene combination useful for species discovery and delimitation in Combretaceae (see also Newmaster and Ragupathy 2009, Petit and Excoffier 2009, Ragupathy et al. 2009, Wang et al. 2009, Arca et al. 2012).

Previous studies have shown that core barcodes are very limited in discriminating taxa that are phylogenetically closely related, and suggested that the efficacy of DNA barcodes should be tested within a phylogenetic context (Clement and Donoghue 2012). We tested this using subgenera and sections of the family Combretaceae. Our evaluation of the discriminatory power of the core barcodes at subgeneric level revealed a striking difference in the performance between the two *Combretum* subgenera, *Combretum* and *Cacoucia*. The difference noted for the discriminatory power of the core barcodes between the two subgenera may reflect differences in their evolutionary history. Indeed, the latest dated phylogeny of Combretaceae indicated that members of the subgenus *Cacoucia* are represented with longer terminal branches than those in subgenus *Combretum* (Maurin 2009).

While we found poor performance at sectional level, for example, in *Angustimarginata*, *Macrostigmatea* and *Conniventia,* this result is not unexpected due to a very low genetic variation one could expect within clades (see Ennos et al. 2005, Clement and

Donoghue 2012). However, the addition of *trnH-psbA* to the core barcodes results in a drastic increase of identification rate at both subgenus and sectional levels, validating the utility of *trnH-psbA* to discriminate even closely related species, except for section *Macrostigmatea* (Newmaster and Ragupathy 2009, Petit and Excoffier 2009, Ragupathy et al. 2009, Wang et al. 2009, but see Zhang et al. 2012, Arca et al. 2012, Clement and Donoghue 2012).

The result for section *Macrostigmatea* reflects earlier tangle cited in previous studies regarding its composition (Stace 1980, Maurin et al. 2010, Jordaan et al. 2011). In our analysis, we included *Spathulipetala* within section *Macrostigmatea* based on suggestions from recent molecular evidence (Maurin et al. 2010). Morphological studies separate these two sections, *Spathulipetala* and *Macrostigmatea* (Stace 1980, Jordaan et al. 2011). Section *Spathulipetala* comprises two members, *Combretum zeyheri* Sond. and *C. mkuzense* J.D.Carr and Retief, which occur in the same geographical location and show close morphological similarity in their fruits (Jordaan et al. 2011). The inclusion of *C. mkuzense,* in this section has been controversial, with some authors (Exell 1978, Stace 1980) advocating for a tentative placement pending further investigation. However, recent molecular study shows close relationship between these two species (*Combretum zeyheri* and *C. mkuzense*) (Maurin et al. 2010), which gives support to earlier morphological treatment. On the other hand, the taxonomy of section *Macrostigmatea* appears to pose fewer challenges as compared to *Spathulipetala*. A recent molecular study (Maurin et al. 2010) suggests lumping of these two sections, *Spathulipetala* and *Macrostigmatea* as members appear embedded in one clade with a high bootstrap support of 100%. Earlier, Exell (1978) had reported that the sections are closely related, as they share similarities in scale size, scale fragmentation into fruit walls and fruit size.

Based on our results, the unclear taxonomy reported for section *Macrostigmatea*, is reflected, indicating a need for further molecular analyses involving more taxa and gene sequences to correctly determine members of this section. Our results also support the proposal of Exell (1978) to lump these two sections. The low performance of the core + *trnH-psbA* in fully discriminating the different species within this section is a strong indicator of the close phylogenetic similarity of the species. Our results indicate not only the utility of DNA barcoding data for discriminating species, but also to detect species that require further molecular analyses.

Conclusions

Our analysis indicates that the poor performance of the core barcodes at family level could not be generalised to lower levels: the core barcodes perform poorly in some sections but shows strong discriminatory power in others. Such findings may indicate that the success of DNA barcodes in discriminating closely related species at least in plants may correlate with the evolutionary distinctiveness of the group tested and, as recently indicated, (see Clement and Donoghue 2012) it may also possibly reflects different bio-

geographic history between clades of the taxonomic group Combretaceae. Overall, we propose the core + *trnH-psbA* as the best barcode for the family Combretaceae.

Acknowledgements

We thank the Government of Canada through Genome Canada and the Ontario Genomics Institute (2008-OGI-ICI-03), The International Development Research Centre (IDRC), Canada and the University of Johannesburg for financial support and various local and international authorities granting us plant collections permits. We thank three anonymous reviewers for providing valuable comments on an earlier draft of the manuscript.

References

Arca M, Hinsinger DD, Cruaud C, Tillier Bousquet J, Frascaria-Lacoste N (2012) Deciduous Trees and the Application of Universal DNA Barcodes: A Case Study on the Circumpolar Fraxinus. PLoS ONE 7: e34089. doi: 10.1371/journal.pone.0034089

Brown R (1810) Prodromius florae novae liollandiae et insulae van Diemen J. Johnson and Company London.

Brown SDJ, Collins RA, Boyer S, Lefort MC, Malumbres-Olarte J, Vink CJ (2012) Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources 12: 562–565. doi: 10.1111/j.1755-0998.2011.03108.x

Burgess KS, Fazekas AJ, Kesanakurti PR, Graham SW, Husband BC, Newmaster SG, Percy DM, Hajibabaei M, Barrett SCH (2011) Discriminating plant species in a local temperate flora using the *rbc*L plus *mat*K DNA barcode. Methods of Ecology and Evolution 2: 333–340. doi: 10.1111/j.2041-210X.2011.00092.x

CBOL Plant Working Group (2009) A DNA Barcode for land plants. Proceedings of the National Academy of Sciences of the USA 106: 12794–12797. doi: 10.1073/pnas.0905845106

Chen SL, Yao H, Han JP, Liu C, Song JY, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leo C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5(1): e8613. doi: 10.1371/journal.pone.0008613

China Plant BOL Group, Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL, Yang JB, Fu CX, Zeng CX, Yang HF, Zhu YJ, Sun YS, Chen SY, Zhao L, Wang K, Yang T, Duan GW (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proceedings of the National Academy of Sciences of the USA 108: 19641–19646. doi: 10.1371/journal.pone.0008613

Clement WL, Donoghue MJ (2012) Barcoding success as a function of phylogenetic relatedness in *Viburnum*, a clade of woody angiosperms. BMC Evolutionary Biology 12: 73.

Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. American Journal of Botany 89: 132–144. doi: 10.3732/ajb.89.1.132

Dahlgren R, Thorne RF (1984) The Order Myrtales: Circumscription, variation, and relationships. Annals of the Missouri Botanical Garden 71: 633–699. doi: 10.2307/2399158

Doyle JJ, Doyle JL (1987) A rapid isolation procedure for small amounts of leaf tissue. Phytochemical Bulletin 19: 11–15.

Edgar R (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797. doi: 10.1093/nar/gkh340

El Ghazlai GEB, Tsuji S, El Ghazaly G, Nilsson S (1998) Combretaceae R. Brown. World Pollen Spore Flora 21: 1–40.

Ennos RA, French GC, Hollingsworth PM (2005) Conserving taxonomic complexity. Trends in Ecology & Evolution 20: 164–168. doi: 10.1016/j.tree.2005.01.012

Exell AW (1978) *Combretum*. In: Launert E (Ed) Flora Zambesiaca: London. Flora Zambesiaca Managing Committee 4: 100–183.

Exell AW, Stace CA (1966) Revision of the Combretaceae. Boletim da Sociedade Broteriana 40: 5–26.

Fay MF, Bayer C, Alverson WS, De Bruijn A, Chase MW (1998) Plastid *rbcL* sequence data indicate a close affinity between *Diegodendron* and *Bixa*. Taxon 47: 43–50. doi: 10.2307/1224017

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3: e2802. doi: 10.1371/journal.pone.0002802

Fazekas AJ, Kesanakurti R, Burgess KS, Percy DM, Graham SW, Barrett SCH (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular Ecology Resources 9: 130–139. doi: 10.1111/j.1755-0998.2009.02652.x

Franzini PZN, Dippenaar-Schoeman AS, Yessoufou K, Van der Bank FH (2013) Combined analyses of genetic and morphological data indicate more than one species of *Cyrtophora* (Araneae: Araneidae) in South Africa. International Journal of Modern Biology Research 1: 21–34.

Gao T, Yao H, Song JY, Liu C, Zhu YJ, Ma X, Pang X, Xu H, Chen S (2010) Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. Journal of Ethnopharmacology 130: 116–121. doi: 10.1016/j.jep.2010.04.026

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proceedings of the National Academy of Sciences of the USA 103: 9. doi: 10.1073/pnas.0510466103

Hebert PD, Stoeckle MY, Zemlak TS, Franci CM (2004) Identification of Birds through DNA Barcodes. PLoS Biology 2: e312. doi: 10.1371/journal.pbio.0020312

Hebert PDN, DeWaard JR, Landry JF (2010) DNA barcodes for 1/1000 of the animal kingdom. Biology Letters 6: 359–362. doi: 10.1098/rsbl.2009.0848

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42: 182–192.

Hollingsworth ML, Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources 9: 439–457. doi: 10.1111/j.1755-0998.2008.02439.x

Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proceedings of the National Academy of Sciences of the USA 108: 19451–19452. doi: 10.1073/pnas.1116812108

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS ONE 6: e19254. doi: 10.1371/journal.pone.0019254

Huang YL, Shi SH (2002) Phylogenetics of Lythraceae *sensu lato:* a preliminary analysis based on chloroplast *rbcL* gene, *psaA-ycf3* spacer and nuclear rDNA internal transcribed spacer (ITS) sequences. International Journal of Plant Sciences 163: 215–225. doi: 10.1086/338392

Jordaan M, Van Wyk AE, Maurin O (2011) Generic status of *Quisqualis* (Combretaceae), with notes on the taxonomy and distribution of *Q. parviflora*. Bothalia 41: 161–169.

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. PLoS ONE 2: e508. doi: 10.1371/journal.pone.0000508

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the USA 102: 8369–8374. doi: 10.1073/pnas.0503123102

Liu C, Shi L, Xu X, Li H, Xing H, Liang D, Jiang K, Pang X, Song J, Chen S (2012) DNA Barcode Goes Two-Dimensions: DNA QR Code Web Server. PLoS ONE 7: e35146. doi: 10.1371/journal.pone.0035146

Maurin O (2009) A phylogenetic study of the family Combretaceae with emphasis on the genus *Combretum* in Africa. PhD Thesis. University of Johannesburg.

Maurin O, Chase MW, Jordaan M, Van der Bank M (2010) Phylogenetic relationships of Combretaceae inferred from nuclear and plastid DNA sequence data: implications for generic classification. Botanical Journal of the Linnean Society 162: 453–476. doi: 10.1111/j.1095-8339.2010.01027.x

Maurin O, Van Wyk AE, Jordaan M, Van der Bank M (2011) A new species of *Combretum* section *Ciliatipetala* (Combretaceae) from southern Africa, with a key to the regional members of the section. South African Journal of Botany 77: 105–111. doi: 10.1016/j.sajb.2010.06.003

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Meier R, Zhang G, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Systematic Biology 57: 809–813. doi: 10.1080/10635150802406343

Meyer CP, Paulay G (2005) DNA barcoding: Error rates based on comprehensive sampling. PLoS Biology 3: e422. doi: 10.1371/journal.pbio.0030422

Muellner AN, Schaefer H, Lahaye R (2011) Evaluation of candidate DNA barcoding loci for economically important timber species of the mahogany family (Meliaceae). Molecular Ecology Resources 11: 450–60. doi: 10.1111/j.1755-0998.2011.02984.x

Newmaster SG, Ragupathy S (2009) Testing plant barcoding in a sister species complex of pantropical Acacia (Mimosoideae, Fabaceae). Molecular Ecology Resources 9: 172–180. doi: 10.1111/j.1755-0998.2009.02642.x

Olmstead RG, Michaels HJ, Scott KM, Palmer JD (1992) Monophyly of the Asteridae and identification of their major lineages inferred from DNA sequences of *rbcL*. Annals of the Missouri Botanical Garden 2: 49–265.

Pang X, Song J, Zhu Y, Xu H, Huang LF, Chen SL (2011) Applying plant DNA barcodes for Rosaceae species identification. Cladistics 27: 165–170. doi: 10.1111/j.1096-0031.2010.00328.x

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. Trends in Ecology & Evololution 24: 386–393. doi: 10.1016/j.tree.2009.02.011

Pettengill JB, Neel MC (2010) An evaluation of candidate plant DNA barcodes and assignment methods in diagnosing 29 species in the genus *Agalinis* (Orobanchaceae). American Journal of Botany 97: 1381–1406. doi: 10.3732/ajb.0900176

Ragupathy S, Newmaster SG, Murugesan M, Balasubramaniam V (2009) DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the Western Ghats in southern India. Molecular Ecology Resources 9: 164–171. doi: 10.1111/j.1755-0998.2009.02641.x

Ren BQ, Xiang XG, Chen ZD (2010) Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers. Molecular Ecolology Resources 10*:* 594–605. doi: 10.1111/j.1755-0998.2009.02815.x

Roy S, Tyagi A, Shulka V, Kumar A, Singh UM, Chaudhary LB, Datt B, Bag SK, Singh PK, Nair NK, Husain T, Tuli R (2010) Universal plant DNA barcode loci may not work in complex groups: a case study with Indian *Berberis* species. PLoS ONE 5: e13674. doi: 10.1371/journal.pone.0013674

Sang T, Crawford DJ, Stuessy TF (1997) Chloroplast DNA phylogeny, reticulate evolution and biogeography of *Paeonia* (Paeoniaceae). American Journal of Botany 84: 1120–1136. doi: 10.2307/2446155

Stace CA (1980) The significance of the leaf epidermis in the taxanomy of the Combretaceae: conclusions. Botanical Journal of the Linnean Society 81: 327–339. doi: 10.1111/j.1095-8339.1980.tb01682.x

Stace CA (2007) Combretaceae. In: Kubitzki K (Ed) The families and genera of vascular plants. Springer, Berlin, 9: 67–82.

Stace CA (2010) Combretaceae. *Terminalia* and *Buchenavia* with Abul-Ridha Alwan. New York Botanical Garden Press (Flora Neotropica Monograph) 107.

Sun Y, Skinner DZ, Liang GH, Hulbert SH (1994) Phylogenetic analysis of sorghum and related taxa using internal transcribed space of nuclear ribosomal DNA. Theoretical and Applied Genetics 89: 26–32. doi: 10.1007/BF00226978

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). 4b10 ed. Sinauer Associates, Sunderland, Massachusetts.

Sytsma KJ, Litt AL, Zjhra ML, Pires JC, Nepokroeff M, Conti E, Walker J, Wilson PG (2004) Clades, clocks and continents: historical and biogeographical analysis of Myrtaceae,

Vochysiaceae and relatives in the Southern Hemisphere. International Journal of Plant Science 165: 85–105. doi: 10.1086/421066

Tan F, Shi S, Zhong Y, Gong X, Wang Y (2002) Phylogenetic relationships of *Combretoideae* (Combretaceae) inferred from plastid, nuclear gene and spacer sequences. Journal of Plant Resources 115: 475–481. doi: 10.1007/s10265-002-0059-1

Van der Bank HF, Greenfield R, Daru BH, Yessoufou K (2012) DNA barcoding reveals micro-evolutionary changes and river system level phylogeographic resolution of African Silver catfish, *Schilbe intermedius* (Actinopterygii: Siluriformes: Schilbeidae) from seven populations across different African river systems. Acta Ichthyologica et Piscatoria 42: 307–320. doi: 10.3750/AIP2012.42.4.04

Wang Q, Yu QS, Liu JQ (2011) Are nuclear loci ideal for barcoding plants? A case study of genetic delimitation of two sister species using multiple loci and multiple intraspecific individuals. Journal of Systematics and Evolution 49: 182–188. doi: 10.1111/j.1759-6831.2011.00135.x

Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J (2010) DNA barcoding of the Lemnaceae, a family of aquatic monocots. BMC Plant Biology 10: 205. doi: 10.1186/1471-2229-10-205

Wang Y, Tao X, Liu H, Chen X, Qiu Y (2009) A two-locus chloroplast (cp) DNA barcode for indetification of different species in *Eucalyptus*. Acta Horticulturae Sinica 36: 1651–1658.

White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (Eds) PCR Protocols: a guide to methods and applications. Academic Press, New York, USA, 315–322.

Wilcox TP, Zwickl DJ, Heath TA, Hillis DM (2002) Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Molecular Phylogenetics and Evolution 25: 361–371. doi: 10.1016/S1055-7903(02)00244-0

Yang JB, Wang YP, Möller M, Gao LM, Wu D (2012) Applying plant DNA barcodes to identify species of *Parnassia* (Parnassiaceae). Molecular Ecology Resources 2: 267–75. doi: 10.1111/j.1755-0998.2011.03095.x

Yang Y, Berry BE (2011) Phylogenetics of the Chamaesyce clade (*Euphorbia*, Euphorbiaceae): Reticulate evolution and long-distance dispersal in a prominent C4 lineage. American Journal of Botany 98: 1486–1503. doi: 10.3732/ajb.1000496

Zhang CY, Wang FY, Hai-Fei Y, Gang HH, Xue CM, Jun G (2012) Testing DNA barcoding in closely related groups of *Lysimachia* L. (Myrsinaceae). Molecular Ecology Resources 12: 98–108. doi: 10.1111/j.1755-0998.2011.03076.x

Appendix 1

Supplementary Table S1. (doi: 10.3897/zookeys.365.5728.app1) File format: Microsoft Excell file (xls).

Explanation note: Full names, voucher information, GenBank and BOLD accession numbers for taxa used in this study. A dash (—) indicates DNA regions not sampled and DNA sequences obtained from GenBank are underlined. Voucher specimens are deposited in the following herbaria: JRAU, University of Johannesburg (UJ), Johannesburg, South Africa; MO, Missouri Botanical Garden, St Louis, USA.

Citation: Gere J, Yessoufou K, Daru BH, Mankga LT, Maurin O, van der Bank M (2013) Incorporating *trnH-psbA* to core DNA barcodes improves significantly species discrimination within southern African Combretaceae. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 127–147. doi: 10.3897/zookeys.365.5728 Supplementary Table S1. doi: 10.3897/zookeys.365.5728.app1

Appendix 2

Supplementary Figure S1. (doi: 10.3897/zookeys.365.5728.app2) File format: Microsoft Word file (docx).

Explanation note: One of most parsimonious trees obtained from the combined plastid and nuclear data (*rbc*La, *mat*K, *trnH-psbA*, and nrITS) set. Clades highlighted indicate the sections that were identified from the MP tree obtained from barcoding gene regions. Bootstrap percentages above 50% are shown above the branches.

Citation: Gere J, Yessoufou K, Daru BH, Mankga LT, Maurin O, van der Bank M (2013) Incorporating *trnH-psbA* to core DNA barcodes improves significantly species discrimination within southern African Combretaceae. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 127–147. doi: 10.3897/zookeys.365.5728 Supplementary Figure S1. doi: 10.3897/zookeys.365.5728.app2

Appendix 3

Supplementary Figure S2. (doi: 10.3897/zookeys.365.5728.app3) File format: Microsoft Word file (docx).

Explanation note: One of most parsimonious trees with branch tips collapsed from the combined plastid and nuclear data (*rbc*L, *mat*K, *psaA-ycf3*, *trnH-psbA*, and nrITS) set. Clades highlighted indicate sections that were identified from the MP tree obtained from barcoding gene regions. Above the branches are Bayesian posterior probability (PP) values (> 0.5) and below are bootstrap percentages above 50%.

Citation: Gere J, Yessoufou K, Daru BH, Mankga LT, Maurin O, van der Bank M (2013) Incorporating *trnH-psbA* to core DNA barcodes improves significantly species discrimination within southern African Combretaceae. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 127–147. doi: 10.3897/zookeys.365.5728 Supplementary Figure S2. doi: 10.3897/zookeys.365.5728.app3

ZooKeys 365: 149–174 (2013)
doi: 10.3897/zookeys.365.6202
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

DNA barcoding and the differentiation between North American and West European *Phormia regina* (Diptera, Calliphoridae, Chrysomyinae)

Kurt Jordaens[1,2], Gontran Sonet[3], Yves Braet[4], Marc De Meyer[1], Thierry Backeljau[2,3], Frankie Goovaerts[2], Luc Bourguignon[4], Stijn Desmyter[4]

1 *Royal Museum for Central Africa, Department of Biology (JEMU), Leuvensesteenweg 13, 3080 Tervuren, Belgium* **2** *University of Antwerp, Evolutionary Ecology Group, Groenenborgerlaan 171, 2020 Antwerp, Belgium* **3** *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (JEMU), Vautierstraat 29, 1000 Brussels, Belgium* **4** *National Institute of Criminalistics and Criminology, Vilvoordsesteenweg 100, 1120 Brussels, Belgium*

Corresponding author: *Kurt Jordaens* (kurt.jordaens@africamuseum.be)

Academic editor: *Z. T. Nagy* | Received 3 September 2013 | Accepted 6 December 2013 | Published 30 December 2013

Citation: Jordaens K, Sonet G, Braet Y, De Meyer M,Backeljau T, Goovaerts F, Bourguignon L, Desmyter S (2013) DNA barcoding and the differentiation between North American and West European *Phormia regina* (Diptera, Calliphoridae, Chrysomyinae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 149–174. doi: 10.3897/zookeys.365.6202

Abstract

Phormia regina (the black fly) is a common Holarctic blow fly species which serves as a primary indicator taxon to estimate minimal post mortem intervals. It is also a major research model in physiological and neurological studies on insect feeding. Previous studies have shown a sequence divergence of up to 4.3% in the mitochondrial COI gene between W European and N American *P. regina* populations. Here, we DNA barcoded *P. regina* specimens from six N American and 17 W European populations and confirmed a mean sequence divergence of ca. 4% between the populations of the two continents, while sequence divergence within each continent was a ten-fold lower. Comparable mean mtDNA sequence divergences were observed for COII (3.7%) and cyt *b* (5.3%), but mean divergence was lower for 16S (0.4–0.6%). Intercontinental divergence at nuclear DNA was very low (≤ 0.1% for both 28S and ITS2), and we did not detect any morphological differentiation between N American and W European specimens. Therefore, we consider the strong differentiation at COI, COII and cyt *b* as intraspecific mtDNA sequence divergence that should be taken into account when using *P. regina* in forensic casework or experimental research.

Keywords
Black fly, COI, COII, cyt *b*, 16S, 28S, ITS2

Introduction

Forensic entomology uses the larval and pupal developmental stages of insects sampled on a corpse to estimate a minimum post-mortem interval (PMImin) of the corpse (Amendt et al. 2004, 2007). This requires i) detailed and accurate knowledge of the developmental rate of the species of forensic interest under different temperature conditions (Charabidze 2012), and ii) identification tools by which the different immature insect stadia can be identified (Catts 1992). Blowflies (family Calliphoridae) are among the most common insects found on dead bodies shortly after death. The species differ in their developmental times and have therefore a high potential for the accurate estimation of the PMImin. Unfortunately, several forensically important blow fly species can hardly be distinguished morphologically, especially in the larval and pupal stages (e.g. Catts 1992). To improve the success and reliability of identifications, a number of molecular techniques and tools have been explored to identify forensically important species (Wells and Stevens 2008, reviewed in Jordaens et al. in press).

Currently, the most popular molecular method for organismal identification is DNA barcoding, which was promoted by Hebert et al. (2003a, b) as a standardized molecular identification tool for all animals. It refers to establishing species-level identifications by sequencing a fragment of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene, the "DNA barcode", into a taxonomically unknown specimen and performing comparisons with a reference library of barcodes of well-identified species. COI barcodes (and other fragments of COI) indeed have been successfully applied in the identification of many calliphorid species (e.g. Wallman and Donnellan 2001, Wells and Sperling 2001, Nelson et al. 2007, Wells and Williams 2007, Harvey et al. 2008, Desmyter and Gosselin 2009, DeBry et al. 2013). Yet, COI fails to unambiguously discriminate among several calliphorid species pairs (e.g. Nelson et al. 2007, see also the Discussion) and the use of alternative identification tools (e.g. other genes) could be necessary to acquire correct identifications.

The monophyly of Calliphoridae has been questioned for many years (e.g. Griffiths 1982) and paraphyly or polyphyly was suggested by a morphology-based parsimony analysis (Rognes 1997). Nonmonophyly was also found in a molecular phylogenetic analysis of the Calyptratae with Calliphoridae being polyphyletic with respect to the Tachinidae and Rhinophoridae. Within this 'calliphorid-tachinid-rhinophorid' clade, the subfamily Chrysomyinae was para- or polyphyletic (Kutty et al. 2010). The Chrysomyinae comprises two tribes, Chrysomyini and Phormiini, of which the Phormiini has three genera (Table 1). *Phormia regina* (Meigen, 1826) (black fly) is the only species in the monotypic genus *Phormia*. It is a Holarctic blow fly species that is commonly found on human or animal faeces (Coffey 1966) and that is frequently found on corpses. It therefore serves as a primary species to estimate the PMImin (e.g. Byrd

Table 1. Taxonomy of the subfamily Chrysomyinae (family Calliphoridae) with indication of the number of DNA sequences (the number of haplotypes is given in parentheses) for each of the species used in this study (numbers combined from this study and GenBank) and for each of the gene fragments studied. No. ind. = number of individuals; No. hapl. = number of haplotypes; No. spp. = number of species.

	Genus/species	**COI**	**COII**	**16S**		**cyt *b***	**ITS2**	**28S**
				251 bp	350 bp			
Chrysomyini	*Chloroprocta* Wulp, 1896							
	Chl. idioidea (Robineau-Desvoidy, 1830)	2(2)					1(1)	1(1)
	Chrysomya Robineau-Desvoidy, 1830							
	C. albiceps (Wiedemann, 1819)	3(2)	1(1)				2(1)	2(1)
	C. bezziana Villeneuve, 1914	5(2)	1(1)			10(6)	2(1)	2(2)
	C. cabrerai Kurahashi & Salazar, 1977	1(1)						
	C. chani Kurahashi, 1979	1(1)					11(2)	
	C. chloropyga (Wiedemann, 1818)	1(1)						2(2)
	C. defixa (Walker, 1856)	1(1)						
	C. flavifrons (Aldrich, 1925)	3(2)	1(1)				4(2)	
	C. greenbergi Wells & Kurahashi, 1996	1(1)						
	C. incisularis (Macquart, 1851)	9(2)	2(2)				1(1)	
	C. latifrons (Malloch, 1927)	6(2)	1(1)				5(1)	
	C. megacephala (Fabricius, 1794)	79(11)	28(7)	66(31)	20(3)	2(2)	42(3)	4(2)
	C. nigripes Aubertin, 1932	9(7)	3(3)				7(1)	
	C. norrisi James, 1971	1(1)	1(1)					
	C. pacifica Kurahashi, 1991	1(1)					1(1)	
	C. pinguis (Walker, 1858)	7(4)	1(1)				14(2)	
	C. putoria (Wiedemann, 1830)	2(2)	1(1)			1(1)	2(1)	
	C. rufifacies (Macquart, 1843)	25(10)	45(9)	10(5)	1(1)		14(1)	2(2)
	C. saffranea (Bigot, 1877)	7(2)	1(1)				8(2)	
	C. semimetallica (Malloch, 1927)	11(5)	3(2)				10(2)	
	C. thanomthini Kurahashi & Tumrasvin, 1977	1(1)						
	C. varipes (Macquart, 1851)	7(6)	6(2)				1(1)	
	C. villeneuvi Patton, 1922						7(1)	
	Cochliomyia Townsend, 1915							
	Co. hominivorax (Coquerel, 1858)	78(73)	65(62)			2(1)	90(24)	2(1)
	Co. macellaria (Fabricius, 1775)	3(3)	1(1)				1(1)	4(1)
	Compsomyiops Townsend, 1918							
	Com. calipes (Bigot, 1877)	1(1)	1(1)					
	Com. fulvicrura (Robineau-Desvoidy, 1830)			1(1)	1(1)			1(1)
	Hemilucilia Brauer, 1895							
	H. segmentaria (Fabricius, 1805)	1(1)					1(1)	1(1)
	H. semidiaphana (Rondani, 1850)	1(1)					1(1)	1(1)
	Paralucilia Brauer & Bergenstamm, 1891							
	Pa. paraensis (Mello, 1969)	1(1)						
	Trypocalliphora Peus, 1960							
	T. braueri (Hendel, 1901)	1(1)						
Phormiini	*Phormia* Robineau-Desvoidy, 1830							
	P. regina (Meigen, 1826)	48(20)	30(9)	15(2)	15(2)	17(10)	36(2)	38(2)
	ProtoPhormia Townsend, 1908							

	Genus/species	COI	COII	16S		cyt *b*	ITS2	28S
				251 bp	350 bp			
	Pr. terraenovae (Robineau-Desvoidy, 1830)	17(7)	1(1)	2(2)	1(1)		1(1)	4(2)
	Protocalliphora Hough, 1899							
	Pro. azurea (Fallen, 1817)		2(2)		1(1)	1(1)		1(1)
	Pro. occidentalis Whitworth, 2003		1(1)					
	Pro. sialia Shannon & Dobroscky, 1924		1(1)	1(1)				
	Protocalliphora sp.			1(1)				
Total no. ind.		339	194	95	39	32	263	66
Total no. hapl.		180	108	42	9	20	55	21
Total no. spp.		36		20	6	6	5	24

and Allen 2001). Further, the species also plays an important role in secondary myasis in cattle (e.g. Francesconi and Lupi 2012) and is used in maggot therapy (Knipling and Rainwater 1937).

Phormia regina is a highly mobile species that is abundant in North American areas with cool spring and fall temperatures and in warmer areas, but then at higher altitudes (Hall 1948, Brundage et al. 2011). The developmental time of *P. regina* seems highly variable and could be influenced by a number of environmental variables (Kamal 1958, Greenberg 1991, Anderson 2000, Byrd and Allen 2001, Nabity et al. 2007, Núñez-Vázquez et al. 2013). Using amplified fragment length polymorphisms (AFLP), Picard and Wells (2009) studied the population genetic structure of N American *P. regina* and found that the N American populations were panmictic but with significant temporal genetic differences within populations, even over short periods of time. They therefore suggested that part of the variation in developmental times and growth curves that was observed in laboratory studies is not only due to local environmental (i.e. laboratory) conditions, but also to differences in the genetic composition of the laboratory stocks. This finding is important for forensic sciences since it shows that forensically relevant ecological data from one population (i.e. from a forensic case) cannot be extrapolated to other populations (i.e. to other forensic cases). Interestingly, Desmyter and Gosselin (2009) found a 4.2% sequence divergence at a 304 bp COI fragment between N American and W European specimens. Subsequently, Boehme et al. (2012) found a similar sequence divergence (range: 3.5%–4.31%) at the COI barcodes between N American and W European *P. regina* specimens.

Because high COI sequence divergences are often indicating species level differentiation (e.g. Hebert et al. 2003a, b), the strong COI differentiation between N American and W European *P. regina* specimens calls for a taxonomic re-assessment. We therefore studied DNA sequence variation in mitochondrial and nuclear DNA, and examined morphological differentiation between N American and W European populations of *P. regina* to i) provide additional DNA barcodes for *P. regina*, ii) examine molecular differentiation between N American and W European specimens in other genes, and iii) assess whether the COI differentiation is correlated with morphological differentiation. The taxonomy of *P. regina* is then re-evaluated in the light of these results.

Material and methods

Specimen collection and morphological examination

Sixty-one adult individuals of *P. regina* were captured at several localities in N America (Indiana, Texas, Virginia, Washington, Wyoming) and W Europe (Belgium, France, Germany) and stored in > 70% ethanol (Appendix 1 - Supplementary Table 1). The individuals were qualitatively scored for the color of 11 external characters (Table 2). In addition, we dissected the male copulatory organs of five W European and five N American individuals to study the general shape of the penis, cerci and surstyli (Figure 1).

DNA sequence analysis

DNA was extracted from on one or two legs. The remaining parts of the vouchers are kept at the NICC (National Institute of Criminalistics and Criminology – Brussels, Belgium) as pinned material. Genomic DNA was extracted using the NucleoSpin Tissue kit (Macherey-Nagel). A fragment of 721 bp from the 5'-end of the COI gene, including the standard barcode region (Hebert et al. 2003a,b), was amplified using primer pair TY-J-1460 and C1-N-2191 (Sperling et al. 1994, Wells and Sperling 2001). Five other DNA markers were sequenced for a more limited set of samples (Appendix 1 - Supplementary Table 1). Fragments of the mitochondrial 16S ribosomal RNA (16S), cytochrome *c* oxidase subunit II (COII), and cytochrome *b* (cyt *b*) genes, and of the nuclear ribosomal internal transcribed spacer 2 (ITS2) and fragment D1–D2 of the 28S ribosomal RNA (28S) were amplified using primer pairs 16Sf.dip/16Sr.dip (Kutty

Table 2. Color scoring of eleven external morphological characters of adult W European and N American *Phormia regina*.

Character	W Europe and N America
calypters	white
first spiraculum	white to yellow
thoracic dorsum	metallic green-bluish to dark green
scutellum	dark green
legs	black
abdomen	metallic green-bluish
facial ridge	red-brown
gena	black
postgena	black
first antennal segment	dark-brown to black
second antennal segment	white-grey

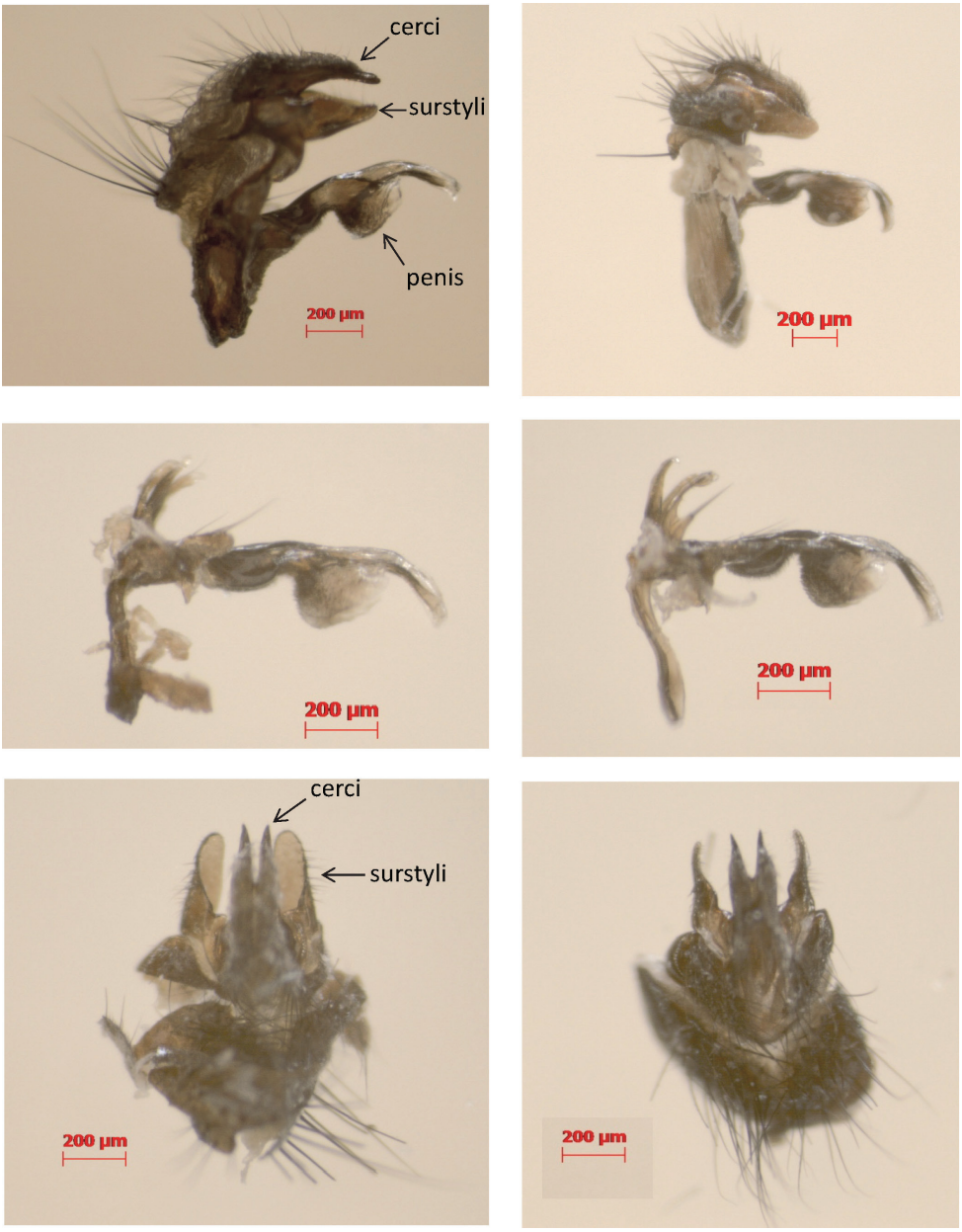


Figure 1. Lateral (top) and dorsal (bottom) view of the male copulatory organs of *Phormia regina* from W Europe (left) and N America (right) with a detail of the penis (middle).

et al. 2007), C2-J-3138/TK-N-3775 (Wells and Sperling 2001), CB1-SE/PDR-WR04 (Ready et al. 2009), ITS2F.dip/ITS2R (Song et al. 2008) and D1F/D2R (Stevens and Wall 2001), respectively.

Each 25 µl PCR reaction was prepared using 1 × PCR buffer, 0.2 mM dNTPs, 0.4 µM of each primer, 2.0 mM $MgCl_2$, 0.5 U of Taq DNA polymerase (Platinum®, Invitrogen), 2-4 µl DNA template (DNA was stored in 100 µl of elution buffer) and enough mQ-H_2O to complete the total PCR reaction volume. The thermal cycler program consisted of an initial denaturation step of 4 min at 94 °C, followed by 30-40 cycles of 45-60 s at 94 °C, 30–60 s at a fragment depending annealing temperature and 90 s at 72 °C; with a final extension of 7 min at 72 °C. The annealing temperatures were 45 °C for COI and COII, 48 °C for 16S and cyt *b*, 50 °C for ITS-2 and 55 °C for 28S. PCR products were cleaned using the NucleoFast96 PCR® kit (Macherey-Nagel) and bidirectionally sequenced on an ABI 3130 Genetic Analyzer (Applied Biosystems) using the BigDye® Terminator Cycle Sequencing Kit v3.1. Together with the *P. regina* specimens we also collected several *ProtoPhormia terraenovae* specimens that were also sequenced to increase the number of material for comparison (Appendix 1 - Supplementary Table 1). Sequences were assembled in SeqScape v2.5 (Applied Biosystems) and deposited in GenBank under accession numbers KF908069–KF908124 (COI), KF908126–KF908152 (COII), KF908153–KF908169 (cyt *b*), KF908054–KF908068 (16S), KF908170–KF908203 (ITS2), and KF908204–KF908237 (28S).

Phormiini and its sister clade Chrysomyini form the Chrysomyinae (Singh and Wells 2011a, b). We therefore downloaded from GenBank (and for all genes) all available sequences (at 11 July 2013) of the Phormiini (genera *Phormia*, *Proto-Phormia* and *Protocalliphora*) and of the Chrysomyini (genera *Chloroprocta*, *Chrysomya*, *Cochliomyia*, *Compsomyiops*, *Hemilucilia*, *Paralucilia* and *Trypocalliphora*) to allow comparison with closely related taxa (Table 1). Sequences were aligned in MAFFT v7 (Katoh and Standley 2013). Sequences with > 5 ambiguous positions were discarded and each dataset was trimmed to equal sequence length (Table 3). The 16S dataset was trimmed at 251 bp and at 350 bp to yield a higher number of Chrysomyinae haplotypes for the latter dataset (i.e. 22 vs. 42 unique haplotypes; six species in the ingroup for both datasets). Alignments are available as fasta files in the online Appendix 2 text file. Unique sequences (haplotypes) were selected in DAMBE5 (Xia 2013). Nucleotide sequence divergences within and between species (based on the haplotypes) were calculated using the uncorrected p-distances in MEGA v5.05 (Tamura et al. 2011). For these calculations we excluded haplotypes that were not identified to the species level (one *Protocalliphora* sp. for COI) or that were most likely identification errors (for details see the Results). MEGA v5.05 was also used to construct Neighbour-Joining (NJ) trees (Saitou and Nei 1987) using the p-distances with complete deletion of positions with ambiguities and alignment gaps (indels). Relative branch support was evaluated with 1000 bootstrap replicates (Felsenstein 1985). In all analyses, several *Lucilia* spp. or *Calliphora* spp. sequences from GenBank were added as outgroups, and for COI we also used *L. sericata* NICC0390 as outgroup (GenBank accession number KF908125). Author names of all species are provided in Table 1.

Table 3. Description of the *Phormia regina* and other Chrysomyinae DNA sequences (including those retrieved from GenBank) for each of the gene fragments.

Marker	COI	COII	16S		cyt *b*	ITS2	28S
			251 bp	350 bp		(without indels)	(without indels)
Fragment size (bp)	655	472	251	350	512	380 (224)	633 (592)
Phormia regina							
Total							
No of sequences	50	30	15	15	17	36	37
No of haplotypes	20	9	2	4	10	4	2
North America (NA)							
No of sequences	27	27	11	11	10	25	23
No of haplotypes	14	7	1	3	7	1	2
Mean intra-NA distances (%)	0.004	0.004	-	0.004	0.005	-	0.002
SE	0.001	0.002	-	0.003	0.002	-	0.002
min. – max.	0.002–0.008	0.002–0.006	-	0.003–0.006	0.002–0.008	-	0.002
Europe (EU)							
No of sequences	23	3	4	4	7	11	14
No of haplotypes	6	2	1	1	3	4(2)	1
Mean intra-EU distances (%)	0.003	0.002	-	-	0.002	0.002	-
SE	0.001	0.002	-	-	0.007	0.002	-
min. – max.	0.002–0.008	0.002	-	-	0.002–0.010	0.002	-
Mean p-distance between NA and EU	0.04	0.037	0.004	0.006	0.053	0.001	0.001
SE	0.007	0.008	-	0.003	0.009	0.001	0.001
min. – max.	0.036–0.044	0.034–0.042	0.004	0.005–0.009	0.047–0.061	0–0.004	0–0.002
Other Chrysomyinae							
Mean intraspecific p-distance	0.005	0.014	0.028	0.014	0.003	0.008	0.003
SE	0.009	0.014	0.009	-	0.002	0.005	0.004
min. – max.	0–0.042	0–0.037	0.018–0.036	0.014	0.002–0.005	0.004–0.015	0–0.010
Mean interspecific p-distance	0.066	0.046	0.038	0.023	0.079	0.085	0.007
SE	0.005	0.005	0.006	0.004	0.007	0.011	0.002
min. – max.	0.011–0.113	0.002–0.135	0.03–0.075	0.023–0.057	0.073–0.141	0.009–0.166	0–0.015

Results

Morphology

We did not detect morphological differences between N American and W European *P. regina* specimens in the 11 external color characters that we scored (Table 2). Also the male copulatory organs of W European and N American *P. regina* specimens were indistinguishable (Figure 1).

DNA sequence analysis

Basic information of the different datasets can be found in Table 3. There was only high bootstrap support for the monophyly of Chrysomyinae, Phormiini or Chrysomyini with 28S and a sister group relationship of *P. regina* and *Pr. terraenovae* with ITS2. Yet, for all fragments, except for 28S, there was high bootstrap support for the monophyly of *P. regina* (Figures 2–4 and Appendix 1 - Supplementary Figures 1–3).

COI: The COI NJ-tree showed two supported clades within *P. regina* (Figure 2). One clade (EU = Europe) comprised six haplotypes from Europe (23 specimens sequenced), while the other clade (NA = North America) comprised 14 haplotypes from N America (27 specimens sequenced). The seven NA haplotypes available in GenBank clustered within the NA clade. The mean p-distance between the EU and NA *P. regina* haplotypes was 0.04 ± 0.007 (Table 3). Sequence divergence in *P. regina* within each continent was approximately a ten-fold lower, viz. EU: 0.003 ± 0.001 – NA: 0.004 ± 0.001.

The mean p-distances between Chrysomyinae species pairs were: between three *Protocalliphora* spp.: 0.05 ± 0.006, 23 *Chrysomya* taxa: 0.06 ± 0.005 (the three *C. megacephala* specimens with GenBank accession numbers KC135924, KC135925 and KC135926 were treated as a different taxon from the other *C. megacephala* specimens because of a strong sequences divergence, viz. mean p-distance = 0.089 ± 0.01; see Figure 2), *Co. macellaria – Co. hominivorax*: 0.068 ± 0.009, and *H. semidiaphana – H. segmentaria*: 0.078 ± 0.001. The mean intra- and interspecific p-distances between all Chrysomyinae species (excluding *P. regina*) were 0.005 ± 0.009 and 0.066 ± 0.005, respectively (Table 3).

COII: The two EU and seven NA haplotypes of *P. regina* (from 30 specimens) formed two strongly supported clades (Figure 3) separated by mean p-distance of 0.037 ± 0.008 (Table 3). The three COII sequences from GenBank (from NA specimens) had the same haplotype as our NA specimens. Sequence divergence in *P. regina* within each continent was approximately a ten-fold lower, viz. EU: 0.002 ± 0.002 – NA: 0.004 ± 0.002 (Table 3). The mean p-distance between the 14 *Chrysomya* taxa was 0.059 ± 0.007. We considered *C. megacephala*_FJ153270 and *C. rufifacies*_FJ839395 as misidentifications, and *C. rufifacies*_AY842670_AY842671 to be different from the other *C. rufifacies* individuals given the high sequence divergence (viz. mean p-distance = 0.10 ± 0.013). The mean p-distance between *Co. macellaria* and *Co. hominivorax* was 0.048 ± 0.009. The mean intra- and interspecific p-distances among all Chrysomyinae species (excluding *P. regina*) were 0.014 ± 0.014 and 0.046 ± 0.005, respectively (Table 3).

Cyt *b*: The three EU and seven NA haplotypes of *P. regina* (from 17 specimens) formed two strongly supported clades (Figure 4) with a mean p-distance of 0.053 ± 0.009 between these two clades (Table 3). There were no cyt *b* sequences of *Phormia* in GenBank. Sequence divergence in *P. regina* within each continent was approximately a ten-fold lower, viz. EU: 0.002 ± 0.007 – NA: 0.005 ± 0.002 (Table 3). The mean p-distance between the three *Chrysomya* species was 0.046 ± 0.005. The mean intra- and

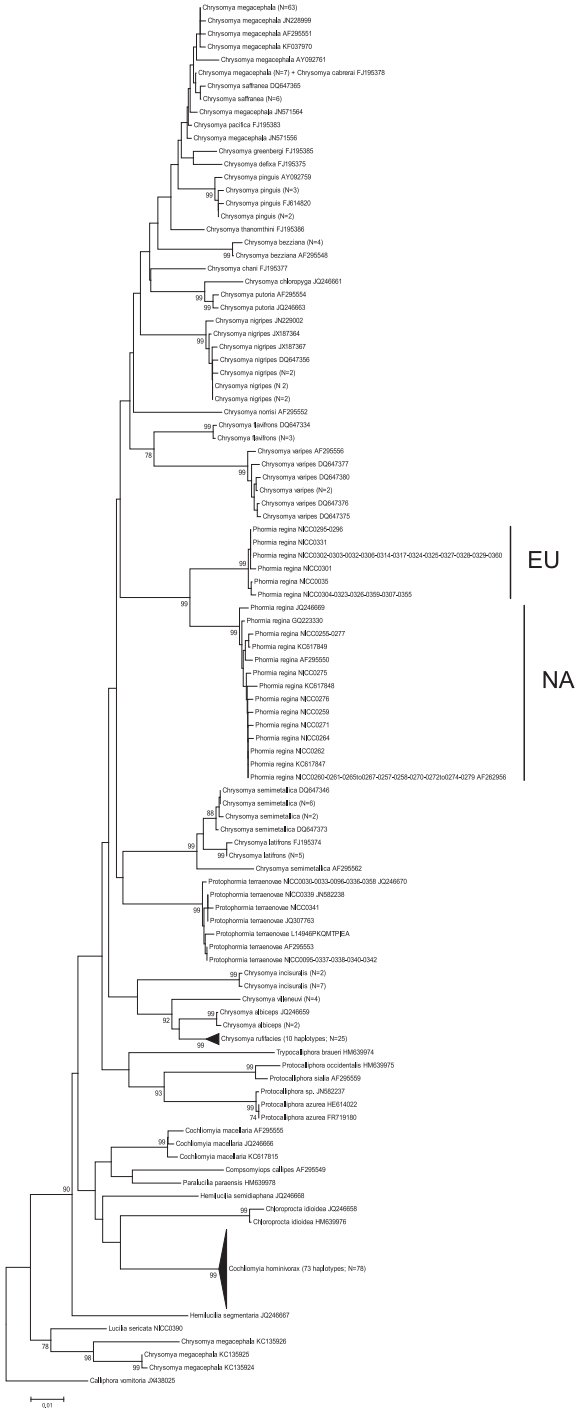


Figure 2. Neighbour-Joining tree (p-distances) of a 655 bp fragment of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene. Bootstrap values ≥ 70% are shown at the nodes. N gives the number of specimens of that haplotype. EU = *P. regina* haplotypes from W Europe; NA = *P. regina* haplotypes from N America.

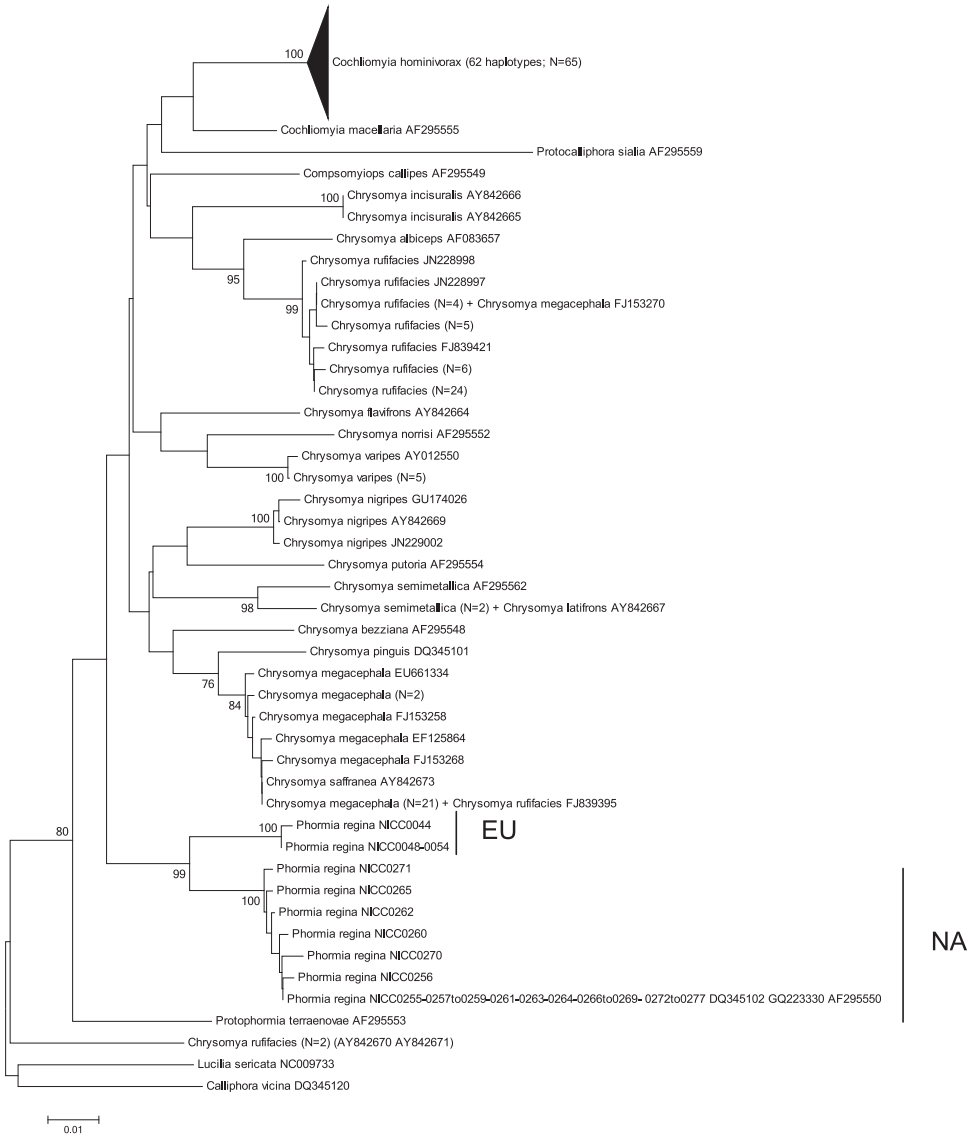


Figure 3. Neighbour-Joining tree (p-distances) of a 472 bp fragment of the mitochondrial cytochrome *c* oxidase subunit II (COII) gene. Bootstrap values ≥ 70% are shown at the nodes. N gives the number of specimens of that haplotype. EU = *P. regina* haplotypes from W Europe; NA = *P. regina* haplotypes from N America.

interspecific p-distances among all Chrysomyinae species (excluding *P. regina*) were 0.003 ± 0.002 and 0.079 ± 0.007, respectively (Table 3).

16S: For the 350 bp dataset, the three NA 16S haplotypes (from 15 specimens) (mean within NA p-distance = 0.004 ± 0.003; Table 3) formed a well-supported clade, and formed a monophyletic group with the single EU haplotype (Supplementary Figure 1A). The mean p-distance between the NA and EU haplotypes was 0.006 ± 0.003.

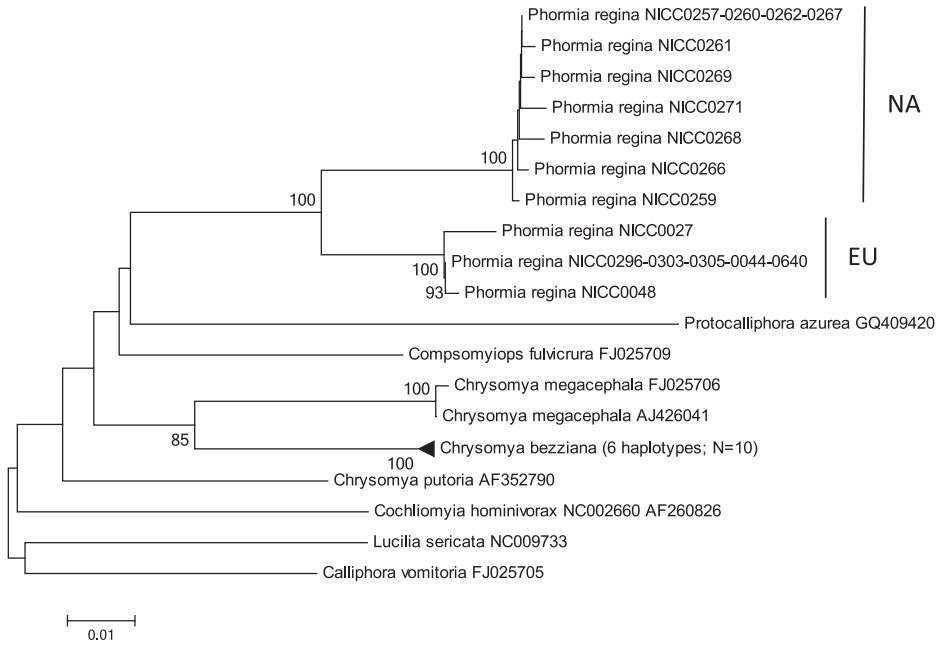


Figure 4. Neighbour-Joining tree (p-distances) of a 512 bp fragment of the mitochondrial cytochrome *b* (cyt *b*) gene. Bootstrap values ≥ 70% are shown at the nodes. N gives the number of specimens of that haplotype. EU = *P. regina* haplotypes from W Europe; NA = *P. regina* haplotypes from N America.

The mean p-distance between *C. megacephala* and *C. rufifacies* was 0.040 ± 0.009. The mean intra- and interspecific p-distances among all Chrysomyinae species (excluding *P. regina*) were 0.014 and 0.023 ± 0.004.

For the 251 bp dataset, all eleven NA specimens had the same haplotype with a p-distance of 0.004 to the EU haplotype (four specimens) (Supplementary Figure 1B). The mean p-distance between *C. megacephala* and *C. rufifacies* was 0.059 ± 0.012. The mean intra- and interspecific p-distances among all Chrysomyinae species (excluding *P. regina*) were 0.028 ± 0.009 and 0.038 ± 0.006, respectively (Table 3).

ITS2: Excluding indels, all *P. regina* specimens (36 specimens) had the same haplotype (Supplementary Figure 2), except for *P. regina* NICC0302 that had a C instead of a T at position 219 of the alignment (p-distance = 0.003). *Phormia regina* NICC0640 had a deletion at position 201, and *P. regina* NICC0048 had an insertion of a G at position 270 of the alignment. Both specimens were from the same locality (Liège – Belgium) in W Europe. The p-distance between *Co. hominivorax* and *Co. macellaria* was 0.008 ± 0.001, that between *H. segmentaria* and *H. semidiaphana* was 0.106 ± 0.018, and the mean p-distance among 16 *Chrysomya* species was 0.085 ± 0.010. The mean intra- and interspecific p-distances among all Chrysomyinae species (excluding *P. regina*) were 0.008 ± 0.005 and 0.085 ± 0.011, respectively (Table 3).

28S: All 37 *P. regina* specimens had the same haplotype, except for *P. regina* JQ246614 from N America that had an AG insertion at positions 460-461 of the

alignment (Supplementary Figure 3). One haplotype of *Pr. terraenovae* (three specimens with GenBank accession numbers AJ300142, JQ307780 and JQ246615) only differed by two indels from haplotype JQ246614 of *P. regina* (at positions 408 and 460-461) (the other *Pr. terraenovae* haplotype differed at more positions). The mean p-distance between *Co. macellaria* and *Co. hominivorax* was 0.005, that between *Pro. azurea* and *Pro. sialia* was zero [an indel at position 439 (A) in *Pro. azurea*) of the alignment], and that between *H. semidiaphana* and *H. segmentaria* was 0.013. The mean p-distance among the six *Chrysomya* species was 0.006 ± 0.002. The mean intra- and interspecific p-distances among all Chrysomyinae species (excluding *P. regina*) were 0.003 ± 0.004 and 0.007 ± 0.002, respectively (Table 3).

Discussion

Desmyter and Gosselin (2009) and Boehme et al. (2012) found a mean sequence divergence of approximately 4% within a 304 bp and the barcoding COI region between N American and W European *P. regina*, respectively. We confirmed this COI divergence with newly sequenced material. Such a strong divergence at COI is common among insect species (e.g. Park et al. 2011a, b, Webb et al. 2012, Ng'endo et al. 2013). Moreover, we here show a similar degree of divergence at two other mtDNA genes, viz. COII (3.7%) and cyt *b* (5.3%). The 'within-continent' divergence in *P. regina* was very low (0.2-0.5% for the three genes) and comparable to the intraspecific differentiation of other Chrysomyinae (0.5% for COI, 1.4% for COII, 0.3% for cyt *b*). Hence, the high between-continent mtDNA differentiation, and low within-continent mtDNA divergence may hint at a taxonomic difference between the N American and W European populations. In order to evaluate this suggestion, we included all publicly available GenBank sequences from species of the subfamily Chrysomyinae for the four mtDNA and two nDNA gene fragments that we sequenced. The combined study of mtDNA and nDNA has proven valuable to disentangle the taxonomy of other calliphorid species (e.g. Nelson et al. 2007, Sonet et al. 2012).

On the one hand, our results show that the mean p-distance of other intrageneric interspecific comparisons (COI: 5–6.8%, COII: 4.8-5.9%, cyt *b*: 4.6%, 16S (251 bp): 5.9%), or among other Chrysomyinae species in general (COI: 6.6%, COII: 4.6%, cyt *b*: 7.9%, 16S (251 bp): 3.8%), are higher than the mean p-distances between N American and W European *P. regina* at the four mtDNA fragments (COI: 4%, COII: 3.7%, cyt *b*: 5.3%, 16S: 0.6%). For cyt *b* the NA-EU differentiation in *P. regina* is higher than that observed within other Chrysomyinae species (0.3%) yet still below the minimum interspecific p-distance (7.3%). On the other hand, for COI and COII, the NA-EU differentiation in *P. regina* is higher than the intraspecific differentiation in other Chrysomyinae species and well within the range of interspecific p-distances within Chrysomyinae. Yet, the low interspecific p-distance between some Chrysomyinae species may be due to misidentifications or may be the result of a natural process (e.g. hybridization, incomplete lineage sorting). Likewise, the high intraspecific variation within some species may be indicative of cryptic diversity (see further).

North American and W European *P. regina* were not differentiated at both nDNA fragments, and at the mtDNA 16S (< 1%), whereas interspecific p-distances in Chrysomyinae in general are substantial for ITS2 (8.5%) and 16S (3.8%). Moreover, the NA-EU differentiation in *P. regina* at these genes was even lower than the minimum intraspecific differentiation within other Chrysomyinae. This suggests that the variation at these genes in *P. regina* is intraspecific variation. Finally, we could neither detect color differences in 11 external characters, nor in the general shape of the male copulatory organs between N American and W European specimens. Evidently, a statistical analysis of more specimens (from a wider range of the species' distribution) is necessary to reliably assess within and among population variation at these (and eventually other) morphological characters. For the time being, we consider the high differentiation at COI, COII and cyt *b*, but the low (16S, nDNA) or lack of (morphological) differentiation, as indicative of substantial intraspecific mtDNA sequence divergence, rather than as a species-level differentiation.

Our findings may have important implications for the use of *P. regina* in forensic and other scientific fields. Indeed, it has been suggested that the high variation in developmental times and growth curves of *P. regina* (e.g. Byrd and Allen 2001 and references therein) is partly due to differences in the population genetic structure (Picard and Wells 2009) and that therefore ecological data obtained from one population should not be generalized or extrapolated to other populations (Byrne et al. 1995). Interestingly, Marchenko (2001) reports a mean accumulated degree-days (from egg to adult) of 148 °C (lower development temperature: 11.4 °C) for Russian/Lithuanian *P. regina,* whereas a mean accumulated degree-days of 162 °C (lower development temperature: 11.16 °C) was found for N American *P. regina* (Yves Braet, unpublished preliminary results). Hence, the strong mtDNA divergence between N American and W European *P. regina* requires a sound comparison of the ecology of populations from both continents, especially since *P. regina* is a key species in the study of the physiology and neurology of insect feeding (e.g. Haselton et al. 2009, Larson and Stoffolano 2011, Ishida et al. 2012). Moreover, if locally diverged populations differ in their developmental biology, then this may affect the estimate of PMImin.

Intraspecific mtDNA divergence in other Chrysomyinae species is sometimes also high, viz. 4.3% for COI in *C. megacephala*, and 2.2%, 2.6% and 3.7% for COII in *C. megacephala*, *C. semimetallica* and *C. rufifacies*, respectively. Whereas these high intraspecific divergences may be due to hybridization/introgression or incomplete lineage sorting, they may also point to misidentifications. Obviously these issues are problematic if DNA barcoding of animals is only based on COI, as advocated by Hebert et al. (2003a, b). For instance, three *C. megacephala* specimens (KC135924, KC139925, KC135926) have a remarkably high p-distance of 8% with the other *C. megacephala* haplotypes and it would be advisable to re-identify these specimens. Also *C. semimetallica* shows much more intraspecific sequence variation (mean p-distance = 0.011 ± 0.003) as compared to other Chrysomyinae species but

at the same time the species has a low mean interspecific p-distance with *C. albiceps* (p-distance = 0.017 ± 0.004).

Although there is no doubt that COI is a useful tool for the identification of forensically important Chrysomyinae species (Wells and Sperling 2001, Nelson et al. 2007, Wells and Williams 2007, Desmyter and Gosselin 2009, Boehme et al. 2012) not all species can be identified with COI. For instance, there is very low mean interspecific p-distance of 0.006 ± 0.002 between *C. megacephala* (excluding the three aforementioned haplotypes), *C. cabrerai*, *C. saffranea* and *C. pacifica* (the first two even share a haplotype) (see also Harvey et al. 2008). Therefore, other genes (or gene fragments) might help to overcome the shortcomings of the sole use of COI as molecular identification tool. We here showed that also COII may be a good DNA barcode marker in the Chrysomyinae. Indeed, the mean interspecific p-distance at COII is 4.6%, whereas the mean intraspecific distance is much lower (1.4%). Yet, the amount of Chrysomyinae COII data that is currently available in public libraries such as GenBank (194 sequences representing 108 haplotypes from 20 species), is rather limited compared to the amount of COI data (339 sequences representing 180 haplotypes from 36 species) (Table 1). Moreover, the problems inherent to misidentifications and introgression also apply to COII (or any other DNA marker). For instance, *C. megacephala* FJ153270 shares a haplotype within the *C. rufifacies* clade, and *C. rufifacies* FJ839395 shares a haplotype within the *C. megacephala* clade. Also other species share haplotypes such as *C. semimetallica* and *C. latifrons*. The other two mtDNA fragments (cyt *b* and 16S) cannot yet be evaluated as DNA barcode markers because of insufficient sequence data (cyt *b*: 32 sequences representing 20 haplotypes of five species; 16S: 39 sequences representing nine haplotypes of six species) (Table 1), but both have been shown to discriminate sufficiently between other dipteran species of forensic interest (Vincent et al. 2000, Li et al. 2010).

So far, the forensically important species within the Chrysomyinae belong to the genera *Chrysomya*, *Cochliomyia*, *Paralucilia*, *ProtoPhormia* and *Phormia*. A number of COI reference datasets of these species are available (e.g. Wallman and Donnellan 2001, Wells and Sperling 2001, Nelson et al. 2007, Wells and Williams 2007, Harvey et al. 2008, Desmyter and Gosselin 2009, Boehme et al. 2012) and they seem to work well to identify most forensically important species. Yet, it is important to also include species without a clear forensic interest in (local) reference databases because this will improve the assessment of species boundaries which, in turn, may help to reach a stable taxonomy.

In conclusion, we observed substantial differentiation between N American and W European *P. regina* at the mtDNA genes COI, COII and cyt *b*, but not at the 16S rDNA and the nDNA genes ITS2 and 28S. Moreover, we neither detected any morphological differentiation between specimens from both continents. We therefore consider the strong mtDNA divergence between specimens from both continents as intraspecific variation. This differentiation has to be taken into account when using *P. regina* in forensic casework or physiological studies. Finally, the use of COII as a DNA barcode marker in the Chrysomyinae seems to perform as good as the standard COI barcode region.

Acknowledgements

We wish to thank Françoise Hubrecht for her support, Knut Rognes for his help with the literature, and Sofie Vanpoucke (NICC) for her help in making the pictures of the genitalia. We thank Jens Amendt, Richard Zehner and Benoît Vincent for providing part of the W European *Phormia* material, and Neal Haskell, Jefferey Tomberlin and Jeffrey Wells for collecting part of the N American *Phormia* specimens. The comments of one referee improved the manuscript considerably. This work was done in the context of FWO research network W0.009.11N "Belgian Network for DNA Barcoding". JEMU is funded by the Belgian Science Policy Office (Belspo).

References

Amendt J, Krettek R, Zehner R (2004) Forensic entomology. Naturwissenschaften 91: 51–65. doi: 10.1007/s00114-003-0493-5

Amendt J, Campobasso CP, Gaudry E, Reiter C, LeBlanc HN, Hall MJR (2007) Best practice in forensic entomology: standards and guidelines. International Journal of Legal Medicine 121: 90–104. doi: 10.1007/s00414-006-0086-x

Anderson GS (2000) Minimum and maximum development rate of some forensically important Calliphoridae (Diptera). Journal of Forensic Science 45: 824–832.

Boehme P, Amendt J, Zehner R (2012) The use of COI barcodes for molecular identification of forensically important fly species in Germany. Parasitology Research 110: 2325–2335. doi: 10.1007/s00436-011-2767-8

Brundage A, Bros S, Honda JY (2011) Seasonal and habitat abundance and distribution of some forensically important blow flies (Diptera: Calliphoridae) in Central California. Forensic Science International 212: 115–120.

Byrd JH, Allen JC (2001) The development of the black blow fly, *Phormia regina* (Meigen). Forensic Science International 120: 79–88. doi: 10.1016/S0379-0738(01)00431-5

Byrne AL, Camann MA, Cyr TL, Catts EP, Espelie KE (1995) Forensic implications of biochemical differences among geographic populations of the black blow fly, *Phormia regina*. Journal of Forensic Sciences 40: 372–377.

Catts EP (1992) Problems in estimating the postmortem interval in death investigations. Journal of Agricultural Entomology 9: 245–255.

Charabidze D (2012) Necrophagous insects and forensic entomology. Annales de la Société Entomologique de France 48: 239–252. doi: 10.1080/00379271.2012.10697773

Coffey MD (1966) Studies on the association of flies (Diptera) with dung in southeastern Washington. Annals of the Entomological Society of America 59: 207–218.

DeBry RW, Timm A, Wong ES, Stamper T, Cookman C, Dahlem GA (2013) DNA-based identification of forensically important *Lucilia* (Diptera: Calliphoridae) in the continental United States. Journal of Forensic Sciences 58: 73–78. doi: 10.1111/j.1556-4029.2012.02176.x

Desmyter S, Gosselin M (2009) COI sequence variability between Chrysomyinae of forensic interest. Forensic Science International 3: 89–95. doi: 10.1016/j.fsigen.2008.11.002

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783–791. doi: 10.2307/2408678

Francesconi F, Lupi O (2012) Myasis. Clinical Microbiology Reviews 25: 79–105. doi: 10.1128/CMR.00010-11

Greenberg B (1991) Flies as forensic indicators. Journal of Medical Entomology 28: 565–577.

Griffiths GCD (1982) On the systematic position of *Mystacinobia* (Diptera: Calliphoridae). Memoirs of the Entomological Society, Washington 10: 70–77.

Hall DG (1948) The blowflies of North America. Thomas Say Foundation, Baltimore, MD.

Harvey ML, Gaudieri S, Villet MH, Dadour IR (2008) A global study of forensically significant calliphorids: implications for identification. Forensic Science International 177: 66–67. doi: 10.1016/j.forsciint.2007.10.009

Haselton AT, Downer KE, Zylstra J, Stoffolano JG (2009) Serotonin inhibits protein feeding in the blow fly, *Phormia regina* (Meigen). Journal of Insect Behavior 22: 452–463. doi: 10.1007/s10905-009-9184-1

Hebert PDN, Ratnasingham S, DeWaard JR (2003a) Barcoding animal life: cytochrome *c* oxidase subunit I divergences among closely related species. Proceedings of the Royal Society of London B (Supplement) 270: S96–S99. doi: 10.1098/rsbl.2003.0025

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003b) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Ishida Y, Nagae T, Azuma M (2012) A water-specific aquaporin is expressed in the olfactory organs of the blowfly, *Phormia regina*. Journal of Chemical Ecology 38: 1057–1061. doi: 10.1007/s10886-012-0157-z

Jordaens K, Sonet G, Desmyter S (in press) IX. Aperçu des méthodes moléculaires d'identification d'insectes d'intérêt forensique. In: Charabidze D, Gosselin M (Eds) Insectes, Cadavre & Scène de Crime – Principes et Applications de l'Entomologie Médico-légale. De Boeck, Belgium.

Kamal AS (1958) Comparative study of thirteen species of sarcosaprophagous Calliphoridae and Sarcophagidae (Diptera). I. Bionomics. Annals of the Entomological Society of America 51: 261–270.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30: 772–780. doi: 10.1093/molbev/mst010

Knipling EF, Rainwater HT (1937) Species and incidence of dipterous larvae concerned in wound myiasis. Journal of Parasitology 23: 451–455. doi: 10.2307/3272391

Kutty SN, Bernasconi MV, Sifner F, Meier R (2007) Sensitivity analysis, molecular systematics and natural history evolution of Scathophagidae (Diptera: Cyclorrhapha: Calyptratae). Cladistics 23: 64–83. doi: 10.1111/j.1096-0031.2006.00131.x

Kutty SN, Pape T, Wiegmann BM, Meier R (2010) Molecular phylogeny of the Calyptratae (Diptera: Cyclorrhapha) with an emphasis on the superfamily Oestroidea and the posi-

tion of Mystacinobiidae and McAlpine's fly. Systematic Entomology 35: 614–635. doi: 10.1111/j.1365-3113.2010.00536.x

Larson K, Stoffolano JG (2011) Effect of high and low concentrations of sugar solutions fed to adult male, *Phormia regina* (Diptera: Calliphoridae), on 'bubbling' behavior. Annals of the Entomological Society of America 104: 1399–1403. doi: 10.1603/AN11029

Li X, Cai JF, Guo YD, Wu KL, Wang JF, Liu QL, Wang XH, Chang YF, Yang L, Lan LM, Zhong M, Wang X, Song C, Liu Y, Li JB, Dai ZH (2010) The availability of 16S rRNA for the identification of forensically important flies (Diptera: Muscidae) in China. Tropical Biomedicine 27: 155–166.

Marchenko MI (2001) Medicolegal relevance of cadaver entomofauna for the determination of the time of death. Forensic Science International 120: 89–109. doi: 10.1016/S0379-0738(01)00416-9

Nabity PD, Higley LG, Heng-Moss TM (2007) Light-induced variability in development of forensically important blow fly *Phormia regina* (Diptera: Calliphoridae). Journal of Medical Entomology 44: 351–358. doi: 10.1603/0022-2585(2007)44[351:LVIDOF]2.0.CO;2

Nelson LA, Wallman JF, Dowton M (2007) Using COI barcodes to identify forensically and medically important blowflies. Medical and Veterinary Entomology 21: 44–52. doi: 10.1111/j.1365-2915.2007.00664.x

Ng'endo RN, Osiemo ZB, Brandl R (2013) DNA barcodes for species identification in the hyperdiverse ant genus *Pheidole* (Formicidae: Myrmicinae). Journal of Insect Science 13: 27. doi: 10.1673/031.013.2701

Núñez-Vázquez C, Tomberlin J, Cantú-Sifuentes M, García-Martínez O (2013) Laboratory development and field validation of *Phormia regina* (Diptera: Calliphoridae). Journal of Medical Entomology 50: 252–260. doi: 10.1603/ME12114

Park DS, Suh SJ, Hebert PDN, Oh HW, Hong KJ (2011a) DNA barcodes for two scale insect families, mealybugs (Hemiptera: Pseudococcidae) and armored scales (Hemiptera: Diaspididae). Bulletin of Entomological Research 101: 429–434. doi: 10.1017/S0007485310000714

Park DS, Foottit R, Maw E, Hebert PDN (2011b) Barcoding bugs: DNA-based identification of the True Bugs (Insecta: Hemiptera: Heteroptera). PLoS ONE 6: e18749. doi: 10.1371/journal.pone.0018749

Picard CJ, Wells JD (2009) Survey of the genetic diversity of *Phormia regina* (Diptera: Calliphoridae) using amplified fragment length polymorphisms. Journal of Medical Entomology 46: 664–670. doi: 10.1603/033.046.0334

Ready PD, Testa JM, Wardhana AH, Al-Izzi M, Khalaj M, Hall MJR (2009) Phylogeography and recent emergence of the Old World screwworm fly, *Chrysomya bezziana*, based on mitochondrial and nuclear gene sequences. Medical and Veterinary Entomology 23: 43–50. doi: 10.1111/j.1365-2915.2008.00771.x

Rognes K (1997) The Calliphoridae (blowflies) (Diptera: Oestroidea) are not a monophyletic group. Cladistics 13: 27–66. doi: 10.1111/j.1096-0031.1997.tb00240.x

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

Singh B, Wells JD (2011a) Chrysomyinae (Diptera: Calliphoridae) is monophyletic: a molecular systematic analysis. Systematic Entomology 36: 415–420. doi: 10.1111/j.1365-3113.2011.00568.x

Singh B, Wells JD (2011b) Molecular systematics of the Calliphoridae (Diptera: Oestroidea): Evidence from one mitochondrial and three nuclear genes. Journal of Medical Entomology 50: 15–23. doi: 10.1603/ME11288

Sonet G, Jordaens K, Braet Y, Desmyter S (2012) Why is the molecular identification of the forensically important blowfly species *Lucilia caesar* and *L. illustris* (family Calliphoridae) so problematic? Forensic Science International 223: 153–159. doi: 10.1016/j.forsciint.2012.08.020

Song Z, Wang X, GeQiu Liang G (2008) Species identification of some common necrophagous flies in Guangdong province, southern China based on the rDNA internal transcribed spacer 2 (ITS2). Forensic Science International 175: 17–22. doi: 10.1016/j.forsciint.2007.04.227

Sperling FA, Anderson GS, Hickey DA (1994) A DNA-based approach to the identification of insect species used for postmortem interval estimation. Journal of Forensic Sciences 39: 418–427.

Stevens JR, Wall R (2001) Genetic relationships between blowflies (Calliphoridae) of forensic importance. Forensic Science International 120: 116–123. doi: 10.1016/S0379-0738(01)00417-0

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Vincent S, Vian JM, Carlotti MP (2000) Partial sequencing of the mitochondrial cytochrome *b* oxidase subunit I gene: A tool for the identification of European species of blow flies for postmortem interval estimation. Journal of Forensic Sciences 45: 820–823.

Wallman JF, Donnellan SC (2001) The utility of mitochondrial DNA sequences for the identification of forensically important blowflies (Diptera: Calliphoridae) in southeastern Australia. Forensic Science International 120: 60–67. doi: 10.1016/S0379-0738(01)00426-1

Webb JM, Jacobus LM, Funk DH, Zhou X, Kondratieff B, Geraci CJ, DeWalt RE, Baird DJ, Richard B, Phillips I, Hebert PDN (2012) A DNA barcode library for North American Ephemeroptera: progress and prospects. PLoS ONE 7: e38063. doi: 10.1371/journal.pone.0038063

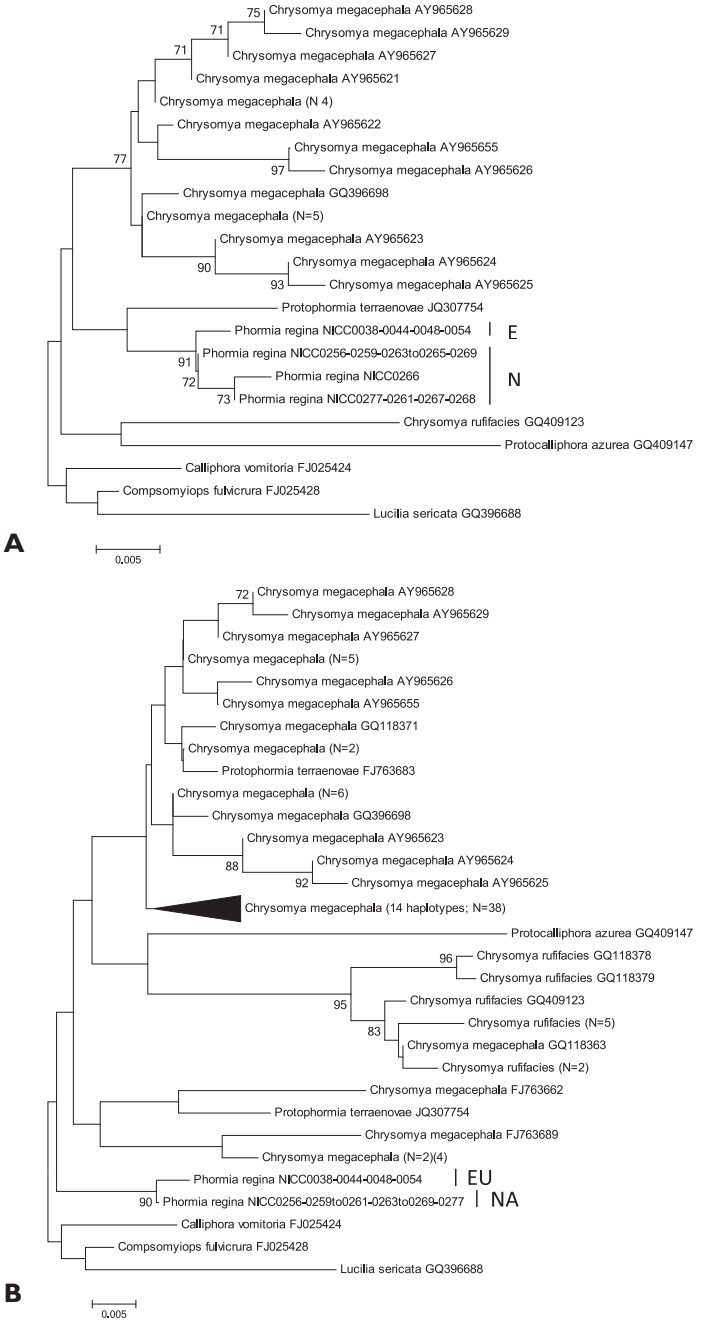
Wells JD, Sperling FAH (2001) DNA-based identification of forensically important Chrysomyinae (Diptera: Calliphoridae). Forensic Science International 120: 110–115. doi: 10.1016/S0379-0738(01)00414-5

Wells JD, Stevens JR (2008) Application of DNA-based methods in forensic entomology. Annual Review in Entomology 53: 103–120. doi: 10.1146/annurev.ento.52.110405.091423

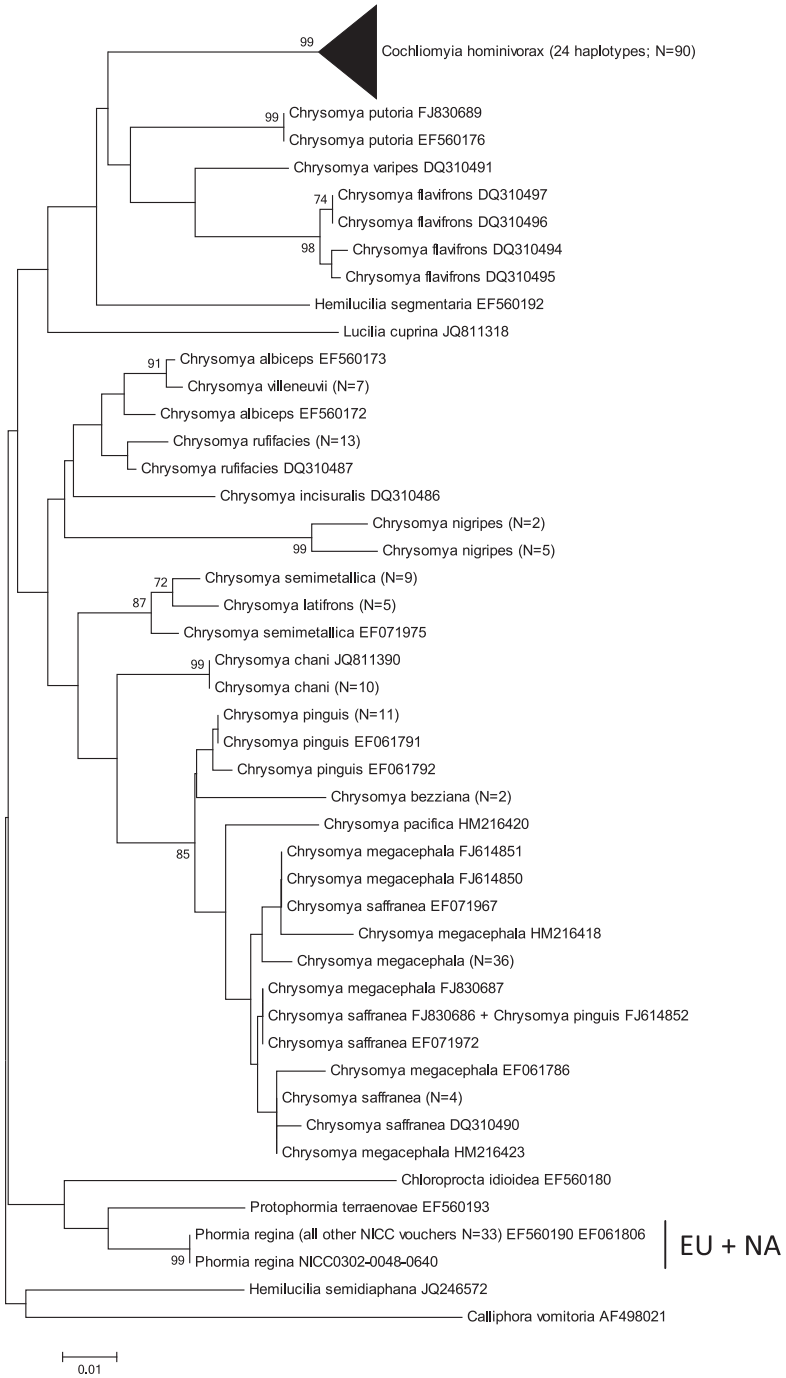
Wells JD, Williams DW (2007) Validation of a DNA-based method for identifying Chrysomyinae (Diptera: Calliphoridae) used in a death investigation. International Journal of Legal Medicine 121: 1–8. doi: 10.1007/s00414-005-0056-8

Xia X (2013) DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Molecular Biology and Evolution 30: 1720–1728. doi: 10.1093/molbev/mst064

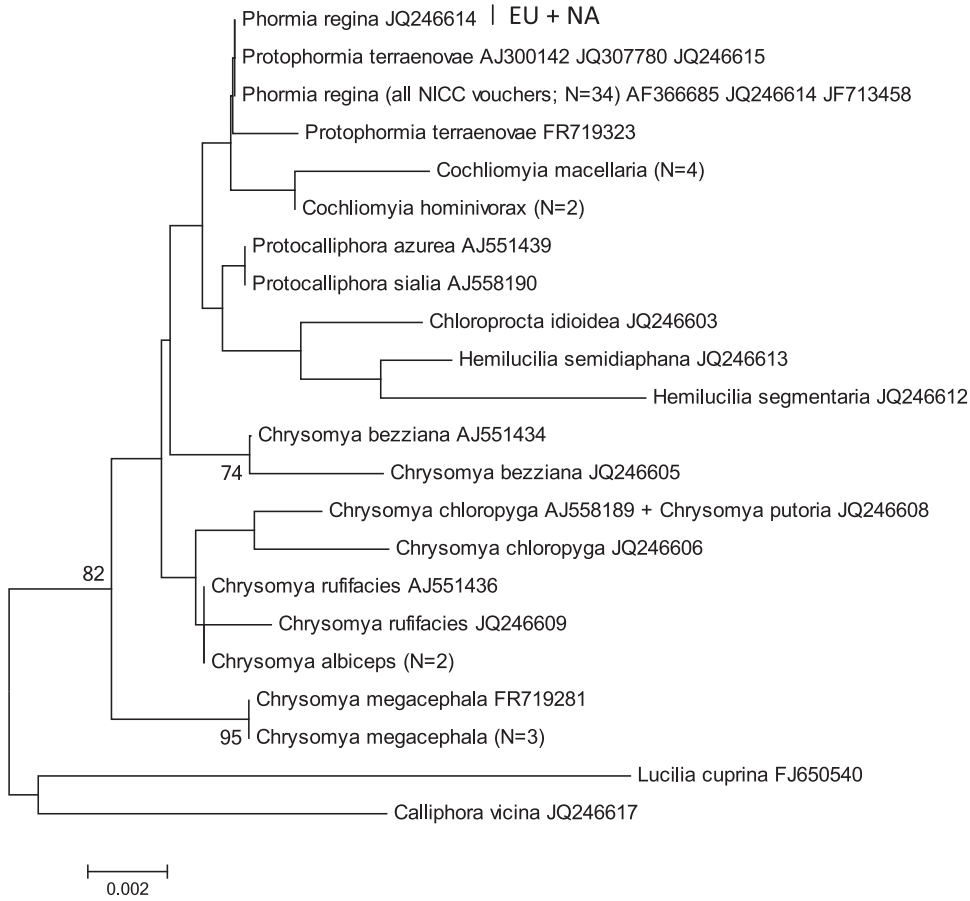
Appendix 1



Supplementary figure 1. Neighbour-Joining tree (p-distances) of a 350 bp (**A**) and of a 251 bp (**B**) fragment of the mitochondrial 16S gene. Bootstrap values ≥ 70% are shown at the nodes. N gives the number of specimens of that haplotype. EU = *P. regina* haplotypes from W Europe; NA = *P. regina* haplotypes from N America.



Supplementary figure 2. Neighbour-Joining tree (p-distances) of a 404 bp (229 bp without indels) fragment of the nuclear internal transcribed spacer 2 (ITS2). Bootstrap values ≥ 70% are shown at the nodes. N gives the number of specimens of that haplotype. EU = *P. regina* haplotypes from W Europe; NA = *P. regina* haplotypes from N America.



Supplementary figure 3. Neighbour-Joining tree (p-distances) of a 633 bp fragment of the nuclear 28S gene. Bootstrap values ≥ 70% are shown at the nodes. N gives the number of specimens of that haplotype. EU = *P. regina* haplotypes from W Europe; NA = *P. regina* haplotypes from N America.

Supplementary table 1. Sampling localities, voucher numbers and GenBank numbers of the *Phormia regina* and *Protophormia terraenovae* that were sequenced in this study.

Species	continent/ country	country/ state	city/county	latitude/longitude	voucher no.	GenBank accession no.					
						COI	COII	16S	cyt *b*	ITS2	28S
Phormia regina	Europe	Belgium	Andrimont	50°36'36"N, 5°54'36"E	NICC 0323	KJ908102					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0324	KF908103					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0325	KF908104					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0326	KF908105					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0327	KF908106					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0328	KF908107					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0329	KF908108					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0331	KF908109					
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0332					KF908199	KF908234
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0334					KF908200	KF908235
			Andrimont	50°36'36"N, 5°54'36"E	NICC 0336					KF908201	
			Auderghem	50°49'05"N, 4°24'41"E	NICC 0032	KF908069					
			Custine	49°41'41"N, 3°49'28"E	NICC 0314	KF908100					
			Genk	50°57'13"N, 5°29'56"E	NICC 0038			KF908054			
			Genk	50°57'13"N, 5°29'56"E	NICC 0355	KF908110					
			Hastière	50°13'14"N, 4°49'39"E	NICC 0027				KF908153	KF908171	
			Laeken	50°53'10"N, 4°22'36"E	NICC 0044		KF908126	KF908055	KF908154	KF908172	
			Liège	50°37'49"N, 5°33'17"E	NICC 0048		KF908127	KF908056	KF908155	KF908173	KF908205
			Liège	50°37'49"N, 5°33'17"E	NICC 0638						KF908202
			Liège	50°37'49"N, 5°33'17"E	NICC 0640				KF908169	KF908203	KF908236
			Liège	50°37'49"N, 5°33'17"E	NICC 0641						KF908237
			Messancy	49°35'31"N, 5°48'54"E	NICC 0317	KF908101					
			Schaerbeek	50°51'34"N, 4°22'25"E	NICC 0035	KF908070					
			Schoonaarde	51°00'17"N, 4°00'05"E	NICC 0359	KF908111					
			Schoonaarde	51°00'17"N, 4°00'05"E	NICC 0360	KF908112					
			Steendorp	51°07'25"N, 4°14'49"E	NICC 0054		KF908128	KF908057			KF908206
			Toernich	49°39'01"N, 5°45'07"E	NICC 0024					KF908170	KF908204

Species	continent/ country	country/ state	city/county	latitude/longitude	voucher no.	GenBank accession no.					
						COI	COII	16S	cyt *b*	ITS2	28S
		France	Sarreguemines	49°06'50"N, 7°04'18"E	NICC 0295	KF908092					KF908227
			Sarreguemines	49°06'50"N, 7°04'18"E	NICC 0296	KF908093			KF908166	KF908195	KF908228
		Germany	Frankfurt	50°06'41"N, 8°40'49"E	NICC 0301	KF908094				KF908196	KF908229
			Frankfurt	50°06'41"N, 8°40'49"E	NICC 0302	KF908095				KF908197	KF908230
			Frankfurt	50°06'41"N, 8°40'49"E	NICC 0303	KF908096			KF908167		
			Frankfurt	50°06'41"N, 8°40'49"E	NICC 0304	KF908097					
			Frankfurt	50°06'41"N, 8°40'49"E	NICC 0305				KF908168		KF908231
			Frankfurt	50°06'41"N, 8°40'49"E	NICC 0306	KF908098				KF908198	KF908232
			Frankfurt	50°06'41"N, 8°40'49"E	NICC 0307	KF908099					KF908233
	USA	Indiana	Rensselaer Co.	40°56'12"N, 87°09'03"W	NICC 0275	KF908088	KF908150				KF908222
			Rensselaer Co.	40°56'12"N, 87°09'03"W	NICC 0276	KF908089	KF908151			KF908192	KF908223
			Rensselaer Co.	40°56'12"N, 87°09'03"W	NICC 0277	KF908090	KF908152	KF908068		KF908193	KF908224
			Rensselaer Co.	40°56'12"N, 87°09'03"W	NICC 0278					KF908194	KF908225
			Rensselaer Co.	40°56'12"N, 87°09'03"W	NICC 0279	KF908091					KF908226
		Texas	Brazos	32°39'41"N, 98°07'19"W	NICC 0265	KF908080	KF908140	KF908063			KF908212
			Brazos	32°39'41"N, 98°07'19"W	NICC 0266	KF908081	KF908141	KF908064	KF908161	KF908184	KF908213
			Brazos	32°39'41"N, 98°07'19"W	NICC 0267	KF908082	KF908142	KF908065	KF908162		KF908214
			Brazos	32°39'41"N, 98°07'19"W	NICC 0268		KF908143	KF908066	KF908163	KF908185	KF908215
			Brazos	32°39'41"N, 98°07'19"W	NICC 0269		KF908144	KF908067	KF908164	KF908186	KF908216
		Virginia	Pr. Williams Co.	38°31'20"N, 77°17'22"W	NICC 0260	KF908075	KF908135		KF908158	KF908179	
			Pr. Williams Co.	38°31'20"N, 77°17'22"W	NICC 0261	KF908076	KF908136	KF908060	KF908159	KF908180	KF908209
			Pr. Williams Co.	38°31'20"N, 77°17'22"W	NICC 0262	KF908077	KF908137		KF908160	KF908181	KF908210
			Pr. Williams Co.	38°31'20"N, 77°17'22"W	NICC 0263	KF908078	KF908138	KF908061		KF908182	
			Pr. Williams Co.	38°31'20"N, 77°17'22"W	NICC 0264	KF908079	KF908139	KF908062		KF908183	KF908211
		Washington	Snohomish Co.	47°54'46"N, 122°05'53"W	NICC 0270	KF908083	KF908145			KF908187	KF908217
			Snohomish Co.	47°54'46"N, 122°05'53"W	NICC 0271	KF908084	KF908146		KF908165	KF908188	KF908218
			Snohomish Co.	47°54'46"N, 122°05'53"W	NICC 0272	KF908085	KF908147			KF908189	KF908219
			Snohomish Co.	47°54'46"N, 122°05'53"W	NICC 0273	KF908086	KF908148			KF908190	KF908220
			Snohomish Co.	47°54'46"N, 122°05'53"W	NICC 0274	KF908087	KF908149			KF908191	KF908212

Species	continent/ country	country/ state	city/county	latitude/longitude	voucher no.	GenBank accession no.					
						COI	COII	16S	cyt *b*	ITS2	28S
		Wyoming	Park Co.	44°31'52"N, 108°57'40"W	NICC 0255	KF908071	KF908130			KF908174	
			Park Co.	44°31'52"N, 108°57'40"W	NICC 0256		KF908131	KF908058		KF908175	
			Park Co.	44°31'52"N, 108°57'40"W	NICC 0257	KF908072	KF908132		KF908156	KF908176	
			Park Co.	44°31'52"N, 108°57'40"W	NICC 0258	KF908073	KF908133			KF908177	KF908207
			Park Co.	44°31'52"N, 108°57'40"W	NICC 0259	KF908074	KF908134	KF908059	KF908157	KF908178	KF908208
Protophormia terraenovae	Europe	Belgium	Andrimont	50°36'36''N, 5°54'36''E	NICC 0030	KF908113					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0095	KF908115					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0096	KF908116					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0336	KF908117					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0337	KF908118					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0338	KF908119					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0339	KF908120					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0340	KF908121					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0341	KF908122					
			Andrimont	50°36'36''N, 5°54'36''E	NICC 0342	KF908123					
			Auderghem	50°49'05''N, 4°24'41''E	NICC 0033	KF908114					
			Auderghem	50°49'05''N, 4°24'41''E	NICC 0358	KF908124					

Appendix 2

Text file with the alignments for all the gene fragments studied. (doi: 10.3897/zookeys.365.6202.app2) File format: Text file (txt).

Explanation note: Text file with the alignments (fasta format) for all the gene fragments studied (COI, COII, cyt *b*, 16S (251 bp), 16S (350 bp), ITS2 and 28S, respectively).

Citation: Jordaens K, Sonet G, Braet Y, De Meyer M, Backeljau T, Goovaerts F, Bourguignon L, Desmyter S (2013) DNA barcoding and the differentiation between North American and West European *Phormia regina* (Diptera, Calliphoridae, Chrysomyinae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 149–174. doi: 10.3897/zookeys.365.6202 Text file with the alignments for all the gene fragments studied. doi: 10.3897/zookeys.365.6202.app2

ZooKeys 365: 175–196 (2013)
doi: 10.3897/zookeys.365.5879
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

DNA barcodes identify Central-Asian *Colias* butterflies (Lepidoptera, Pieridae)

Juha Laiho[1], Gunilla Ståhls[2]

1 *Persövägen 148, FI-10600 Ekenäs, Finland* **2** *Finnish Museum of Natural History, Zoological museum, PO Box 17, FI-00014 University of Helsinki, Finland*

Corresponding author: *Gunilla Ståhls* (gunilla.stahls@helsinki.fi)

Academic editor: *T. Backeljau* | Received 1 July 2013 | Accepted 7 September 2013 | Published 30 December 2013

Citation: Laiho J, Ståhls G (2013) DNA barcodes identify Central-Asian *Colias* butterflies (Lepidoptera, Pieridae). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 175–196. doi: 10.3897/zookeys.365.5879

Abstract

A majority of the known *Colias* species (Lepidoptera: Pieridae, Coliadinae) occur in the mountainous regions of Central-Asia, vast areas that are hard to access, rendering the knowledge of many species limited due to the lack of extensive sampling. Two gene regions, the mitochondrial COI 'barcode' region and the nuclear ribosomal protein RpS2 gene region were used for exploring the utility of these DNA markers for species identification. A comprehensive sampling of COI barcodes for Central-Asian *Colias* butterflies showed that the barcodes facilitated identification of most of the included species. Phylogenetic reconstruction based on parsimony and Neighbour-Joining recovered most species as monophyletic entities. For the RpS2 gene region species-specific sequences were registered for some of the included *Colias* spp. Nevertheless, this gene region was not deemed useful as additional molecular 'barcode'. A parsimony analysis of the combined COI and RpS2 data did not support the current subgeneric classification based on morphological characteristics.

Keywords

Barcoding, COI, *Colias*, Central-Asia, RpS2

Introduction

The use of a standardized gene region, i.e. a 650 bp fragment of the 5'-region of the mitochondrial cytochrome *c* oxidase subunit I (hereafter COI), as a DNA barcode (Hebert et al. 2003), to facilitate identification of biological specimens, as well as for calling attention to possible new species has generated a steadily increasing number of DNA barcoding studies of invertebrates (Taylor and Harris 2012), and particularly of Lepidoptera (see www.lepbarcoding.org). While the utility of DNA barcoding as an investigative tool has gained much support, there still remain a number of problems related to the use of a single DNA sequence as a taxon barcode. Several studies on Lepidoptera have shown that species may be polymorphic and/or share haplotypes (Nice et al. 2002, Wahlberg et al. 2003, Elias et al. 2007, Schmidt and Sperling 2008), so that identifications may become less reliable. Additionally, it has been shown that incomplete lineage sorting or mitochondrial introgression could obscure the delimitation of closely related taxa (Tautz et al. 2003, Zakharov et al. 2009). Using one or a few specimens as representatives of a species indeed provides us with little information about their intraspecific variation, particularly for widely distributed species (e.g. Funk and Omland 2003, Seberg et al. 2003, Sperling 2003).

The genus *Colias*

The butterfly genus *Colias* Fabricius, 1807 is a genus of the family Pieridae (subfamily Coliadinae), comprising about 85 species. Most of its species have a limited distribution in the Arctic and Alpine regions of the Holarctic realm, but two species occur in the Afrotropical and seven are known from the Neotropical regions (Verhulst 2000). A few species are widely distributed and common, such as the Palaearctic *C. erate* (Esper, 1805) and *C. croceus* (Geoffroy, 1785), and the Nearctic *C. eurytheme* Boisduval, 1852 and *C. philodice* Godart, 1819. As a consequence, these taxa are frequently used in ethological, ecological and genetic research (e.g. Pollock et al. 1998, Wang and Porter 2004, Porter and Levin 2010). *Colias erate* and *C. croceus* are a species pair where only typical specimens can be reliably distinguished morphologically, and members of these species are known to frequently hybridize (e.g. Dinca et al. 2011 and references therein). Lukhtanov et al. (2009) indicated that mitochondrial introgression was a likely explanation for the shared barcodes they registered between these sympatric taxa. The Nearctic taxa *C. eurytheme* and *C. philodice* are broadly sympatric sister species that hybridize frequently and that likely share a significant portion of their genomes through introgression (e.g. Wang and Porter 2004, Porter and Levin 2010). Verhulst (2000) illustrated hybrid individuals of six species of *Colias* from the Palaearctic region, including *C. croceus*.

The Central-Asian mountainous regions harbour nearly half of all *Colias* species. The distribution, ecology and taxonomy are still incompletely documented for most of these species, mainly due to their remote occurrences (Verhulst 2000). Central-Asian *Colias* species occurring in remote mountainous areas that are hard to access have been

far less studied than their North American or European congeners. An important part of the older material that exists in museum collections worldwide (e.g. from Tibet) originates from early collecting expeditions in the late 19th and early 20th centuries. Important material was, however, also collected within the former Soviet Union during 20th century. Fieldwork in Central-Asia has subsequently become less complicated, and thus new material is again available for research. As a result of this, new species such as *Colias aegidii* Verhulst, 1990 and *Colias adelaidae* Verhulst, 1991, have been described, as well as a number of new subspecies. Despite an increasing research effort on Central-Asian *Colias* species there are as yet no published studies on their phylogenetic relationships.

The first contribution to the species classification of *Colias* was given by Berger (1986), who used a few morphological characters to establish a comprehensive subgeneric classification, comprising the subgenera *Colias* Fabricius, 1807, *Neocolias* Berger, 1986, *Eucolias* Berger, 1986, *Eriocolias* Watson, 1895, *Palaeocolias* Berger, 1986, *Similicolias* Berger, 1986, *Scalidoneura* Butler, 1869 and *Paracolias* Berger, 1986. Later, Ferris (1993) used 84, mainly morphological, characters to reconstruct a phylogeny of all North American *Colias* species known at that time, which was the first species phylogeny within the genus *Colias*. The first contribution to the knowledge of the molecular phylogenetic relationships of the North American *Colias* species was made by Pollock et al. (1998), who studied a number of *Colias* species using a 333 bp sequence fragment of the mtDNA COI gene. They found some small differences between species classified in the subgenera *Neocolias* and *Eriocolias*, thus supporting Berger's (1986) separation of *Neocolias* from *Eriocolias*. Pollock et al. (1998) also noted that even though *Colias* is a speciose genus, this was not mirrored in the COI sequence diversity. Wheat and Watt (2008) studied the molecular phylogenetic relationships of North American *Colias* taxa using mitochondrial gene sequences (ribosomal 12S and 16S rRNA, Leu2 and Val tRNA and COI + II). Their results showed that the COI sequences only allowed identification of some of the taxa supported by the full data set used in their study. The results of their study further suggested that species radiations within *Colias* are comparatively young as compared with those of related pierid butterflies, since molecular divergences among species were small. Based on molecular data Brunton (1998) studied the phylogenetic relationships of the 12 *Colias* species occurring in Europe. He recovered three monophyletic groups largely corresponding to geographical distributions. He concluded that the Scandinavian species appeared to be the oldest in Europe, sharing a common ancestor with *Colias* species from the USA. According to Brunton (1998) the European *Colias* species radiated from Scandinavia to the rest of Europe forming an eastern clade and a western clade. As with Pollock et al. (1998), the results did not agree with Berger's (1986) subgeneric classification.

The aim of the present study was to test the usefulness of COI barcodes for species identification of a broad representation of Central-Asian *Colias* species, including nine *Colias* species overlapping with Lukhtanov et al.'s (2009) study, and 19 species not previously barcoded. In addition, we wanted to elucidate the informativeness of the RpS2 gene region that Wahlberg and Wheat (2008) found informative for lepidopter-

an phylogenetic relationships. We tested the nuclear ribosomal protein gene RpS2 as a potential complementary barcode region for *Colias* and for use in a combined analysis with COI for testing the current subgeneric classification of the species in the present study. We also contrasted our COI barcodes against a larger set of COI barcodes of *Colias* taxa available from GenBank (GB).

Materials and methods

Study area and taxon sampling

This study includes material from the mountain regions of Kirgizistan, Tadzhikistan, northern Afghanistan, northern Pakistan and India (e.g. mountain ranges Tian Shan, Hindu Kush, Karakorum, Himalaya) and the mountain regions in the Chinese provinces Qinghai, Gansu, Sichuan, Yunnan and the autonomous regions Tibet and Xinjiang Uygur. The *Colias* fauna of these Central-Asian regions comprises about 34 species (Verhulst 2000) while the species number for Central Asia in broad sense is over 40 species.

The taxon sampling aimed to cover as many of the *Colias* species from this area as possible. Additionally, a few *Colias* species occurring in adjacent territories (e.g. Buryatia) were also available for molecular study. Whenever possible, several individuals of each species were analysed to assess intraspecific variation. The available specimens used for molecular study consisted of a total of 56 adult specimens covering 27 species of Central-Asian *Colias* and two *Colias* species from adjacent territories (Table 1). The specimens are preserved as DNA voucher specimens and labelled accordingly, to be deposited in the collections of the Zoological Museum of Finnish Museum of Natural History, Helsinki, Finland (MZH) (DNA voucher specimens MZH_JL1-JL71). Species identifications were verified by JL based on easily recognizable diagnostic characters using the monograph by Verhulst (2000), while the taxonomy is according to Grieshuber and Lamas (2007). Additionally, we used 35 COI barcode sequences (17 species) of Palaearctic *Colias* species obtained from GB, as listed in Table 2.

Laboratory methods

Total genomic DNA was extracted form 2-5 legs of dried, pinned butterfly specimens using NucleoSpin® Tissue Kit (Machery-Nagel), according to manufacturer's protocols, and resuspended in 50 µl ultrapure water.

The primer pair LCO-1490 (5'-GGTCAACAAATCATAAAGATATTGG-3') and HCO-2198 (5'-TAAACTTCAGGGTGACCAAAAAATCA-3') (Folmer et al. 1994) was used to amplify a ca. 650 bp fragment of the mitochondrial COI gene. The polymerase chain reactions (PCR) were done under the following parameters: initial heating 95 °C for 2 min, following 30 cycles of 94 °C for 30 s, 49 °C for 30 s and 72 °C for 2 min, followed by a final extension of 72 °C for 7 min. The primer pair RpS2

Table 1. List of specimens used for molecular analyses including GenBank accession numbers.

Species	Sex	Locality and date	Lab code	COI accession number	RpS2 accession number
subgenus *Colias* Fabricius, 1807					
Colias hyale (Linnaeus, 1758) *irkutskana* Stauder, 1923	male	Russia, SW Transbaikalia, Buryatia, Selenga river district, Gusinoye Ozero village env., steppe rivulet valley, 7.6.2003	MZH_JL35	HE775142	HE775198
Colias hyale (Linnaeus, 1758) *irkutskana* Stauder, 1923	male	Russia, SW Transbaikalia, Buryatia, Selenga river district, Gusinoye Ozero village env., steppe rivulet valley, 7.6.2003	MZH_JL44	HE775143	HE775199
subgenus *Eriocolias* Berger, 1986					
Colias adelaidae adelaidae Verhulst, 1991	male	China, Gansu, Xia-He, 3400 m, 35°11'N, 102°31'E, 25.6.2004	MZH_JL61	HE775187	HE775243
Colias alpherakii alpherakii Staudinger, 1882	female	Kyrgyzstan, Alai mts., 4 km SE Tengizbai pass, 3400 m, 3.7.2001	MZH_JL37	HE775169	HE775225
Colias alpherakii alpherakii Staudinger, 1882	female	Kyrgyzstan, Alai mts., 4 km SE Tengizbai pass, 3400 m, 3.7.2001	MZH_JL51	HE775180	HE775236
Colias berylla berylla Fawcett, 1904	male	China, S Tibet, Himalaya Mts., Lablungla pass, 4800 m, 18–22.7.2001	MZH_JL48	HE775178	HE775234
Colias berylla berylla Fawcett, 1904	male	China, Tibet, Lhodak, 4600 m, 15.7.2002	MZH_JL55	HE775182	HE775238
Colias christophi christophi Grum Grshimailo, 1885	female	Tadjikistan, Turkestanskyi Mts., Kumbel pass, 3000 m, July 2002	MZH_JL45	HE775175	HE775231
Colias christophi helialaica Schulte, 1988	male	Kyrgyzstan, Alai Mts., W end of Tengizbai pass, 3700 m, 5–6.7.2001	MZH_JL67	HE775192	HE775246
Colias cocandica cocandica Erschoff, 1874	male	Kyrgyzstan, Suusamyr Mt. r., Alabel pass, 3200 m, 10.7.2002	MZH_JL43	HE775174	HE775230
Colias cocandica hinducucica Verity, 1911	male	Tajikistan, E Pamir, Ak-Buura Mts., 4250 m,14–15.7.2003	MZH_JL34	HE775168	HE775224
Colias cocandica pljushtchi Verhulst, 2000	male	Kyrgyzstan, Sary Dzhaz riv. bas., Kaindy-Ketta mts., Tashkoro village, 3000 m 10.7.2003	MZH_JL19	HE775160	HE775216
Colias eogene C. et R. Felder, [1865] *elissa* Grum Grshimailo, 1890	male	Kyrgyzstan, W end of Tengizbai pass, 3700 m, 5–6.7.2001	MZH_JL1	HE775144	HE775200
Colias eogene C. et R. Felder, [1865] *elissa* Grum Grshimailo, 1890	male	Kyrgyzstan, W end of Tengizbai pass, 3700 m, 5–6.7.2001	MZH_JL40	HE775171	HE775227
Colias fieldii Ménétriés, 1855 *chinensis* Verity, 1909	male	China, Sichuan, Zhangia, 3000 m, 32°47'N, 103°36'E, 6.6.2002	MZH_JL50	HE775179	HE775235
Colias fieldii Ménétriés, 1855 *chinensis* Verity, 1909	female	China, Gansu, Shin-Long-Shan, 2800 m, 35°48'N, 103°59'E, 29.6.2004	MZH_JL60	HE775186	HE775242
Colias grumi grumi Alphéraky, 1897	female	China, Gansu, Altun Shan, road from Aksay to Danjing pass, 2500–2800 m, 22–23.7.2002	MZH_JL54	HE775197	-

Species	Sex	Locality and date	Lab code	COI accession number	RpS2 accession number
Colias heos heos (Herbst, 1792)	male	Russia, SW Transbaikalia, Buryatia, Selenga river district, Gusinoye Ozero village env., steppe rivulet valley, 1.7.2003	MZH_JL39	HE775170	HE775226
Colias heos heos (Herbst, 1792)	male	Russia, SW Transbaikalia, Buryatia, Selenga river district, Gusinoye Ozero village env., steppe rivulet valley, 1.7.2003	MZH_JL46	HE775176	HE775232
Colias lada lada Grum Grshimailo, 1891	male	China, Sichuan, Maningano surr., 31°56'N, 99°12'E, 4500 m, 15.6.2002	MZH_JL7	HE775150	HE775206
Colias lada lada Grum Grshimailo, 1891	male	China, Sichuan, Maningano surr., 31°56'N, 99°12'E, 4500 m, 15.6.2002	MZH_JL27	HE775165	HE775221
Colias ladakensis Felder, 1865 *seitzi* Bollow, 1939	male	China, SW Tibet, Himalaya Mts., 100km W Paryang, 4650–5000 m, 13.6.2004	MZH_JL4	HE775147	HE775203
Colias ladakensis Felder, 1865 *seitzi* Bollow, 1939	male	China, SW Tibet, Himalaya Mts., 100km W Paryang, 4650–5000 m, 13.6.2004	MZH_JL57	HE775183	HE775239
Colias marcopolo marcopolo Grum Grshimailo, 1888	male	Tadjikistan, E Pamir, Dunkeldyk Lake, 4400 m, 25.7.2003	MZH_JL30	HE775166	HE775222
Colias marcopolo marcopolo Grum Grshimailo, 1888	male	Tadjikistan, E Pamir, Dunkeldyk Lake, 4400 m, 25.7.2003	MZH_JL33	HE775167	HE775223
Colias marcopolo marcopolo Grum Grshimailo, 1888	male	Tadjikistan, E Pamir, Dunkeldyk Lake, 4400 m, 25.7.2003	MZH_JL41	HE775172	HE775228
Colias montium montium Oberthür, 1886	male	China, Sichuan, Maningano surr., 31°55'N, 99°12'E, 4000 m, 9–18.6.2004	MZH_JL59	HE775185	HE775241
Colias nebulosa Oberthür, 1894 *sungpani* Bang-Haas, 1927	male	China, Sichuan, Maningano surr., 31°56'N, 99°12'E, 4500 m, 15.6.2002	MZH_JL9	HE775152	HE775208
Colias nebulosa Oberthür, 1894 *sungpani* Bang-Haas, 1927	male	China, Sichuan, Maningano surr., 31°56'N, 99°12'E, 4500 m, 15.6.2002	MZH_JL24	HE775162	HE775218
Colias nebulosa Oberthür, 1894 *sungpani* Bang-Haas, 1927	male	China, Sichuan, Maningano surr., 31°56'N, 99°12'E, 4500 m, 15.6.2002	MZH_JL26	HE775164	HE775220
Colias nina Fawcett, 1904 *hingstoni* Riley, 1923	male	China, SW Tibet, Himalaya Mts., 60 km S Saga, 4600–5000 m, 7–8.6.2004	MZH_JL53	HE775181	HE775237
Colias nina Fawcett, 1904 *hingstoni* Riley, 1923	male	China, SW Tibet, Himalaya Mts., Lablongla pass, 4800 m, 5.6.2004	MZH_JL58	HE775184	HE775240
Colias regia regia Grum Grshimailo, 1887	male	Kyrgyzstan, Kaindy-Ketta Mt. r., Kumar pass, 3200 m, 12.7.2003	MZH_JL8	HE775151	HE775207
Colias regia regia Grum Grshimailo, 1887	male	Kyrgyzstan, Kaindy-Ketta Mt. r., Kumar pass, 3200 m, 12.7.2003	MZH_JL42	HE775173	HE775229
Colias romanovi romanovi Grum Grshimailo, 1885	male	Kyrgyzstan, Alai mts., 4 km SE Tengizbai pass, 3400 m, 7–8.7.2001	MZH_JL3	HE775146	HE775202

Species	Sex	Locality and date	Lab code	COI accession number	RpS2 accession number
Colias romanovi romanovi Grum Grshimailo, 1885	male	Kyrgyzstan, Alai mts., 4 km SE Tengizbai pass, 3400 m, 7–8.7.2001	MZH_JL47	HE775177	HE775233
Colias sieversi sieversi Grum Grshimailo, 1887	male	Tadjikistan, Peter I Mts., Ganishob, 2400 m, 17.6.2004	MZH_JL70	HE775195	-
Colias sifanica sifanica Grum Grshimailo, 1891	male	China, Gansu, Xia-He, 3400 m, 35°11'N, 102°31'E, 25.6.2004	MZH_JL11	HE775154	HE775210
Colias sifanica sifanica Grum Grshimailo, 1891	male	China, Gansu, Xia-He, 3400 m, 35°11'N, 102°31'E, 25.6.2004	MZH_JL64	HE775189	HE775245
Colias staudingeri Alphéraky, 1881 *pamira* Grum Grshimailo, 1890	male	Kyrgyzstan, Zaalaisky (Transalai) Mts., Altyn Dara river, 3000 m, 25.7.2000	MZH_JL2	HE775145	HE775201
Colias staudingeri Alphéraky, 1881 *pamira* Grum Grshimailo, 1890	male	Kyrgyzstan, Zaalaisky (Transalai) Mts., Altyn Dara river, 3000 m, 25.7.2000	MZH_JL13	HE775156	HE775212
Colias staudingeri Alphéraky, 1881 *pamira* Grum Grshimailo, 1890	male	Kyrgyzstan, Zaalaisky (Transalai) Mts., Altyn Dara river, 3000 m, 25.7.2000	MZH_JL23	HE775161	HE775217
Colias stoliczkana stoliczkana Moore, 1882	male	India, Jammu Kashmir, Ladakh Range, Markha Valley, Ganda Pass, 4600 m, 12.7.2001	MZH_JL15	HE775158	HE775214
Colias thisoa Ménétriés, 1832 *aeolides* Grum Grshimailo, 1890	male	Kyrgyzstan, Sary Dzhaz riv. bas., Kaindy-Ketta mts., Tashkoro village, 3000 m, 10.7.2003	MZH_JL10	HE775153	HE775209
Colias thisoa Ménétriés, 1832 *aeolides* Grum Grshimailo, 1890	female	Kyrgyzstan, Sary Dzhaz riv. bas., Kaindy-Ketta mts., Tashkoro village, 3000 m, 10.7.2003	MZH_JL17	HE775159	HE775215
Colias thisoa Ménétriés, 1832 *aeolides* Grum Grshimailo, 1890	female	Kyrgyzstan, Sary Dzhaz riv. bas., Kaindy-Ketta mts., Tashkoro village, 3000 m, 10.7.2003	MZH_JL25	HE775163	HE775219
Colias thrasibulus thrasibulus Fruhstorfer, 1910	male	China, W Tibet, Mandhata Mt., 4900 m, 15–16.7.2003	MZH_JL14	HE775157	HE775213
Colias tibetana tibetana Riley, 1922	male	China, Tibet, Himalaya Mts., Nyalam, 4200 m, 8.7.2003	MZH_JL6	HE775149	HE775205
Colias tibetana tibetana Riley, 1922	male	China, SW Tibet, Himalaya Mts., Nyalam, 3700–4200 m, 28–30.6.2004	MZH_JL63	HE775188	HE775244
Colias wanda wanda Grum Grshimailo, 1907	male	China, Qinghai, 20km NW of Zhidoi City, 4700–5000 m, 16.7.2000	MZH_JL66	HE775191	-
Colias wanda wanda Grum Grshimailo, 1907	male	China, S. Tibet, Cona, 4500–4700 m, 24–25.6.2004	MZH_JL69	HE775194	-
Colias wiskotti Staudinger, 1882 *draconis* Grum Grshimailo, 1891	male	Uzbekistan, Chandalas Mts., Chakmksh village, 2600 m, 27.6.2004	MZH_JL71	HE775196	-
Colias wiskotti Staudinger, 1882 *hofmannorum* Eckweiler, 2000	male	Iran, Khorasan, 75km SE of Birjand, 2200 m, 18–20.5.2002	MZH_JL68	HE775193	-
Colias wiskotti Staudinger, 1882 *separata* Grum Grshimailo, 1888	male	Kyrgyzstan, Alai mts., 4km SE Tengizbai pass, 3400 m, 3.7.2001	MZH_JL65	HE775190	-
subgenus *Eucolias* Berger, 1986					

Species	Sex	Locality and date	Lab code	COI accession number	RpS2 accession number
Colias tyche tyche (de Boeber, 1812)	male	Russia, East Siberia, Lake Baikal, Khamar-Daban Mts., Slyudyanka river, taiga, 800 m, 14.6.2003	MZH_JL5	HE775148	HE775204
Colias tyche tyche (de Boeber, 1812)	male	Russia, East Sayan, Buryatia, Mondy env., Huruma river, 1500 m, 6.6.2002	MZH_JL12	HE775155	HE775211

Table 2. List of *Colias* GenBank samples of the COI barcode used in this study.

Species	GenBank accession number
Colias alpherakii	FJ663407
Colias christophi	FJ663409
Colias chrysotheme elena	FJ663410
Colias chrysotheme elena	FJ663411
Colias croceus	EF457737
Colias croceus	FJ663412
Colias croceus	GU688507
Colias croceus	HQ004279
Colias croceus	HQ004282
Colias eogene	FJ663415
Colias eogene	FJ663416
Colias erate amdensis	EF457736
Colias erate poliographus	EF457735
Colias erate poliographus	EU583852
Colias erate poliographus	GU372561
Colias fieldii	EF584859
Colias hyale	FJ663418
Colias hyale	FJ663421
Colias hyale	HQ004297
Colias hyperborea	EF457739
Colias marcopolo	FJ663422
Colias marcopolo	FJ663423
Colias myrmidone	HQ004303
Colias phicomone	HM393178
Colias regia	FJ663427
Colias tamerlana mongola	FJ663424
Colias tamerlana mongola	FJ663425
Colias tamerlana mongola	FJ663426
Colias thisoa thisoa	FJ663429
Colias tyche	FJ663430
Colias wiskotti chrysoptera	FJ663431
Colias wiskotti chrysoptera	FJ663432
Colias wiskotti chrysoptera	FJ663433
Colias wiskotti wiskotti	FJ663435
Colias wiskotti wiskotti	FJ663436

nF (5'-ATCWCGYGGTGGYGATAGAG-3') and RpS2 nR (5'-ATGRGGCTTKCCRATCTTGT-3') (Wahlberg and Wheat 2008) was used to amplify a ca. 400 bp fragment of the nuclear RpS2 gene. The PCR were carried out following the PCR cycling profile described in Wahlberg and Wheat (2008): initial heating 95 °C for 7 min, 40 cycles of 95 °C for 30 s, 50 °C for 30 s, 72 °C for 2 min, and a final extension period of 72 °C for 10 min. Sequencing of the double-stranded PCR product was carried out on an ABI PRISM® 377 Automated Sequencer (Applied Biosystems) following manufacturer's recommendations. All PCR primers were used for sequencing. Sequences were inspected and edited using Sequence Navigator® (Applied Biosystems).

Sequence analysis

We analysed and clustered our sequence data using parsimony and Neighbour-Joining (NJ) of K2P-distances. We used parsimony and NJ for our newly generated COI sequence dataset, NJ for RpS2 sequences, parsimony for the concatenated COI and RpS2 sequences, and, finally, NJ for the combined COI sequences generated in this study and those in GB. All trees were rooted using *Papilio glaucus* (family Papilionidae) and *Aporia crategi* (Pieridae, subfamily Pierinae) as outgroup taxa.

Parsimony analysis was performed using NONA (Goloboff 1999) and spawn with the aid of Winclada (Nixon 2002), using a heuristic search algorithm with 1000 random addition replicates (mult*1000), holding 10 trees per round (hold/10), max trees set to 10,000 and applying TBR branch swapping. All base positions were treated as equally weighted characters. Nodal support was assessed with bootstrap resampling (1000 replicates) using Winclada (Nixon 2002). MEGA5 (Tamura et al. 2011) was used for NJ clustering using 1000 bootstrap replicates. The Kimura 2-parameter model was used for NJ clustering of the COI sequences, while the Tamura-Nei model with gamma distributed rates was chosen for the RpS2 sequences.

Results

Sequences

We obtained a 643 bp COI barcode for 56 *Colias* specimens, and a 409 bp fragment of RpS2 was obtained for 49 specimens (Table 1). A+T content of the COI sequences was 69.22%, and of the RpS2 45.0%. There were 115 parsimony informative sites for COI and 39 for RpS2.

Uncorrected pairwise divergences between ingroup taxa ranged between 1.09 and 4.09% (mean 2.77%) for COI and 0.0–1.7% (mean 1.0%) for RpS2. GenBank accession numbers are given in Table 1. Intraspecific uncorrected distances were up to 1.09% (in *C. thisoa*) for COI, with specimens of most species differing by less than 4 nucleotide changes.

Identification: COI vs. RpS2

The parsimony analysis of the new COI sequences yielded four equally parsimonious trees (CI = 0.59, RI = 0.75) the strict consensus tree of which is presented in Figure 1. The NJ tree is presented in Figure 2.

The majority of the species could be identified with COI alone, as no COI haplotypes were shared between species. Both parsimony and NJ trees recovered 25 (out of 28) species as monophyletic groups (Figures 1–2). Neither *Colias cocandica*, nor *C. nebulosa* formed monophyletic entities, as their sequences were scattered over various parts of the trees. The two samples of *C. tyche* were not recovered as sister taxa, for sample MZH_JL5 appeared as sister taxon of *C. heos*. The overall topologies of the parsimony and NJ trees were identical, except for the placement of *C. thrasibulus*. Parsimony placed the taxon as sister to a clade of five taxa (Figure 1), while NJ placed it as sister to *C. romanovi* (Figure 2). The external morphology of *C. thrasibulus* is rather different from that of *C. romanovi*, while some similarities can be found between *C. thrasibulus* and *C. nina*, *C. ladakensis*, *C. tibetana* and *C. cocandica* (Figure 1). Only 17 of the 39 parsimony informative sites of RpS2 were variable among the 49 ingroup members. NJ only recovered few species as separate lineages due to the shallow divergences (Figure 3). The information content of this gene region is best interpreted as a character-based diagnostic table, as suggested by DeSalle et al. (2005). This gene region yielded species specific (diagnostic) haplotypes for 11 species out of 33 (Table 3).

Analysis of the concatenated COI + RpS2 data

The parsimony analysis of COI + RpS2 yielded nine trees of length 560 steps (CI = 0.63, RI = 0.72), the strict consensus tree of which is shown in Figure 4. *Colias cocandica*, *C. nebulosa* and *C. tyche* were not monophyletic and *C. thrasibulus* had the same position as in the COI cladogram (Figure 1).

Analysis of all the COI sequences

The strict consensus cladogram for all the available COI data resolved the taxa in the same positions as in the tree of the new COI sequences only. For ten species of the present study sequences were also available from GB. Sequences of most species clustered together as monophyletic entities, except for *C. nebulosa*, *C. cocandica*, *C. tyche* and *C. regia.* For *C. regia* the GB sequence (GB accession no FJ663427) did not cluster together with our sequences. The GB barcodes of *C. erate* and *C. croceus* were shared by these two taxa.

Neither the Himalayan and south Tibetan adjacent mountain *Colias* fauna (*berylla*, *ladakensis*, *nina*, *stoliczkana, thrasibulus*, *tibetana*), nor the east Tibetan,

Table 3. Species haplotypes for 17 variable positions of RpS2 for Central-Asian *Colias* species (RpS2 data matrix positions no 14, 152, 170, 176, 189, 191, 194, 195, 218, 284, 287, 302, 341, 353, 356, 365, 380).

Haplotype	**positions of RpS2**
MZH_JL35_hyale	TCCCCGGGTCCATTTTC
MZH_JL44_hyale	TCCCCGGGTCCATTTTC
MZH_JL02_staudingeri	TCCTCGAGTTCAAATCC
MZH_JL13_staudingeri	TCCTCGAGTTCAAATCC
MZH_JL23_staudingeri	TCCTCGAGTTCAAATCC
MZH_JL43_cocandica_cocandica	TCCCCGAGTTCAAATCC
MZH_JL41_marcopolo	TACCCGAGTTCAAAACC
MZH_JL30_marcopolo	TACCCGAGTTCAAAACC
MZH_JL07_lada	TCCCAAAAGTCGATTCC
MZH_JL27_lada	TCCCAAAAGTCGATTCC
MZH_JL25_thisoa	TCCCAAAAGTCGATTCC
MZH_JL10_thisoa	TCCCAAAAGTCGATTCC
MZH_JL17_thisoa	TCCCAAAAGTCGATTCC
MZH_JL05_tyche	TCCCAAAAGTCGATTCC
MZH_JL12_tyche	TCCCAAAAGTCGTTTCC
MZH_JL39_heos	TCCCAAAAGTCGATTCC
MZH_JL46_heos	TCCCAAAAGTCGATTCC
MZH_JL53_nina	TCCCAAAAGTCGATTCC
MZH_JL58_nina	CCCCCGAAGTCGATTCC
MZH_JL11_sifanica	TCCCCGAGGTCGWTTCC
MZH_JL64_sifanica	TCTCCGAGGTCGATTCC
MZH_JL57_ladakensis	TCCCCGAGGTCGATTCC
MZH_JL06_tibetana	TCCTCGAGGTTATTTCC
MZH_JL09_nebulosa	TCCTCGAGGTTATTTCC
MZH_JL26_nebulosa	TCCTCGAGGTTATTTCC
MZH_JL14_thrasibulus	TCCTCGAGGTTATTTCC
MZH_JL01_eogene	TCCTCGAGGTTATTTCT
MZH_JL04_ladakensis	TCTCCGAGGTTATTTCC
MZH_JL15_stoliczkana	TCTCCGAGGTTGTTTCT
MZH_JL19_cocandica_pljushtchi	TCCTCGAGTTCATTTCC
MZH_JL34_cocandica_hinducucia	TCCTCGAGTTCATTTCC
MZH_JL03_romanovi	TCCTCGAGTTCATTTCC
MZH_JL08_regia	TCCCCGAGTTCATTTCT
MZH_JL42_regia	TCCCCGAGTTCATTTCT
MZH_JL47_romanovi	CCCTCGAGTTCATTTCC
MZH_JL51_alpherakii	TCCCCGAGTTCATTTCC
MZH_JL37_alpherakii	CACCCGAGTTCATTTCC
MZH_JL67_christophi_christophi	TCCTCGAGTTCATTTCC
MZH_JL45_christophi_kali	TCCTCGAGTTCGTTTCC
MZH_JL40_eogene	TCCTCGAGGTTGTTTCT
MZH_JL24_nebulosa	TCCTCGAGGTCGTTTCC

Haplotype	positions of RpS2
MZH_JL59_montium	CCCTCGAGGTTGTTTCC
MZH_JL61_adelaidae	TCCTCGAGGTCGTTTCC
MZH_JL60_fieldii	TCCTCGAGGTTATTTCC
MZH_JL50_fieldii	TCCTCGAGGTTATTTCT
MZH_JL33_marcopolo	TCCCCGAGGTCATTACT
MZH_JL63_tibetana	TCCTCGAGGTTATWTCC
MZH_JL48_berylla	TCCCCGAGGTCGAATCC
MZH_JL55_berylla	TCCCCGAGGTCGAATCC

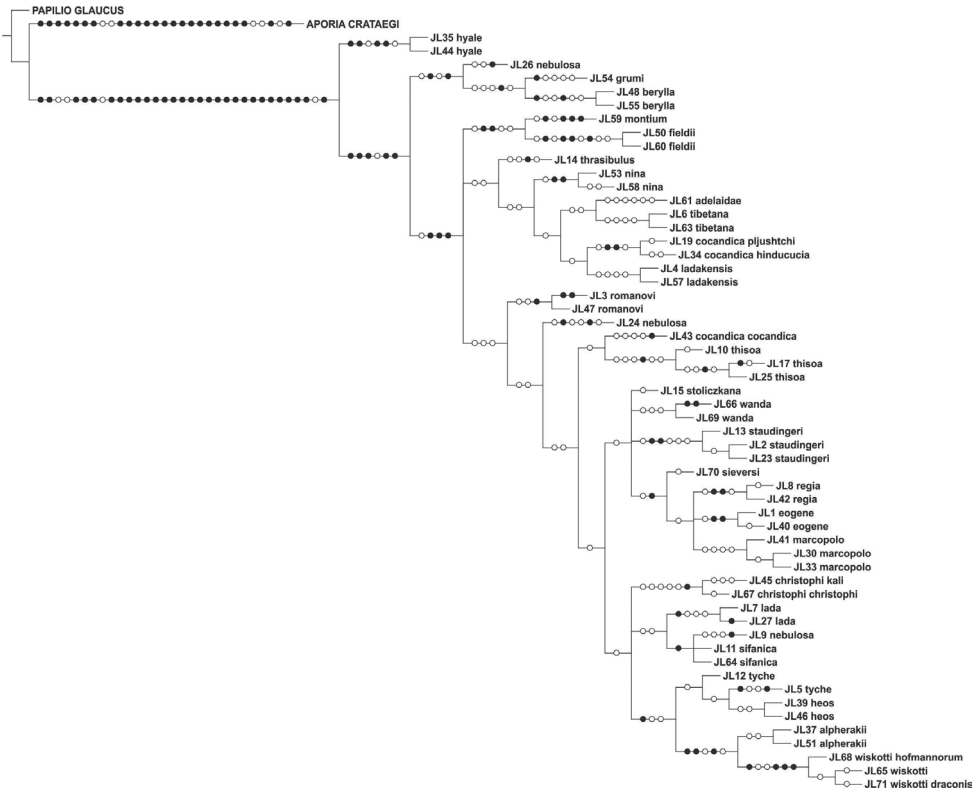


Figure 1. Strict consensus cladogram of *Colias* COI sequences obtained in this study.

Qinghai, Gansu and Sichuan species aggregates (*adelaidae, grumi*, *lada*, *montium*, *nebulosa, sifanica, wanda*) were resolved as species clusters similar to the Tian Shan, Pamir and Hindukush species.

Several COI haplotypes were noted for a few species, even among specimens obtained from the same locality (e.g. *C. staudingeri* and *C. thisoa*). Taxa not resolved as monophyletic clusters were the species *C. cocandica* and *C. nebulosa*. All the included subspecies of *C. cocandica* (*C. c. cocandica*, *C. c. pljutshtshi* and *C. c. hinducucia*) showed distinct COI sequences, with *cocandica cocandica* as most different.

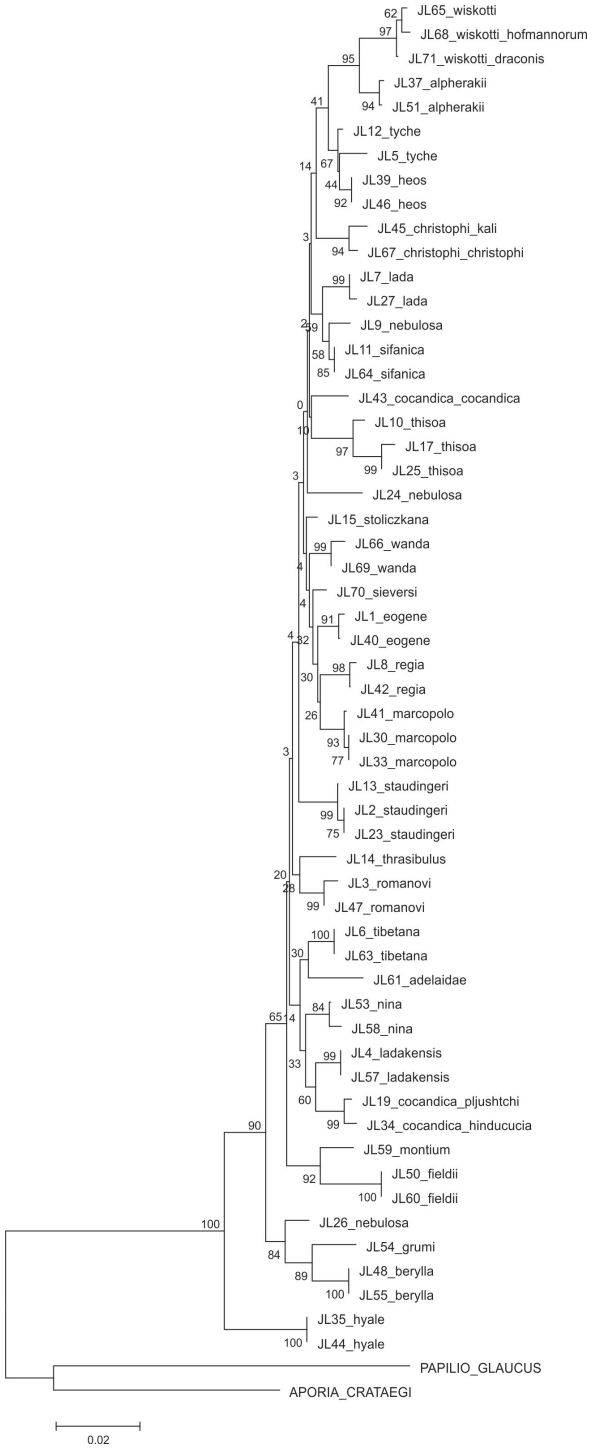


Figure 2. Neighbour-Joining tree using the K2P-parameter model for the COI sequences obtained in this study.

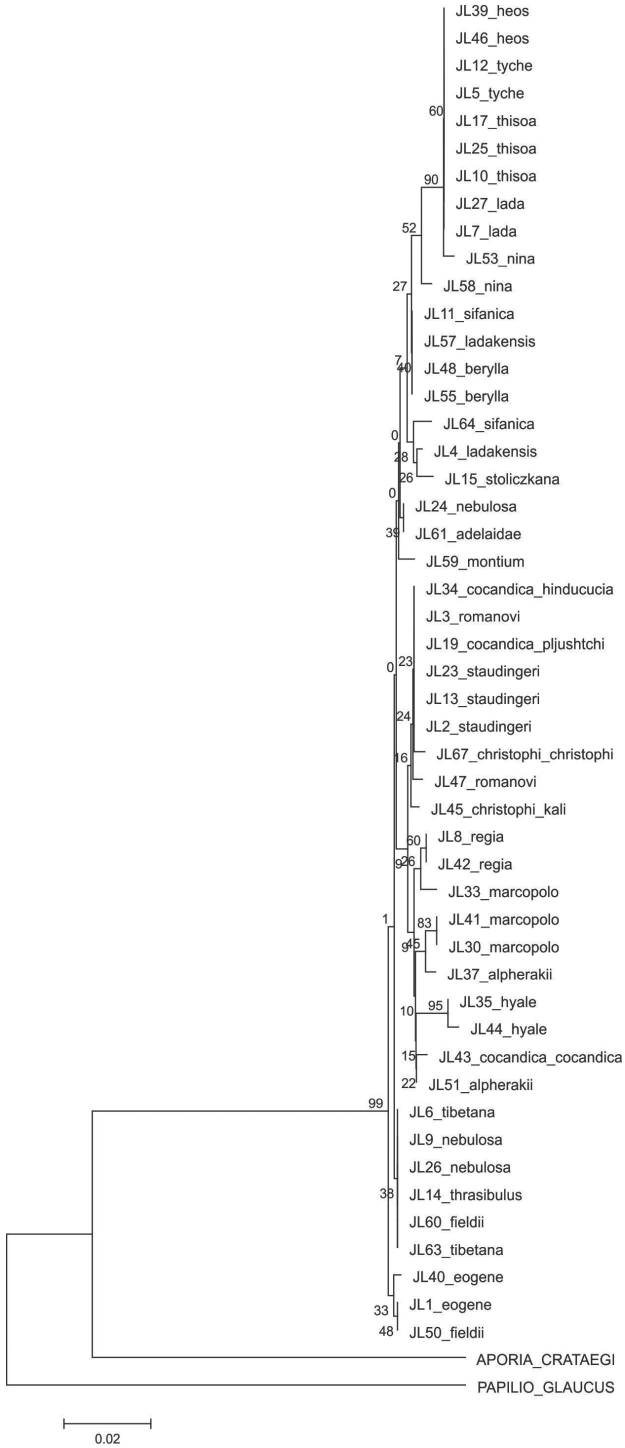


Figure 3. Neighbour-Joining tree using the Tamura-Nei model with gamma distributed rates for the RpS2 sequences.

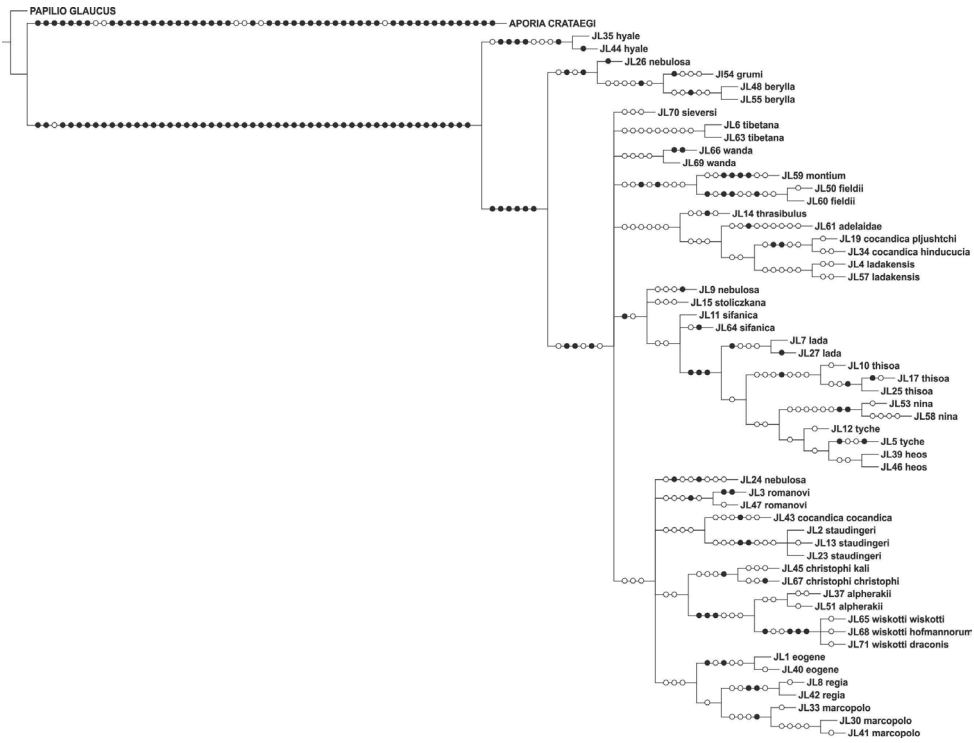


Figure 4. Strict consensus cladogram of the concatenated data set of COI + RpS2.

Discussion

Barcoding

Lukhtanov et al. (2009) tested the utility of COI barcodes for Central Asian butterflies by sampling specimens from a considerable geographical range. They observed that this substantially increased intraspecific variation reducing the interspecific divergences ("barcoding gap"), but that this did not hamper species identification. The present study shows that most *Colias* taxa form monophyletic entities that can be identified with COI data alone. The RpS2 gene region showed identical sequences in *cocandica pljutshtshi* and *cocandica hinducucia* (Table 3, Figure 3), differing by only three nucleotides from *cocandica cocandica*. Based on the molecular data the recognition of these subspecies is not or weakly supported.

The fact that the three *C. nebulosa* samples were scattered over different parts of the COI tree might be the result of a laboratory contamination due to carry over between samples. The *C. nebulosa* samples were collected on the same day and in the same place. *C. nebulosa* is morphologically distinct from other *Colias* species, excluding possible misidentification. The RpS2 data, however, could point to two morphologically cryptic species in sympatry (samples MZH_JL24 vs. MZH_JL9 and MZH_JL26), so that the different COI barcodes might represent numts, despite no apparent 'signs' (no

indels). This discrepancy between morphology and DNA sequence data emphasises the necessity to use multiple samples to detect this sort of challenging issues.

Even though *C. cocandica* and *C. nebulosa* did not form monophyletic groups our results show that COI barcodes are useful for (1) identifying Palaearctic and Central-Asian *Colias*, (2) pointing to a possible cryptic species, and (3) highlighting the necessity to further investigate the question on the subspecific rank of *C. cocandica cocandica*.

The utility of RpS2 as a species barcode for *Colias* spp. is clearly more limited, since e.g. *C. heos*, *C. lada*, *C. nina*, *C. thisoa* of the subgenus *Eriocolias* and *C. tyche* (subgenus *Eucolias*) have identical sequences (Table 3, Figure 3). Still, RpS2 yielded species specific (diagnostic) haplotypes for 11 species of the subgenus *Eriocolias* and for *C. hyale* (subgenus *Colias* s.str.).

Congruence with traditional classification: analysis of concatenated COI + RpS2

The strict consensus tree was more resolved than either of the trees resulting from separate analyses of the gene regions (Figure 4).

Although the concatenated data did not resolve the phylogenetic relationships among all *Colias* species, some observations can be made. The majority of the species confined to the adjacent Tian Shan, Pamir and Hindukush mountain ranges form a well supported clade. This includes *C. eogene*, *C. regia*, *C. romanovi*, *C. marcopolo*, *C. staudingeri*, *C. christophi*, *C. alpherakii* and *C. wiskotti*. Yet, *C. sieversi*, which also occurs in these mountain ranges (Peter I and Khozratishoh mountains), was not included in this clade. *C. sieversi* is morphologically most similar to *C. alpherakii*, thus showing another case of disagreement between morphological and DNA sequence data. *C. thisoa,* too, lives in the aforementioned mountain ranges, but it has a wider distribution, stretching from Turkey to the Altai Mountains. A third taxon, *C. c. cocandica*, is considered closely related to *C. tamerlana* (e.g. Verhulst 2000), a species occurring in southern Siberia and Mongolia. Thus, the origin of *C. thisoa* and *C. c. cocandica* may differ from that of the species confined to the Tian Shan, Pamir and Hindukush mountain range. One sample of *C. cocandica* (MZH_JL43) was placed within this "mountain" clade, while the other two samples appeared as sister taxa to the Himalayan species *C. ladakensis*. As with *C. sieversi*, our DNA data disagree with the morphological characters, but it should be noted that this clade is not well supported. Conversely, two morphologically similar Himalayan species, viz. *C. nina* and *C. ladakensis,* were assigned to different clades. In the COI + RpS2 tree they were placed in different, more encompassing species clusters (Figure 4), in the COI NJ tree they were joined with *C. c. pljutshtshi* and *C. c. hinducucia* (Figure 2), while the COI cladogram resolved these taxa together with *C. adelaidae*, *C. tibetana*, *C. c. pljutshtshi* and *C. c. hinducucia* (Figure 1).

The analyses did not support the monophyly of the subgenera *Eucolias* and *Eriocolias* sensu Berger (1986). The *Eucolias* species *C. tyche* was not resolved as a separate monophyletic lineage, but was resolved into *Eriocolias*. This is congruent with the results of Pollock et al. (1998) and Brunton (1998). Only the the subgenus *Colias*, here represented by *C. hyale*, is supported as a distinct lineage, placed as sister to all other *Colias* sp.

Barcodes of Palaearctic *Colias* spp.

The parsimony (Figure 5) and NJ analyses (Figure 6) of the larger matrix of Palaearctic COI barcodes (total COI) recovered the same species clusters, but some of the species show different placements (e.g. *C. thisoa*, *C. christophi*). This is not surprising as all internal nodes are very shallow. The samples of *C. tyche* and *C. hyperborea* show very low sequence difference, morphologically these taxa are different, and they largely share the same distribution area. An example of species that share the same distribution and that exhibit clear morphological similarities, and which as such were resolved as sister species in both analyses, includes *C. wiskotti* and *C. alpherakii*. Identification of Palaearctic *Colias* based on COI barcodes is in most cases possible, since shared haplotypes were recorded only for *C. erate* and *C. croceus*.

Intraspecific variation is notable between some of the recognized subspecies, both among our own samples and those downloaded from GB. The intraspecific variation can partly be explained by morphologically clearly distinct subspecies, such as those of *C. wiskotti,* or by specimens from widely different localities, such the different specimens of *C. hyale* (sample FJ663418 from Russia, FJ663421 from Kazakhstan, HQ004297 from Romania and MZH_JL35 and MZH_JL44 from SW Transbaikalia). However, notable intraspecific variation also occurs within populations, such as *C. thisoa aeolides* with all samples originating from the same locality and date, but the limited sampling prevents conclusions on the reasons for this. It is apparent that the understanding of intraspecific variability of the COI barcode for *Colias* is presently very limited.

The combined COI data of our sequences and sequences downloaded from GB include species belonging to one additional subgenus, *Neocolias*, represented by *C. myrmidone and C. erate*. Only the subgenus *Colias*, represented by *C. hyale*, is well supported as distinct lineage. Yet, one specimen of *C. hyale* (FJ663419) clustered together with *C. erate* (*Neocolias*) and *C. croceus* (*Eriocolias*). The other subgenera were not resolved as clades according to present classification, in agreement with our results for the combined analysis.

Our findings generally support COI as a species specific barcode for *Colias*, but we also highlight the necessity of including multiple individuals of species in molecular barcoding studies. Problematic 'cases' of widely divergent barcodes or conflicting morphological and molecular 'signals' are found in most if not all barcoding studies, and this study makes no exception.

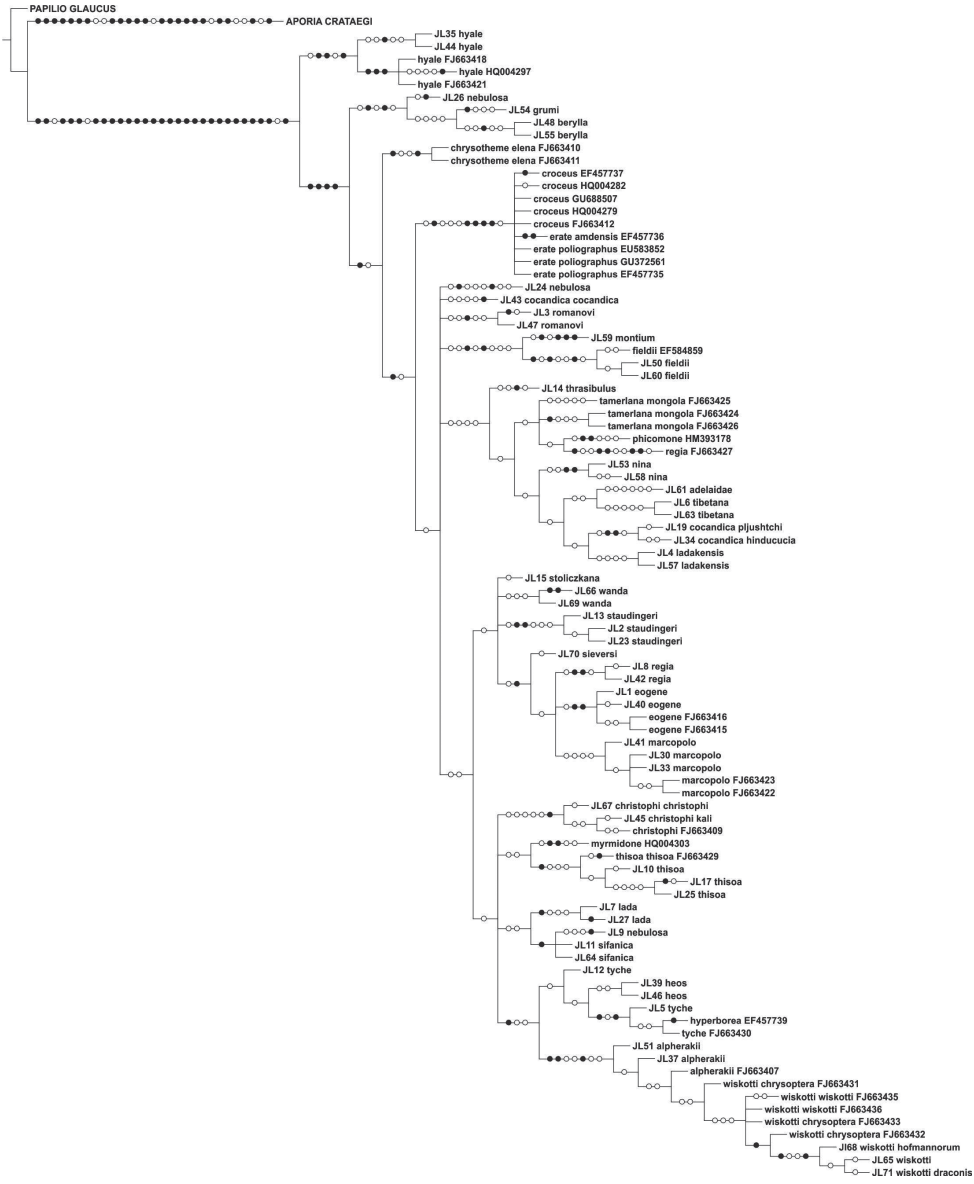


Figure 5. Strict consensus cladogram of COI sequences for Palaearctic *Colias* taxa.

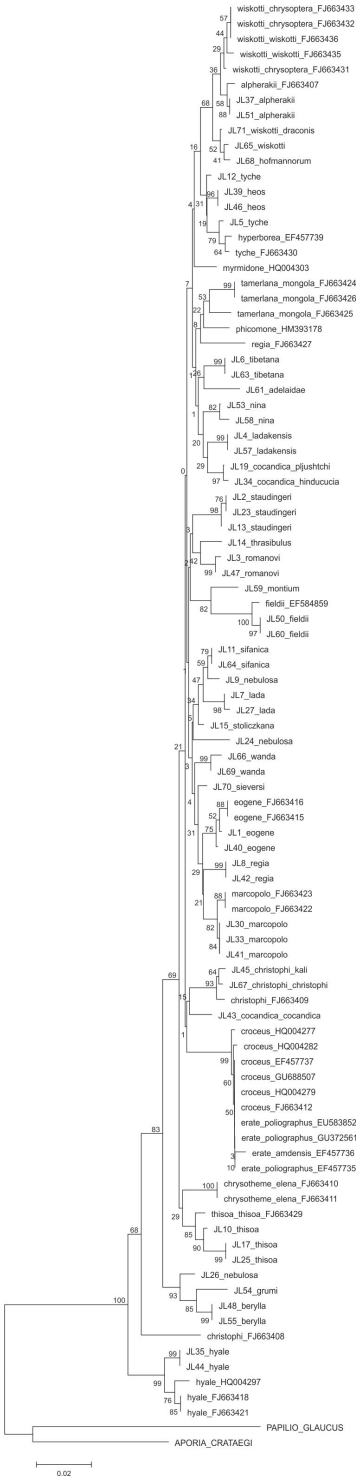


Figure 6. Neighbour-Joining tree using the K2P-model of COI sequences for Palaearctc *Colias* taxa.

Acknowledgements

JL thanks the Societas Entomologica Helsingforsiensis for support for the DNA work.

References

Back W, Miller MA, Opler PA (2011) Genetic, phenetic, and distributional relationships of nearctic *Euchloe* (Pieridae, Pierinae, Anthocharidini). Journal of the Lepidopterists' Society 65: 1–14.

Berger LA (1986) Systématique du genre *Colias* F. supplément Lambillionea 7–8.

Brunton CFA (1998) The evolution of ultraviolet patterns in European *Colias* butterflies (Lepidoptera: Pieridae): a phylogeny using mitochondrial DNA. Heredity 80: 611–616. doi: 10.1046/j.1365-2540.1998.00336.x

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2008) DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservacion Guanacaste, Costa Rica. Proceedings of the National Academy of Sciences of the USA 105: 6350–6355. doi: 10.1073/pnas.0712181105

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1905–1916.

Dinca V, Zakharov EV, Hebert PDN, Vila R (2011) Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. Proceedings of the Royal Society B 278: 347–355. doi: 10.1098/rspb.2010.1089

Elias M, Hill RI , Willmott KR, Dasmahapatra KK, Brower AVZ, Mallet J, Jiggins CD (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proceedings of the Royal Society B 274: 2881–2889. doi: 10.1098/rspb.2007.1035

Ferris CD (1993) Reassessment of the *Colias alexandra* group, the Legume-feeding species, and preliminary cladistic analysis of the North American *Colias* (Pieridae: Coliadinae). Bulletin of the Allyn Museum 138: 1–91.

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. Annual Review of Ecology and Systematics 34: 397–423. doi: 10.1146/annurev.ecolsys.34.011802.132421

Goloboff P (1999) NONA computer program. Published by the author, Tucuman, Argentina.

Grieshuber J, Lamas G (2007) A synonymic list of the genus *Colias* Fabricius, 1807. Mitteilungen der Münchner Entomologischen Gesellschaft 97: 131–171.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–322.

Linares MC, Soto-Calderón ID, Lees DC, Anthony NM (2009) High mitochondrial diversity in geographically widespread butterflies of Madagascar: a test of the DNA barcoding approach. Molecular Phylogenetics and Evolution 50: 485–495. doi: 10.1016/j.ympev.2008.11.008

Lukhtanov VA, Sourakov A, Zakharov EV, Hebert PDN (2009) DNA barcoding Central Asian butterflies: Increasing geographical dimension does not significantly reduce the success of

species identification. Molecular Ecology Resources 9: 1302–1310. doi: 10.1111/j.1755-0998.2009.02577.x

Nice CC, Fordyce JA, Shapiro AM, Ffrench-Constant R (2002) Lack of evidence for reproductive isolation among ecologically specialized lycaenid butterflies. Ecological Entomology 27: 702–712. doi: 10.1046/j.1365-2311.2002.00458.x

Nixon KC (2002) Winclada Version 1.00.08. Published by the author, Ithaca, New York. http://www.cladistics.com

Pollock DD, Watt WB, Rashbrook VK, Iyengar EV (1998) Molecular phylogeny for *Colias* butterflies ant their relatives (Lepidoptera: Pieridae). Annals of the Entomological Society of America 91: 524–531.

Porter AH, Levin EJ (2010) Parallel evolution in sympatric, hybridizing species: performance of *Colias* butterflies on their introduced host plants. Entomologia Experimentalis et Applicata 124: 77–99. doi: 10.1111/j.1570-7458.2007.00553.x

Schmidt BC, Sperling FA (2008) Widespread decoupling of mtDNA variation and species integrity in *Grammia* tiger moths (Lepidoptera: Noctuidae). Systematic Entomology 33: 613–634. doi: 10.1111/j.1365-3113.2008.00433.x

Seberg O, Humphries CJ, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends in Ecology & Evolution 18: 63–65. doi: 10.1016/S0169-5347(02)00059-9

Sperling F (2003) Butterfly species and molecular phylogenies. P. 431–458. In: Boggs CL, Watt WB, Ehrlich PR. Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago, 756 pp.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. Trends in Ecology & Evolution 18: 70–74. doi: 10.1016/S0169-5347(02)00041-1

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12: 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Verhulst J (2000) Les *Colias* du Globe – Monograph of the Genus *Colias*. Coecke and Evers, Keltern, 308 pp.

Wahlberg N, Oliveira R, Scott JA (2003) Phylogenetic relationships of *Phyciodes* butterfly species (Lepidoptera: Nymphalidae): complex mtDNA variation and species delimitations. Systematic Entomology 28: 257–273. doi: 10.1046/j.1365-3113.2003.00212.x

Wahlberg N, Wheat CW (2008) Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of Lepidoptera. Systematic Biology 57: 231–242. doi: 10.1080/10635150802033006

Wang B, Porter AH (2004) An AFLP-based interspecific linkage map of sympatric, hybridizing *Colias* butterflies. Genetics 168: 215–225. doi: 10.1534/genetics.104.028118

Wheat CW, Watt WB (2008) A mitochondrial-DNA-based phylogeny for some evolutionary-genetic model species of *Colias* butterflies (Lepidoptera, Pieridae). Molecular Phylogenetics and Evolution 47: 893–902. doi: 10.1016/j.ympev.2008.03.013

Zakharov EV, Lobo NF, Nowak C, Hellmann JJ (2009) Introgression as a likely cause of mtDNA paraphyly in two allopatric skippers (Lepidoptera: Hesperiidae). Heredity 102: 590–599. doi: 10.1038/hdy.2009.26

ZooKeys 365: 197–213 (2013)
doi: 10.3897/zookeys.365.5670
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

DNA barcoding as a complementary tool for conservation and valorisation of forest resources

Angeliki Laiou[1], Luca Aconiti Mandolini[1], Roberta Piredda[1], Rosanna Bellarosa[1], Marco Cosimo Simeone[1]

1 *Department of Agriculture, Forests, Nature and Energy (DAFNE) - Università degli Studi della Tuscia, via S. Camillo de' Lellis, 01100 Viterbo, Italy*

Corresponding author: *Marco Cosimo Simeone* (mcsimeone@unitus.it)

Academic editor: *Z. T. Nagy* | Received 27 May 2013 | Accepted 6 December 2013 | Published 30 December 2013

Citation: Laiou A, Mandolini LA, Piredda R, Bellarosa R, Simeone MC (2013) DNA barcoding as a complementary tool for conservation and valorisation of forest resources. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 197–213. doi: 10.3897/zookeys.365.5670

Abstract

Since the pre-historic era, humans have been using forests as a food, drugs and handcraft reservoir. Today, the use of botanical raw material to produce pharmaceuticals, herbal remedies, teas, spirits, cosmetics, sweets, dietary supplements, special industrial compounds and crude materials constitute an important global resource in terms of healthcare and economy. In recent years, DNA barcoding has been suggested as a useful molecular technique to complement traditional taxonomic expertise for fast species identification and biodiversity inventories. In this study, *in situ* application of DNA barcodes was tested on a selected group of forest tree species with the aim of contributing to the identification, conservation and trade control of these valuable plant resources.

The "core barcode" for land plants (*rbc*L, *mat*K, and trnH-psbA) was tested on 68 tree specimens (24 taxa). Universality of the method, ease of data retrieval and correct species assignment using sequence character states, presence of DNA barcoding gaps and GenBank discrimination assessment were evaluated. The markers showed different prospects of reliable applicability. *Rbc*L and trnH-psbA displayed 100% amplification and sequencing success, while *mat*K did not amplify in some plant groups. The majority of species had a single haplotype. The trnH-psbA region showed the highest genetic variability, but in most cases the high intraspecific sequence divergence revealed the absence of a clear DNA barcoding gap. We also faced an important limitation because the taxonomic coverage of the public reference database is incomplete. Overall, species identification success was 66.7%.

This work illustrates current limitations in the applicability of DNA barcoding to taxonomic forest surveys. These difficulties urge for an improvement of technical protocols and an increase of the number of sequences and taxa in public databases.

Keywords
DNA barcoding, Forest Biodiversity, Medicinal and Aromatic plants, Conservation

Introduction

Forests figure prominently among the world's most important ecosystems. The importance of trees in sustaining biodiversity and habitat stability, as well as to provide a large variety of environmental services is well acknowledged. Nevertheless, the increasing human impact, the recent environmental decay, and the on-going climate change are among the main factors affecting forest communities, especially at local and regional scales within the Mediterranean basin (FOREST EUROPE, UNECE and FAO 2011). In the meantime, international market pressures call for higher quality standards. One way to convince decision-makers of the importance of conserving wild plants and habitats is to demonstrate their economic potential (Kathe 2006). The socio-economic contribution of forests to livelihood and the impact of their use on the environment are essential components of modern concepts for sustainable forest management (Arnold and Perez 2001).

Temperate and boreal forests are a traditional source, not only for timber, but also for many products that have been extracted from forests for millennia, including resin, tannin, fodder, litter, medical plants, fruits, nuts, roots, mushrooms, seeds, honey, ornamentals and exudates. Today there is an institutional rediscovery of the value of forest products and services other than timber, and the total value of Non-Wood Goods (NWGs) reported in Europe has almost tripled since 2007 (FOREST EUROPE, UNECE and FAO 2011).

Besides wood trade, Mediterranean woody flora includes numerous valuable species used as ornamentals or for secondary products processing and marketing (edibles, industrial and medicinal compounds). The option of stimulating the production of non-timber forest products has long been considered promising (Arnold and Perez 2001, Wunder 2001), and it is well illustrated in the case of Medicinal and Aromatic Plants (MAPs). In many Euro-Mediterranean countries MAPs resources are still unknown or overlooked (Lange 2006). In other countries, the necessary plant materials (roots, bark, leaves, fruits and seeds) are generally collected and sold by local people to traders and to the industry. Final products are then purchased by international exporters (WHO 2003). Forest overexploitation, product forgery and misidentifications are common risks, with the latter two usually occurring as a result of morphologically indistinguishable materials, species with similar common names, or intentional substitution of economically valuable materials by inexpensive specimens. At the same time, plant misidentification and forgery are serious threats to human health (Vanherweghem et al. 1993, Barthelson et al. 2006, Sundus 2008). The identification of herbal medicinal materials using traditional, organoleptic and chemical methods can be difficult, particularly for processed materials of a plant (Govindaraghavan et al. 2012). Also plant germplasm (seeds and seedlings) purchased for the establishment of MAPs orchards,

afforestation programs, and ornamentals, may be difficult to recognize. Therefore, an accurate, universal, stable and specific method allowing non-specialists to identify the source species from a tiny amount of tissue is needed.

Molecular technology is considered a reliable alternative tool for the identification of plant species (e.g. Savolainen et al. 2000) and DNA barcoding is the latest move towards the generation of universal standards (Kane and Cronk 2008). A DNA barcode is a universally accepted short DNA sequence allowing the prompt and unambiguous identification of species (Savolainen et al. 2005), promoted for a variety of biological applications (Hollingsworth et al. 2011), including biodiversity inventories (Costion et al. 2011, de Vere et al. 2012), the identification of medicinal plants (Heubl et al. 2010), of natural health products (Wallace et al. 2012), and of tree species listed in the Convention on International Trade of Endangered Species (Muellner et al. 2011).

Based on the relative ease of amplification, sequencing, multi-alignment and the amount of variation displayed (sufficient to discriminate among sister species without affecting their correct assignation through intraspecific variation), three plastid loci are currently used in plants: *rbc*L (a universal but slowly evolving coding region), *mat*K (a relatively fast evolving coding region) and trnH-psbA (a rapidly evolving intergenic spacer) (CBOL Plant Working Group 2009). More recently, the nuclear ribosomal internal transcribed spacer (ITS) has also been suggested as an efficient barcoding locus for complex plant groups (Hollingsworth et al. 2011).

Tree taxa have peculiar biological, evolutionary and taxonomic features that are likely to constitute a challenge to species recognition through DNA barcodes, viz. the generally low mutation rate of the plastid DNA, their ability to hybridize, and their narrowly defined species limits (Petit and Hampe 2006). Nevertheless, DNA barcoding has proven its utility in several detailed studies of tree genera (Newmaster et al. 2008, Newmaster and Ragupathy 2009, Kress et al. 2009, 2010, Ren et al. 2010, Roy et al. 2010, Liu et al. 2011). In this study, *in situ* application of DNA barcoding was applied to a number of indigenous and introduced tree species in the Mediterranean area, with medicinal, ornamental, edible, industrial and conservation relevance. Taxa were analysed with the core barcode for land plants (*rbc*L, *mat*K, and trnH-psbA); ease and success to achieve correct species identification were evaluated based on the relative efficiency of each marker, data quality and representation in the GenBank/EMBL database. Our final objective is to provide a contribution to the future assemblage of a regional data/species inventory in the Mediterranean area for adequate identification, conservation and trade control of these valuable resources.

Materials and methods

Plant material and molecular analyses

Sixty eight trees belonging to 24 species (ten genera, nine families) were sampled in the wild (Italy, Greece and adjacent areas) and/or Botanic Gardens (Table 1). Plants were

Table 1. Sample list.

Familia	Species		Relevance	No. of samples
Pinaceae	*Cedrus*	*atlantica*	Ornamental/afforestation	3
		deodara	Ornamental/afforestation	3
		libani	Ornamental/afforestation/conservation	3
Rosaceae	*Crataegus*	*monogyna*	Medicinal/ornamental	3
		oxyacantha	Medicinal/ornamental	2
		azarolus	Food industry/conservation	4
	Sorbus	*aria*	/	3
		aucuparia	Ornamental/conservation	2
		domestica	Medicinal/food industry	3
		torminalis	Valuable wood industry	3
Sapindaceae	*Aesculus*	*hippocastanus*	Medicinal/ornamental	3
		indica	/	3
Oleaceae	*Fraxinus*	*ornus*	Medicinal/food industry	5
		angustifolia	/	3
		excelsior	/	2
Adoxaceae	*Sambucus*	*nigra*	Medicinal	5
		ebulus	/	2
		racemosa	/	1
Passifloraceae	*Passiflora*	*incarnata*	Medicinal/ornamental	2
		edulis	Food industry	1
Lythraceae	*Punica*	*granatum*	Medicinal/food industry/ornamental	4
Rhamnaceae	*Ziziphus*	*jujuba*	Medicinal/food industry	3
Aquifoliaceae	*Ilex*	*aquifolium*	Medicinal/ornamental/conservation	4
		latifolia	/	1

identified directly in the field. Herbarium specimens and lyophilized green tissues of the collected material were vouchered and preserved at the Mediterranean Forest DNA bank of the University of Tuscia (www.Medna-bank.eu).

DNA extractions were performed with the DNeasy Plant Minikit (QIAGEN), following the manufacturer's instructions. The universal applicability of the technical analyses was considered a prerequisite for exploring the DNA barcoding potential in a practical floristic case study: uniform PCR procedures were thus performed for all taxa and barcoding loci. Genomic DNAs (ca. 40 ng) were amplified with RTG PCR beads (GE Healthcare) in 25 µl final volume according to the manufacturer's protocol. Thermocycling conditions were as follows: 94 °C for 3 min, followed by 35 cycles of 94 °C for 30 s, 53 °C for 40 s and 72 °C for 40 s, with a final extension step of 10 min at 72 °C. Primers for the investigated barcoding region are shown in Table 2. MatK1F/2R oligos were used in *Cedrus* (Wang et al. 1999). PCR products were cleaned with Illustra DNA/Gel Band Purification Kit (GE Healthcare). Standard aliquots were submitted to Macrogen Inc. (http://www.macrogen.com) for sequencing. Electropherograms were edited with CHROMAS 2.3 (http://www.technelysium.com.au) and checked visually.

Table 2. Primers list.

Marker region	Primers	Reference
***rbc*L**	Fw - ATGTCACCACAAACAGAAAC	Kress et al. (2005)
	Rev - TCGCATGTACCTGCAGTAGC	
trnH-psbA	Fw - CGCGCATGGTGGATTCACAATCC	Shaw et al. (2007)
	Rev - GTTATGCATGAACGTAATGCTC	
***mat*K_Kim**	Fw - CGTACAGTACTTTTGTGTTTACGAG	Kim (unpublished)
	Rev - ACCCAGTCCATCTAAATCTTGGTTC	
***mat*K1F/2R**	Fw - GAACTCGTCGGATGGAGTG	Wang et al. (1999)
	Rev - TAAACGATCCTCTCATTCACGA	

Bioinformatics tools

Sequences were aligned with MEGA5 (Tamura et al. 2011) and checked by eye. Haplotypes were defined with BLASTClust v2.2.20 (http://toolkit.tuebingen.mpg.de/blastclust) with the following command line: blastclust -i infile -o outfile -p F -L1 -bT -S100, thus requiring to cluster together only sequences with 100% identity and length coverage. All the species presenting single haplotypes were considered efficiently discriminated; those displaying at least one haplotype in common with another species were considered precluded to discrimination.

Species discrimination power of the investigated loci was also assessed using the genetic distance approach, to evaluate whether the amount of variation displayed was sufficient to discriminate sister species without affecting their correct assignation through intraspecific variation. This approach is at the basis of the "barcoding gap" definition, i.e. the assumption that the amount of sequence divergence within species is smaller than that between species. Uncorrected p-distance matrices of sequence divergences within and among congeneric species were calculated for each gene fragment and for the two joined markers (*rbc*L + trnH-psbA), with MEGA5. All the species presenting a minimum interspecific distance value higher than their maximum intraspecific distance were considered successfully discriminated (Meyer et al. 2008).

Finally, we simulated a barcode identification scenario using each sequence as an unknown query and GenBank (http://www.ncbi.nlm.nih.gov) as global reference database. The NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy) was screened to assess the presence of the investigated species set in GenBank, relatively to markers under study. The identification ability of every single marker was evaluated using the megaBLAST algorithm (http://blast.ncbi.nlm.nih.gov) with default parameters and adjusted to retrieve 5000 sequences. A query sequence was considered as successfully identified if the top Bit-score obtained in GenBank matched the name of the species (Ross et al. 2008). Identification success was only inferred for species/sequences represented in GenBank. When more than one species shared a top Bit-Score or the species scored lower, the result was considered an identification failure.

Results

Markers' main features

Optimal amplification rates were obtained with *rbc*L and trnH-psbA which produced clear, single-banded PCR products from all 68 investigated samples (136 sequences; 100% efficiency). MatK was not consistently amplified in the Pinaceae and Rosaceae (44.1% of the investigated dataset) and thus it was not included in further analyses. All *rbc*L electropherograms were easily read and analysed. Conversely, the very long polynucleotide repeats in the trnH-psbA regions of *Sambucus* sp. made subsequent traces hardly readable. Consequently, in this genus the entire sequences were completed by joining partial bidirectional reads (Kress and Erickson 2007). The alignment of *rbc*L sequences was straightforward with a consensus of 688 bp (no indels found). The trnH-psbA sequences varied greatly in length, ranging from 396 (*Sorbus* and *Crataegus* spp.) to 622 bp (*Cedrus* spp.). Numerous gaps were observed in this region. An indel of 45 bp turned out to be diagnostic to discriminate the two *Aesculus* species, an indel of 55 bp discriminated *Fraxinus ornus* from *F. excelsior* and *F. angustifolia*, one of 66 bp discriminated *Sambucus ebulus* from *S. racemosa* and other indels (20-22 bp) were diagnostic for *Sorbus torminalis* and *Cedrus deodara*. Shorter gaps (1-19 bp) were detected intraspecifically in all species except in *Punica*, *Ziziphus* and *Ilex*. All sequences have been deposited in GenBank under accession numbers HG765031-HG765098 (*rbc*L), and HG764963-HG765030 (trnH-psbA).

Markers' discrimination ability

The alignment–free method implemented in BLUSTClust produced for each marker the haplotypes shown in Table 3. Based on the uniqueness of sequence character states, trnH-psbA generated a total of 43 haplotypes, 35 of which could be ascribed to single species. Common haplotypes were displayed by 14 individuals of the following species pairs, thus preventing their discrimination: *Fraxinus angustifolia - F. excelsior* (three samples), *Crataegus monogyna - C. oxyacantha* (four samples), *Sorbus aucuparia - S. domestica* (two samples), *Ilex aquifolium - I. latifolia* (five samples). Consequently, trnH-psbA discrimination ability was 79.4% of the investigated plants, corresponding to 66.7% of the species in the total dataset, 63.6% considering only those genera in which at least one species pair was sampled.

RbcL displayed a much lower sequence differentiation (with a total of 31 haplotypes, 12 of which were shared between species). No haplotypes were shared among species from different genera. The two-marker combination did not improve markedly the discrimination efficacy displayed by trnH-psbA alone.

In this study, the two potential DNA barcodes displayed different levels of intra- and interspecific distances. With *rbc*L, all intraspecific uncorrected p-distances were zero, except in *Cedrus atlantica* (0.0014), *Sorbus aria* (0.0014), *S. aucuparia* (0.0028),

Table 3. Haplotypes generated by BLASTClust in the investigated dataset with both markers and their combination. Shaded: species where unique haplotypes (either single or in combination) were detected.

Species		Samples	Unique haplotypes			Inter-species shared haplotypes		
			*rbc*L	trnH-psbA	Combined	*rbc*L	trnH-psbA	Combined
Cedrus	*atlantica*	3	2	2	2	/	/	/
	deodara	3	1	1	1	/	/	/
	libani	3	1	1	1	/	/	/
Crataegus	*monogyna*	3	/	/	/	1	1	1
	oxyacantha	2	/	1	1	1	1	1
	azarolus	4	/	2	2	1	/	/
Sorbus	*aria*	3	1	3	3	/	/	/
	aucuparia	2	1	1	1	1	1	1
	domestica	3	/	1	1	1	1	1
	torminalis	3	1	1	1	/	/	/
Aesculus	*hippocastanus*	3	1	2	2	/	/	/
	indica	3	1	3	3	/	/	/
Fraxinus	*ornus*	5	2	4	5	1	/	/
	angustifolia	3	/	1	1	1	1	1
	excelsior	2	/	/	/	1	1	1
Sambucus	*nigra*	5	1	4	4	1	/	/
	ebulus	2	1	2	2	1	/	/
	racemosa	1	1	1	1	/	/	/
Passiflora	*incarnata*	2	2	2	2	/	/	/
	edulis	1	1	1	1	/	/	/
Punica	*granatum*	4	1	1	1	n.d.	n.d.	n.d.
Ziziphus	*jujuba*	3	1	1	1	n.d.	n.d.	n.d.
Ilex	*aquifolium*	4	/	/	/	1	1	1
	latifolia	1	/	/	/	1	1	1
Total		**68**	**19**	**35**	**36**	**12**	**8**	**8**

Crataegus monogyna (0.0028), and *Sambucus ebulus* (0.004). Zero interspecific distances were detected between individuals belonging to *Sorbus aucuparia* and *S. domestica*, among the three *Crataegus* species, the three *Fraxinus* species, between *Sambucus nigra* and *S. ebulus*, and between the two *Ilex* species. Conversely, no intraspecific sequence variation was found at trnH-psbA in *Cedrus deodara*, *C. libani*, *Sorbus torminalis*, *Crataegus monogyna*, *C. oxyacantha*, *Fraxinus angustifolia*, *Sambucus racemosa*, *Passiflora edulis*, *Punica granatum*, *Ziziphus jujuba* and the two *Ilex* species. Interspecific genetic differences produced by this marker exhibited values higher than zero (0.0018–0.0298) only in five species belonging to *Cedrus, Aesculus* and *Passiflora* genera, and in *Fraxinus ornus* and *Sambucus racemosa*.

The values of the maximum intra- and minimum interspecific sequence divergence of the two combined barcoding loci are shown in Table 4 (all interspecific distances involve congeneric species). In agreement with data based on the single markers, non-overlapping intra- and interspecific distances were observed in a few species groups. As

Table 4. Values of maximum inter- and minimum intraspecific uncorrected p-distances resulting from the combination of *rbc*L + trnH-psbA sequences, and relative barcoding gaps calculated in 24 forest tree taxa; n.d. = not determined; * = no sister species included in the dataset; ** = *taxa* with single accession. Shaded: species where a barcoding gap was detected.

	Samples	Max. Intrasp. distance	Min Intersp. distance	Barcoding gap
Cedrus atlantica	3	0.0015	0.0015	0
Cedrus deodara	3	0	0.0015	0.0015
Cedrus libani	3	0	0.0023	0.0023
Sorbus aria	3	0.002898554	0.000950571	- 0.0019
Sorbus aucuparia	2	0.0058	0	- 0.0058
Sorbus domestica	3	0.0009	0	- 0.0009
Sorbus torminalis	3	0	0.0009	0.0009
Crataegus azarolus	3	0.0009	0	- 0.0009
Crataegus monogyna	2	0.0019	0	- 0.0019
Crataegus oxyacantha	4	0	0	0
Aesculus hippocastanus	3	0	0.0064	0.0064
Aesculus indica	3	0	0.0064	0.0064
Fraxinus ornus	5	0.00568	0.00284	- 0.0028
Fraxinus angustifolia	3	0.0036	0	- 0.0036
Fraxinus excelsior	2	0	0	0
Sambucus nigra	5	0.0017	0	- 0.0017
Sambucus ebulus	2	0.0101	0	- 0.0101
*Sambucus racemosa***	1	n.d.	0.0142	n.d.
Passiflora incarnata	2	0.02397	0.01588	- 0.0081
*Passiflora edulis***	1	n.d.	0.0158	n.d.
*Punica granatum**	4	0	n.d.	n.d.
*Ziziphus jujuba**	3	0	n.d.	n.d.
Ilex aquifolium	4	0	0	0
*Ilex latifolia***	1	n.d.	0	n.d.

such, barcoding gaps were observed in *Cedrus deodara* and *C. libani*, *Sorbus torminalis*, and the two *Aesculus* species. All remaining taxa displayed equal (e.g. in *Cedrus atlantica*) or higher values of intra- than interspecific divergence (e.g. in *Passiflora incarnata*, *Fraxinus ornus*, *Sorbus aria*). Several species showed sequences involving zero interspecific divergence (e.g. *Sorbus domestica*, *S. aucuparia*, *Fraxinus excelsior*, *F. angustifolia*, *Sambucus nigra*, *S. ebulus*, *Crataegus* spp.). The lack of additional conspecific samples did not allow a comparison with the high levels of interspecific divergences shown by two species (*Passiflora edulis* and *Sambucus racemosa*). These results suggest that there is a barcoding gap in only five out of 19 analyzed species, corresponding to 26.3% of our dataset (taxa with only one individual/species or one species/genus excluded).

The NCBI Taxonomy database screening revealed that all the species in our dataset were represented by *rbc*L and trnH-psbA marker sequences in the database, except for *Aesculus indica*, *Cedrus libani* (neither marker), *Crataegus azarolus* and *Sorbus domestica* (only *rbc*L present).

When BLASTed to GenBank, all our *rbc*L sequences were identified by the reference sequences at the genus level (87.5% of total taxa), or even at the species level (41.6%). Genus misidentification occurred in the three *Crataegus* species, for which genera *Cotoneaster*, *Pyrus*, *Piracantha*, *Amelanchier*, *Chaenomeles* (all belonging to the Rosaceae family) and *Crataegus* were also the best match. In contrast, correct genus and species identifications were obtained for *Ilex aquifolium*, *Passiflora incarnata* and *P. edulis*, *Punica granatum*, *Ziziphus jujuba*, *Sambucus nigra*, *Sorbus torminalis*, *Cedrus atlantica* and *C. deodara*.

TrnH-psbA was outperformed by *rbc*L, since none of the *Sorbus* sequences (four species) matched the right genus, and only eight species (33.3%) were correctly identified (*Fraxinus ornus*, *Passiflora incarnata*, *Punica granatum*, *Ziziphus jujuba*, *Sambucus racemosa*, *Cedrus atlantica* and *C. deodara*). All other samples shared the highest score with other species (e.g. *Aesculus hippocastanum* with *A. turbinata*, *Fraxinus excelsior* with *F. angusitfolia*, *Sambucus nigra* with *S. racemosa*, *Crataegus monogyna* with several other species), or even hit the wrong species (e.g. *Ilex aquifolium*, *Sambucus ebulus*, *Crataegus oxyacantha*). The four taxa not represented in GenBank (*Cedrus libani*, *Aesculus indica*, *Creataegus azarolus* and *Sorbus domestica*) were assigned to the correct genus. As a final result, only 11 species were correctly identified by the two locus-combination corresponding to 55% of the investigated species having a reference in GenBank (45.8% of the total species set). A summary of the correct species identifications achieved with the three discrimination methods used in the present study is shown in Table 5. Thirteen species (54.2% of our dataset) were identified by at least two methods. Only two species (*Cedrus deodara* and *Sorbus torminalis*) were identified with the three methods, whereas the absence of conspecific GenBank references prevented the same full identification for *Cedrus libani* and *Aesculus indica*. In contrast, six species (corresponding to three species pairs and totalling 25% of our dataset) appeared unidentifiable with any method: *Crataegus monogyna*, *C. oxyacantha*, *Sorbus aucuparia*, *S. domestica*, *Fraxinus angustifolia*, *F. excelsior*. Two species (*Crataegus azarolus* and *Sorbus aria*) were discriminated only by means of sequence specificity but received no confidence by any of the other two approaches (the former was absent in GenBank).

Discussion

Marker applicability

In our dataset, the *rbc*L + trnH-psbA combination showed the highest amplification and sequencing success (100%), whereas *mat*K showed a much lower success (55.9%). Specifically, the currently most adopted primers set for Angiosperms (*mat*K_KIM) failed in the amplification of the Rosaceae, and *mat*K1F/2R primers, suggested for the Pinaceae, failed to amplify *Cedrus* sp. In addition, *mat*K also revealed severe difficulties in the amplification and/or sequencing steps in the genera *Berberis* (Berberidaceae), *Vitex* (Rhamnaceae), *Cercis* (Leguminosae) and *Ginkgo* (Ginkgoaceae), in the ongoing

Table 5. Summary of the species identification success achieved with *rbc*L + trnH-psbA and the three discrimination methods in the present study: occurrence of unique haplotypes in the total species set, genetic distances among and within congeneric species, correct species match in the GenBank database. Green: correct identification; red: non confident/wrong identification; shaded = not determined (no intra- or interspecific samples investigated); a = species absent in GenBank with either one or both markers.

Species		Identification success		
		Haplotype specificity	Min. inter- > max. intra-specific distance	GenBank correct match
Cedrus	*atlantica*	√	-	√
	deodara	√	√	√
	libani	√	√	a
Crataegus	*monogyna*	-	-	-
	oxyacantha	-	-	-
	azarolus	√	-	a
Sorbus	*aria*	√	-	-
	aucuparia	-	-	-
	domestica	-	-	a
	torminalis	√	√	√
Aesculus	*hippocastanus*	√	√	-
	indica	√	√	a
Fraxinus	*ornus*	√	-	√
	angustifolia	-	-	-
	excelsior	-	-	-
Sambucus	*nigra*	√	-	√
	ebulus	√	-	-
	racemosa	√	n.d.	√
Passiflora	*incarnata*	√	-	√
	edulis	√	n.d.	√
Punica	*granatum*	√	n.d.	√
Ziziphus	*jujuba*	√	n.d.	√
Ilex	*aquifolium*	-	-	√
	latifolia	-	n.d.	-
Efficacy		**66.7%**	**26.3%**	**55%**

prosecution of this work. The lack of universality of *mat*K was already reported by e.g. Kress and Erickson (2007), Fazekas et al. (2008), Ford et al. (2009), De Mattia et al. (2012). MatK_KIM, (Kim, unpublished) is still considered the primer set with the highest match for eudicots, while *mat*K1F/2R was efficiently used in a comprehensive study across Pinaceae (Wang et al. 1999). Dunning and Savolainen (2010) also noted that *mat*K_KIM is not the best choice for Rosaceae and rather suggested the use of specific primer sets. The difficulty of defining the best primer choice for *mat*K in Conifers was already faced by e.g. Li et al. (2011) and Armenise et al. (2012). When applied to international trade and safe use of medicinal plants, *mat*K yielded 54.0% of amplification efficiency in Chen et al. (2010), whereas Kool et al. (2012) produced

PCR products for less than 30% of the specimens, and sequencing success was only 10% in Wallace et al. (2012).

In contrast, trnH–psbA provided better discrimination than *mat*K in many diverse tree genera such as *Alnus* (Roy et al. 2010), *Ficus* (Ren et al. 2010), *Quercus* (Simeone et al. 2013), and more generally in Angiosperms (Pang et al. 2012). Nevertheless, *mat*K is still recommended by the CBOL Plant working Group (2009) as the first option to rely on in terms of sequence variability. We therefore suggest that an efficient barcoding workflow should include a first preliminary screening with *mat*K universal primer set(s) and then, depending to the amplification results, to select trnH-psbA as an additional marker to *rbc*L. Alternatively, a simple and clear morphological trait may be included in the analysis or address the search for the most appropriate *mat*K primer set based on the biological group under study (Bruni et al. 2012, Dunning and Savolainen 2010).

Species identification and discrimination

The BLUSTClust analysis yielded a 66.7% species discrimination, which is a bit lower but still in line with the general limit acknowledged for land plants when markers from a single genetic linkage group are used (ca. 70%; CBOL Plant Working Group 2009). In agreement, similar percentages (68–71%) were obtained in broader taxonomic investigations in forests of North and meso-America (Fazekas et al. 2008, Gonzalez et al. 2009), although by use of a different way to assess species identification success (i.e. support for species monophyly through barcodes). Our barcoding data, dedicated to woody plants sampled in a different ecological zone, approach Piredda et al. (2011), who reported 73% efficiency in a floristic investigation of the Italian tree flora by means of sequence specificity; nevertheless, more intraspecific diversity and more species pairs were surveyed in the present work.

The highest identification success was achieved with the analysis based on the uniqueness of sequence character states, where some parts in the haplotypes (especially some trnH-psbA indels) appeared diagnostics for certain species. However, more data are required to confirm these diagnostic sequence features. Yet, if confirmed, these features may be important in view of the generally low interspecific divergences we observed. Conversely, the analysis with the barcoding gaps suggests that such a discrimination approach may yield a lower efficiency, at least with trnH-psbA, since the uncorrected p-distance analysis removed all indels. A further complication we encountered was constituted by the high intraspecific divergences (e.g. in *C. atlantica*) and the sharing of haplotypes among congeneric species (e.g. in *Sorbus*, *Crataegus*, *Fraxinus*, *Sambucus*). All these results challenge the application of DNA barcoding with *rbc*L + trnH-psbA in the taxa investigated here. This is the more so as GenBank also showed a low identification efficiency and sometimes lead to erroneous identifications, most often due to the limited number of available reference sequences and their sometimes very high intraspecific divergences. Little and

Stevenson (2007) and Ross et al. (2008) found that BLAST (and other similarity methods) can give accurate identifications on GenBank (see also de Vere et al. 2012 and Pang et al. 2012), although some distorted results, in inverse proportion to the number of reference sequences per species in the databases, may render these approaches inappropriate. Ideally, a reference library should provide multiple samples from unambiguously identified species or taxa, and cover intraspecific variability and closely related species to evaluate the degree of divergences among barcodes. Unfortunately, the reference list in the GenBank database is still far from complete. The small numbers of available sequences per species and for either marker prevented us from confidently retrieving correct species names in *Aesculus hippocastanum*, *Fraxinus excelsior*, *Ilex latifolium*, *Crataegus monogyna* (highest scores shared with other congenerics). Moreover, it induced us to assign a query to the wrong species, as in the cases of *Aesculus indica* (*A. pavia*), *Fraxinus angustifolia* (*F. excelsior*), *Passiflora edulis* (*P. incarnata*), *Sambucus ebulus* (*S. adnata*), *Crataegus azarolus* and *C. oxyacantha* (*C. monogyna*), *Cedrus libani* (*C. deodara*), and the four *Sorbus* species. Clearly, a consistent enrichment of the reference databases is a priority for future applications of DNA barcoding.

DNA barcoding of medicinal and aromatic plants

DNA barcoding is a substantial improvement of our capacity to document the existing biodiversity. It is also a powerful research complement for human socio-economics, safety, trade control, frauds discovery and detection of forgeries in plant commercial products (Newmaster and Ragupathy 2010). Kool et al. (2012), for example, were able to document 18 misidentifications and eight forgeries among 111 samples of medicinal plants in a local market in Marrakech (Morocco).

The Mediterranean woody flora comprises numerous valuable species used as ornamentals or for secondary products processing and marketing (edibles, essential oils, medicinal compounds). Field identification, authentication and certification of germplasm and raw materials are a major concern. As such, our results on *Cedrus* support previous findings that members of Pinaceae can be efficiently barcoded with *rbc*L + trnH-psbA (at least at a regional scale; Armenise et al. 2012). Cedars involve four different extant species: the three more highly diffused and with great ornamental, ecological and cultural relevance were here discriminated, while *Cedrus brevifolia*, a highly protected, rare endemic surviving in only one population on Troodos Mountains (Cyprus), still awaits further investigations. We also found specific haplotypes for the highly important and largely cultivated *Punica granatum*. In this case as well, further investigations involving the only other species of genus *Punica* (*Punica protopunica*, a rare endemic of the Socotra Island, Yemen, very similar in morphology, production of fruits and secondary metabolites) would eventually provide new tools for its conservation and management.

On the other hand, we confirm the difficulties previously encountered in barcoding *Fraxinus* (Arca et al. 2012) and the extensive interspecific haplotype sharing in *Crataegus* (Fineschi et al. 2005) and *Sorbus* (Robertson et al. 2010). For instance, Burgess et al. (2011) were able to discriminate only one out of four *Crataegus* species with five barcoding markers. Indeed, these genera are likely to be as refractory to barcoding as other woody groups including oaks (Piredda et al. 2011) and willows (von Crautlein et al. 2011). Low mutation rates, incomplete lineage sorting and hybridization are the most reported causes (Hollingsworth et al. 2011). However, we were able to discriminate *Fraxinus ornus*, a very important medicinal and industrial plant, and *Crataegus azarolus*, a protected fruit tree, historically used for a number of medicinal purposes. Conversely, we were unable to discriminate the *Crataegus monogyna - C. oxyacantha* species pair (see also Bruni et al. 2012), but this has little practical importance since both hawthorns are equally used for the same medicinal purposes. Very promising data were collected on *Sorbus aria* and *S. torminalis*, *Ilex aquifolium*, *Aesculus Hippocastanum*, *Passiflora* and *Ziziphus jujuba*, suggesting that an efficient barcoding could be achieved on these species, at least at regional scales. In contrast, *Sambucus* sp. showed a large intraspecific divergence and require further investigations on larger datasets. More recently, the nuclear ribosomal ITS (especially the ITS2 portion) has been suggested as an efficient barcoding locus for complex plant groups (Chen et al. 2010). However, Kool et al. (2012) could not use this marker in 45% of their dataset because of the low amplification and sequencing efficacy detected and fungal contamination, particularly in the root material. Therefore, this marker still appears not completely devoid of some pitfalls and certainly will require an improvement of current protocols.

Conclusion

Recently, an outstanding research interest towards DNA barcoding of regional floras with biological and/or economical relevance has spread. In the present work, we lay the foundations towards DNA barcoding applications of important woody plant genera in the Mediterranean basin, such as *Cedrus*, *Aesculus*, *Ilex, Passifllora, Punica, Sambucus*, *Sorbus*, *Ziziphus*. All these genera include valuable taxa for multiple natural and economic purposes, and combine with similar DNA barcoding investigations performed on Euro-Mediterranean forested land in recent years (Piredda et al. 2011, von Crautlein et al. 2011, Armenise et al. 2012, Simeone et al. 2013). Gathered results expose limitations of DNA barcoding, most of which are due to (1) the imperfect discrimination ability of the markers and methods currently in use, (2) the biological peculiarities of some genera, and (3) the low taxonomic coverage of the reference databases. Future technological advances, additional markers and larger sample sets at different geographical scales (from continental to local) are therefore auspicated to improve current protocols and identification success for the practical conservation and valorisation of forest natural resources.

References

Arca M, Hinsinger DD, Cruaud C, Tillier A, Bousquet J, Frascaria-Lacoste N (2012) Deciduous trees and the application of universal DNA barcodes: A case study on the circumpolar *Fraxinus*. PLoS ONE 7: e34089. doi: 10.1371/journal.pone.0034089

Armenise L, Simeone MC, Piredda R, Schirone B (2012) Validation of DNA barcoding as an efficient tool for taxon identification and detection of species diversity in Italian conifers. European Journal of Forest Research 131: 1337–1353. doi: 10.1007/s10342-010-0420-1

Arnold JEM, Ruiz Perez M (2001) Can non-timber forest products match tropical forest conservation and development objectives? Ecological Economics 39: 437–447. doi: 10.1016/S0921-8009(01)00236-1

Barthelson RA, Sundareshan P, Galbraith DW, Woosley RL (2006) Development of a comprehensive detection method for medicinal and toxic plant species. American Journal of Botany 93: 566–574. doi: 10.3732/ajb.93.4.566

Bruni I, De Mattia F, Martellos S, Galimberti A, Savadori P, Casiraghi M, Nimis PL, Labra M (2012) DNA barcoding as an effective tool in improving a digital plant identification system: A case study for the area of Mt. Valerio, Trieste (NE Italy). PLoS ONE 7: e43256. doi: 10.1371/journal.pone.0043256

Burgess KS, Fazekas AJ, Kesanakurti PR, Graham SW, Husband BC, Newmaster SG, Percy DM, Hajibabaei M, Barrett SCH (2011) Discriminating plant species in a local temperate flora using the rbcL+matK DNA barcode. Methods in Ecology and Evolution 2011 2: 333–340. doi: 10.1111/j.2041-210X.2011.00092.x

CBOL Plant Working Group (2009) CBOL approves matK and rbcL as the BARCODE regions for Land Plants, Statement by the Executive Committee, Consortium for the Barcode of Life. Proceedings of the National Academy of Sciences of the USA 106: 12794–12797. doi: 10.1073/pnas.0905845106

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5: e8613. doi: 10.1371/journal.pone.0008613

Costion C, Ford A, Cross H, Crayn D, Harrington M, Lowe A (2011) Plant DNA Barcodes Can Accurately Estimate Species Richness in Poorly Known Floras. PLoS ONE 6: e26841. doi: 10.1371/journal.pone.0026841

De Mattia F, Gentili R, Bruni I, Galimberti A, Sgorbati S, Casiraghi M, Labra M (2012) A multi-marker DNA barcoding approach to save time and resources in vegetation surveys. Botanical Journal of the Linnean Society 169: 518–529. doi: 10.1111/j.1095-8339.2012.01251.x

de Vere N, Rich TCG, Ford CR, Trinder SA, Long C, Moore CW, Satterthwaite D, Davies H, Allainguillaume J, Ronca S, Tatarinova T, Garbett H, Walker K, Wilkinson MJ (2012) DNA Barcoding the Native Flowering Plants and Conifers of Wales. PLoS ONE 7: e37945. doi: 10.1371/journal.pone.0037945

Dunning LT, Savolainen V (2010) Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. Botanical Journal of the Linnean Society 164: 1–9. doi: 10.1111/j.1095-8339.2010.01071.x

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH (2008) Multiple multilocus DNA Barcodes from the plastid genome discriminate plant species equally well. PLoS ONE 3: e2802. doi: 10.1371/journal.pone.0002802

Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, Newmaster SG, Hajibabaei M, Husband BC (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular Ecology Resources 9: 130–139. doi: 10.1111/j.1755-0998.2009.02652.x

Fineschi S, Salvini D, Turchini D, Pastorelli R, Vendramin GG (2005) *Crataegus monogyna* Jacq. and *C. laevigata* (Poir.) DC. (Rosaceae, Maloideae) display low level of genetic diversity assessed by chloroplast markers. Plant Systematics and Evolution 250: 187–196. doi: 10.1007/s00606-004-0228-x

Ford CS, Ayres KL, Toomey N, Haider N, Van Alphen Stahl J, Kelly LJ, Wikstrom N, Hollingsworth PM, Duff RJ, Hoot SB, Cowan RS, Chase MW, Wilkinson MJ (2009) Selection of candidate coding DNA barcoding regions for use on land plants. Botanical Journal of the Linnean Society 159: 1–11. doi: 10.1111/j.1095-8339.2008.00938.x

FOREST EUROPE, UNECE and FAO (2011) State of Europe's Forests 2011. Status and Trends in Sustainable Forest Management in Europe. Ministerial Conference on the Protection of Forests in Europe.

Global Strategy for Plant Conservation (2002) Convention on Biological Diversity: Global Strategy for Plant Conservation, Montreal.

Gonzalez MA, Baraloto C, Engel J, Mori SA, Petronelli P, Riera B, Roger A, Thebaud C, Chave J (2009) Identification of Amazonian Trees with DNA Barcodes. PLoS ONE 4: e7483. doi: 10.1371/journal.pone.0007483

Govindaraghavan S, Hennell JR, Sucher NJ (2012) From classical taxonomy to genome and metabolome: Towards comprehensive quality standards for medicinal herb raw materials and extracts. Fitoterapia 83: 979–988. doi: 10.1016/j.fitote.2012.05.001

Heubl G (2010) New aspects of DNA-based authentication of Chinese medicinal plants by molecular biological techniques. Planta Medica 76: 1963–1974. doi: 10.1055/s-0030-1250519

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a Plant DNA barcode. PLoS ONE 6: e19254. doi: 10.1371/journal.pone.0019254

Kane NC, Cronk Q (2008) Botany without borders: barcoding in focus. Molecular Ecology 17: 5175–5176. doi: 10.1111/j.1365-294X.2008.03972.x

Kathe W (2006) Revision of the Guidelines on the conservation of medicinal plants by WHO, IUCN, WWF AND TRAFFICR. In: Bogers J, Craker LE, Lange D (Eds) Medicinal and Aromatic Plants. Springer, the Netherlands, 109–120. doi: 10.1007/1-4020-5449-1_8

Kool A, de Boer HJ, Krűger A, Rydberg Å, Abbad A, et al. (2012) Molecular Identification of Commercialized Medicinal Plants in Southern Morocco. PLoS ONE 7: e39459. doi: 10.1371/journal.pone.0039459

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS ONE 2: e508. doi: 10.1371/journal.pone.0039459

Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences of the USA 106: 18621–18626. doi: 10.1073/pnas.0909820106

Lange D (2006) International trade in medicinal and aromatic plants. In: Bogers RJ, Craker LE, Lange D (Eds) Medicinal and Aromatic Plants. Springer, Netherlands, 155–170.

Li Y, Gao L-M, Poudel RC, Li D-Z, Forrest A (2011) High universality of matK primers for barcoding gymnosperms. Journal of Systematics and Evolution 49: 169–175. doi: 10.1111/j.1759-6831.2011.00128.x

Little DP, Stevenson DW (2007) A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. Cladistics 23: 1–21. doi: 10.1111/j.1096-0031.2006.00126.x

Liu J, Moller M, Gao LM, Zhang DQ, Zhuki DE (2011) DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. Molecular Ecology Resources 11: 89–100. doi: 10.1111/j.1755-0998.2010.02907.x

Meyer R, Zhang GY, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Systematic Biology 57: 809–813. doi: 10.1080/10635150802406343

Muellner AN, Schaefer H, Lahaye R (2011) Evaluation of candidate DNA barcoding loci for economically important timber species of the mahogany family (Meliaceae). Molecular Ecology Resources 11: 450–460. doi: 10.1111/j.1755-0998.2011.02984.x

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J (2008) Testing candidate plant barcode regions in the Myristicaceae. Molecular Ecology Resources 8: 480–490. doi: 10.1111/j.1471-8286.2007.02002.x

Newmaster SG, Ragupathy S (2009) Testing plant barcoding in a sister species complex of pantropical *Acacia* (Mimosoideae, Fabaceae). Molecular Ecology Resources 9: 172–180. doi: 10.1111/j.1755-0998.2009.02642.x

Pang X, Liu C, Shi L, Liu R, Liang D, Li H, Cherny SS, Chen S (2012) Utility of the trnH–psbA intergenic spacer region and its combinations as plant DNA barcodes: A meta-analysis. PLoS ONE 7: e48833. doi: 10.1371/journal.pone.0048833

Petit RJ, Hampe A (2006) Some Evolutionary Consequences of Being a Tree. Annual Review of Ecology, Evolution, and Systematics 37: 187–214. doi: 10.1146/annurev.ecolsys.37.091305.110215

Piredda R, Simeone MC, Attimonelli M, Bellarosa R, Schirone B (2011) Prospects of barcoding the Italian wild dendroflora: oaks reveal severe limitations to tracking species identity. Molecular Ecology Resources 11: 72–83. doi: 10.1111/j.1755-0998.2010.02900.x

Ren BQ, Xiang XG, Chen ZD (2010) Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers. Molecular Ecology Resources 10: 594–605. doi: 10.1111/j.1755-0998.2009.02815.x

Robertson A, Rich TCG, Allen MA, Houston L, Roberts C, Bridle JR, Harris SA, Hiscock SJ (2010) Hybridization and polyploidy as drivers of continuing evolution and speciation in *Sorbus.* Molecular Ecology Resources 19: 1675–1690. doi: 10.1111/j.1365-294X.2010.04585.x

Ross HA, Murugan S, Li WLS (2008) Testing the reliability of genetic methods of species identification via simulation. Systematic Biology 57: 216–230. doi: 10.1080/10635150802032990

Roy S, Tyagi A, Shukla V, Kumar A, Singh UM, Chaudhary LB, Datt B, Bag SK, Singh PK, Nair NK, Husain T, Tuli R (2010) Universal Plant DNA Barcode Loci May Not Work in Complex Groups: A Case Study with Indian *Berberis* Species. PLoS ONE 5: e13674. doi: 10.1371/journal.pone.0013674

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, de Bruijn AY, Sullivan S, Qiu Y-L (2000) Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences. Systematic Biology 49: 306–362. doi: 10.1093/sysbio/49.2.306

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1850–1811. doi: 10.1098/rstb.2005.1730

Simeone MC, Piredda R, Papini A, Vessella F, Schirone B (2013) Application of plastid and nuclear markers to DNA barcoding of Euro – Mediterranean oaks (*Quercus*, Fagaceae): problems, prospects and phylogenetic implications. Botanical Journal of the Linnean Society 172: 478–499. doi: 10.1111/boj.12059

Sundus T (2008) Authentication of medicinal plant material by DNA fingerprinting. World Review of Science, Technology and Sustainable Development 5: 151–160. doi: 10.1504/WRSTSD.2008.018558

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Vanherweghem J-L, Tielemans C, Abramowicz D, Depierreux M, Vanhaelen-Fastre R, Vanhaelen M, Dratwa M, Richard C, Vandervelde D, Verbeelen D, Jadoul M (1993) Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs. The Lancet 341: 387–391. doi: 10.1016/0140-6736(93)92984-2

von Crautlein M, Korpelainen H, Pietilainen M, Rikkinen J (2011) DNA barcoding: a tool for improved taxon identification and detection of species diversity. Biodiversity Conservation 20: 373–380. doi: 10.1007/s10531-010-9964-0

Wallace LJ, Boilard SMAL, Eagle SHC, Spall JL, Shokralla S, Hajibabaei M (2012) DNA barcodes for everyday life: Routine authentication of Natural Health Products. Food Research International 49: 446–452. doi: 10.1016/j.foodres.2012.07.048

Wang XR, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE (1999) Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*L, *mat*K, *rpl*20-*rps*18 spacer and *trn*V intron sequences. American Journal of Botany 86: 1742–1753. doi: 10.2307/2656672

WHO (2003) Guidelines on good agricultural and collection practices (GACP) for medicinal plants. World Health Organization, Geneva.

Wunder S (2001) Poverty alleviation and tropical forests – what scope for synergies? World Development 29: 1817–1833. doi: 10.1016/S0305-750X(01)00070-5

ZooKeys 365: 215–233 (2013)
doi: 10.3897/zookeys.365.5730
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Efficacy of the core DNA barcodes in identifying processed and poorly conserved plant materials commonly used in South African traditional medicine

Ledile T. Mankga[1], Kowiyou Yessoufou[1], Annah M. Moteetee[1], Barnabas H. Daru[1], Michelle van der Bank[1]

1 *African Centre for DNA Barcoding, Department of Botany and Plant Biotechnology, University of Johannesburg, PO Box 524, Auckland Park 2006, Johannesburg, South Africa*

Corresponding author: *Ledile T. Mankga* (lmankga@yahoo.com)

Academic editor: *Z. T. Nagy* | Received 1 June 2013 | Accepted 25 October 2013 | Published 30 December 2013

Citation: Mankga LT, Yessoufou K, Moteetee AM, Daru BH, van der Bank M (2013) Efficacy of the core DNA barcodes in identifying processed and poorly conserved plant materials commonly used in South African traditional medicine. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 215–233. doi: 10.3897/zookeys.365.5730

Abstract

Medicinal plants cover a broad range of taxa, which may be phylogenetically less related but morphologically very similar. Such morphological similarity between species may lead to misidentification and inappropriate use. Also the substitution of a medicinal plant by a cheaper alternative (e.g. other non-medicinal plant species), either due to misidentification, or deliberately to cheat consumers, is an issue of growing concern. In this study, we used DNA barcoding to identify commonly used medicinal plants in South Africa. Using the core plant barcodes, *mat*K and *rbc*La, obtained from processed and poorly conserved materials sold at the muthi traditional medicine market, we tested efficacy of the barcodes in species discrimination. Based on genetic divergence, PCR amplification efficiency and BLAST algorithm, we revealed varied discriminatory potentials for the DNA barcodes. In general, the barcodes exhibited high discriminatory power, indicating their effectiveness in verifying the identity of the most common plant species traded in South African medicinal markets. BLAST algorithm successfully matched 61% of the queries against a reference database, suggesting that most of the information supplied by sellers at traditional medicinal markets in South Africa is correct. Our findings reinforce the utility of DNA barcoding technique in limiting false identification that can harm public health.

Keywords

Core DNA barcodes, medicinal plants, species identification, South Africa

Introduction

Traditional medicine is regarded as the most famous health care system in the world (WHO 2002), likely because of its accessibility and popularity. Currently, over 80% of human population around the globe relies on medicinal plants for their daily fight for better health (WHO 2002). In Africa, access to modern medical treatment is very limited largely due to lack of facilities or, when hospitals exist; their services are unaffordable for the majority. As a result, medicinal plants are extensively used to meet people's needs for health care (Staden 1999, Hostettman et al. 2000, WHO 2002, Fyhrquist 2007, Koduru et al. 2007).

South Africa has a rich tropical and temperate flora, harbouring approximately 24,000 species, which account for more than 10% of the world's vascular plants (Germishuizen and Meyer 2003). Of this unique diversity, approximately 3000 species (~13%) are used as medicines, with a large number of them exported to other countries even outside Africa (Van Wyk et al. 1997).

In the recent past, harvesting medicinal plants was the domain of trained traditional healers, well known for their skills as herbalists or diviners who respected customary conservation practices (Cunningham 1993). Today, however, the gathering and trading of medicinal plants is no longer restricted to traditional healers but has entered informal commercial sectors of the South African economy, resulting in an increase in the number of herbal gatherers and traders (Dold and Cocks 2002). Mander (1998) recorded more than 100,000 traditional healers in South Africa. For example, in the Province of KwaZulu-Natal alone, between 20,000 and 30,000 people, mainly women, make their living from trade of non-timber forest products, particularly medicinal plants (Mander 1998). This intensive gathering of plants from the wild poses a serious threat to South Africa's rich biodiversity (Dold and Cocks 2002), increases risk of extinction (Hoareau and DaSilva 1999) and leads to scarcity of commonly used medicinal plants (Cunningham 1991, Mander 1997, 1998, Dold and Cocks 2002). Species such as *Ocotea bullata* (Burch.) Baill., *Warburgia salutaris* (G. Bertol.) Chiov. and *Bowie volubilis* Harv. ex Hook. f., which were once abundant, are now threatened with extinction due to over-harvesting in the wild (www.redlist.sanbi.org). In addition, some species such as *Cassine transvaalensis* (Burtt Davy) Codd, and *Erythrophleum lasianthum* Corbishley, are now becoming threatened also due to over-harvesting in the wild (Fennel et al. 2004). Given the increasing pressure on medicinal plants, there is a need for increasing commitment towards efficient controls and better practices that can help preserve medicinal plant diversity in South Africa.

To reach this objective, the primary step requires a reliable tool for accurate plant identification. Traditional plant identification is based on morphological characteristics, which can be problematic especially for medicinal plants that are mainly traded as dried or processed barks, dried leaves, roots, and stems (Figure 1) in popular markets known in South Africa as muthi market. As such, traded medicinal plants are devoid of identification diagnostics making morphologically-based identification non applicable

(Dold and Cocks 2002). Also, medicinal plants cover a broad range of taxa, which may be phylogenetically less related but morphologically very similar. Such similarity between species may lead to misidentification and inappropriate use (Chen et al. 2010). This is of high concern as it may cause fatalities especially given that several medicinal plants are poisonous (Watt and Breyer-Brandwijk 1962, Van Wyk et al. 2002, Bruni et al. 2010). For instance, WHO (2004) reported in Hong Kong, fourteen cases of accidental substitution of the roots of *Gentiana* and *Clematis* species with that of *Podophyllum hexandrum* Royle for their antiviral qualities due to similarity in the morphological features of their roots. Similarly, a serious case of cardiac arrhythmias was reported as a side effect, caused by the accidental substitution of plantain (*Plantago major* L.; used as dietary supplements) with *Digitalis lanata* Ehrh. (used for heart conditions; WHO 2004). In the early 2000's, large quantities of misidentified plantains were shipped to more than 150 manufacturers, distributors and retailers in the United States over a period of two years (WHO 2004). Another case of misidentification was in India, where mustard oil was accidentally contaminated with seeds of *Argemone mexicana* L., resulting in an epidemic of dropsy (WHO 2004). The misidentification of these seeds could have been avoided if there had been proper quality control of source materials (WHO 2004).

Given such alarming situations of misidentification, developing techniques to assist and support traditional plant identification (e.g. assigning dried barks, roots or leaves to species) is an urgent matter not only to preserve biodiversity and traditional knowledge attached to each plant (Yessoufou 2005) but also to secure human health (Chen et al. 2010). From this perspective, we propose that the use of DNA barcoding can assist in distinguishing species and assigning unidentified individuals or any plant organs or materials to species level (Kress et al. 2005, Kress and Erickson 2008, Lahaye et al. 2008, Kesanakurti et al. 2011). DNA barcoding is the use of a short gene sequence from a standardised region of the genome that could – in principle – distinguish between even closely related species (Hebert et al. 2004, Lahaye et al. 2008, Kesanakurti et al. 2011, Van der Bank et al. 2012). Ideally, DNA barcoding studies use fresh or well-preserved materials as sources of DNA. However, this is not always practical in many situations where DNA is already degraded because materials are either already processed or poorly preserved. Such situations include diet analyses (Huang 1972), ancient DNA studies (Pääbo et al. 2004), specimen identification from environmental DNA samples (Gratz 2004) and medicinal materials in muthi markets.

Two DNA regions were recently proposed as core barcodes, *rbc*La and *mat*K (CBOL 2009) with their identification efficacy estimated at 70–80% for land plants. The efficacy of DNA barcodes has rarely been evaluated for plant materials that are poorly stored or already processed; to our knowledge only one recent study has evaluated this with regards to animals where the discriminatory power of a mini-barcode was assessed in processed materials (Boyer et al. 2012). In this study, we focus on poorly conserved and processed medicinal plant materials sold in a South African muthi market with specific emphasis on commonly used plants. First, we constructed a DNA

barcode library for these medicinal plants using fresh materials. Second, we bought poorly conserved and processed materials sold at the muthi market, and tested the efficacy of the core barcodes in assigning these processed materials to their species using the DNA barcode library as the reference.

Material and methods

Taxon sampling

A total of 108 species belonging to 55 plant families were identified as commonly used medicinal plants in South Africa based on a literature survey (Hutchings et al. 1996, Van Wyk et al. 1997, Van Wyk and Gericke 2000) (see Appendix). We collected these plants from several localities in four Provinces in South Africa: Gauteng, Limpopo, Mpumalanga, and the Western Cape. Our sampling comprised 185 specimens (see Appendix). Collection details, taxonomy, voucher numbers, GPS coordinates, field pictures, and sequence data (*mat*K and *rbc*La) are archived online on the Barcode of Life Data Systems (BOLD) (www.boldsystems.org). The voucher specimens for all the taxa as well as GenBank and BOLD accession numbers are listed in the Appendix.

In addition, we included in this study, plant materials bought from the Faraday muthi market (henceforth muthi samples) in Johannesburg, South Africa. A muthi market is a popular market where trade and services in African traditional medicines are provided to the general public. Materials sold in this market include various plant parts such as dried or fresh leaves, seeds, barks, and roots, etc. (Figure 1). These materials are sometimes in poorly stored and/or processed states (e.g. powder). In total, we included 18 additional muthi samples in our sampling and recorded their vernacular names (mainly in isiZulu) as provided by the sellers. It was not possible to assign scientific names to the samples at the time of purchase as they were in poor condition or had already been processed.

DNA extraction, amplification, sequencing and alignment

Of the 108 species collected from the wild, leaf samples of 37 species were sent to the Canadian Centre for DNA Barcoding (CCDB) in Canada, where total DNA was extracted, the two core DNA barcodes (*mat*K and *rbc*La) were amplified and sequenced according to CCDB protocols. The sequencing for the remaining 71 species was done at the African Centre for DNA Barcoding (ACDB) in South Africa. The 18 muthi samples were also processed and sequenced at the ACDB.

DNA extraction followed the 2× CTAB method (Doyle and Doyle 1987). Polyvinyl pyrolidone (2% PVP) was added to reduce the effect of high polysaccharide concentration in the samples. After precipitating the DNA with 100% ethanol, it was stored at -20 °C for a minimum of two weeks (Fay et al. 1998). DNA extracts were

Figure 1. Examples of medicinal herbs bought at Faraday muthi market in Johannesburg **A** different medicinal herbs in bags **B** Seeds of *Entada rheedii* (tindili) **C** mixed herbs (fembo) **D** A twig of *Adenia gummifera* (mphinde umshaye) **E** Barks of *Vachellia* sp. (umkhanya-kute) **F** Bulb of *Boophane disticha* (umqotho) **G** mixed herbs **H** *Myrothamnus flabellifolius* (vuka) **I** Barks of *Vachellia* sp. (umkhanya-kute) **J** *Sarcostemma viminale* (ube nam) **K** Plant of *Clivia* sp. (mayime) **L** *Stangeria eriopus* (imfingo) **M** mixed herbs (isihlalakahle) **N** Tuber (umbonsi) **O** *Helichrysum* sp. (impepo) and **P** Twigs of *Synadenium cupulare* (umdletshane). Names in brackets are vernacular names in isiZulu.

purified using QIAquik silica columns (Qiagen Inc., Hilden, Germany) according to the manufacturers' protocol.

For both genes, PCR amplification was performed using ReadyMix Mastermix (Advanced Biotechnologies, Epson, Surrey, UK). We added 3.2% bovine serum albumin (BSA) to all reactions to serve as stabilizer for enzymes, to reduce problems with secondary structure, and improve annealing (Palumbi 1996). PCR amplification was performed using either the 9800 Fast Thermal Cycler or the GeneAmp PCR System 9700 machines. PCR programs used are as follows: (a) for *rbc*La, pre-melt at 94 °C for 60 sec, denaturation at 94 °C for 60 s, annealing at 48 °C for 60 s, extension at 72 °C for 60 s (for 28 cycles), followed by a final extension at 72 °C for 7 min, and (b) for *mat*K, the protocol consisted of pre-melt at 94 °C for 3 min, denaturation at 94 °C for 60 sec, annealing at 52 °C for 60 s, extension at 72 °C for 2 min (for 30 cycles), final extension at 72 °C for 7 min.

Cycle sequencing reactions were carried out in a GeneAmp PCR System 9700 thermal cycler using the ABI PRISM® BigDye® Terminator v3.1 (Applied Biosystems,

Inc., California, USA). Cycle sequencing products were precipitated in ethanol and sodium acetate to remove excess dye terminators. Then suspended into 10 µl HiDi formamide (ABI) before sequencing on a ABI 3130 *xl* Genetic Analyzer (ABI).

Complementary strands were assembled and edited using Sequencher v3.1 (Gene Codes, Ann Arbor, Michigan, USA). All the sequences generated at ACDB and CCDB including those retrieved from BOLD were aligned manually in PAUP* v4.0b10 (Swofford 2002).

Data analyses

All analyses were conducted in the R package Spider (Brown et al. 2012). Only species for which sequences of both genes (*rbc*La and *mat*K) were available were included in the analyses. First, we evaluated K2P-interspecific and intraspecific genetic distances using Wilcoxon's sum rank test and the significance of the differences between both distances was tested. Second, we determined the genetic distance suitable as threshold with which to test the efficacy of the DNA regions in assigning sequences to species. Third, we tested the identification efficacy used medicinal plants using three distance-based methods: best close match (Meier et al. 2006), near neighbour, and species identification methods used by BOLD (www.boldsystems.org). The best close match and near neighbour analyses measure the identification efficacy by searching for the closest individuals; the former focuses on a single nearest neighbour match, whereas the latter considers all matches within a specific threshold. The BOLD species identification method performed species delimitation based on a distance cut-off of 1%.

We then evaluated the ability of the core DNA barcodes in assigning poorly conserved or already-processed plant materials to species. For this test, the barcoding technique was applied on all 18 muthi samples. Our procedure here consisted of two steps. The first involved the use of vernacular names (in isiZulu) for the muthi samples to identify their scientific names based on Hutchings et al. (1996). The second step was based on the BLAST algorithm implemented in the BOLD identification system (www.boldsystems.org/index.php/IDS_OpenIdEngine) for *mat*K and *rbc*L sequences. The BLAST algorithm measures the efficiency of species identification against a global data repository such as BOLD or GenBank (Munch et al. 2008). The program takes a query of the sequence and matches it against the database selected by the user. The E-value and maximum identity are two statistics that can be used to measure the efficiency of species identification. The results are reported in a rank list whereby the closer the hit is to 100% and the E-value to 0, the better the identification efficiency. The DNA sequences generated from the 18 poorly conserved and degraded muthi samples were BLASTed against the reference database of medicinal plants available on the BOLD system. For additional evidence to the BLAST test, we included the sequences of muthi samples (as queries) in the database of DNA matrix

generated for all medicinal plants, and reconstructed a maximum parsimony (MP) phylogeny based on the combined DNA matrix. Our objective here was to trace on the phylogeny, the positions of muthi samples (our queries) along the phylogenetic tree. Support for the groupings was analysed using bootstrapping. Maximum parsimony analysis was performed using PAUP* v4.0b10 (Swofford 2002). Tree searches were done using heuristic searches with 1000 random sequence additions but keeping only 10 trees. Tree bisection-reconnection was performed with all character transformations treated as equally likely i.e. Fitch parsimony (Fitch 1971). Bootstrap resampling (Felsenstein 1985) was done also in PAUP* v4.0b10 (Swofford 2002). Node support was assessed based on the following scale: BS 50–74% (weak bootstrap support) and 75-100% for strong support (Hillis and Bull 1993, Murphy et al. 2001, Daru et al. 2013).

Results

Based on genetic divergence, *rbc*La exhibits the lowest mean interspecific distance (0.08); in contrast, *mat*K exhibits the highest mean interspecific distance, which almost doubles that of *rbc*La + *mat*K (0.22 versus 0.119 respectively). From the genetic variation test based on K2P-distance for *mat*K, we found that interspecific distance was significantly higher than intraspecific ($inter_{median}$ = 0.232 vs. $intra_{median}$ = 0.00; Wilcoxon sum rank test, $p < 0.001$; Table 1), indicating that a barcode gap exists for *mat*K. Also, a similar pattern was found for *rbc*La, high significant difference between inter- and intraspecific distances ($inter_{median}$ = 0.07 vs. $intra_{median}$ = 0.001, $p < 0.001$). We also found that when *rbc*La and *mat*K were combined the interspecific distance was significantly higher than intraspecific distance ($inter_{median}$ = 0.12 vs $intra_{median}$ = 0.00, $p < 0.001$). Furthermore, our analyses indicate that a clear barcode gap exist between the range of intra- versus interspecific distances for all regions (Figure 2).

The Tajima's K index of sequence was divergence measured as the mean number of substitutions per nucleotide which indicates that *mat*K had the lowest sequence divergence (3%) whereas *rbc*La and *rbc*La + *mat*K had similar divergence indices of 6% and 5% respectively.

We calculated the optimised genetic distance (threshold) with which the discriminatory power for different gene regions was evaluated. Apart from *rbc*La for which the optimised threshold was lower than 1%, both *mat*K and *rbc*La + *mat*K had optimised thresholds greater than 1% (i.e. 1.44% and 1.25% respectively). Using these cut-offs, we then evaluated the discriminatory power of different regions. We found that the combination *rbc*La + *mat*K provided the best discriminatory power based on the near neighbour and the best close match methods (96% and 97% respectively, Table 2). However, using the BOLD identification criteria, the discriminatory power of the combined regions dropped to 85% which is close to 86% for *mat*K alone but higher than that of *rbc*La (76%). Also, the application of

Table 1. Summary statistics indicating the range and means of intra- and interspecific distances for the gene regions and combination tested.

DNA regions	Numbers of sequences	Sequence length	K	Range inter	Mean inter (±SD)	Range intra-	Mean intra (±SD)	Threshold (%)
*rbc*La	141	552	0.06	0–0.16	0.080±0.022	0–0.004	0.0002±0.0007	0.63
*mat*K	140	915	0.03	0–0.51	0.220±0.066	0–0.012	0.0008±0.0022	1.44
*rbcL*a+*mat*K	140	1467	0.05	0–0.33	0.119±0.035	0–0.109	0.0039±0.0196	1.25

Table 2. Efficacy of DNA barcodes in identification of commonly used medicinal plants in South Africa.

DNA regions	Near Neighbour		BOLD (1%)				Best close match			
	False (%)	True (%)	Ambiguous (%)	Correct (%)	Incorrect (%)	No ID (%)	Ambiguous (%)	Correct (%)	Incorrect (%)	No ID (%)
*rbc*La	5	95	23	76	0	1	6	93	1	0
*mat*K	7	93	10	86	1	3	4	92	1	3
*rbc*La + *mat*K	4	96	11	85	1	3	0	97	2	1

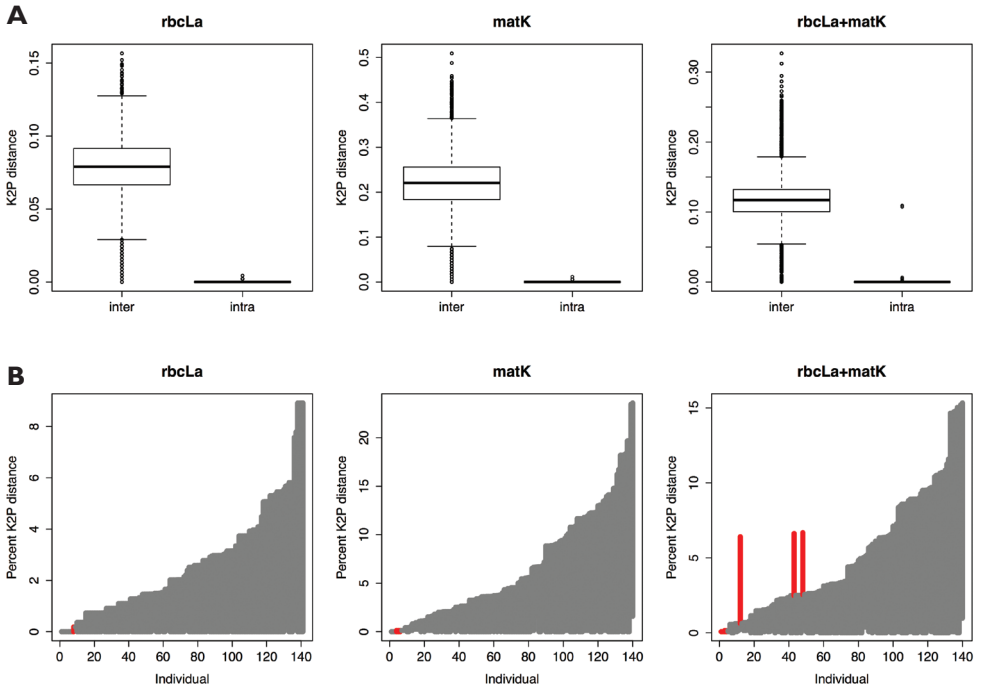


Figure 2. Evaluation of barcode gaps in *mat*K, *rbc*La and *rbc*La + *mat*K for commonly used medicinal plants of South Africa. **A** Boxplots indicate the genetic variation between interspecific distance and intraspecific distance; the boxplots clearly shows significant differences between inter- and intraspecific distances for all gene regions tested ($P < 0.001$; see text) **B** Lineplot of the barcode gap for the commonly used plants in South African medicine. For each gene region, the grey lines correspond to the furthest intraspecific distance (bottom of line value), and the closest interspecific distance (top of line value). The red lines show where this relationship is reversed, i.e. cases where there is no barcode gap.

BOLD identification criteria results in higher proportion of ambiguous identification: *rbc*La (23%), *mat*K (10%) and *rbc*La + *mat*K (11%). Conversely, the best close match method had the lowest proportion of ambiguous identification (i.e. 0–7%) for all regions tested.

We then BLASTed (compared) the sequences for the 18 poorly conserved and degraded muthi samples against the BOLD identification system. Two muthi samples proved difficult to amplify whereas the amplification was successful for the 16 remaining muthi samples (Table 3). Of the 16 samples, the BLAST test was successful for 11 samples (61%), indicating that the scientific names recovered from BLAST test matched perfectly the scientific names expected based on vernacular names. However, we found mismatches for five samples. These results were also indicated on the MP phylogeny presented in Figure 3.

Table 3. Comparison of BLAST results against common and scientific names for the muthi samples. – indicates specimens for which PCR failed. **?** indicates specimens for which common names or scientific names could not be found in the available literature. IUCN redlist obtained from http://redlist.sanbi.org

Common names from "muthi" market	Common names from literature (in isiZulu; Hutchings et al. 2006)	Scientific names (Hutchings et al. 2006)	APG III (2009) Family	IUCN red list	BLAST sequence similarity - BOLD %	Do BLAST results match the correct scientific names?
1. *impepo*	*impepo*	*Helichrysum* sp.	Asteraceae	-	100	True
2. *isihlalakahle*	?	*Harworthia limifolia*	Asphodelaceae	Vulnerable (VU)	-	Amplification failed
3. *fembo*	?	?	?	-	-	Amplification failed
4. *mkhanya kute*	*umkhanya-kute, umdlovune, umhlofunga, umhlosinga*	*Vachellia xanthophloea*	Fabaceae	Least Concern (LC)	98	True
5. *tindili*	*umbhone, tindili*	*Entada rheedii*	Fabaceae	Least Concern (LC)	99	True
6. *ubhonsi*	?	*Mappia racemosa*	Icacinaceae	Vulnerable (VU)	89	False
7. *ukhanyakute*	*umkhanya-kute, umdlovune, umhlofunga, umhlosinga*	*Vachellia xanthophloea*	Fabaceae	Least Concern (LC)	99	True
8. *umlilo*	*uzililo, ililo elikhulu*	*Stapelia gigantea*	Apocynaceae	Least Concern (LC)	100	True
9. *malilisa*	?	*Holarrhena pubescens*	Apocynaceae	Least Concern (LC)	100	False
10. *umhlahlawehlathi*	*umhluswane*	*Cissus reniferus*	Vitaceae	-	97	False
11. *mphinde umshaye*	?	*Adenia gummifera*	Passifloraceae	-	-	False
12. *ube nam*	*umbelebele, umpelepe*	*Sarcostemma viminale*	Asclepiadaceae	Least Concern (LC)	100	False
13. *isikholokhota*	*isikholokhoto*	*Sanseviera hyacinthiodes*	Asparagaceae	Least Concern (LC)	99	True
14. *mayime*	*umayime*	*Clivia miniata*	Amaryllidaceae	Vulnerable (VU)	99	True
15. *umdletshane*	*umbulele, umdletshane*	*Synadenium cupulare*	Euphorbiaceae	Least Concern (LC)	100	True
16. *vuka*	*vuka*	*Myrothamnus flabellifolius*	Myrothamnaceae	DDT	100	True
17. *umqotho*	*incotha, incwadi*	*Boophane disticha*	Amaryllidaceae	Declining	100	True
18. *imfingo*	*imfingo*	*Stangeria eriopus*	Stangeriaceae	Vulnerable (VU)	100	True

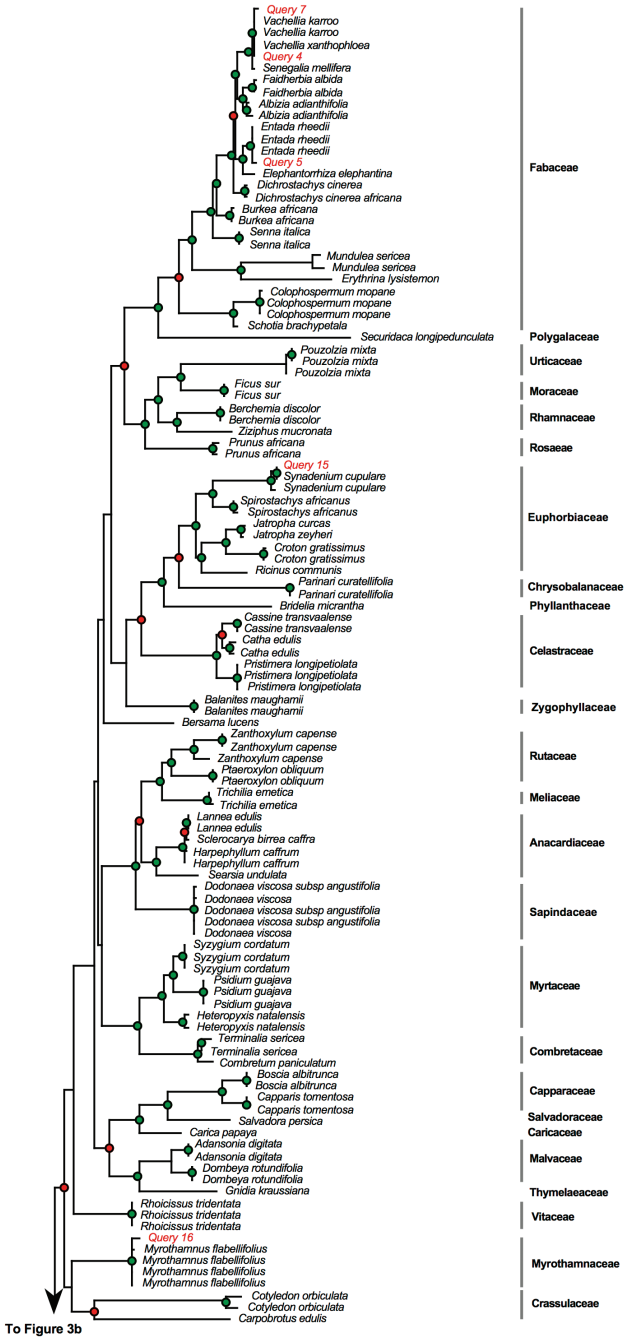


Figure 3. Phylogram obtained from the maximum parsimony analysis of *mat*K with muthi samples included as "query". Green dots indicate well-supported nodes (bootstrap support > 74%) and red dots indicate low bootstrap support (BS < 74%).Phylogram obtained from the maximum parsimony analysis of *mat*K with muthi samples included as "query". Green dots indicate well-supported nodes (bootstrap support > 74%) and red dots indicate low bootstrap support (BS < 74%).

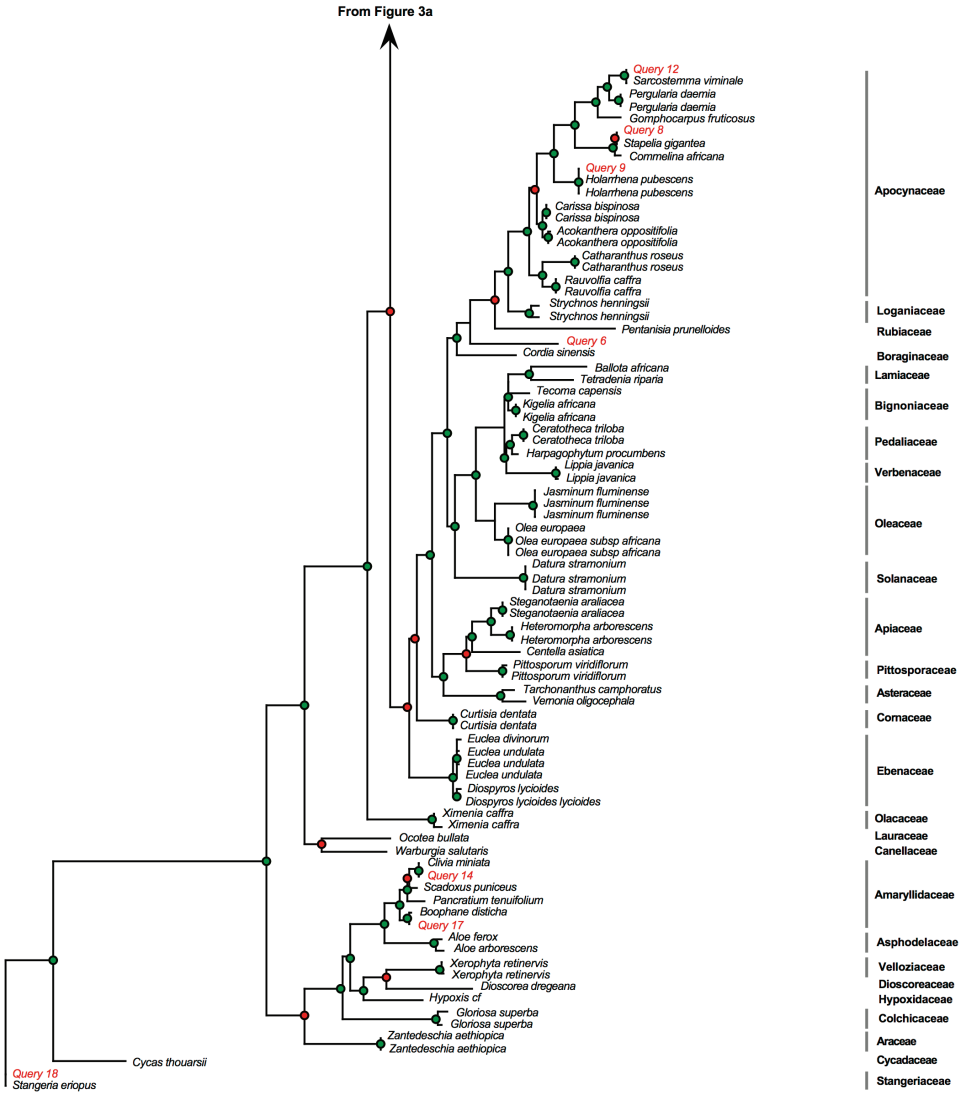


Figure 3. Continued.

Discussion

The efficiency of a good barcode relies fundamentally on its ability to distinguish between closely related species. This is achieved only when there is enough genetic differentiation between rather than within species, i.e. when interspecific distance is significantly higher than intraspecific distance (Hebert et al. 2004, Savolainen et al. 2005, Lahaye et al. 2008). We tested this expectation on commonly used medicinal plants using *mat*K and *rbc*La. We found that both regions (*mat*K and *rbc*La) exhibit a

significant barcode gap, suggesting that they should be efficient in assigning processed medicinal plants to species level. Further, the performance of each gene was very high for single and core barcodes (76–97%) but highest for the core under near neighbour and best close match methods. Overall, the core barcodes proves reliable in identifying commonly used medicinal plants of South Africa.

In several studies, the discriminatory power of the core barcodes has been questioned (Hollingsworth et al. 2009, Pettengill and Neel 2010, Roy et al. 2010, Wang et al. 2010, Clement and Donoghue 2012, Liu et al. 2012). These studies mainly focused on closely related species or single lineages. A recent study with a similar objective to ours also discounts the potential of the core barcodes in discriminating Chinese medicinal plants (Chen et al. 2010). The authors found a more reliable discriminatory power of 92.7% for ITS2 at the genus and species level from different plant families and closely related taxa. In our study, we did not include ITS2, but we found a similar power of 85% to 96% for the core barcodes (*mat*K and *rbc*La) in the context of South African commonly used medicinal plants. Chen et al. 2010 included 400 samples belonging to 326 species in 98 families covering dicots, monocots, gymnosperms and ferns of Chinese medicinal plants. Such broad sampling likely increased the probability of high proportion of closely related species, resulting in the low performance of the core barcodes in their study. However, our sampling size is limited to only commonly used medicinal plants (~108 species), and this restriction likely increases the chance of having less related species, leading to a higher performance we found for the core barcodes.

We further tested, the performance of the core barcodes by evaluating their identification efficacy on 18 medicinal plant products bought at the Faraday muthi market in Johannesburg, South Africa. The sequences generated from these 18 plant materials were BLASTed against the reference library on BOLD database system. Given that the plant materials sold at the muthi market were poorly conserved (dried, processed, etc.), we expected a very low percentage of DNA recovery and amplification. Possible explanation for the five samples that yielded false identification, and the two that failed are that the samples could be a mixture of leaves from multiple species. Such limitation could be overcome using individual sequencing of all components of mixed DNA samples based on high throughput sequencing techniques e.g. pyrosequencing technology, which is capable of simultaneously detecting many thousands of different sequences in a mixed sample, without the need for sub-cloning (Margulies et al. 2005).

Another possibility for the amplification failure observed in our study for some samples could be attributable to a bad post harvest condition of preservation, which may result in DNA degradation. Again, such limitation could be overcome through the search of a 'mini-barcode' (Meusnier et al. 2008, Boyer et al. 2012). The technique of sliding window analysis is now available for that purpose and has been proven reliable (Boyer et al. 2012). Given that medicinal plants are often poorly conserved or processed materials, the chance of successful extraction and amplification of long DNA fragments (> 200 bp) is very low (Meusnier et al. 2008, Boyer et al. 2012). As such, a

search for shorter and informative fragment is necessary if we are to verify the identity of commonly used medicinal plants which are generally devoid of morphological features. Furthermore, we found some mismatch in species identification by the BLAST algorithm and the corresponding species based on vernacular names. Although, South African medicinal plants are well documented (e.g. Hutchings et al. 1996, Van Wyk et al. 1997), it remains highly likely that the mismatch might not be an artefact of erroneous claims from plant sellers, but presumably due to the variation of names used for the same plants across different ethnic groups.

The continual removal of medicinal plants from the wild has become worrisome in southern Africa (Setshogo and Mbereki 2011). Therefore, understanding the scarcity and popularity of plants at the muthi market is the starting point for conservation and evaluating threatened species (Williams et al. 2000, Setshogo and Mbereki 2011). For instance, Williams et al. (2000) mentioned *Helichrysum* sp. as being scarce and threatened in the future because of its popularity and demand at the muthi markets. The harvesting of the whole plant, bulb, tuber or roots before the seeds germinate damages the plant more than harvesting only leaves, seeds, bark or fruits (as seen in Figure 1). Although only about 22% of the muthi samples are currently threatened with extinction (Table 3), continual over-exploitation in the wild might eventually change the status for currently non-threatened species to threatened category. Therefore, there is an urgent need to conserve medicinal plants by cultivating them at home gardens (Setshogo and Mbereki 2011).

In conclusion, our analyses indicate that most of the information supplied by the sellers at the muthi market were correct. This could be due to the fact that we tested only 18 samples. Therefore, it remains possible that if we increase our sample size, we might detect important mismatch between the sellers' claims and the products sold. We also propose a continued effort to increase the barcode library of South African medicinal plants, and in case of difficulties due to degraded materials, a pyro-sequencing technique in tandem with mini-barcodes is necessary. Our suggestions and findings are expected to be of great use in limiting false identification that can harm public health.

Acknowledgements

We thank the Canadian Center for DNA Barcoding for sequencing support. The National Research Foundation (NRF) South Africa and the International Development Research Centre (IDRC), Canada are greatly acknowledged for funding. This project is also partly funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (2008-OGI-ICI-03). Particular thanks go to the sellers at the Faraday muthi market in Johannesburg and Mr Stanley Khumalo for the isiZulu language translation. We thank two anonymous reviewers who provided valuable comments on an earlier draft of the manuscript.

References

Angiosperm Phylogeny Group III (2009) An update of APG classification for the orders and families of flowering plants. Botanical Journal of Linnean Society 161: 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

Boyer S, Brown SD, Collins RA, Cruickshank RH, Lefort M, Malumbres-Olarte J, Wratten SD (2012) Sliding window analyses for optimal selection of mini-barcodes and application to 454-pyrosequencing for specimen identification from degraded DNA. PLoS ONE 7: e38215. doi: 10.1371/journal.pone.0038215

Brown SDJ, Collins RA, Boyer S, Lefort MC, Malumbres-Olarte J, Vink CJ, Cruickshank RH (2012) Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources 12: 562–565. doi: 10.1111/j.1755-0998.2011.03108.x

Bruni I, De Mattia F, Galimberti A, Galasso G, Banfi E, Casiraghi M, Labra M (2010) Identification of poisonous plants by DNA barcoding approach. International Journal of Legal Medicine 124: 595–603. doi: 10.1007/s00414-010-0447-3

CBOL Plant Working Group (2009) A DNA Barcode for land plants. Proceedings of the National Academy of Sciences of the USA 106: 12794–12797. doi: 10.1073/pnas.0905845106

Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo P, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5: e8613. doi: 10.1371/journal.pone.0008613

Clement WL, Donoghue MJ (2012) Barcoding success as a function of phylogenetic relatedness in *Viburnum*, a clade of woody angiosperms. BMC Evolutionary Biology 12: 73. doi: 10.1186/1471-2148-12-73

Cunningham AB (1991) The herbal medicine trade: Resource depletion and environmental management for a hidden economy. In: Preston-whyte E, Rogerson C (Eds) South Africa informal economy, chap. 12. Oxford University Press, Cape Town, 196–206.

Cunningham AB (1993) African medicinal plants: setting priorities at the interface between conservation and primary health care. People and plants working paper 1. UNESCO, Paris. http://unesdoc.unesco.org/images/0009/000967/09670.pdf

Daru BH, Manning JC, Boatwright JS, Maurin O, Maclean N, Schaefer H, Kuzmina M, Van der Bank M (2013) Molecular and morphological analysis of subfamily Alooideae (Asphodelaceae) and the inclusion of *Chortolirion* in *Aloe*. Taxon 62: 62–76. http://www.ingentaconnect.com/content/iapt/tax/2013/00000062/00000001/art00006

Dold AP, Cocks ML (2002) The trade in medicinal plants in the Eastern Cape Province, South Africa. South African Journal of Science 98: 589–597.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19: 11–15.

Fay MF, Bayer C, Alverson WS, De Bruijn A, Chase MW (1998) Plastid *rbcL* sequence data indicate a close affinity between *Diegodendron* and *Bixa*. Taxon 47: 43–50. doi: 10.2307/1224017

Felsenstein J (1985) Confidence levels on phylogenies: an approach using the bootstrap. Evolution 39: 783–791. doi: 10.2307/2408678

Fennel CW, Light ME, Sparg GI, Stafford GI, Van Staden J (2004) Assessing African medicinal plants for efficacy and safety: agricultural and storage practices. Journal of Ethnopharmacology 95: 113–121. doi: 10.1016/j.jep.2004.05.025

Fitch WM (1971) Towards defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20: 406–416. doi: 10.1093/sysbio/20.4.406

Fyhrquist A (2007) Traditional medicinal uses and biological activities of some plant extracts of Africa *Combretum* Loefl., *Terminalia* L. and *Pteleopsis* Engl. species (Combretaceae). PhD thesis, Yliopistopaino, Helsinki.

Germishuizen G, Meyer NL (2003) Plants of southern Africa: An annotated checklist. Strelitzia 14. National Botanical Institute, Pretoria.

Gratz NG (2004) Critical review of the vector status of *Aedes albopictus*. Medical and Veterinary Entomology 18: 215–227. doi: 10.1111/j.0269-283X.2004.00513.x

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) 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 USA 101: 14812–14817. doi: 10.1073/pnas.0406166101

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42: 182–192. doi: 10.1093/sysbio/42.2.182

Hoareau L, DaSilva EJ (1999) Medicinal plants: a re-emerging health aid. Journal of Biotechnology 2: 717–3458. http://www.scielo.cl/pdf/ejb/v2/art02.pdf

Hollingsworth ML, Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources 9: 439–457. doi: 10.1111/j.1755-0998.2008.02439.x

Hostettman K, Marston A, Ndjoko K, Wolfender JL (2000) The potential of African plants as a source of drugs. Current Organic Chemistry 4: 973–1010. doi: 10.2174/1385272003375923

Huang YM (1972) Contributions to the mosquito fauna of Southeast Asia. XIV. The subgenus *Stegomyia* of *Aedes* in Southeast Asia. I- The *Scutellaris* group of species. Contributions of the American Entomological Institute 9: 1–109. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA510169

Hutchings A, Scott AH, Lewis G, Cunnignham AB (1996) Zulu medicinal plants. University of Natal Press, Scottsville, South Africa.

Kesanakurti PR, Fazekas AJ, Burgess KS, Percy DM, Newmaster SG, Graham SW (2011) Spatial patterns of plant diversity below ground as revealed by DNA barcoding. Molecular Ecology 20: 1289–1302. doi: 10.1111/j.1365-294X.2010.04989.x

Koduru S, Grierson DS, Afolayan AJ (2007) Ethnobotanical information of medicinal plants used for the treatment of cancer in the Eastern Cape province, South Africa. Current Science 92: 906–908.

Kress WJ, Erickson DL (2008) A two-locus global DNA barcode for land plants: The coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. PLoS ONE 2: e508. doi: 10.1371/journal.pone.0000508

Kress WJ, Wurdack KJ, Zimmer EA, Weight IA, Jazen DH (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the USA 102: 8369–8374. doi: 10.1073/pnas.0503123102

Lahaye R, Van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the floras of biodiversity hotspot. Proceedings of the National Academy of Sciences of the USA 105: 2923–2928. doi: 10.1073/pnas.0709936105

Liu C, Shi L, Xu X, Li H, Xing H, Liang D, Jiang K, Pang X, Song J, Chen S (2012) DNA barcode goes two-dimensions: DNA QR code web server. PLoS ONE 7: e35146. doi: 10.1371/journal.pone.0035146

Mander M (1997) Medicinal plant marketing in Bushbuckridge and Mpumalanga: A market survey and recommended strategies for sustaining the supply of plants in the region. Unpublished report, Danish Cooperation for Environment and Development, Danish Environment Protection Agency, Strandgade.

Mander M (1998) Marketing of indigenous plants in South Africa. A case study in KwaZulu-Natal. Food and Agriculture Organization, Rome. http://www.fao.org/docrep/w919/w91900.htm

Margulies M, Egholm M, Ahman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactor. Nature 437: 376–380. doi: 10.1038/nature03959

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Meusnier I, Singer GAC, Landry J, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9: 214. doi: 10.1186/1471-2164-9-214

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical signment of DNA sequences using Bayesian phylogenetics. Systematic Biology 57: 750–757. doi: 10.1080/10635150802422316

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348–2351. doi: 10.1126/science.1067179

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. (2004) Genetic analyses from ancient DNA. Annual Review of Genetics 38: 645–79. doi: 10.1146/annurev.genet.37.110801.143214

Palumbi SR (1996) Nucleic acids II: the polymerase chain reaction. In: Hillis DM, Moritz C, Mable BK. Molecular Systematics, Second Edition. Sinauer & Associates Inc. Publishers, Sunderland, 241–246.

Pettengill JB, Neel MC (2010) An evaluation of candidate plant DNA barcodes and assignment methods in diagnosing 29 species in the genus *Agalinis* (Orobanchaceae). American Journal of Botany 97: 1381–1406. doi: 10.3732/ajb.0900176

Roy S, Tyagi A, Shulka V, Kumar A, Singh UM, Chaudhary LB, Datt B, Bag SK, Singh PK, Nair NK, Husain T, Tuli R (2010) Universal plant DNA barcode loci may not work in complex groups: a case study with Indian *Berberis* species. PLoS ONE 5: e13674. doi: 10.1371/journal.pone.0013674

Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopedia of life: an introduction to DNA barcoding. Philosophical Transactions of the Royal Society 360: 1805–1811. doi: 10.1098/rstb.2005.1730

Setshogo MP, Mbereki CM (2011) Floristic diversity and uses of medicinal plants sold by street vendors in Gaborone, Botswana. The African Journal of Plant Science and Biotechnology 5(1): 69–74.

Staden JV (1999) Medicinal plants in southern Africa: utilization, sustainability, conservation – can we change mindsets? Outlook on Agriculture 28: 75–76.

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). 10 Ed. Sinauer Associates, Sunderland, Massachusetts.

Van der Bank HF, Greenfield R, Daru BH, Yessoufou K (2012) DNA barcoding reveals microevolutionary changes and river system-level phylogeographic resolution of seven populations of African silver catfish, *Schilbe intermedius* (Siluriformes, Schilbeidae). Acta Ichthyologica et Piscatoria 42: 307–320. doi: 10.3750/AIP2012.42.4.04

Van Wyk B-E, Gericke N (2000) People's plants: A guide to useful plants of southern Africa. Briza Publications, Pretoria.

Van Wyk B-E, Van Heerden F, Van Oudtshoorn B (2002) Poisonous plants of South Africa. Briza Publications, Pretoria, South Africa.

Van Wyk B-E, Van Oudtshoorn B, Gericke N (1997) First Edition. Medicinal plants of South Africa. Briza Publications, Pretoria.

Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J (2010) DNA barcoding of the Lemnaceae, a family of aquatic monocots. BMC Plant Biology 10: 205. doi: 10.1186/1471-2229-10-205

Watt JM, Breyer-Brandwijk MG (1962) Second Edition. The Medicinal and Poisonous plants of Southern and Eastern Africa. Livingstone, London.

Williams VL, Balkwill K, Witkowski ETF (2000) Unraveling the commercial market for medicinal plants and plant parts on the Witwatersrand, South Africa. Economic Botany 54: 310–327. doi: 10.1007/BF02864784

World Health Organization (2002) WHO launches the first global strategy on traditional and alternative medicine. Geneva, Switzerland. http://www.who.int/mediacentre/news/releases/release38/en/

World Health Organization (2004) Medicinal plants – guidelines to promote patient safety and plant conservation for a US$ 60 billion industry. http://www.who.int/mediacentre/news/notes/2004/np3/en/

Yessoufou K (2005) Ecological and ethnobotanical research on *Irvingia gabonensis* and *Blighia sapida* in Plateau Province, Eastern Benin. MSc thesis, University of Abomey-Calavi, Benin.

Appendix

List of taxa with voucher information. (doi: 10.3897/zookeys.365.5730.app) File format: Microsoft Word file (docx).

Explanation note: List of taxa with voucher information, BOLD and GenBank accession numbers for each DNA region. English common names have been chosen. In cases where English common names were unavailable, names in native languages were used.

Citation: Mankga LT, Yessoufou K, Moteetee AM, Daru BH, van der Bank M (2013) Efficacy of the core DNA barcodes in identifying processed and poorly conserved plant materials commonly used in South African traditional medicine. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 215–233. doi: 10.3897/zookeys.365.5730 List of taxa with voucher information. doi: 10.3897/zookeys.365.5730.app

ZooKeys 365: 235–244 (2013)
doi: 10.3897/zookeys.365.5905
www.zookeys.org

SHORT COMMUNICATION

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Using DNA barcoding to differentiate invasive *Dreissena* species (Mollusca, Bivalvia)

Jonathan Marescaux[1], Karine Van Doninck[1]

1 *Laboratory of Evolutionary Genetics and Ecology, Research Unit in Environmental and Evolutionary Biology, Department of Biology, University of Namur, 61 rue de Bruxelles, 5000 Namur, Belgium*

Corresponding author: *Jonathan Marescaux* (jonathan.marescaux@unamur.be)

Academic editor: *M. De Meyer* | Received 2 July 2013 | Accepted 16 October 2013 | Published 30 December 2013

Citation: Marescaux J, Van Doninck K (2013) Using DNA barcoding to differentiate invasive *Dreissena* species (Mollusca, Bivalvia). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 235–244. doi: 10.3897/zookeys.365.5905

Abstract

The zebra mussel (*Dreissena polymorpha*) and the quagga mussel (*Dreissena rostriformis bugensis*) are considered as the most competitive invaders in freshwaters of Europe and North America. Although shell characteristics exist to differentiate both species, phenotypic plasticity in the genus *Dreissena* does not always allow a clear identification. Therefore, the need to find an accurate identification method is essential. DNA barcoding has been proven to be an adequate procedure to discriminate species. The cytochrome *c* oxidase subunit I mitochondrial gene (COI) is considered as the standard barcode for animals. We tested the use of this gene as an efficient DNA barcode and found that it allow rapid and accurate identification of adult *Dreissena* individuals.

Keywords

COI, zebra mussel, quagga mussel, barcoding gap, RFLP

Introduction

Biological invasions are a topical issue in today's world since they are the biggest threat to biodiversity after habitat destruction. The first, and probably the biggest, problem for scientists is to deal with widely divergent perceptions of the criteria defining "invasive" species (Colautti and MacIsaac 2004). In the management and policy field, such

species are defined as "alien species whose introduction does, or is likely to, cause economic or environmental harm or harm to human health" (Invasive Species Advisory Committee 2006). By cons, from a strict scientific point of view, an invasive species is "an exotic species that present a tendency to spread with high densities from its point of introduction" (Vermeij 1996, Beisel and Lévêque 2010). A second problem for both scientists and managers is to rapidly characterize a new invasion.

The zebra mussel (*Dreissena polymorpha* (Pallas, 1771) and the quagga mussel (*Dreissena rostriformis bugensis* Andrusov, 1897) are invasive freshwater bivalves in Europe and North America (Mills et al. 1996, Son 2007). Both species are native to the Ponto-Caspian area (Son 2007) and have major negative ecological and economic impacts such as biofouling and food web alteration (Sousa et al. 2013). Several studies have shown that the newly introduced quagga mussel can often dominate well-established zebra mussel populations within only a few years and even outcompete it in some cases (Wilson et al. 2006, Heiler et al. 2012). Wilke et al. (2010) showed that, in addition to the well-known zebra and quagga mussels, two others *Dreissena* species native to the Balkans (*D. presbensis* (Kobelt, 1915) and *D. blanci* Westerlund, 1890) begin to expand in the area and may be potentially invasive in Europe.

Although *Dreissena* specialists may discriminate adults of the different species based on internal and external shell features (Pathy and Mackie 1993, Mills et al. 1996, Sablon et al. 2010), this task remains difficult for managers. It becomes even more problematic when identifying larvae, which is the most invasive form of *Dreissena* (Marescaux et al. 2012a, b). For example, the invasion of the Meuse River in Belgium by the quagga mussel remained undetected because Belgian national agencies never made the distinction with the zebra mussel. Therefore, tools for rapid identification of both adult specimens and larvae are needed in order to detect newly invaded habitats. DNA barcoding has been proven to be an effective method both for species detection and to assign new specimens to already identified species (Hebert et al. 2003a, Birky et al. 2010). Here we amplified part of the cytochrome *c* oxidase subunit I (COI) mitochondrial gene, the most-widely utilized gene for animal DNA barcoding (Consortium for the Barcode of Life 2013) and we tested four delimitation metrics to differentiate *Dreissena* species. We also demonstrate that restriction fragment length polymorphism (RFLP) could be used as an inexpensive method to distinguish between zebra and quagga mussel.

Methods

Samples collection

Dreissena samples were collected in the Meuse River (see Marescaux et al. 2012a, b for sampling protocol and locations). The mussels were collected in the littoral zone of the river bank from stones which were picked up manually from a depth of 30–40 cm.

COI sequencing

Total genomic DNA was extracted from 241 *Dreissena* individuals using the «DNeasy Blood and Tissue» kit (Qiagen) according to manufacturer guidelines. To minimize cost, DNA extraction with the CTAB (hexadecyltrimethylamoniumbromide) protocol proposed by Winnepenninckx et al. (1993) could also be used. A fragment of 654 base pairs (bp) of the COI mitochondrial gene was amplified using universal primers (Folmer et al. 1994). Amplifications were performed in 25 µl total volume including 0.5 or 1 µl of gDNA, 1× GoTaq Green reaction buffer (Promega), 200 µM of dNTPs (Promega), 0.5 µM of both primers and 0.1 U of GoTaq DNA polymerase (Promega). PCR cycling conditions were as follows: an initial step of 94 °C for 4 min, followed by 30 cycles of 94 °C for 45 s, 45 °C for 45 s and 72 °C for 45 s, and then a final extension of 72 °C for 10 min. DNA sequencing was performed by the Genoscreen Company (France). Sequences were visualized and aligned using BioEdit v7.0.5.3 (Hall 1998).

Phylogenetic analysis

Sequences were collapsed into unique haplotypes using DnaSP (Librado and Rozas 2009). In order to determine the number of *Dreissena* species in the Meuse River we tested three barcoding methods: (i) the "Operational Taxonomic Units" (OTU) (Hebert et al. 2003a), (ii) the "Automatic Barcode Gap Discovery" (ABGD) (Puillandre et al. 2012), and (iii) the "K/θ method" (4 × rule) (Birky et al. 2010). The K/θ method specifies that if the genetic distance between clusters is higher than 4 times the genetic distance within the cluster then species are distinct (Birky et al. 2010, Tang et al. 2012). Neighbour-Joining (NJ) trees and matrix of pairwise distances were calculated using the Kimura 2-parameter (K2P) model and were generated using MEGA4 in order to define OTU's (Tamura et al. 2007). Sequences found in GenBank (Table 1) were used to construct a haplotype network using Network v4.6 (Bandelt et al. 1999).

Restriction fragment length polymorphism analysis (RFLP)

Using the *restriction map* application (http://www.bioinformatics.org/sms2/rest_map.html), we selected two endonucleases to differentially cut the COI gene of *Dreissena* species: Hinf I and Nla III. We also tested two other enzymes used in previous studies: Nla IV (Baldwin et al. 1996) and Scr FI (Claxton et al. 1998).

Restriction analysis of the amplified 654 bp COI fragment was carried out on each dreissenid haplotype (using individuals from the Meuse River). For each haplotype, the RFLP was performed in 31 µl total volume including 10 µl of PCR reaction mixture, 18 µl of distilled water, 2 µl of buffer (supplied by the manufacturer with the enzyme), and 1 µl of enzyme. Digests were incubated at 37 °C for 3 hours and then loaded on 2% agarose gels.

Table 1. GenBank accession numbers and localities of *Dreissena* spp. sequences included in the network analysis.

GenBank	Taxon	Location
DQ840122	*Dreissena polymorpha polymorpha*	Black and Caspian Seas
DQ840125	*Dreissena polymorpha polymorpha*	Liman, Caspian Sea
DQ840123	*Dreissena polymorpha polymorpha*	Caspian Sea
DQ840121	*Dreissena polymorpha polymorpha*	Black and Caspian Seas
EF414493	*Dreissena polymorpha*	Turkey
U47653	*Dreissena polymorpha*	Lake Ontario
AF474404	*Dreissena polymorpha*	Poland
EU484441	*Dreissena polymorpha*	Lake Superior
EU484437	*Dreissena polymorpha*	Lake Superior
EU484448	*Dreissena polymorpha*	Lake Superior
EU484444	*Dreissena polymorpha*	Lake Superior
AM748997	*Dreissena polymorpha*	Italy
AM748986	*Dreissena polymorpha*	Germany
AM748977	*Dreissena polymorpha*	Italy
U47651	*Dreissena bugensis*	Lake Ontario
U47650	*Dreissena bugensis* var. *profunda*	Lake Ontario
DQ840132	*Dreissena bugensis*	Black Sea
EF080861	*Dreissena rostriformis bugensis*	Hollandsch Diep
AF495877	*Dreissena bugensis*	Ukraine
AF479637	*Dreissena bugensis*	Ukraine
AM748999	*Dreissena polymorpha*	Germany

Results

Sequencing of the 654 bp COI fragment revealed seven haplotypes among the 241 *Dreissena* individuals. The OTU method revealed, by a NJ tree, two clusters separated by a genetic distance of 18.5% (Figure 1a), which is higher than the 3% threshold typically used for species delimitation with COI (Hebert et al. 2003b). This first analysis, therefore, suggests the occurrence of two species. We obtained the same results with the ABGD method. Indeed, the K2P-distances show two distinct clusters (Figure 1b). One cluster formed by haplotype 1 and 2, and a second cluster containing the five other haplotypes, all corresponding to those separated in the tree. Moreover, the genetic distances within our two clusters (0.6% and 0.2%, respectively) are four times lower than the genetic distance between them (18.5%) (Figure) confirming the presence of two *Dreissena* species.

Our network (Figure 2) revealed that haplotypes 1 and 2 (Q1 and Q2) cluster with *D. r. bugensis* and the five other haplotypes (Z1 to Z5) cluster with *D. polymorpha*. This, together with the three barcoding methods which each identified two clusters, shows that both *D. polymorpha* and *D. r. bugensis* species occur in the Meuse River.

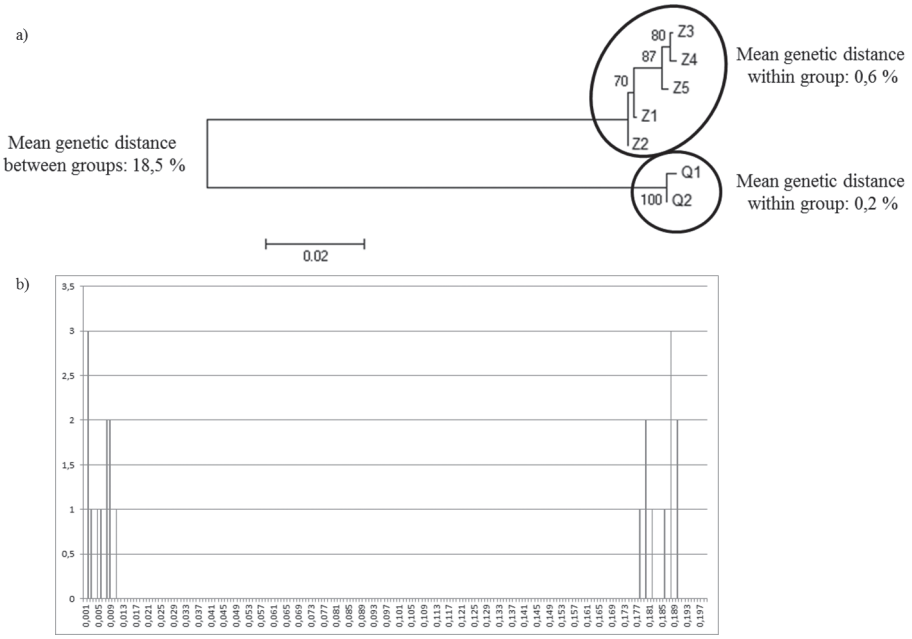


Figure 1. Barcoding analysis based on a fragment of 654 base pairs of the COI gene. **a**) NJ analysis of K2P-pairwise distances **b**) "barcoding gap" method based on the K2P-pairwise distance.

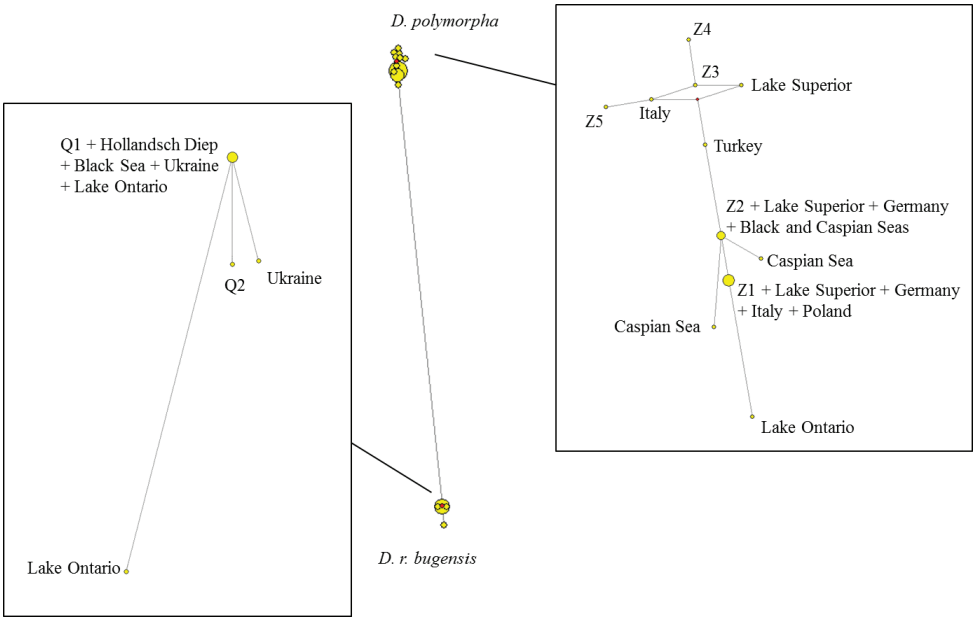


Figure 2. Haplotype networks based on a fragment of 654 base pairs of the COI gene. Our seven haplotypes are labelled: Q1 and Q2 for haplotypes 1 and 2 (belonging to *D. r. bugensis*) / Z1 to Z5 for the 5 other haplotypes (belonging to *D. polymorpha*).

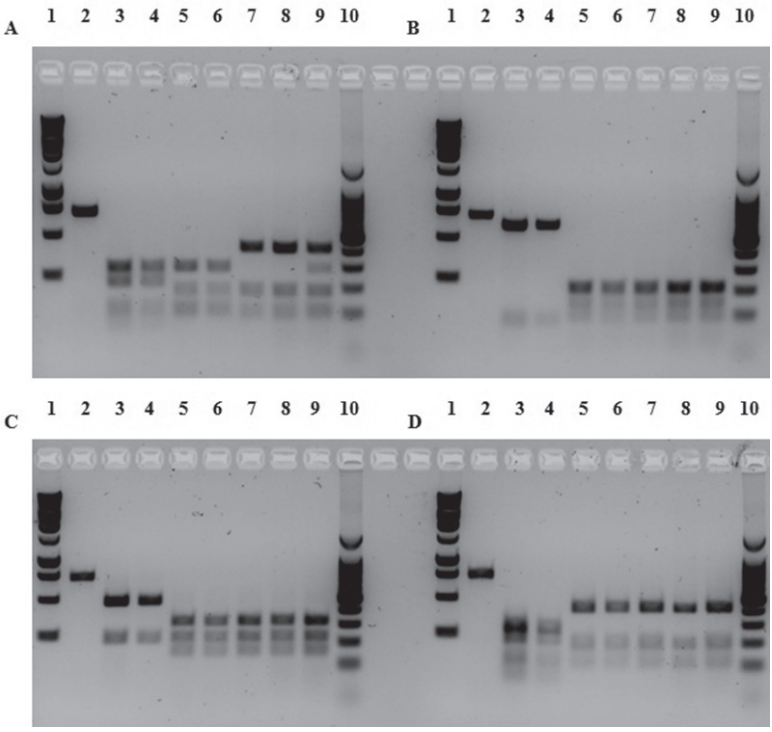


Figure 3. RFLP analysis of the COI gene to distinguish *Dreissena rostriformis bugensis* (Q haplotype) and *Dreissena polymorpha* (Z haplotype) using the endonucleases (**A**) Nla IV (**B**) Hinf I (**C**) Nla III and (**D**) Scr FI. Lane 1, 1-kb ladder; lane 2, non-digested fragment of quagga mussel; lane 3, Q1 haplotype; lane 4, Q2 haplotype; lane 5, Z1 haplotype; lane 6, Z2 haplotype; lane 7, Z3 haplotype; lane 8, Z4 haplotype; lane 9, Z5 haplotype; lane 10, 100-bp ladder.

Digestion profiles for each haplotype are illustrated in Figure 3. Each of the four endonucleases tested, yielded distinct restriction patterns between both *Dreissena* species. Digestion with Nla IV produced four fragments in quagga mussels (Q haplotype) of approximately 70, 79, 211, and 294 bp and three distinct patterns for the zebra mussel (Z haplotype): haplotype Z1 and Z2 (91, 120, 150, and 293 bp), haplotype Z3 and Z4 (91, 150, and 413 bp), and haplotype Z5 (91, 150, 200, and 413 bp). We suggest here that the 200 bp fragment of the haplotype Z5 is an artefact, as confirmed by the restriction map, since the summed fragment lengths do not add up to the expected 654 bp. We infer that haplotype Z5 has the same pattern as haplotype Z3 and Z4. Digestion with Hinf I produced two fragments in quagga mussels of approximately 73 and 581 bp and five fragments in zebra mussels of approximately 31, 101, 114, 195, and 213 bp. The small fragments can not be distinguished on the gel but the difference between quagga and zebra is clear. Digestion with Nla III produced two fragments in quagga mussels of approximately 193 and 461 bp and three fragments in zebra mussels of approximately 193, 319, and 335 bp. Digestion with Scr FI produced five fragments in quagga mussels of approximately

42, 53, 120, 171, and 268 bp and three fragments in zebra mussels of approximately 95, 152, and 407 bp. The digestion pattern for the quagga mussel using the endonuclease Scr FI is not clearly defined (smear) since the five fragments are very short.

Discussion

On September 9 2013, the European Commission has published a proposal for a Regulation on the prevention and management of the introduction and spread of invasive alien species. This proposal highlights three types of interventions: prevention, early warning and rapid response, and then management of invasive species (European Commission 2013). In this context, rapid identification methods are needed to detect invasive species in periodic surveys, e.g. inspection of ballast water. We showed in previous work (Marescaux et al. 2012a, b) that visual identification and morphometric analyses are not always sufficient to differentiate both zebra and quagga mussel probably due to phenotypic plasticity. This is particularly true for larval identification. In addition, two other *Dreissena* species may become invasive and should be detected promptly.

In order to help managers and national agencies, we propose here the use of the COI mitochondrial gene as a barcode to discriminate *D. polymorpha* and *D. r. bugensis*. Moreover, it is possible to conduct a RFLP analysis on this gene to obtain results without sequencing cost. This method could also easily be applied to *D. presbensis* and *D. blanci* since the COI gene have already been sequenced by Albrecht et al. (2007) and Wilke et al. (2010) and sequences are available on GenBank (accession numbers EF414478–EF414492, EF414496, HM209829–HM210081). We showed that the endonuclease Nla IV, previously used by Baldwin et al. (1996), presents different restriction patterns for the zebra mussel haplotype and not a clear distinction between some zebra mussel haplotypes (Z1 and Z2) and the quagga mussel haplotypes. Therefore, we do not recommend the use of this enzyme to discriminate between quagga and zebra mussel. The three other endonucleases tested during this study present a clear distinction between both species despite the fact that a smear appears using endonucleases Hinf I and Scr FI. Moreover, Nla III and Scr FI will produce a unique RFLP banding pattern for *D. blanci* and *D. presbensis* different from those observed in the zebra and quagga mussel.

This study is the first step of an extensive phylogeographical analysis on the invasion of Western Europe by the dreissenids. Further experiments will be needed to assess potential risks of both zebra and quagga mussels on native biodiversity in Western European rivers, e.g. predation on phytoplankton, infestation on native bivalves and alteration of macro-invertebrate communities.

Acknowledgements

Special thanks to Emilie Etoundi and Doctor Xiang Li for the help with the delimitation metrics. We also thank two reviewers and the editor for their helpful comments

and critical reading of this manuscript. This study received financial support from the University of Namur. Jonathan Marescaux is funded by a PhD grant from the Belgian National Fund for Scientific Research (FRS-FNRS).

References

Albrecht C, Schultheiß R, Kevrekidis T, Streit B, Wilke T (2007) Invaders or endemics? Molecular phylogenetics, biogeography and systematics of *Dreissena* in the Balkans. Freshwater Biology 52: 1525–1536. doi: 10.1111/j.1365-2427.2007.01784.x

Baldwin BS, Black M, Sanjur O, Gustafson R, Lutz RA, Vrijenhoek RC (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.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution 16: 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Beisel JN, Lévêque C (2010) Introductions d'espèces dans les milieux aquatiques: Faut-il avoir peur des invasions biologiques? Editions Quae, 232 pp.

Birky CW, Adams J, Gemmel M, Perry J (2010) Using population genetic theory and DNA sequences for species detection and identification in asexual organisms. PLoS ONE 5: e10609. doi: 10.1371/journal.pone.0010609

Claxton WT, Boulding EG (1998) A new molecular technique for identifying field collections of zebra mussel (*Dreissena polymorpha*) and quagga mussel (*Dreissena bugensis*) veliger larvae applied to eastern Lake Erie, Lake Ontario, and Lake Simcoe. Canadian Journal of Zoology 76: 194–198.

Colautti RI, MacIsaac HJ (2004) A neutral terminology to define 'invasive species'. Diversity and Distribution 10: 135–141. doi: 10.1111/j.1366-9516.2004.00061.x

Consortium for the Barcode of Life (2013) Identifying species with DNA barcoding. http://www.barcodeoflife.org/

European Commission (2013) Proposal for a regulation of the European parliament and of the council on the prevention and management of the introduction and spread of invasive alien species. http://ec.europa.eu/environment/nature/invasivealien/index_en.htm

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.

Hall TA (1998) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Hebert PDN, Ratnasingham S, deWaard JR (2003b) Barcoding animal life: cytochrome c oxidase subunit I divergences among closely related species. Proceedings of the Royal Society of London B 270 (Supplement): S96-S99. doi: 10.1098/rsbl.2003.0025

Heiler KCM, Brandt S, Albrect C, Hauffe T, Wilke T (2012) A new approach for dating introduction events of the quagga mussel (*Dreissena rostriformis bugensis*). Biological Invasions 14: 1311–1316. doi: 10.1007/s10530-011-0161-1

Invasive Species Advisory Committee (2006) Invasive species definition clarification and guidance white paper. Washington, D.C. http://www.invasivespecies.gov/global/ISAC/ISAC_documents/ISAC%20Definititions%20White%20Paper%20%20-%20FINAL%20VERSION.pdf

Librado P, Rozas J (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452. doi: 10.1093/bioinformatics/btp187

Marescaux J, Bij de Vaate A, Van Doninck K (2012a) First records of *Dreissena rostriformis bugensis* (Andrusov, 1897) in the Meuse River. BioInvasions Records 1: 119–124. doi: 10.3391/bir.2012.1.2.05

Marescaux J, Molloy DP, Giamberini L, Albrecht C, Van Doninck K (2012b) First records of the quagga mussel, *Dreissena rostriformis bugensis* (Andrusov, 1897), in the Meuse River within France. BioInvasions Records 1: 273–276. doi: 10.3391/bir.2012.1.4.05

Mills EL, Rosenberg G, Spidle AP, Ludyanskiy M, Pligin Y, May B (1996) A review of the biology and ecology of the quagga mussel (*Dreissena bugensis*), a second species of freshwater dreissenid introduced to North America. American Zoologist 36: 271–286. doi: 10.1093/icb/36.3.271

Pathy DA, Mackie GL (1993) Comparative shell morphology of *Dreissena polymorpha*, *Mytilopsis leucophaeata* and the "quagga" mussel (Bivalvia: Dreissenidae) in North America. Canadian Journal of Zoology 71: 1012–1023. doi: 10.1139/z93-135

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology 21: 1864–1877. doi: 10.1111/j.1365-294X.2011.05239.x

Sablon R, Vercauteren T, Jacobs P (2010) De quaggamossel (*Dreissena rostriformis bugensis* (Andrusov, 1897)), een recent gevonden invasieve zoetwatermossel in Vlaanderen. Antenne 4: 32–36.

Son MO (2007) Native range of the zebra mussel and quagga mussel and new data on their invasions within the Ponto-Caspian Region. Aquatic Invasions 2: 174–184. doi: 10.3391/ai.2007.2.3.4

Sousa R, Novais A, Costa R, Strayer DL (2013) Invasive bivalves in fresh waters: impacts from individuals to ecosystems and possible control strategies. Hydrobiologia published online on 22 January 2013. doi: 10.1007/s10750-012-1409-1

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599. doi: 10.1093/molbev/msm092

Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D (2012) The widely used small subunit I8S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences of the USA 109: 16208–16212. doi: 10.1073/pnas.1209160109

Vermeij GJ (1996) An agenda for invasion biology. Biological Conservation 78: 3–9. doi: 10.1016/0006-3207(96)00013-4

Wilke T, Schultheiß R, Albrecht C, Bornmann N, Trajanovski S, Kevrekidis T (2010) Native *Dreissena* freshwater mussels in the Balkans: in and out of ancient lakes. Biogeosciences 7: 3051–3065. doi: 10.5194/bg-7-3051-2010

Wilson KA, Howell ET, Jackson DA (2006) Replacement of zebra mussels by quagga mussels in the Canadian nearshore of Lake Ontario: the importance of substrate, round goby abundance, and upwelling frequency. Journal of Great Lakes Research 32: 11–28. doi: 10.3394/0380-1330(2006)32[11:ROZMBQ]2.0.CO;2

Winnepenninckx B, Backeljau T, De Wachter R (1993) Extraction of high molecular weight DNA from molluscs. Trends in Genetics 9: 407.

ZooKeys 365: 245–261 (2013)
doi: 10.3897/zookeys.365.5787
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol

Jeremy A. Miller[1,2,3], Kevin K. Beentjes[1], Peter van Helsdingen[4], Steven IJland[5]

1 *Naturalis Biodiversity Center, Postbus 9517, 2300 RA Leiden, the Netherlands* **2** *Department of Entomology, California Academy of Sciences, 55 Music Concourse Drive, Golden Gate Park, San Francisco, California 94118, USA* **3** *Plazi, Zinggstrasse 16, Bern, Switzerland* **4** *European Invertebrate Survey – Nederland, Postbus 9517, 2300 RA Leiden, the Netherlands* **5** *Gabriel Metzustraat 1, 2316 AJ Leiden, the Netherlands*

Corresponding author: *Jeremy A. Miller* (jeremy.miller@naturalis.nl)

Academic editor: *T. Backeljau* | Received 15 June 2013 | Accepted 6 October 2013 | Published 30 December 2013

Citation: Miller JA, Beentjes KK, van Helsdingen P, IJland S (2013) Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 245–261. doi: 10.3897/zookeys.365.5787

Abstract

We report initial results from an ongoing effort to build a library of DNA barcode sequences for Dutch spiders and investigate the utility of museum collections as a source of specimens for barcoding spiders. Source material for the library comes from a combination of specimens freshly collected in the field specifically for this project and museum specimens collected in the past. For the museum specimens, we focus on 31 species that have been frequently collected over the past several decades. A series of progressively older specimens representing these 31 species were selected for DNA barcoding. Based on the pattern of sequencing successes and failures, we find that smaller-bodied species expire before larger-bodied species as tissue sources for single-PCR standard DNA barcoding. Body size and age of oldest successful DNA barcode are significantly correlated after factoring out phylogenetic effects using independent contrasts analysis. We found some evidence that extracted DNA concentration is correlated with body size and inversely correlated with time since collection, but these relationships are neither strong nor consistent. DNA was extracted from all specimens using standard destructive techniques involving the removal and grinding of tissue. A subset of specimens was selected to evaluate nondestructive extraction. Nondestructive extractions significantly extended the DNA barcoding shelf life of museum specimens, especially small-bodied species, and yielded higher DNA concentrations compared to destructive extractions. All primary data are publically available through a Dryad archive and the Barcode of Life database.

Keywords
DNA extraction, DNA concentration, DNA preservation, DNA degradation, independent contrasts, museum

Introduction

The DNA barcoding enterprise has demonstrated its utility for contributing to studies of both well-known and poorly-known taxonomic communities. Studies of diverse tropical arthropods often include many species without formal names (e.g. Smith et al. 2005, Janzen et al. 2009). DNA barcode sequences in conjunction with morphological data are a potent combination for a wide range of biodiversity applications (Dayrat 2005, Will et al. 2005, Goldstein and DeSalle 2011, Riedel et al. 2013). The focus of this research is to develop a DNA barcode library for a well-known fauna: Dutch spiders. The list of spider species recorded from the Netherlands, which stands as of this writing at 644, has been extensively documented and periodically updated through the Fauna Europaea database (Helsdingen 1999, 2013). The specimens necessary to build such a library come from collections, either fresh material or natural history museums. The national natural history collection for the Netherlands is curated at the Naturalis Biodiversity Center. We investigated how a variety of factors (time since collection, body size, phylogenetic distance) influence the success of DNA barcode sequencing. Our goal is to characterize which specimens in the collection are or are not likely to yield a successful DNA barcode sequence, and to use this knowledge to efficiently build a barcode library based on a combination of fresh and museum specimens.

A collection like Naturalis makes large numbers of spider specimens accessible for research, including many rare species. Traditional natural history museums like Naturalis store collections in cool, dark environments to keep specimens preserved over long periods of time. However, these conditions are inadequate to completely prevent degradation of specimen DNA. Spider collections are typically preserved in 70-80% ethanol. At these concentrations, ethanol has oxidative and hydrolytic effects that can degrade DNA over time (Vink et al. 2005). DNA degradation eventually proceeds to the point that the standard animal DNA barcode locus, a ~650 base pair region of the mitochondrial cytochrome *c* oxidase subunit I gene (COI), fails to amplify using basic protocols. It may still be possible to sequence part or all of the DNA barcode region by amplifying a series of short sections and reassembling them (Van Houdt et al. 2010, Andersen and Mills 2012, Zuccon et al. 2012), but this approach requires a substantial increase in time and resources devoted per specimen.

Freshly collected specimens present fewer technical obstacles to successful DNA barcode sequencing. Obtaining and processing samples requires some time and effort. Sample contents are influenced by a wide range of factors, including weather, season, and collecting methodology. So perhaps beyond some common species, one cannot predict with certainty which species will be represented in the samples.

Fresh and museum collections have complementary strengths and weaknesses when it comes to the efficient development of a DNA barcode library. Initially, field work generates fresh specimens of many species in need of barcoding. As the DNA barcode library grows, it eventually becomes increasingly difficult to find fresh specimens of species that have not been barcoded previously. This may be true even while the number of barcoded species is substantially lower than the number of species known from the Netherlands. This may be the time to turn to the museum collection and specifically target species that have eluded current field work. However, natural history museums are a resource for the global research community and activities that can damage museum specimens, including DNA extraction, should be undertaken with consideration that the anticipated research value will outweigh any specimen degradation. To this end, we have investigated barcode sequencing success rates as a function of years since collection, considering both destructive and nondestructive DNA extraction methods. Species representing a variety of spider lineages and a range of body sizes were included.

Methods

Fresh collections

Spiders were collected from several locations in the Netherlands. Collecting methods included beating or sweeping vegetation, sifting leaf litter, and hand collecting. 70% Ethanol was used as a preservative. Samples were kept at -20 °C when not being worked on. Specimens were identified by taxonomic experts on the Dutch spider fauna and exemplars were selected for DNA barcoding.

Museum collection

31 frequently collected species were selected (Figure 2). For the 199 and 200, 1-4 specimens of each species were selected per decade, and 1–2 specimens per decade were selected as available going back to 1950. This was supplemented with 1-3 fresh or museum specimens from 2010–2012. Specimens collected using pitfall traps were avoided because the preservative formalin, commonly used in pitfalls, damages DNA (Gurdebeke and Maelfait 2002). However, historical specimen data labels may not always indicate when specimens were collected using formalin pitfalls. All 31 time series species yielded DNA barcode sequences for at least some specimens, indicating that sequencing failures could not be attributed to a lack of primer specificity.

The Naturalis spider collection has been kept (along with most of the Naturalis collection) in a 60 m collection tower since 1998. Conditions are controlled and monitored, with temperature maintained between 17–18 °C and relative humidity 50–55%. We have been unable to find data on conditions prior to the move to the

tower. Specimens are kept in cotton-stoppered glass vials; up to several dozen vials are kept together submerged in 70% ethanol within a larger jar. This is intended to keep ethanol concentration stable.

DNA barcode sequencing

Initial source tissue for both fresh and museum specimens was a single leg, removed from the specimen and ground using a sterile blade in a 1.2 ml eppendorf tube, then incubated for three hours in lysis buffer with proteinase K. For second round extractions from selected museum specimens, DNA was extracted by placing the entire specimen (minus one leg consumed by destructive extraction) directly (without grinding) in lysis buffer with proteinase K for the three hour incubation step. After incubation, the specimen was returned to ethanol and the extraction continued using the lysis buffer solution. This caused negligible to slight further damage to the specimen (Rowley et al. 2007, Paquin and Vink 2009). These two methods are referred to in this paper as destructive and nondestructive extraction, respectively. Some of the larger species (*Araneus quadratus* Clerck, 1757, *Tegenaria atrica* C. L. Koch, 1843, *Dolomedes plantarius* Clerck, 1757) could not be fit into the extraction tubes without damage and were excluded from the nondestructive extraction portion of the study.

Extractions proceeded using the Thermo Scientific KingFisher Flex magnetic bead extraction robot at the Naturalis Biodiversity Center DNA barcoding facility using the Macherey-Nagel NucleoMag 96 Tissue kit. To obtain the standard animal DNA barcode fragment of the mitochondrial COI gene (Hebert et al. 2003), PCR was performed using the primers LCO1490 (5'-GGTCAACAAATCATAAAGATATTGG-3') (Folmer et al. 1994) and Chelicerate Reverse 2 (5'-GGATGGCCAAAAAATCAAAATAAATG-3') (Barrett and Hebert 2005). PCR reactions contained 18.75 µl mQ, 2.5 µl 10 × PCR buffer CL, 1.0 µl 25 mM of each primer, 0.5 µl 2.5 mM dNTPs and 0.25 µl 5U Qiagen Taq. PCR was performed using an initial denaturation step of 180 s at 94 °C, followed by 40 cycles of 15 s at 94 °C, 30 s at 50 °C and 40 s at 72 °C, and finishing with a final extension of 300 s at 72 °C and pause at 12 °C. Sequencing was performed by Macrogen (http://www.macrogen.com) or BaseClear (http://www.baseclear.com/). For all barcoded specimens, sequences, images, and collection data were uploaded to the Barcode of Life Data Systems (BOLD; http://www.boldsystems.org/) in the project NLARA "Araneae of the Netherlands". DNA concentration was assessed using 1.5 µl samples of genomic DNA extract run through a NanoDrop ND-1000 Spectrophotometer (www.nanodrop.com/).

Correlates of sequencing success and failure

We used independent contrasts (Felsenstein 1985, Garland et al. 1992) to investigate species body size and phylogenetic distance as factors that might explain the oldest suc-

cessful sequence from the 31 frequently collected species. The independent contrasts method factors out the phylogenetic non-independence of species so that correlations between two continuous variables can be validly tested on a collection of species. Each species was scored for body size and years since collection for the oldest successful DNA barcode sequence. Male and female body sizes were taken from the literature (Roberts 1985, 1987, Nentwig et al. 2013) and averaged. A single exemplar sequence representing each focal species was taken from the freshest available specimen. We generated a Neighbour-Joining tree in DAMBE (Xia and Xie 2001; F84 model, 10,000 random addition steps). We used the PDAP package in Mesquite (Midford et al. 2010, Maddison and Maddison 2011) to perform independent contrasts analysis. Other statistical analyses ($\log_{10}$ transformation, Pearson's r correlation, ANOVA and χ^2) were performed using PAST (Hammer et al. 2001).

The amount of tissue taken from each specimen for destructive DNA extraction was not quantified or controlled for and was substantially different among the species in the study. We therefore investigated the role of DNA concentration. We looked for a relationship between 1) body size and 2) years since collection against DNA concentration (ng/µl) and DNA barcode sequencing success rates for specimens included in the time series study based on both destructive and nondestructive extraction.

Recent collections covered a broader set of species than the time series study. Tree-based methods like independent contrasts are not applicable to this dataset because species that failed to produce a DNA barcode sequence could not be included in the tree. We searched the BOLD databases for sequences to represent these species, but a substantial number (9 of 14) are currently not available. Body size was calculated as for the time series species.

Data resources

All occurrence data for specimens included in this study are available as part of a Dryad (http://datadryad.org/) data package (doi: 10.5061/dryad.q08). Occurrence data are presented as a tab delimited text file with Darwin Core fields (http://darwincore.googlecode.com/svn/trunk/terms/index.htm), plus custom fields for recording destructive and nondestructive sequencing success, DNA sequences, DNA concentration data, and hyperlinks to records on BOLD (http://www.boldsystems.org/). Also included in the Dryad data package is a KML file that can be opened using Google Earth (http://earth.google.com/) to display an interactive map plotting Dutch spider specimens included in this study. Click on placemarks to reveal specimen data and, where available, a hyperlink to sequence data for that specimen on BOLD (http://www.boldsystems.org/). The Dryad data package also includes all sequence data for this study in fasta format, two Nexus files generated using Mesquite (Maddison and Maddison 2011) for the independent contrasts analyses, and Appendix - Figure S1 illustrating correlations based on independent contrasts analyses.

Results

We obtained DNA barcode sequences for 145 spider species (91.2% of the 159 species attempted) based on 452 fresh and museum specimens (Figure 1A). Sequences ranged from 510 to 658 bp (mean 650.1). The 14 species attempted that failed to yield a DNA barcode were *Clubiona subtilis* L. Koch, 1867 (Clubionidae); *Harpactea hombergi* (Scopoli, 1763) (Dysderidae); *Haplodrassus silvestris* (Blackwall, 1833) (Gnaphosidae); *Cnephalocotes obscurus* (Blackwall, 1834), *Dismodicus elevatus* (C.L. Koch, 1838), *Entelecara congenera* (O. Pickard-Cambridge, 1879), *Erigone dentipalpis* (Wider, 1834), *Gnathonarium dentatum* (Wider, 1834), *Gongylidium rufipes* (Linnaeus, 1758), *Macrargus rufus* (Wider, 1834), *Walckenaeria antica* (Wider, 1834) (Linyphiidae); *Arctosa leopardus* (Sundevall, 1833) (Lycosidae); *Pholcus phalangioides* (Fuesslin, 1775) (Pholcidae); and *Pachygnatha listeri* Sundevall, 1830 (Tetragnathidae).

For fresh specimens (collected 2010 or later), the overall sequencing success rate was 90.6%. For specimens collected between 2000 and 2009, the success rate drops slightly to 78.4%. For specimens collected in the 199, sequencing success drops to 59.2%, then to 35.3% for specimens collected in the 198, then to around 20% for specimens collected in the 197 and 196, and finally 12.5% for specimens collected in the 195 (Figures 1, 2).

When genetic distance is accounted for using independent contrasts, we found a significant positive correlation between body size and years since collection for successful DNA barcode sequences (Appendix - Figure S1). Using our protocol and a single long run PCR, the standard DNA barcode sequences can be obtained from larger spider species for a longer period of time compared to smaller spider species. This relationship holds regardless of whether we consider only data from destructive extractions (R^2 = 0.39, F (1, 29) = 18.87, p = 1.56E-4) or all extractions (R^2 = 0.23, F (1, 29) = 8.43, p = 6.99E-3) despite the fact that three of the species were too large to include in the nondestructive extraction portion of the study.

Body size is correlated with DNA concentration based on data from destructive extractions (r (281) = 0.30, p = 2.31E-03); this relationship is not evident for the smaller dataset based on non-destructive extractions (r (130) = 0.05, p = 0.61). Years since collection is correlated with DNA concentration based on data from the non-destructive extractions (r (130) = 0.20, p = 0.02) but not the destructive extractions (r (281) = 0.01, p = 0.92). In all cases, the dependent variable was $\log_{10}$ transformed. Nondestructive extractions did yield significantly higher concentrations compared to destructive extractions (Figures 3, 4; one-way ANOVA, $p < 0.05$ whether considering only extracts that produced a barcode sequence (F (1, 159) = 120.2, p = 3.45E-18), extracts that failed (F (1, 232) = 184.1, p = 295E-28), or all extracts measured (F (1, 395) = 305.7, p = 4.19E-48). In all cases, concentration values were $\log_{10}$ transformed. Note that nondestructive samples all had one leg removed (consumed for destructive samples); we don't know what effect this might have had on barcoding success since the space left by the removed leg leading to the interior of the prosoma may have facilitated the extraction.

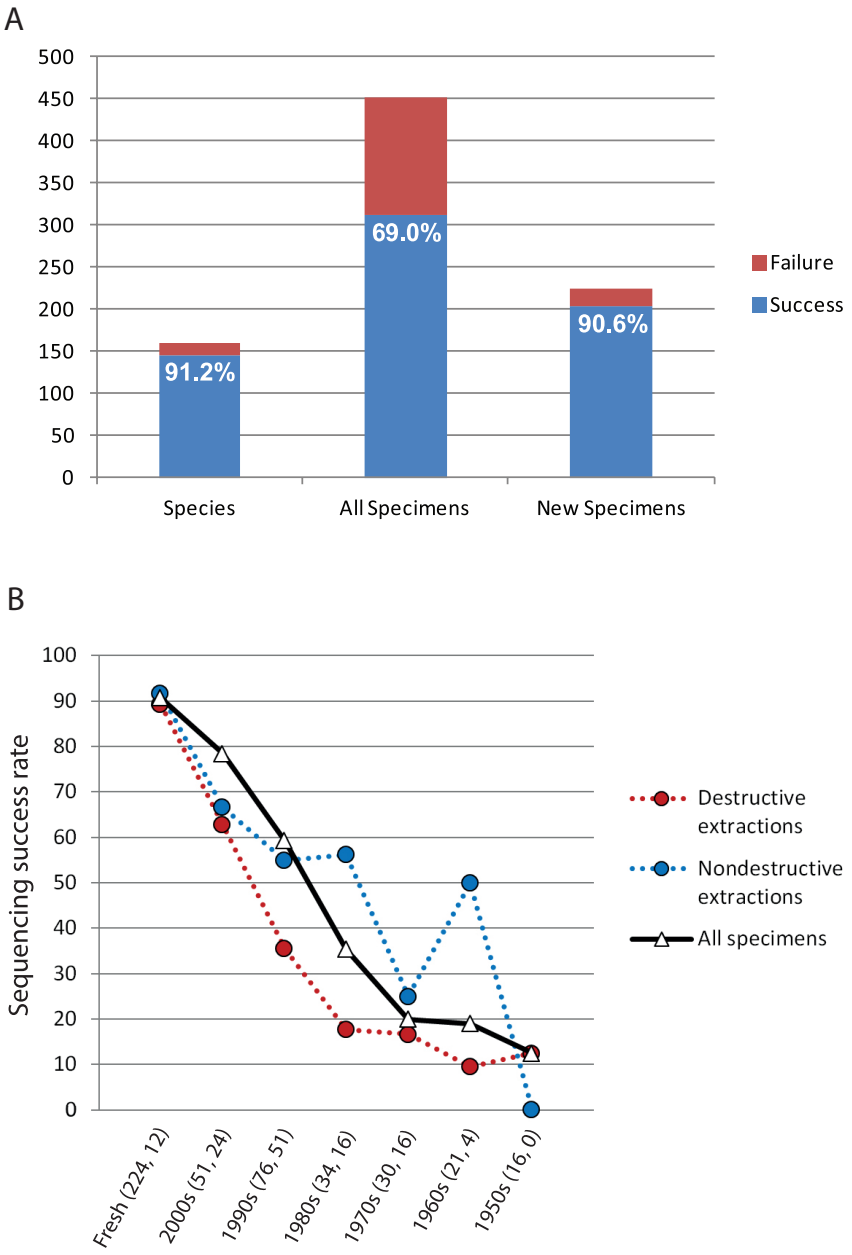


Figure 1. A Sequencing success profile for specimens included in this study. Data are species attempted, all specimens in the study including the time series, and fresh specimens collected in 2010 or later. Success expressed as a percentage appears on the blue (success) portion of each bar **B** Sequencing success rates for fresh (collected 2010 or later) and older specimens grouped by decade. Data given for all extractions regardless of method, and also partitioned into destructive and nondestructive extraction methods. Total number of specimens attempted and the subset of specimens attempted using nondestructive extraction given in parentheses. Note that the relatively high success rate for nondestructive extractions of specimens from the 196 is based on two successes out of four attempts.

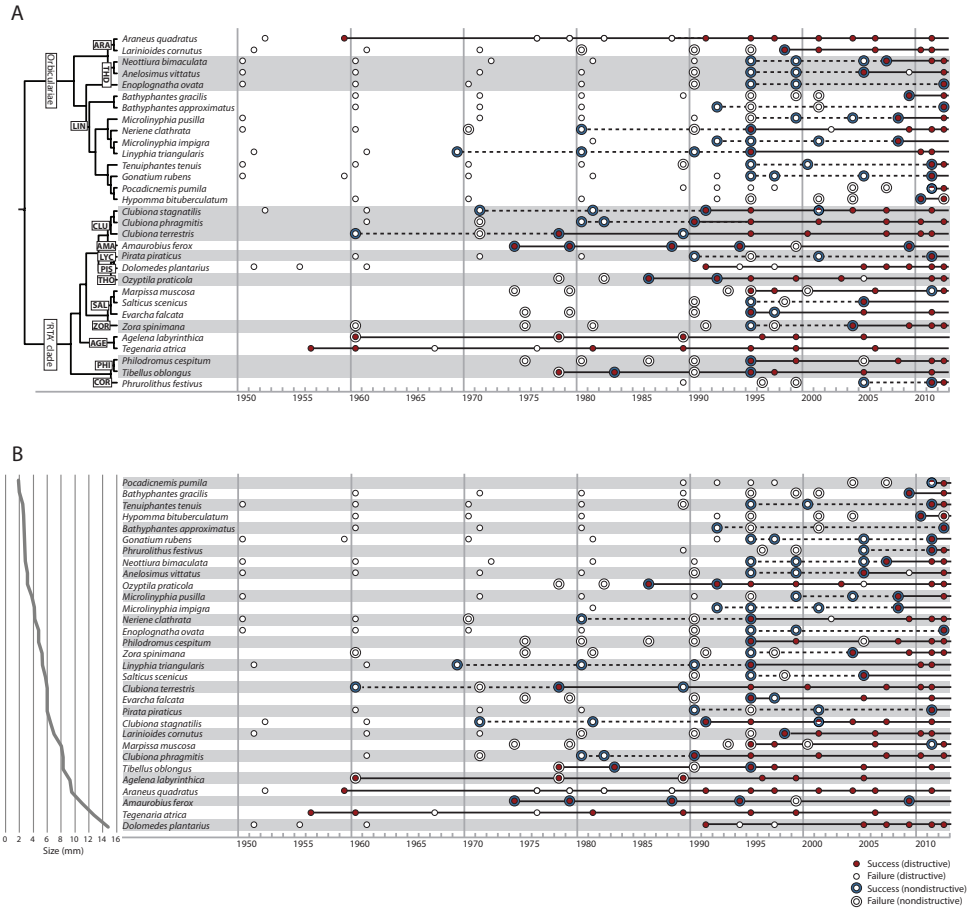


Figure 2. Sequencing success for the time series study of 31 spider species frequently collected in the Netherlands. Data for each species arranged horizontally along a time axis (year of collection). Small circles represent standard destructive extraction; outer circle represents nondestructive extraction. Red small circle and blue outer circle indicate successful sequencing, unfilled circles represent failed attempts; half-filled circle indicates mixed success among multiple specimens for that species and year. Solid horizontal lines extend from the present to the oldest successful DNA barcode based on destructive extraction for each species; where nondestructive extraction yielded successful DNA barcode from older specimens, this is indicated by a dashed line. Data are arranged according to a Neighbour-Joining tree (**A**) or by species body size (**B**). Spider families and major lineages (Orbiculariae and 'RTA' clade) are indicated in **A**. **AGE** Agelenidae **AMA** Amaurobiidae **ARA** Araneidae **CLU** Clubionidae **COR** Corinnidae **LIN** Linyphiidae **LYC** Lycosidae **PHI** Philodromidae **PIS** Pisauridae **SAL** Salticidae **THD** Theridiidae **THO** Thomisidae **ZOR** Zoridae.

Of 123 samples where both destructive and nondestructive extraction methods were tried, 38 produced successful barcodes using destructive extraction and 85 produced successful barcodes using nondestructive extraction. Of the 38 successful destructive extraction barcodes, 32 (84.2%) were also successful using nondestructive extraction while 6 (15.8%) failed. Of the 85 unsuccessful destructive barcodes, 38 (44.7%) were successful using nondestructive extraction while the remaining 47 failed

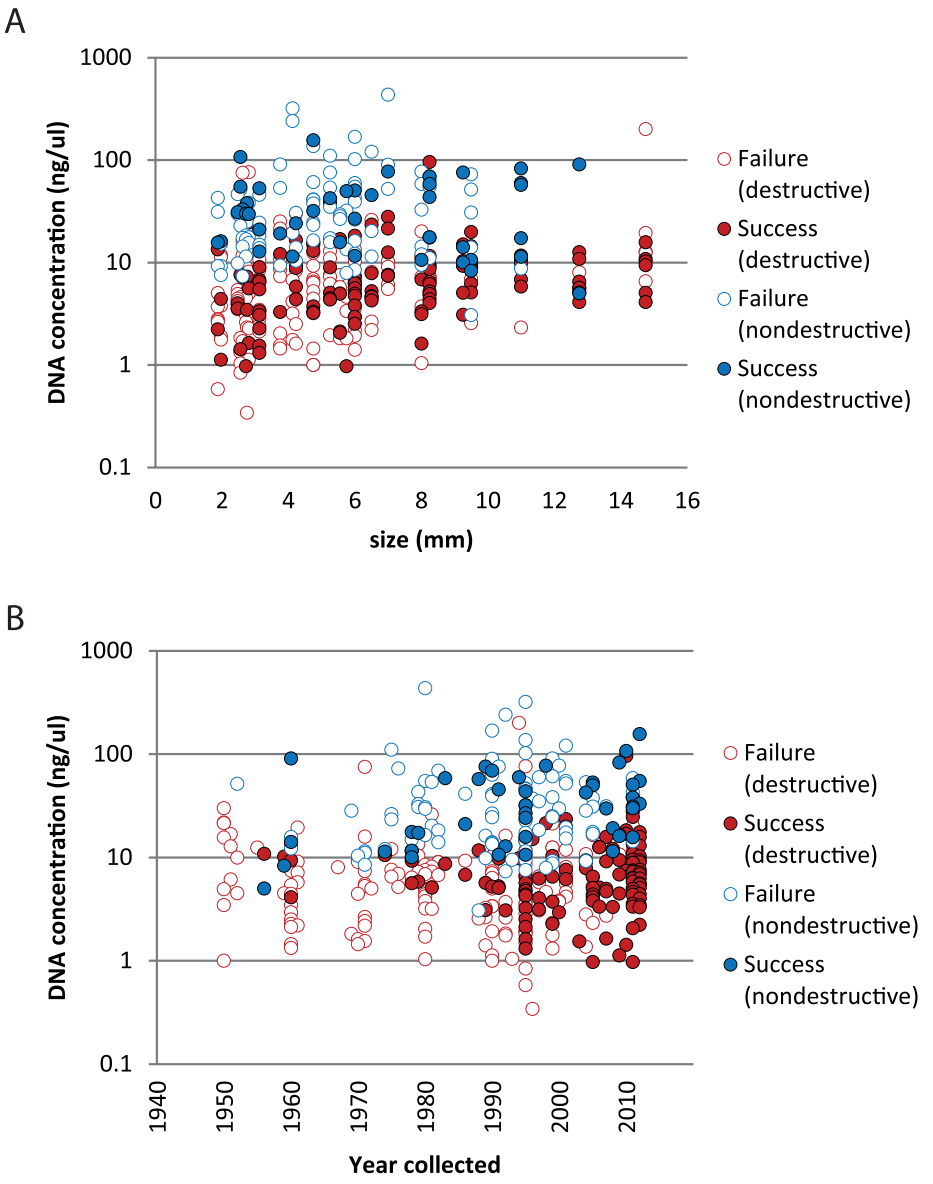


Figure 3. DNA concentration ($\log_{10}$ transformed) for specimens in the time series study that yielded or failed to yield a successful DNA barcode sequence arranged by **A** body size **B** year collected. Successes (filled circles) and failures (while circles) partitioned into destructive (red) and nondestructive (blue) DNA extraction methods.

using both methods. So although nondestructive extraction failed in about 15% of the cases where destructive sampling was successful, nondestructive extraction was significantly better at yielding successful barcode sequences, particularly when destructive extraction failed (χ^2 (2, N = 123) = 16.71, p = 0.0002).

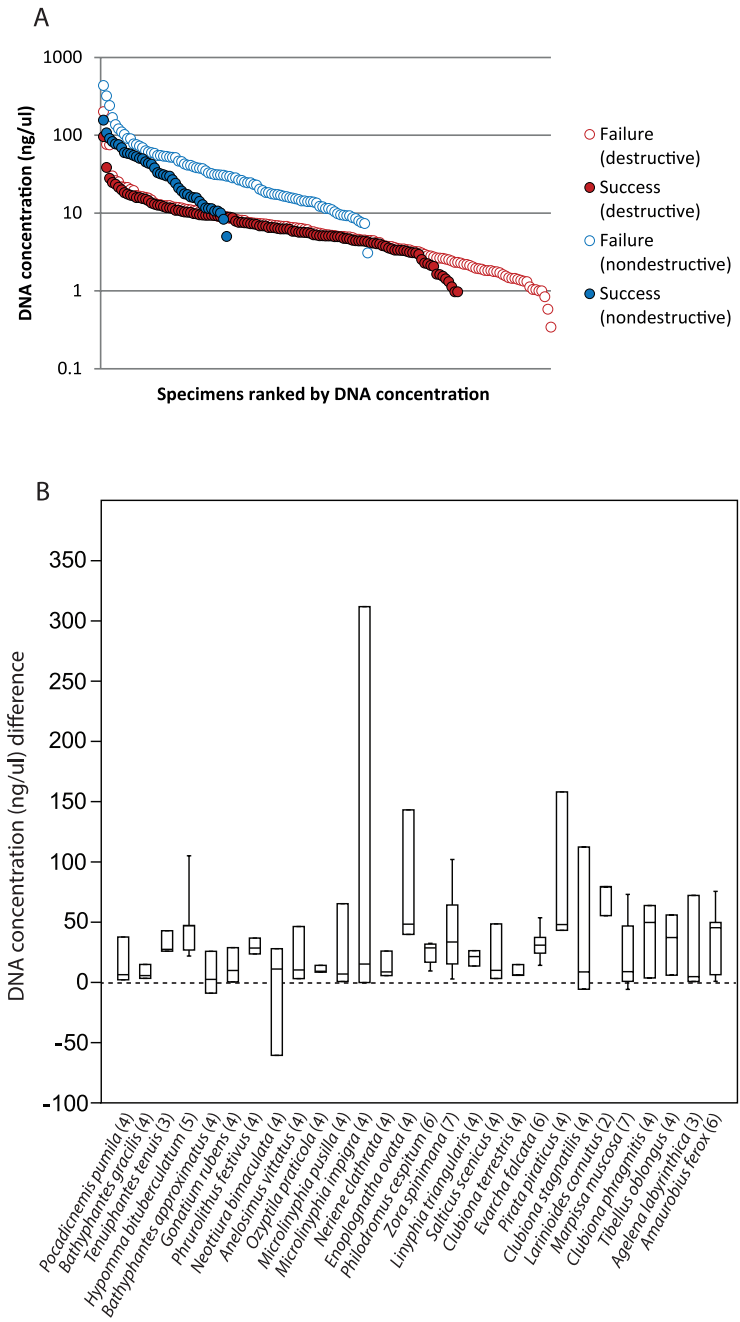


Figure 4. A DNA concentration ($\log_{10}$ transformed) for specimens in the time series study that yielded or failed to yield a successful DNA barcode sequence ranked by DNA concentration; symbols as in Figure 3 **B** Box plot showing difference in DNA concentration for specimens extracted using both destructive and nondestructive methods; species arranged by size (*Araneus quadratus*, *Tegenaria atrica*, and *Dolomedes plantarius* excluded). Sample size in parentheses, boxes are 25–75% quartiles bisected by the median, whisker lines indicate minimum/maximum values (where $n > 4$).

The combination of destructive and nondestructive extractions extended the DNA barcoding shelf life of the species in our study over destructive extraction alone by an average of 9.3 years. The nondestructive portion of our study was not comprehensive, involving only 123 (44.6%) of the specimens and 28 (90.3%) of the species in the time series study. The oldest successful barcode specimen was on average 6.7 years older for the nondestructive extraction data compared to the destructive extractions. The oldest successful barcode template was from a nondestructive extraction in 17 of the 28 species compared (60.7%); the oldest successful barcode template came from a destructive extraction in only 3 of the species (10.7%). However, for one of these species (*Agelena labyrinthica* (Clerck, 1757)) the nondestructive extraction never produced a successful barcode sequence while the destructive extractions were effective for every specimen attempted (n = 6) going back to 1960. In *Marpissa muscosa* (Clerck, 1757), destructive extractions were also much more effective than nondestructive extractions (Figure 2).

Discussion

Failure rates for DNA barcode sequencing rise with time since collection, but body size is a significant factor. For freshly collected specimens overall, body size is not a predictor of sequencing success or failure (Figure 5A). But larger species have a longer DNA barcoding shelf life than smaller species under museum collection conditions, at least using a single pair of primers to amplify the entire ~650 base pair region in one reaction. This may be explained in part by the finding that concentration of extracted DNA is correlated with specimen size and inversely correlated with specimen age, but this relationship is neither strong nor consistently found. The dominant protocol for spider DNA barcoding and other Sanger sequencing involves the removal of tissue from the specimen, typically from one or more legs. Our data suggest that nondestructive extraction techniques can significantly improve the chances of obtaining a DNA barcode sequence. Considering only the commonly applied destructive extraction technique, small spiders are useful for only a few years while those with a body size of around 3 mm or more have a modest chance of yielding a barcode sequence for about 20 years after collection. But with judicious application of nondestructive extraction, spiders from museum collections with a body length of 4 mm or less have a modest chance of yielding a DNA barcode sequence from a single PCR reaction for about 15 years since collection while spiders above this size can yield barcode sequences for a considerably longer time. For some of the larger species, we did not include specimens old enough to fail to produce DNA barcodes, so their real shelf life may be even longer than indicated here (Figure 2B).

All of the species in the time series study and nearly all the fresh specimens attempted belong to two major sister clades: the Orbiculariae (orb web weavers and their descendents) and the 'RTA' clade (so named for the synapomorphic retrolateral tibial apophysis of the male pedipalp; Coddington and Levi 1991). Together, these clades account for about 83% of described spider diversity (Platnick 2013). Recent

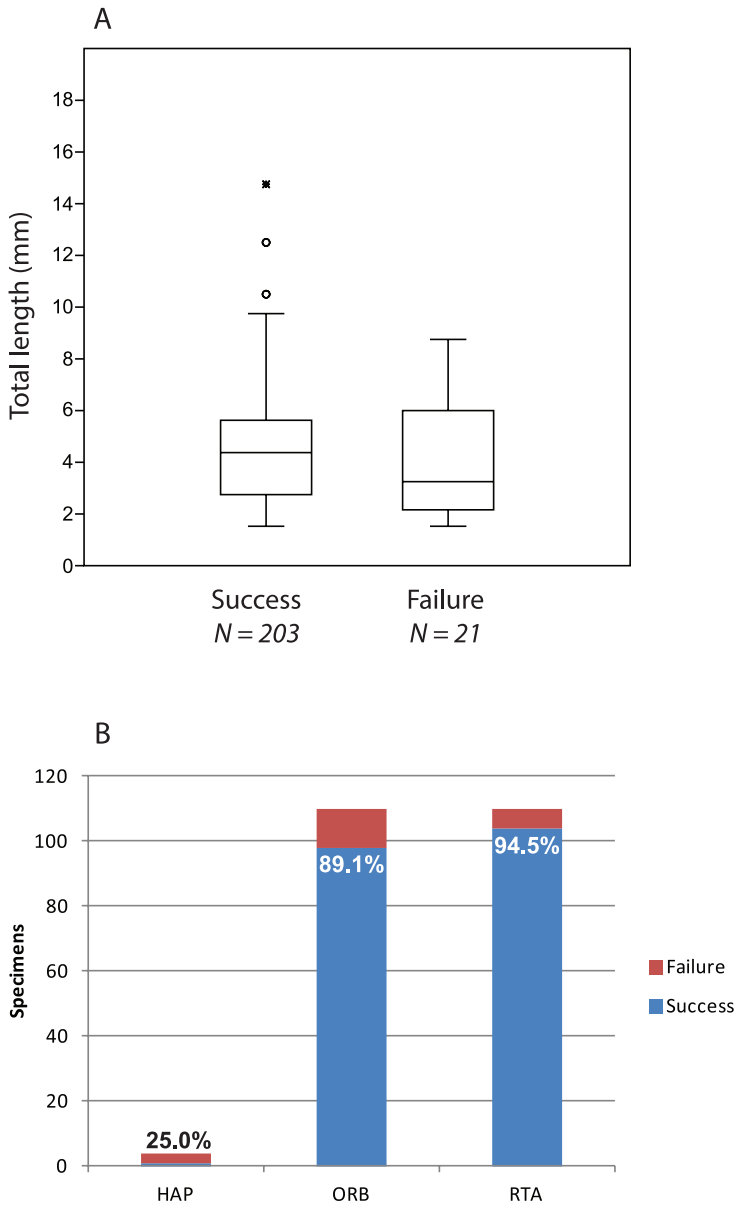


Figure 5. DNA barcode sequencing success for fresh specimens (collected 2010 or later). **A** Specimen body size not significantly different for successful vs. failed DNA barcode sequencing attempts (one-way ANOVA, $F(1, 216) = 1.45$, $p = 0.230$). Boxes are 25–75% quartiles bisected by the median, whisker lines drawn to the largest/smallest data point less than 1.5 times the box height, outliers less than 3 times the box height shown as circles, more than 3 shown as stars. **B** Most of the fresh specimens included in this study belonged to one of two clades: Orbiculariae (ORB) or the 'RTA' clade (RTA); only a handful of specimens represented older phylogenetic branches, such as haplogyne (HAP) spiders; no mygalomorph spiders were included; success expressed as a percentage appears on or above each bar. Success rate for Orbiculariae vs. 'RTA' clade specimens not significantly different ($\chi^2 = 2.18$, d.f. = 2, $N = 220$, $p = 0.337$).

field work found very few representatives of spider lineages that branched off before the origin of the Orbiculariae+'RTA' clade (e.g. Haplogynae and other early branching araneomorphs, or Mygalomorphae, which account for only 20 and 3 of the 644 recorded Dutch spider species respectively; Figure 5B). So results reported here may not be generalizable beyond this major spider lineage. Our data indicate no difference in failure rate for Orbiculariae compared to the 'RTA' clade (χ^2 = (2, N = 220) = 2.18, p = 0.34; Figure 5B).

We found no differences in sequencing success rate by lineage. It may yet be that changes in chemistry (e.g. DNAase, PCR inhibitors), primer binding site sequences, or other heritable characteristics might make some spider lineages more resistant to sequencing than others.

Several recent studies have investigated the relationship between specimen age and DNA barcode sequencing success for museum collections (Van Houdt et al. 2010, Andersen and Mills 2012, Zuccon et al. 2012). These studies include PCR reactions targeting short portions of the DNA barcode region as a way of compensating for the DNA degradation that comes with time. With field collection ongoing, we do not yet know which species available in the museum collection might elude contemporary field work. As field work becomes increasingly inefficient at producing fresh specimens of unbarcoded species, the museum collection may become the only readily available source for certain species. Based on what we have learned through this study about body size and specimen age, we will be able to predict whether standard protocols are likely to produce a successful DNA barcode sequence, or if more refined and targeted methods including PCRs targeting one or more sub-regions of the DNA barcode, should be employed. The success of nondestructive extraction demonstrated here coupled with the need to preserve museum specimens for a variety of research purposes bodes well for museum collections as a source of material for spider barcode libraries, and perhaps other alcohol collections as well.

DNA barcoding spiders in Europe

The initiative to create a library of DNA barcode sequences for Dutch spiders occurs in a broader context. Research teams in several European countries are involved in similar national projects (see http://www.araneae.unibe.ch/barcoding/content/15/Barcoding-of-European-spiders). The synergies anticipated from multiple libraries across Europe and beyond are exciting. As these libraries mature, they will become a reference not only for taxonomic identification, but for assessing intraspecific variation across the region. As barcode sequence data are independent of the morphological characters traditionally used to establish and subsequently recognize species, they will provide a check of species concepts as applied internationally. We may find that some species considered widespread exhibit sufficient sequence variation and geographical structure to warrant further study, or discover a lack of variation in different nominal species that could indicate these species are in fact one. Of the nearly 4,900 spider species recorded

from Europe, more than 2,000 are known from only one country (Helsdingen 2013). It may well be that some portion of this national endemism is an artifact.

The development of a DNA barcode library of European spiders is too large a task for any one research group. Data standards and a community data repository facilitate the reuse and reevaluation of DNA barcode data generated by independent labs (Ratnasingham and Hebert 2007). The increasing adoption by the scientific community of data standards and online resources for data aggregation strengthens both cooperative and adversarial (i.e., independent repeatability) aspects of biodiversity research, contributing to both productivity and rigor (Johnson 2011). As the data become aggregated, inconsistencies will be revealed suggesting possible errors that should be investigated and corrected using an approach that integrates data from all available sources including morphology (Dayrat 2005, Will et al. 2005, Goldstein and DeSalle 2011, Riedel et al. 2013).

Beyond barcoding

In recent years, cost curves for next generation DNA sequencing technologies (NGS) have been falling. As time goes on, it seems inevitable that NGS will become increasingly competitive with traditional Sanger sequencing. NGS approaches are less dependent on long intact DNA fragments compared to the long run Sanger barcoding demonstrated here (Ekblom and Galindo 2011, Lemmon et al. 2012). This suggests that spider collections such as the one at Naturalis may be even richer as a source of data for NGS studies than we found using traditional sequencing.

Acknowledgements

This study was supported by a Naturalis research grant to JM and PvH. Naturalis Rugzakje funding supported JM's participation in a special symposium on spider barcoding at the 2012 European Congress of Arachnology in Ljubljana, Slovenia. Thanks to Matjaž Kuntner (Slovenian Academy of Sciences and Arts), Jason Bond (Auburn University), Wolfgang Nentwig (University of Bern), Christian Kropf (Naturhistorisches Museum Bern), Miquel Arnedo (Universitat de Barcelona), and Menno Schilthuizen (Naturalis Research) for stimulating discussion and insight. Cor Vink (Canterbury Museum) kindly agreed to read and comment on an earlier draft of the manuscript. Miguel Arnedo, Jason Bond, and two anonymous reviewers contributed constructive comments and useful suggestions. Frank Stokvis of the Naturalis DNA Barcoding Facility generated some of the sequence data. Marcel Tuijt (Naturalis Technische Dienst) and René Dekker (Naturalis Director of Collections) provided data on conditions in the Naturalis collection. Thanks to Karen van Dorp (Naturalis Collections Department) for her interest, assistance and cooperation, and to Zoltán Nagy and the FWO research community "Belgian Network for DNA Barcoding" (BeBOL) for organizing this special issue.

References

Andersen JC, Mills NJ (2012) DNA extraction from museum specimens of parasitic Hymenoptera. PLoS ONE 7: e45549. doi: 10.1371/journal.pone.0045549

Barrett RDH, Hebert PDN (2005) Identifying spiders through DNA barcodes. Canadian Journal of Zoology 83: 481–491. doi: 10.1139/Z05-024

Coddington JA, Levi HW (1991) Systematics and evolution of spiders (Araneae). Annual Review of Ecology and Systematics 22: 565–592. http://www.jstor.org/stable/2097274, doi: 10.1146/annurev.es.22.110191.003025

Dayrat B (2005) Towards integrative taxonomy. Biological Journal of the Linnean Society 85: 407–415. doi: 10.1111/j.1095-8312.2005.00503.x

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107: 1–15. doi: 10.1038/hdy.2010.152

Felsenstein J (1985) Phylogenies and the comparative method. American Naturalist 125: 1–15. http://www.jstor.org/stable/2461605

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for the amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3: 294–299. http://www.ncbi.nlm.nih.gov/pubmed/7881515

Garland T, Harvey PH, Ives AR (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology 41: 18–32. doi: 10.1093/sysbio/41.1.18

Goldstein PZ, DeSalle R (2011) Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. Bioessays 33: 135–147. doi: 10.1002/bies.201000036

Gurdebeke S, Maelfait J-P (2002) Pitfall trapping in population genetics studies: finding the right "solution". Journal of Arachnology 30: 255–261. http://www.jstor.org/stable/3706268, doi: 10.1636/0161-8202(2002)030[0255:PTIPGS]2.0.CO;2

Hammer Ø, Harper DAT, Ryan PD (2001) PAST: Paleontological Statistical software package for education and data analysis. Palaeontologia Electronica 4: 9 pp.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Helsdingen PJv (1999) Catalogus van de Nederlandse spinnen (Araneae). Nederlandse Faunistiche Mededelingen 10: 1–189.

Helsdingen PJv (2013) Araneae, Spiders. Fauna Europaea. Version 26. http://www.faunaeur.org/

Janzen DH, Hallwachs W, Blandin P, Burns JM, Cadiou J-M, Chacon I, Dapkey T, Deans AR, Epstein ME, Espinoza B, Franclemont JG, Haber WA, Hajibabaei M, Hall JPW, Hebert PDN, Gauld ID, Harvey DJ, Hausmann A, Kitching IJ, Lafontaine D, Landry J-F, Lemaire C, Miller JY, Miller JS, Miller L, Miller SE, Montero J, Munroe E, Green SR, Ratnasingham S, Rawlins JE, Robbins RK, Rodriguez JJ, Rougerie R, Sharkey MJ, Smith MA, Solis MA, Sullivan JB, Thiaucourt P, Wahl DB, Weller SJ, Whitfield JB, Willmott KR, Wood DM, Woodley NE, Wilson JJ (2009) Integration of DNA barcoding into

an ongoing inventory of complex tropical biodiversity. Molecular Ecology Resources 9 (Suppl. 1): 1–26. doi: 10.1111/j.1755-0998.2009.02628.x

Johnson NF (2011) A collaborative, integrated and electronic future for taxonomy. Invertebrate Systematics 25: 471–475. doi: 10.1071/IS11052

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. Systematic Biology 61: 727–744. doi: 10.1093/sysbio/sys049

Maddison WP, Maddison DR (2011) Mesquite: A Modular System for Evolutionary Analysis. Version 2.75. http://mesquiteproject.org

Midford PE, Garland T, Maddison WP (2010) PDAP Package of Mesquite. Version 1.16

Nentwig W, Blick T, Gloor D, Hänggi A, Kropf C (Eds) (2013) Spiders of Europe. Version 06.2013. http://www.araneae.unibe.ch/

Paquin P, Vink CJ (2009) Testing compatibility between molecular and morphological techniques for arthropod systematics: a minimally destructive DNA extraction method that preserved morphological integrity, and the effect of lactic acid on DNA quality. Journal of Insect Conservation 13: 453–457. doi: 10.1007/s10841-008-9183-0

Platnick N (2013) The world spider catalog, version 13.5. American Museum of Natural History. http://research.amnh.org/iz/spiders/catalog, doi: 10.5531/db.iz.0001

Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (www.barcodinglife.org). Molecular Ecology Notes 7: 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Riedel A, Sagata K, Suhardjono YR, Tänzler R, Balke M (2013) Integrative taxonomy on the fast track – towards more sustainability in biodiversity research. Frontiers in Zoology 10: 15. http://www.frontiersinzoology.com/content/10/1/15, doi: 10.1186/1742-9994-10-15

Roberts MJ (1985) The Spiders of Great Britain and Ireland, Volume 1: Atypidae to Theridiosomatidae. Harley Books, Colchester.

Roberts MJ (1987) The Spiders of Great Britain and Ireland, Volume 2: Linyphiidae and check list. Harley Books, Colchester.

Rowley DL, Coddington JA, Gates MW, Norrbom AL, Ochoa RA, Vandenberg NJ, Greenstone MH (2007) Vouchering DNA-barcoded specimens: test of a nondestructive extraction protocol for terrestrial arthropods. Molecular Ecology Notes 7: 915–924. doi: 10.1111/j.1471-8286.2007.01905.x

Smith MA, Fisher BL, Hebert PDN (2005) DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. Philosophical Transactions of the Royal Society B 360: 1825–1834. doi: 10.1098/rstb.2005.1714

Van Houdt JKJ, Breman FC, Virgilio M, De Meyer M (2010) Recovering full DNA barcodes from natural history collections of tephritid fruitflies (Tephritidae, Diptera) using mini barcodes. Molecular Ecology Resources 10: 459–465. doi: 10.1111/j.1755-0998.2009.02800.x

Vink CJ, Thomas SM, Paquin P, Hayashi CY, Hedin M (2005) The effects of preservatives and temperatures on arachnid DNA. Invertebrate Systematics 19: 99–104. doi: 10.1071/IS04039

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 54: 844–851. doi: 10.1080/10635150500354878

Xia X, Xie Z (2001) DAMBE: Data analysis in molecular biology and evolution. Journal of Heredity 92: 371-373. doi: 10.1093/jhered/92.4.371

Zuccon D, Brisset J, Corbari L, Puillandre N, Utge J, Samadi S (2012) An optimised protocol for barcoding museum collections of decapod crustaceans: a case-study for a 10-40-years-old collection. Invertebrate Systematics 26: 592–600. doi: 10.1071/IS12027

Appendix

Electronic supplementary documents. (doi: 10.3897/zookeys.365.5787.app) File format: WinZip Archive. (zip).

Explanation note: Archive contents:

Figure 1.doc – Correlations based on independent contrasts.

Milleretal2013DutchSpiderBarcodeDwC.txt – Occurrence data for all specimens in the study.

Milleretal2013DutchSpiderBarcode.kml – Occurrence data in KML (keyhole markup language.

Milleretal2013DutchSpiderBarcode.fasta – All sequence data for this study in fasta format.

Milleretal2013DutchSpiderBarcodeDestructiveExtractions.nex – Nexus files generated using Mesquite (http://mesquiteproject.org).

Milleretal2013DutchSpiderBarcodeAllExtractions.nex – Nexus files generated using Mesquite (http://mesquiteproject.org).

Citation: Miller JA, Beentjes KK, van Helsdingen P, IJland S (2013) Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. In: Nagt ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 245–261. doi: 10.3897/zookeys.365.5787 Electronic supplementary documents. doi: 10.3897/zookeys.365.5787.app

ZooKeys 365: 263–278 (2013)
doi: 10.3897/zookeys.365.6070
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Using DNA barcodes for assessing diversity in the family Hybotidae (Diptera, Empidoidea)

Zoltán T. Nagy[1], Gontran Sonet[1], Jonas Mortelmans[2], Camille Vandewynkel[3], Patrick Grootaert[2]

1 *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (JEMU), Rue Vautierstraat 29, 1000 Brussels, Belgium* **2** *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (Entomology), Rue Vautierstraat 29, 1000 Brussels, Belgium* **3** *Laboratoire des Sciences de l'eau et environnement, Faculté des Sciences et Techniques, Avenue Albert Thomas, 23, 87060 Limoges, France*

Corresponding author: *Zoltán T. Nagy* (ztnagy@naturalsciences.be)

Academic editor: *K. Jordaens* | Received 7 August 2013 | Accepted 27 November 2013 | Published 30 December 2013

Citation: Nagy ZT, Sonet G, Mortelmans J, Vandewynkel C, Grootaert P (2013) Using DNA barcodes for assessing diversity in the family Hybotidae (Diptera, Empidoidea). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 263–278. doi: 10.3897/zookeys.365.6070

Abstract

Empidoidea is one of the largest extant lineages of flies, but phylogenetic relationships among species of this group are poorly investigated and global diversity remains scarcely assessed. In this context, one of the most enigmatic empidoid families is Hybotidae. Within the framework of a pilot study, we barcoded 339 specimens of Old World hybotids belonging to 164 species and 22 genera (plus two *Empis* as outgroups) and attempted to evaluate whether patterns of intra- and interspecific divergences match the current taxonomy. We used a large sampling of diverse Hybotidae. The material came from the Palaearctic (Belgium, France, Portugal and Russian Caucasus), the Afrotropic (Democratic Republic of the Congo) and the Oriental realms (Singapore and Thailand). Thereby, we optimized lab protocols for barcoding hybotids. Although DNA barcodes generally well distinguished recognized taxa, the study also revealed a number of unexpected phenomena: e.g., undescribed taxa found within morphologically very similar or identical specimens, especially when geographic distance was large; some morphologically distinct species showed no genetic divergence; or different pattern of intraspecific divergence between populations or closely related species. Using COI sequences and simple Neighbour-Joining tree reconstructions, the monophyly of many species- and genus-level taxa was well supported, but more inclusive taxonomical levels did not receive significant bootstrap support. We conclude that in hybotids DNA barcoding might be well used to identify species, when two main constraints are considered. First, incomplete barcoding libraries hinder

efficient (correct) identification. Therefore, extra efforts are needed to increase the representation of hybotids in these databases. Second, the spatial scale of sampling has to be taken into account, and especially for widespread species or species complexes with unclear taxonomy, an integrative approach has to be used to clarify species boundaries and identities.

Keywords
COI, cryptic species, DNA barcoding, geographic distances, taxonomy

Introduction

With over 11,400 described species, the superfamily Empidoidea represents one of the largest extant lineages of flies (Diptera, Brachycera) (Evenhuis et al. 2007, Pape et al. 2009). According to the most recent systematic revision (Sinclair and Cumming 2006), this superfamily comprises five families: the Atelestidae, Brachystomatidae, Dolichopodidae *sensu lato*, Empididae and Hybotidae. Commonly known as 'dance flies', the Empidoidea most likely originated in the Mesozoic (ca. 150 million years ago, Wiegmann et al. 2003) and now have a nearly cosmopolitan distribution. The high species diversity of this group is matched by high morphological diversity which is also very well expressed in genitalic traits. Genital morphology is still the main decisive diagnostic character used in the morphological identification and subsequent classification.

Studies carried out over the last few decades indicate the family Hybotidae is to be monophyletic (Chvala 1983, Collins and Wiegmann 2002, Sinclair and Cumming 2006, Moulton and Wiegmann 2007). The family includes ca. 2000 described species worldwide (Yang et al. 2007), and typically consists of small-bodied insects (i.e. 1–6 mm in total length). The vast majority of known hybotid species are predators that either hunt their prey in the air (e.g., some Ocydromiinae, Hybotinae) or on the ground (Tachydromiinae). These flies can be easily recognized by a spherical head with distinctive morphology as described by Sinclair and Cumming (2006), the presence of a palpifer, and fore-tibial gland, restriction of the gonocoxal apodeme to the anterolateral margin of the hypandrium, apex of antenna often with long, slender seta-like receptor, laterotergite bare, and R_{4+5} unbranched. Their male genitalia are also distinctive, and spectacular, the hypopygium being often rotated 45–90° to the right, so that the cerci, which are usually located on the dorsal side of the animal are now on the right side of the body. Sinclair and Cumming (2006) classified the Hybotidae into five subfamilies: the Hybotinae, Ocydromiinae, Oedaleinae, Tachydromiinae and Trichininae (the genus *Stuckenbergomyia* Smith, 1969 does not seem to fit into any of these and probably deserves its own subfamily). However, our current understanding of the phylogenetics, taxonomy and natural history of the Hybotidae is limited (Collins and Wiegmann 2002, Moulton and Wiegmann 2007), with several groups being little known (Sinclair and Cumming 2006). In addition, large parts of the distributional range of this family have been poorly explored (e.g. Central Africa, the Oriental region and Neotropics). In some of these regions the diversity of hybotid flies has probably been greatly under-estimated. For instance, Grootaert and Shamshev (2013)

recently described 25 new hybotid species, all of which were collected during a short field expedition in the Democratic Republic of the Congo (D. R. Congo). The current study utilizes new tissue samples from specimens from a range of localities in Europe, Asia and Africa. It has been made possible by extensive field collections carried out by the senior author (P.G.), who has added substantially to material currently available from the Old World.

DNA barcoding (Hebert et al. 2003), based on a ca. 650 bp DNA region from the 5' end of the mitochondrial COI gene, is an easy-to-use molecular approach that allows quick assignment of samples into 'genetic' groups. In situations where a reference database of specimen data and morphospecies identifications exists, this technique can be used for exploring and comparing species limits as defined by morphological vs. DNA-based criteria. Although the family Hybotidae is species-rich, genetic data (i.e., DNA barcode sequences) for this group are surprisingly scarce in public databases, such as GenBank and The Barcode of Life Data Systems – BOLD (Ratnasingham and Hebert 2007). While there are over 500 COI sequences of Empidoidea in GenBank, we could only find a single, correctly classified Hybotidae sequence. Furthermore, although there are several DNA barcoding projects underway in North America that are analyzing large numbers of Hybotid species, sequence data from these studies have yet to be made available to the public. We found four DNA barcode sequences of hybotids in BOLD, but these are from specimens collected in Canada, and therefore fall outside the geographical scope of our study, which is restricted to Old World taxa.

In the current paper, we optimized protocols for DNA barcoding of hybotid flies and performed a preliminary barcoding study on selected genera and species of this group. We hope that this approach will accelerate an inventory of hybotid flies. Here, we investigated the ability of the barcoding data coming from a large sampling of diverse Hybotidae to reveal cryptic species, patterns of geographic variation, and putative new species.

Material and methods

A total of 339 specimens, representing 164 morphospecies of Hybotidae (see Supplementary file 1) were selected and sequenced for this study. All material was collected between 2008 and 2012 in three biogeographic realms: the Palaearctic (Belgium, Portugal, France and Russian Caucasus), Afrotropical (D. R. Congo) and Oriental (Singapore and Thailand) realms. Our outgroup taxon was *Empis tessellata* Fabricius, 1794 (Empidoidea, Empididae), of which two individuals were sequenced. Specimens were collected mainly using sweep nets and Malaise traps, and were initially preserved in 70% ethanol. After identification, all specimens were transferred to 96% ethanol and then stored at 4 °C in order to minimize DNA degradation. Either complete specimens or mid or hind legs were used for total genomic DNA extraction. Immediately prior to extraction, each tissue sample was placed in a microtube and air-dried. DNA extractions were carried out using the NucleoSpin Tissue kit (Macherey-Nagel). We followed

the manufacturer's instructions, but used a longer lysis time (i.e. around 24 hours). After lysis, fly specimens were transferred to absolute ethanol and were put back to the collection. Voucher specimens have been deposited in the entomological collections of the Royal Belgian Institute of Natural Sciences (RBINS).

PCR conditions were optimized by testing primer concentrations of 0.1, 0.2 and 0.4 μM and $MgCl_2$ concentrations of 1.5–2 mM against a gradient of annealing temperatures. The best results were obtained by using the protocol as follows: each reaction (total volume of 25 μl) contained 2–3 μl DNA extract, 0.4 μM of each primer, 0.03 unit/μl Platinum Taq polymerase (Invitrogen), 1 × PCR Buffer, 0.2 mM dNTP, 2 mM $MgCl_2$ and ca. 15 μl of sterile water. The COI region of interest was amplified using the standard animal barcoding primers, LCO1490 and HCO2198 (Folmer et al. 1994), and the primer pair TY-J-1460 and C1-N-2191 (Wells and Sperling 1999), with an annealing temperature of 45 °C and 48 °C, respectively, and 40 PCR cycles. PCR results were assessed using agarose gel electrophoresis and PCR products were purified on NucleoFast 96 PCR Plates (Macherey-Nagel). Sanger sequencing was carried out on an ABI 313l automated capillary sequencer using BigDye v1.1 or v3.1 chemistry (Life Technologies).

DNA sequences were checked and assembled with SeqScape v2.5 (Life Technologies). Neighbour-Joining (NJ) trees based on uncorrected (p) distances were calculated in MEGA5 (Tamura et al. 2011). Non-parametric bootstrapping with 1000 replicates was performed to evaluate branch support. Pairwise divergences at three levels (intraspecific, interspecific and intrageneric, as well as interspecific but not intrageneric) were calculated using R v2.15.2 and the *ape* package (Paradis et al. 2004).

Results

Our sampling covered all the currently accepted subfamilies and tribes of the Hybotidae. At the generic level, we investigated 22 of the 66 known genera (see Table 1 for full details). The DNA sequence data set consisted of 341 COI sequences (339 Hybotidae and two Empididae), each sequence being of 657 bp in length. These sequences were deposited in BOLD and GenBank (BOLD Process IDs EMPID001-13 – EMPID341-13).

An NJ tree without species names and additional sample information is shown in Figure 1 (a fully annotated tree is shown in Supplementary File 2). 'Species-level' groups (i.e., close to or at terminal nodes) were generally well supported, while deep-level groups were not (see Figure 1 and Supplementary File 2). Especially when low intraspecific distance was observed, these groups (considered as molecular operational taxonomic units – MOTUs) often received 100% bootstrap support. At a 1% distance threshold (as it is also used by BOLD), 99% of the clusters (i.e., 70 out of 71 MOTUs) were supported by bootstrap values above 95%. Many recognized species were well resolved and distinguished using the COI data, but we observed a number of problems that are discussed below. Although representatives of more inclusive taxa, such as tribes

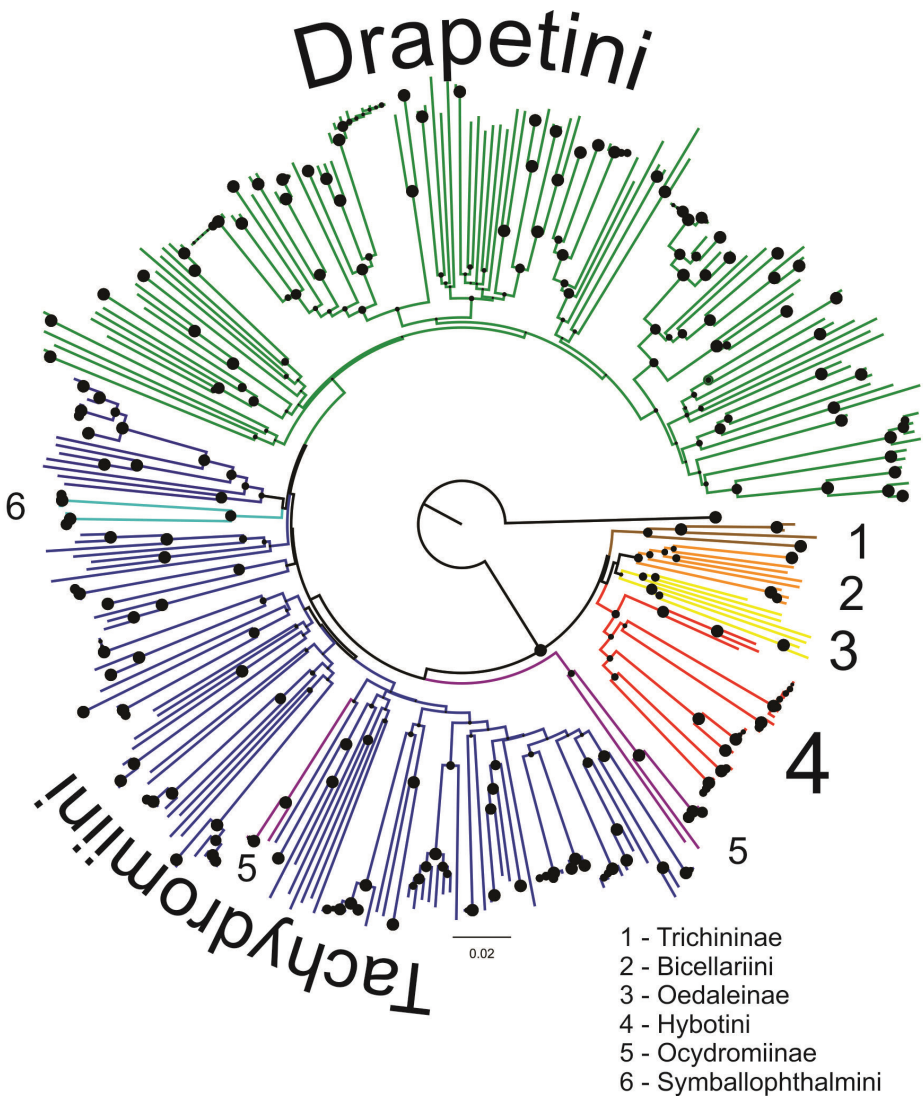


Figure 1. Neighbour-Joining tree representing hybotid diversity of 339 selected samples. The tree was rooted with *Empis tessellata* (Empididae). Circles represent branch supports, bootstrap values are according to circles' size.

and subfamilies (with the exception of the Ocydromiinae and Tachydromiini), were usually recovered in single clusters, these clusters were not supported (bootstrap values < 70%). The only exception is the tribe Symballophthalmini, represented in the data set by just two species, which was supported by a bootstrap value of 87.7%.

For most genera, the number of species represented in our analysis was very limited. Similarly, the number of conspecific sequences was also generally low, ranging between 1–9. Nonetheless, we observed considerable overlaps between intraspecific

Table 1. Global suprageneric systematics of Hybotidae (without the genus *Stuckenbergomyia*) and genera investigated in the current barcoding study.

Subfamily (Tribe)	Number of genera	Investigated genera
Trichininae	2	1 (*Trichina*)
Ocydromiinae	15	3 (*Leptopeza*, *Ocydromia*, *Oropezella*)
Oedaleinae	4	3 (*Allanthalia*, *Euthyneura*, *Oedalea*)
Tachydromiinae		
- Symballophthalmini	1	1 (*Symballophthalmus*)
- Tachydromiini	8	4 (*Ariasella*, *Platypalpus*, *Tachydromia*, *Tachypeza*)
- Drapetini	18	6 (*Chersodromia*, *Crossopalpus*, *Drapetis*, *Elaphropeza*, *Nanodromia*, *Stilpon*)
Hybotinae		
- Bicellariini	13	1 (*Bicellaria*)
- Hybotini	14	3 (*Hybos*, *Syndyas*, *Syneches*)

Table 2. Patterns of intra- and interspecific distances observed in four species-rich genera of our dataset.

Tribe	Genus	No. of species	No. of sequences	No. of haplotypes	Intraspecific distances (%)	Interspecific distances (%)
Tachydromiini	*Platypalpus*	45	98	81	0–16.89	0–18.72
Tachydromiini	*Tachydromia*	12	21	18	0–5.48	1.07–18.11
Drapetini	*Chersodromia*	12	36	26	0–3.04	6.09–15.53
Drapetini	*Elaphropeza*	43	75	68	0–5.48	1.83–19.63

(0–17.2%) and interspecific divergences (0–21.81%). Among congeners, interspecific divergences ranged between 0–19.9%, while we observed higher divergence between samples of different genera (5.85–21.81%). Hence, no barcoding gaps existed between any of these ranks. The ranges of pairwise distances were overall high in the four genera represented by the highest number of samples (Table 2). We observed extensive overlap between intra- and interspecific divergences in both the species-rich genera of Tachydromiini, *Platypalpus* and *Tachydromia*, with less extensive, or no overlap, in the genera *Chersodromia* and *Elaphropeza*, both belonging to the Drapetini (Table 2).

Below, we describe five categories of cases where ranges of intra- and interspecific distances did not seem consistent with the current taxonomy and would require more investigation.

Different patterns of intraspecific divergence in congeneric species

We found that in some congeneric species the levels of sequence variation observed both within populations and between populations in close proximity were low. Contrastingly, other congeneric species showed widely different levels of intraspecific divergence. An interesting case in this respect is the brachypterous *Chersodromia curtipennis* and the fully-winged *C. pontica*, both of which occur on the Taman Peninsula (Krasnodar region of Russia). Samples were taken at various sites on the Taman Peninsula, ranging from

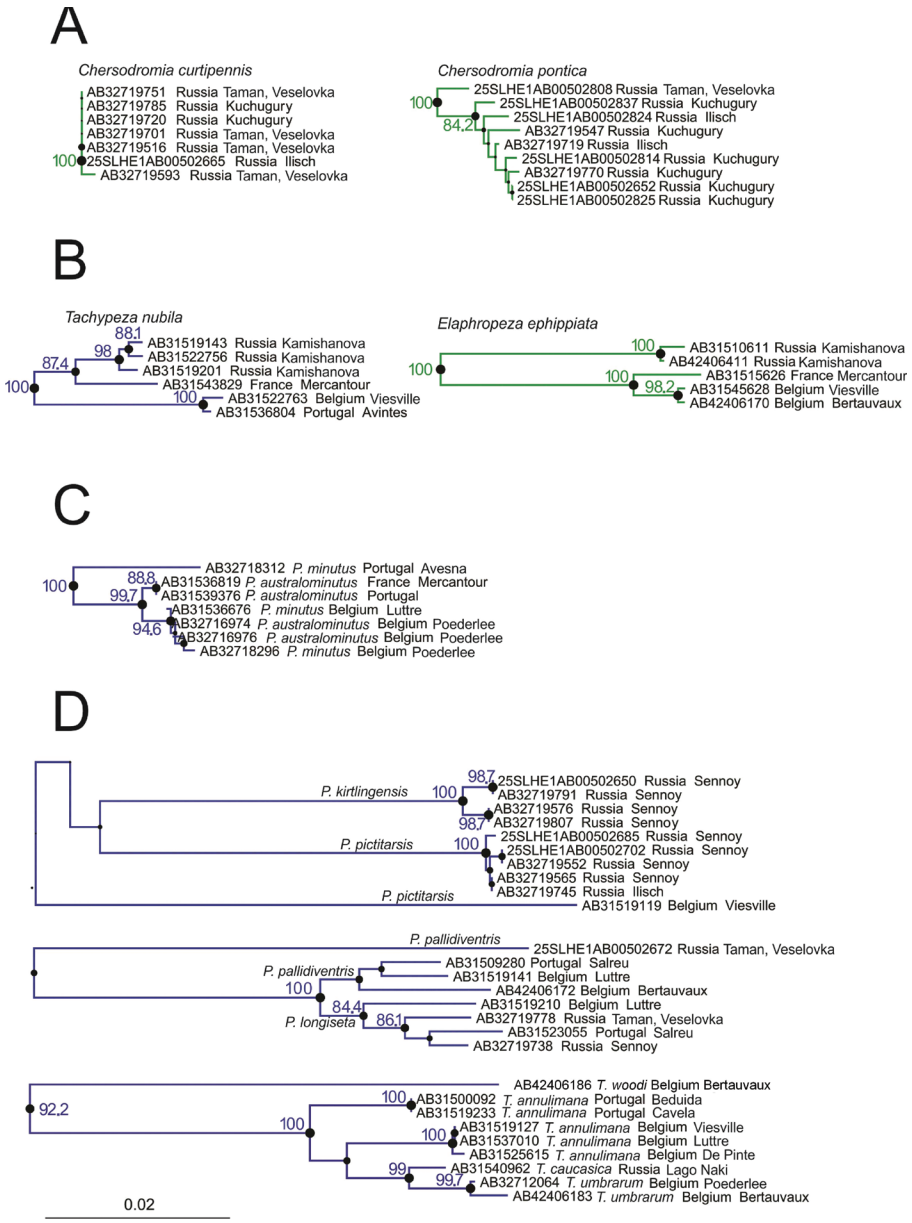


Figure 2. Subtrees showing cases where ranges of intra- and interspecific distances do not seem consistent with the current taxonomy and would require more investigation. See details in text. Circles represent branch supports. Bootstrap values are according to circles' size, bootstrap values are shown in numbers when > 80%.

the North, on the coast of the Sea of Azov, to the South along the Black Sea (Taman: Veselovka). While the brachypterous species showed virtually no genetic variation (uncorrected pairwise divergence was between 0–0.15%), the fully-winged species showed an expressed pattern of divergence with pairwise p-distances of 0–1.37% (Figure 2A).

Variation in intraspecific divergence may be related to spatial distance

Some species sampled across large geographical areas showed high levels of genetic divergence between populations. This is among others the case for two species that are widespread and very common in Europe: *Tachypeza nubila* and *Elaphropeza ephippiata*. While we could not detect any morphological differences (i.e. of the body and the male genitalia) between populations in western European and the Russian Caucasus, intraspecific pairwise genetic divergences ranged between 0.3–3.5% in *T. nubila* and between 0.2–5.48% in *E. ephippiata* (Figure 2B). Unexpectedly large 'intraspecific' divergences may indicate undescribed diversity at the species level. In many cases, large 'intraspecific' divergences were found between specimens from the same locality or from adjacent sites (Table 3, upper part), and examples in this respect include *Platypalpus caucasicus* (Russian Caucasus), *Platypalpus annulipes* (Belgium), *Trichina elongata* (Russian Caucasus), *Bicellaria nigra* (Russian Caucasus), *Tachydromia annulimana* (within Europe), *Elaphropeza nuda* (D. R. Congo) and *Elaphropeza monospina* (Singapore). In a number of other cases, large 'intraspecific' divergences were observed between geographically distant populations (Table 3, lower part); this was observed for *Platypalpus pictitarsis* (Russian Caucasus versus Belgium), *Platypalpus pallidiventris* (Russian Caucasus versus Europe), *Leptopeza flavipes* (Russian Caucasus versus Belgium), *Oedalea zetterstedti* (Russian Caucasus versus Belgium), *Euthyneura myrtilli* (Russian Caucasus versus Europe), *Platypalpus nigritarsis* (Russian Caucasus versus France) and *Tachydromia aemula* (Russian Caucasus versus Portugal). In all of these cases, morphological differences of genitalia (or other diagnostic characters) were not assessed in details, and therefore these divergences may well reflect interspecific differences. Remarkably, no significant differences in divergence ranges were observed between the two types of cases (i.e., associated or not with large spatial distances).

Genetic overlap of putative 'sister' species

Platypalpus minutus and *P. australominutus* are externally very similar except that male genitalia are consistently different (Grootaert 1989): In northern Belgium both species are sympatric and often occur syntopically. More to the South of Belgium and in the South of France mainly *P. australominutus* occurs. The Belgian specimens of these two species could not be distinguished by COI sequences due to shared haplotypes (Figure 2C). However, a specimen from Portugal provisionally identified as *P. minutus* was quite different from the clade *australominutus-minutus* from Belgium.

Complex taxonomy

In Figure 2D, three examples are shown where the unclear taxonomy of the involved species or species complex was reflected in para- or polyphyletic taxa. For example,

Table 3. Range of pairwise p-distances in cases where unexpectedly high 'intraspecific' divergence was observed (> 5%).

Species or species complex	Range of pairwise p-distances (%)
Platypalpus caucasicus	0.46–8.07
Platypalpus annulipes	0–9.80
Trichina elongata	0.91–8.83
Bicellaria nigra	9.44
Tachydromia annulimana	0–10.35
Elaphropeza nuda	0–5.33
Elaphropeza monospina	5.33
Platypalpus pictitarsis	0–10.20
Platypalpus pallidiventris	1.37–10.05
Leptopeza flavipes	0–7.01
Oedalea zetterstedti	7.91
Euthyneura myrtilli	1.52–10.96
Platypalpus nigritarsis	5.33
Tachydromia aemula	5.48

Platypalpus pictitarsis and *P. kirtlingensis* are morphologically very similar. They differ in the colour of the fore leg and the palpus. Barcode sequences showed that both species are genetically different (uncorrected pairwise interspecific divergence was at least 8.37%). In addition, however, a single male of *P. pictitarsis* from Belgium (AB31519119) rendered this taxon polyphyletic. Both external morphology and genitalia of the Belgian and Caucasian *pictitarsis* was the same and a deeper study will be needed to clarify this issue.

Another example involved the sister species *Platypalpus pallidiventris* and *P. longiseta*, which differ morphologically only in a few but distinct characters. Also, these species were genetically closely related except a single specimen of *P. pallidiventris* (25SLHE1AB00502672, see Figure 2D), which rendered this species paraphyletic. This single specimen of *P. pallidiventris* from Caucasus is a female, exhibiting less diagnostic characters than males, and could therefore belong to another species. This observation urges for a more intensive collection and study of these sister species. When we discarded this divergent sequence, both species showed a moderate intraspecific structuring. The bootstrap value supporting the cluster containing both species without the divergent specimen was 100%. In addition, the reciprocal monophyly of both species was supported with bootstrap values of 77.3% and 84.4% for *P. pallidiventris* and *P. longiseta*, respectively.

A third example involved four species (Figure 2D). Originally, a female (AB42406186) of *T. woodi* was identified as *T. annulimana*. However, the considerable divergence at COI between this specimen and all other specimens of *T. annulimana* (10.35%) suggested a misidentification. A reexamination of the specimen revealed that *T. woodi* has the costa between vein Rnd R_{2+3} thickened, an unpublished feature that confirmed the misidentification. *T. caucasica* from Caucasus and *T. um-*

brarum from Belgium, both also belonging to the *annulimana*-group, showed a very low interspecific divergence (1.07-1.52%). This suggests, in combination with the little morphological differences reported between the two species, a very close relationship between the two species and does not exclude that they are conspecific.

Discussion

The barcoding of dipterans commenced relatively early as part of the global DNA barcoding initiative. On the one hand, DNA barcoding performed well in several lineages. However, so far mostly Holarctic dipterans have been investigated where species diversity is overall lower than in tropical biomes. For instance, DNA barcoding proved to work well for Canadian (Cywinska et al. 2006) and Chinese mosquitoes (Wang et al. 2012), Nearctic simuliids (Rivera and Currie 2009) and muscids (Renaud et al. 2012) and so on, but this approach was less extensively tested on tropical taxa. On the other hand, the usefulness of dipteran DNA barcodes in species identification has been criticized (e.g., Meier et al. 2006, Whitworth et al. 2007) and new criteria for specific assignment have been introduced (Meier et al. 2006, best match and best close match criteria). An inherent problem with dipterans is the possibly high amount of unknown diversity on a global scale leading to incomplete databases, the substantial age of some large radiations that is linked to (very) high mitochondrial sequence diversity, and the limited taxonomic expertize on particular groups hampering successful identification. Unfortunately, reference barcode libraries of species-rich taxa are often incomplete. In fact, in many insect groups a few common species are overwhelmed by a high number of rare species (Lim et al. 2011). While common species are likely better represented, rare species are often missing in barcode libraries. This may lead to imbalanced taxon representation. Another critical issue of DNA barcoding is the effect of geographical sampling (Bergsten et al. 2012). Generally, identification success is dropping with increasing spatial scale of sampling, and may pose a real problem for all widespread taxa. In summary, all of these issues make DNA barcoding difficult. Nevertheless, DNA barcoding has been generally advocated as a pragmatic first step in the integrative taxonomic framework, also for problematic taxa (Tan et al. 2010, Nagy et al. 2012).

In the meantime, several dipteran barcoding projects have been started, particularly with respect to medically, forensically or commercially (e.g., related to agriculture) relevant lineages such as mosquitoes, muscids, tephritids and drosophilids (see details at http://boldsystems.org). Hybotids, or more broadly the empidoids, have no known medical, forensic or commercial importance, therefore there are overall much less intensively studied. The current dataset presented herein is a result of a pilot study focusing on Old World hybotid diversity. An overall high sequence divergence was observed in our dataset, which is not surprising in the light of the age and diversification pattern of dance flies (Wiegmann et al. 2003). Although most species could be well distinguished based on a single mitochondrial marker (Figure 1 and Supplementary

file 2), we observed several inconsistencies with current classification and extensive overlaps of intra- and interspecific divergences.

Our limited sampling of specific and subspecific levels with up to nine samples per species did not allow us to perform extensive tests on barcoding performance and species (or genus) delimitations. We are also aware of the potential pitfalls when analyzing taxonomically incomplete datasets. In such cases, a hierarchical sampling should be followed whenever possible. In these cases, the number of sampled genera and more inclusive taxa should be maximized (Zhang et al. 2013). Here, we sampled all subfamilies and tribes, as well as one third of all hybotid genera, but sampling at the specific level remained far below 10%. Simulation of the sampling effect can be performed (e.g., Nagy et al. 2012), and this simulation may give hints about the power of DNA barcoding. Regarding species delimitation, simple methods relying solely on genetic distances are still broadly used, although there are many inherent problems with them (e.g., Meier et al. 2006). First, species delimitation simply based on genetic divergence is difficult to convey and interpret in a "universally acceptable" species concept (Krishnamurthy and Francis 2012). Second, large intraspecific distances and low interspecific distances among closely related species may pose a major problem, and even the use of refined criteria such as best close match (Meier et al. 2006) or *ad hoc* thresholds (Virgilio et al. 2012) might not solve this issue. Therefore, in datasets where intra- and interspecific divergences largely overlap, using distance-based thresholds alone may not work. In these cases, species or species complexes may have to be analyzed individually and also other DNA markers (including nuclear markers) should be considered for species delimitation and perhaps for revising our ideas about species identification.

In the case of recently diverged species, a number of methods have been compared (van Velzen et al. 2012) such as tree-based (Neighbour-Joining or tree-based parsimony), similarity-based (nearest neighbour or BLAST), statistical and diagnostic or character-based (e.g., BLOG: Bertolazzi et al. 2009, DNA-BAR: DasGupta et al. 2005) approaches. Similarity- and character-based methods have been shown to usually outperform tree-based methods (van Velzen et al. 2012), and some studies have found that character-based approaches may work better than distance-based methods (e.g. Bergmann et al. 2013). However, further analytical approaches need to be explored. Irrespective of the approach used, success in species identification can decrease with increasing sampling (Bergsten et al. 2012). Overall, the use of multi-gene markers and coalescent methods seem to be inevitable for efficient species delimitation (see Jörger et al. 2012), but this is clearly beyond the scope of DNA barcoding *sensu stricto*.

Although we focused on problematic or unexpected cases, in most of these examples, DNA barcoding may still be useful, provided that precautions are taken with respect to taxonomic and geographic sampling effects. Moreover, species identification in Hybotidae is based primarily on male terminalia and possibly some of the species concept situations are due to misidentification of females. Also, collecting precise information on collection site, life history, habitat, morphology etc. can very well contribute to the interpretation of the DNA barcoding results. Our finding about intraspecific divergence patterns in the brachypterous vs. the fully-winged species (*Chersodromia cur-*

tipennis and *C. pontica*, respectively) exemplifies this well. The reduced mobility of the brachypterous species is apparently linked to the low intraspecific diversity but mechanisms are still unclear. In many cases where we found unexpectedly large intraspecific divergences (Table 3), we probably deal with undescribed species, and therefore, in fact, with interspecific divergences. Nevertheless, further investigations are necessary to clarify the taxonomic status of the divergent populations. We advocate in-depth investigations involving more diagnostic traits and multi-gene analyses, evaluated in an integrative taxonomic framework (Padial et al. 2010), even if these analyses may take longer time, and cost more (e.g., additional lab work needed to obtain further sequence data).

Conclusions

In the current study, we provided a baseline for further studies on hybotid diversity using a DNA barcoding approach. We provided an optimized lab protocol for routine barcoding. We conclude that DNA barcoding can assist to identify hybotid taxa. Also cryptic species may be revealed by appropriate genetic markers, mostly because the morphological differences are not well assessed. Nevertheless, we emphasize to have an integrative look on barcoding data, and use this approach as a pragmatic first step in taxonomic practice or for biodiversity assessments.

Acknowledgements

We thank Jean-Luc Renneson, Marc Pollet, Pol Limbourg, Alain Drumont (Belgium), Christophe Daugeron (France), Rui Andrade (Portugal), Igor Shamshev and Semen Kustov (Russia) and Adrian Plant (United Kingdom) for their assistance. We also thank three anonymous reviewers for their constructive comments. We appreciate the assistance of Dinarzarde Raheem (Belgium), her comments and corrections improved the text significantly. We also thank the National Parks of Singapore (Singapore) and the University of Kisangani (D. R. Congo). JEMU is financed by the Belgian Science Policy Office (Belspo).

References

Bergmann T, Rach J, Damm S, DeSalle R, Schierwater B, Hadrys H (2013) The potential of distance-based thresholds and character-based DNA barcoding for defining problematic taxonomic entities by CO1 and ND1. Molecular Ecology Resources 13: 1069–1081. doi: 10.1111/1755-0998.12125

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN, Ribera I, Nilsson AN, Barraclough TG, Vogler AP (2012) The ef-

fect of geographical scale of sampling on DNA barcoding. Systematic Biology 61: 851–869. doi: 10.1093/sysbio/sys037

Bertolazzi P, Felici G, Weitschek E (2009) Learning to classify species with barcodes. BMC Bioinformatics 10 (Suppl 14): S7. doi: 10.1186/1471-2105-10-S14-S7

Chvála M (1983) The Empidoidea (Diptera) of Fennoscandia and Denmark. II. General Part. The families Hybotidae, Atelestidae and Microphoridae. Fauna Entomologica Scandinavica 12: 1–279.

Collins KP, Wiegmann BM (2002) Phylogenetic relationships and placement of the Empidoidea (Diptera: Brachycera) based on 28S rDNA and EF– sequences. Insect Systematics and Evolution 33: 421–444. doi: 10.1163/187631202X00226

Cywinska A, Hunter FF, Hebert PDN (2006) Identifying Canadian mosquitoes through DNA barcodes. Medical and Veterinary Entomology 20: 413–424. doi: 10.1111/j.1365-2915.2006.00653.x

DasGupta B, Konwar KM, Măndoiu II, Shvartsman AA (2005) DNA-BAR: distinguisher selection for DNA barcoding. Bioinformatics 21: 3424–3426. doi: 10.1093/bioinformatics/bti547

Evenhuis NL, Pape T, Pont AC, Thompson FC (Eds) (2007) V Biosystematic Database of World Diptera. http://www.diptera.org/Diptera/biosys.htm

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: 294–297.

Grootaert P (1989) Description of a new *Platypalpus* species, closely allied to *P. minutus* Meigen (Diptera Empidoidea Hybotidae) from Europe. Bulletin et Annales de la Société royale belge d'Entomologie 125: 243–250.

Grootaert P, Shamshev I (2013) The flies of the family Hybotidae (Diptera, Empidoidea) collected during the Boyekoli Ebale Congo 2010 Expedition in Democratic Republic of Congo. Zootaxa 3603: 1–61. doi: 10.11646/zootaxa.3603.1.1

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Jörger KM, Norenburg JL, Wilson NG, Schrödl M (2012) Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. BMC Evolutionary Biology 12: 245. doi: 10.1186/1471-2148-12-245

Krishnamurthy PK, Francis RA (2012) A critical review on the utility of DNA barcoding in biodiversity conservation. Biodiversity and Conservation 21: 1901–1919. doi: 10.1007/s10531-012-0306-2

Lim GS, Balke M, Meier R (2011) Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Systematic Biology 61: 165–169. doi: 10.1093/sysbio/syr030

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Moulton JK, Wiegmann BM (2007) The phylogenetic relationships of flies in the superfamily Empidoidea (Insecta: Diptera). Molecular Phylogenetics and Evolution 43:701–713. doi: 10.1016/j.ympev.2007.02.029

Nagy ZT, Sonet G, Glaw F, Vences M (2012) First large-scale DNA barcoding assessment of reptiles in the biodiversity hotspot of Madagascar, based on newly designed COI primers. PLoS ONE 7: e34506. doi: 10.1371/journal.pone.0034506

Padial JM, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. Frontiers in Zoology 7: 16. doi: 10.1186/1742-9994-7-16

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290. doi: 10.1093/bioinformatics/btg412

Pape T, Bickel D, Meier R (2009) Diptera diversity: Status, challenges and tools. Brill, Leiden-Boston, 459 pp.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes 7: 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Renaud AK, Savage J, Adamowicz SJ (2012) DNA barcoding of Northern Nearctic Muscidae (Diptera) reveals high correspondence between morphological and molecular species limits. BMC Ecology 12: 24. doi: 10.1186/1472-6785-12-24

Rivera J, Currie D (2009) Identification of Nearctic black flies using DNA barcodes (Diptera: Simuliidae). Molecular Ecology Resources 9: 224–236. doi: 10.1111/j.1755-0998.2009.02648.x

Sinclair BJ, Cumming JM (2006) The morphology, higher-level phylogeny and classification of the Empidoidea (Diptera). Zootaxa 1180: 1–172. http://www.mapress.com/zootaxa/200/zt0118140.pdf

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Tan DSH, Ang Y, Lim GS, Ismail MRB, Meier R (2010) From 'cryptic species' to integrative taxonomy: an iterative process involving DNA sequences, morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae: Diptera). Zoologica Scripta 39: 51–61. doi: 10.1111/j.1463-6409.2009.00408.x

van Velzen R, Weitschek E, Felici G, Bakker FT (2012) DNA barcoding of recently diverged species: Relative performance of matching methods. PLoS ONE 7: e30490. doi: 10.1371/journal.pone.0030490

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M (2012) Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7: e31581. doi: 10.1371/journal.pone.0031581

Wang G, Li C, Guo X, Xing D, Dong Y, Wang Z, Zhang Y, Liu M, Zheng Z, Zhang H, Zhu X, Wu Z, Zhao T (2012) Identifying the main mosquito species in China based on DNA barcoding. PLoS ONE 7: e47051. doi: 10.1371/journal.pone.0047051

Wells JD, Sperling FA (1999) Molecular phylogeny of *Chrysomia albiceps* and *C. rufifacies* (Diptera: Calliphoridae). Journal of Medical Entomology 36: 222–226. http://www.ncbi.nlm.nih.gov/pubmed/10337087

Whitworth TL, Dawson RD, Magalon H, Baudry E (2007) DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). Proceedings of the Royal Society of London B 274: 1731–1739. doi: 10.1098/rspb.2007.0062

Wiegmann BM, Yeates DK, Thorne JL, Kishino H (2003) Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. Systematic Biology 52: 745–756. doi: 10.1080/10635150390250965

Yang D, Zhang KY, Yao G, Zhang JH (2007) World catalog of Empididae (Insecta: Diptera). China Agricultural University Press, Beijing, 1–599.

Zhang W, Fan X, Zhu S, Zhao H, Fu L (2013) Species-specific identification from incomplete sampling: Applying DNA barcodes to monitoring invasive *Solanum* plants. PLoS ONE 8: e55927. doi: 10.1371/journal.pone.0055927

Appendix 1

Samples used in the current study. (doi: 10.3897/zookeys.365.6070.app1) File format: Microsoft Excel file (xls).

Citation: Nagy ZT, Sonet G, Mortelmans J, Vandewynkel C, Grootaert P (2013) Using DNA barcodes for assessing diversity in the family Hybotidae (Diptera, Empidoidea). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 6070: 263–278. doi: 10.3897/zookeys.365.6070 Samples used in the current study. doi: 10.3897/zookeys.365.6070.app1

Appendix 2

Neighbour-Joining tree representing hybotid diversity of 339 selected samples. (doi: 10.3897/zookeys.365.6070.app2) File format: Adobe PDF (pdf).

Explanation note: Neighbour-Joining tree representing hybotid diversity of 339 selected samples. The tree was rooted with *Empis tessellata* (Empididae). Bootstrap support was estimated with 1000 replicates.

Citation: Nagy ZT, Sonet G, Mortelmans J, Vandewynkel C, Grootaert P (2013) Using DNA barcodes for assessing diversity in the family Hybotidae (Diptera, Empidoidea). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 263–278. doi: 10.3897/zookeys.365.6070 Neighbour-Joining tree representing hybotid diversity of 339 selected samples. doi: 10.3897/zookeys.365.6070.app2

ZooKeys 365: 279–305 (2013)
doi: 10.3897/zookeys.365.5819
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Half of the European fruit fly species barcoded (Diptera, Tephritidae); a feasibility test for molecular identification

John Smit[1], Bastian Reijnen[2], Frank Stokvis[2]

1 *European Invertebrate Survey – the Netherlands, P.O. Box 9517, 2300 RA, Leiden, the Netherlands*
2 *Naturalis Biodiversity Centre, P.O. Box 9517, 2300 RA Leiden, the Netherlands*

Corresponding author: *John Smit* (john.smit@naturalis.nl)

Academic editor: *Z. T. Nagy* | Received 18 June 2013 | Accepted 18 October 2013 | Published 30 December 2013

Citation: Smit J, Reijnen B, Stokvis F (2013) Half of the European fruit fly species barcoded (Diptera, Tephritidae); a feasibility test for molecular identification. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 279–305. doi: 10.3897/zookeys.365.5819

Abstract

A feasibility test of molecular identification of European fruit flies (Diptera: Tephritidae) based on COI barcode sequences has been executed. A dataset containing 555 sequences of 135 ingroup species from three subfamilies and 42 genera and one single outgroup species has been analysed. 73.3% of all included species could be identified based on their COI barcode gene, based on similarity and distances. The low success rate is caused by singletons as well as some problematic groups: several species groups within the genus *Terellia* and especially the genus *Urophora*. With slightly more than 100 sequences - almost 20% of the total - this genus alone constitutes the larger part of the failure for molecular identification for this dataset. Deleting the singletons and *Urophora* results in a success-rate of 87.1% of all queries and 93.23% of the not discarded queries as correctly identified. *Urophora* is of special interest due to its economic importance as beneficial species for weed control, therefore it is desirable to have alternative markers for molecular identification.

We demonstrate that the success of DNA barcoding for identification purposes strongly depends on the contents of the database used to BLAST against. Especially the necessity of including multiple specimens per species of geographically distinct populations and different ecologies for the understanding of the intra- versus interspecific variation is demonstrated. Furthermore thresholds and the distinction between true and false positives and negatives should not only be used to increase the reliability of the success of molecular identification but also to point out problematic groups, which should then be flagged in the reference database suggesting alternative methods for identification.

Keywords
COI, DNA barcoding, reference database

Introduction

Tephritidae, or true fruit flies, are a large group of flies (Diptera) with some 4,500 species described (Norrbom et al.1999). The majority of the species are phytophagous. About 35% of them attack soft fruits, including many commercial crops, and some 250 species are considered mild to severe pests (White and Elson-Harris 1992, McPheron and Steck 1996). On the other hand some 40% attack flower heads of or induce galls on Asteraceae, some of which are considered beneficial for the control of invasive weeds outside their natural range (White et al. 1990, White and Elson-Harris 1992, Turner 1996).

Among the economically important taxa five genera have been listed on the quarantine list of the European Union: *Anastrepha* Schiner, 1868, *Bactrocera* Macquart, 1835, *Ceratitis* Macleay, 1829, *Dacus* Fabricius, 1805 and *Rhagoletis* Loew, 1862 (Annex IAI of the Council Directive 2000/29/EC). Most species within these genera are notoriously difficult to identify, therefore the genera are placed on the quarantine list as a whole, despite the fact that not all are pest species. Interceptions on commercial products almost always concern larvae, which are next to impossible to identify. Moreover the number of species that can attack a specific host plant is unknown and the geographic ranges of many species are poorly documented. Therefore there is a desperate need for an alternative method for unambiguous identification of these Tephritid species, especially among plant protection organizations. Hebert et al. (2003) proposed a molecular identification based on a 658 base pair region sequence of the cytochrome *c* oxidase subunit I gene of the mitochondrial DNA (mtDNA), the so-called DNA barcode region (partial COI or *CoxI* gene). Their proposal for the use of the barcoding gene for a molecular identification system initiated the Consortium of the Barcoding of Life (CBOL) in 2004 (http://www.barcoding.si.edu/AboutCBOL.htm). CBOL´s aim is to explore and develop the potential of DNA barcoding for research as a practical tool for species identification. One of the pilot projects was the Tephritid Barcoding Initiative (TBI) with the ambitious aim of gathering barcodes of some 2000 species of fruit flies, focusing mainly on pest and beneficial species. Several studies have been published over the last decade comparing COI sequence datasets with morphological ones for identification purposes among fruit flies, most of which focused on a single genus or a species group within a genus or at most a few closely related genera (Smith-Caldas et al. 2001, Barr et al. 2006, Boykin et al. 2006, Schutze et al. 2007, Nakahara and Muraj 2008, Virgilio et al. 2008, Kohnen et al. 2009, Zhang et al. 2010, Jackson et al. 2011). Virgilio et al. (2012) are the only ones testing DNA barcoding on an extensive dataset of fruit flies, comparable to ours it contains 602 sequences of 153 species. However, it still covers only a limited part of the family, for all species belong to just 10 genera and all are of the same subfamily.

In our study we chose a different approach: instead of focussing on certain species groups or genera, we sequenced as many European species that we could get a hold of, including multiple specimens from distinct geographical populations for as many species as possible. This generated a dataset containing 555 sequences of half of the European species; 124 of the approximately 240 (Smit 2010), from all three subfamilies that are present on the continent. As a result the feasibility of DNA barcoding as an identification tool could be tested over a wide range of species within the family, meanwhile providing a significant contribution to the COI dataset of the Tephritid barcoding database based on morphologically identified specimens. Additional aims were to shed some light on the amount of inter- versus intraspecific variation over a large dataset of fruit fly species belonging to various tribes from different subfamilies as well as testing the phylogenetic signal within the COI barcoding gene.

Material and methods

Specimen acquisition

Data on the voucher specimens are provided in Appendix. The vast majority of specimens was collected throughout Europe in 2009 (n = 494). Specimens were directly stored in ethanol 96%. Some of the older material, collected before 2009, has been either directly collected in ethanol 96% (n = 23) or was collected with a Malaise trap (ethanol 70%) and later transferred to ethanol 96% (n = 38).

The oldest material included in this study is from 1999, collected in Kyrgyzstan by Valery Korneyev; this material was stored in 70% ethanol until DNA extraction and amplification. Of the 18 specimens collected, only four resulted in full barcode sequences, hence these are the only ones included in the dataset.

We have included up to eight specimens from geographically distinct populations in order to test the intraspecific variation for as many species as possible. However, we were unable to obtain more than one specimen for a number of species, whereas we have included between 9 and 15 specimens for species with uncertain taxonomy due to species complexes or host races (Table 1). For *Chaetostomella cylindrica* (Robineau-Desvoidy, 1830) we included 23 specimens in order to cover as much of the host races as possible (Knio et al. 2007, Smith et al. 2009).

The dataset contains 13 specimens of 11 species originating from Peru, some of which have their congeners among European taxa. These were added to see whether these more distant related taxa have any affect the molecular identification of a dataset of primarily European species. Thus adding a second geographical scale, besides multiple populations per species.

Additionally one outgroup specimen from the closely related family Ulidiidae was used to root the tree: *Ulidia nigripennis* Loew, 1845.

The dataset includes 554 sequences of 135 ingroup species from three different subfamilies and 42 genera and one outgroup sequence.

Table 1. The number of species with their range of specimens included in our dataset.

Specimens per species	No. species
1	41
2–8	78
9–15	15
> 15	1

Table 2. Primer pairs used for amplification of the COI marker.

Primer name	Primer sequence	Length (in bp)
L1490 (Folmer et al. 1994)	5' - GGTCAACAAATCATAAAGATATTGG - 3'	658
H2198 (Folmer et al. 1994)	5' - TAAACTTCAGGGTGACCAAAAAATCA - 3'	
TEP_F2	5' - TAGGAGCAGTAAATTTTAT - 3'	(+H2198) 211
TEP_R2	5' - CAAAAACTTATATTATTTAT - 3'	(+L1490) 241
TEP_F4	5' - ATTATAATTGGAGGATTTGG - 3'	268
TEP_R4	5' - GTAATTCCTGTTGATCGTATATTAAT - 3'	
TEPCOIF	5' - TAAACTTCAGCCATTTAATC - 3'	777
TEPCOIR	5' - TTTTCCTGATTCTTGTCTAA - 3'	

DNA extraction and amplification

One or two legs per specimen were used for genomic DNA extraction using the 96 wells Qiagen DNeasy Blood and Tissue Kit with a modified protocol. Due to the small size of the legs the tissue was manually ground with a disposable pestle in a 1.5 ml tube. The lysate was transferred to 96 well plates. Elution was performed in 50 µl elution buffer. 658 bp products were amplified using PCR primers LCO1490 and HCO2198 (Folmer et al. 1994) in most specimens. Amplification failed in some specimens therefore different primer sets were developed based on the full mitochondrial genomes of *Bactrocera oleae* (Rossi, 1790) (GU108464) and *Ceratitis capitata* (Wiedemann, 1824) (AJ242872) obtained from GenBank. Primers can be found in Table 2, their corresponding positions within the COI region are depicted in Figure 1.

The 25 µl PCR reaction mixes contained 18.75 µl of ddH_2O, 2.5 µl of 10 × CoralLoad PCR Buffer (Qiagen), 1 µl of each primer (10 pM), 1.25 U of Taq DNA Polymerase (Qiagen), 0.5 µl of dNTP's and 1 µl of DNA template. The amplification protocol consisted of 3 min at 94 °C followed by 40 to 50 cycles of 15 s at 94 °C,30 s at 60 °C to 35 °C and 40 s at,72 °C and a final 5 min at 72 °C.

Direct sequencing was performed at Macrogen, Korea on a ABI 3730XL sequencer.

Data analysis

Sequences recovered did not contain any insertions, deletions, or stop codons. 555 specimens representing 136 different species from various geographical locations were

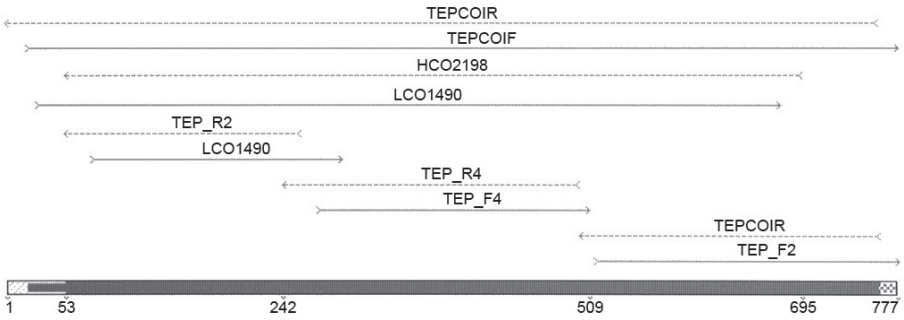


Figure 1. Primer positions within the COI region.

included in the dataset, resulting in a final alignment of 554 ingroup taxa and a single outgroup. Sequences were assembled and adjusted with Sequencher 4.10.1 (Gene Codes Corp.). Bioedit version 7.0.9.0 (Hall 1999) was used to align the sequences and MacClade version 4.08 (Maddison and Maddison 2000) was used to check for stopcodons. All sequence data, additional geographic and ecological data as well as photographs of the specimens were uploaded to the BOLD database, which ID codes are included in Appendix.

Molecular identification

The Neighbour-Joining analyses were performed using MEGA5 (Tamura et al. 2011). Distance analysis was conducted using the Kimura 2-parameter model (K2P) (Kimura 1980), and will simply be referred to as distance. The values given in brackets after the mean distance are ranges. The number of informative nucleotide characters in the dataset was 302. Success of the NJ tree-based identification (NJT) is assessed as described Hebert et al. (2003); i.e., sequences were considered successfully identified as long as they formed species-specific clusters. Species with sequences at multiple positions in the tree were considered misidentifications and singletons were counted as ambiguous. Second we used the revised criteria (NJT_M) as described by Meier et al. (2006); where identification is considered successful when a sequence is found at least one node into a cluster of exclusively conspecific sequences or in a polytomy with conspecifics. Species with sequences at least one node into an allospecific cluster or polytomy of allospecific sequences are considered misidentifications. Singletons, sequences as a sister group to conspecifics as well as sequences within a polytomy with at least one conspecific and allospecific sequence are considered ambiguous.

Additional to the tree-based identification we used an identification based on direct sequence comparison by using each sequence as a query to all other sequences in the dataset. SpeciesIdentifier v1.7.8 (Meier et al. 2006) was used to calculate distances, to find the closest barcode match and to determine the threshold value below which 95% of all intraspecific distances are found. The identification criteria used are 'Best

Match' (BM) and 'Best Close Match' (BCM) as described by Meier et al. (2006). The identification is considered successful in BM when the closest match is from the same species. When the species are different it is considered a misidentification. Several equally good best matches from more than one species is considered ambiguous. In BCM the criteria are the same as BM, but the results have to fall within the 9^h percentile of all intraspecific distances.

Finally we included the "All species barcodes" (ASB) criteria as described by Meier et al. (2006). This analyses uses the same threshold as used in BCM and identifications were only considered successful when all conspecific sequences top the list of best matches. When at least one allospecific sequence is more similar than the least similar conspecific sequence identification is considered ambiguous, if the query is more similar to all sequences from another species it is considered a misidentification.

Virgilio et al. (2012) introduced a method to improve the accuracy of the interpretation of the success-rates by distinguishing between true and false positives and negatives. True positives (TP) are the queries that have been correctly identified and are below the threshold value, false positives (FP) are incorrectly identified and below the threshold value. True negatives (TN) are correctly rejected because they are misidentified and above the threshold value, false negatives (FN) are correctly identified queries that are rejected because their distance is above the threshold value. Distinguishing these categories allows statements on the accuracy ((TP+TN/n.queries), precision (TP/(TP+FP)), overall ID error ((FP+FN)/n.queries) and relative ID error (FP/(TP+FP)), see Virgilio et al. (2012). These values are assessed for the dataset at hand.

Results

DNA extraction and amplification

The DNA of the majority of the specimens could be amplified with the standard PCR primers (Folmer et al. 1994). However, 23 out of the 555 samples needed alternative primers (Table 2). Nearly half only needed one alternative primer (Table 3), whereas others, like the Kyrgyzstan material, needed a cocktail of primers and the amplification protocol needed adjustment as given above.

Sequence alignment and analyses

The data are presented in a Neighbour-Joining tree only (Figure 2) for we are merely interested in a distance-based clustering of species based on similarity of the sequences and not a character based clustering of the sequences. Despite the fact that the NJ tree fits very well to both the morphological phylogenetic tree (Korneyev 2000) as well as the recent molecular ones (Han et al. 2006, Han and Ro 2009) it is stressed here that this tree may not reflect the true phylogenetic tree, because running the

Table 3. The species for which alternative primers have been used for DNA amplification.

Taxon (no specimens)	Probable reason for failure	Used primer(s)	Additional sequences with Folmer et al. (1994)
Acanthiophilus walkeri (1)	DNA degraded, specimen stored in ethanol 70% for 7 years	All	0
Bactrocera oleae (1)	DNA degraded, specimen stored in ethanol 70%	All	1
Plaumannimyia sp. (1)	?	TEPCOI	0
Rhagoletis cerasi (1)	?	TEPCOI	4
Rhagoletis cingulata (3)	Taxon-specific mutation at primer site?	TEPCOI	0
Rhagoletis samojlovitshae (1)	DNA degraded, specimen stored in ethanol 70% for 10 years	All	0
Sphenella marginata (7)	Taxon-specific mutation at primer site?	TEPCOI, TEP_F2, TEP_R2 & Folmer et al. (1994)	0
Tephritis nebulosa (1)	DNA degraded, specimen stored in ethanol 70% for 10 years	All	0
Terellia colon (1)	?	TEPCOI	11
Terellia luteola (1)	DNA degraded, specimen stored in ethanol 70% for 10 years	TEPCOI, TEP_F2, TEP_R2 & Folmer et al. (1994)	1
Trupanea cf. *metoeca* (1)	DNA degraded, specimen stored in ethanol 70% for 2 years	TEPCOI	0
Trypeta artemisiae (2)	?	TEPCOI	1
Ulidia nigripennis (1)	?	TEPCOI	0
Urophora ivannikovi (1)	DNA degraded, specimen stored in ethanol 70% for 10 years	All	0

data through a Maximum Parsimony (MP) and Maximum Likelihood (ML) analyses result in different topologies.

We only focus on the feasibility of DNA barcoding for molecular identification, any probable taxonomic implications of the data generated are not dealt within this paper.

Molecular identification

With some exceptions the COI barcodes in general seem to provide a good molecular marker for identification of European fruit fly species. The mean distances between species was on average 13.2% (0.15–25.27%) whereas within a species this was a mere 0.24% (0–2.80%) (Figure 3). There is no clear barcode-gap for 2.7% of all pairwise comparisons fell between the minimum interspecific distance (0.15%) and the maximum intraspecific distance (2.8%). Among the genera the mean distances were 1.49% (0–8.78%) within and 14.96% (5.92–23.61%) between the genera. The distances between the ingroup genera and the outgroup was 21.18% (17.11–25.72%).

Identification success-rates of all five criteria are given in Table 4. Several species groups within the genus *Terellia* Robineau-Desvoidy, 1830 and apparently none of the species of *Urophora* Robineau-Desvoidy, 1830 could reliably be identified using COI barcodes.

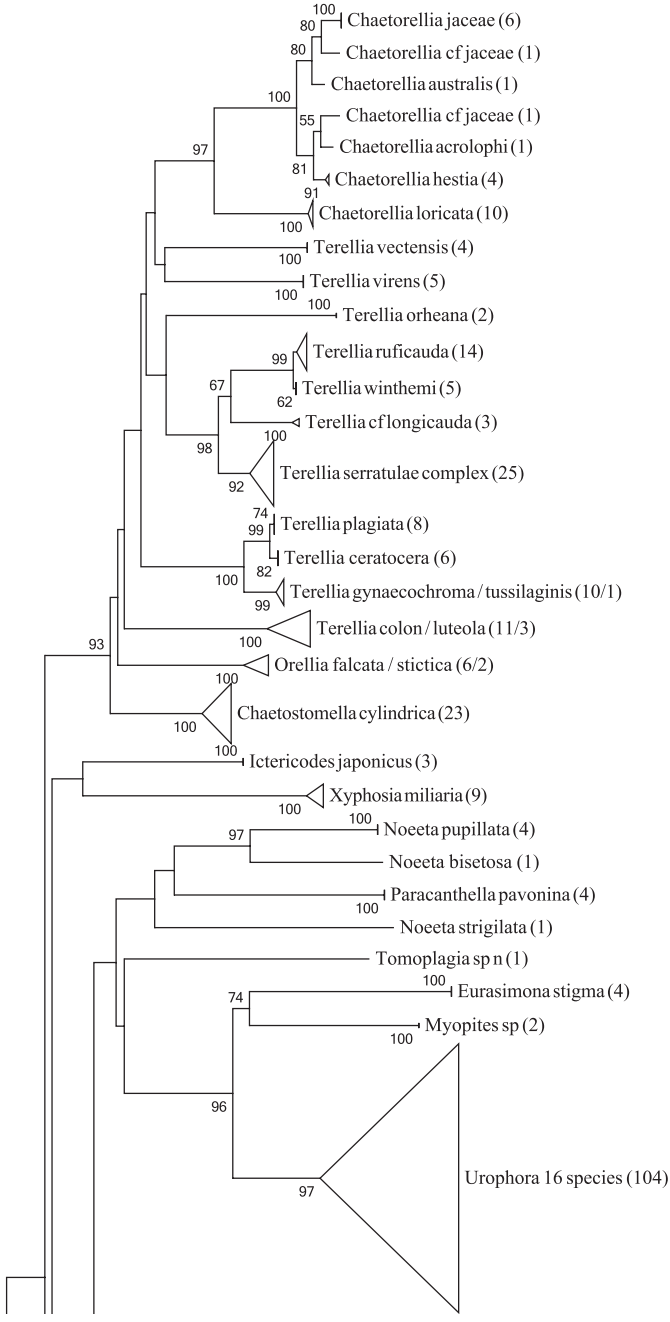


Figure 2. The Neighbour-Joining tree of the entire dataset based on COI barcodes. Terminal branches have been collapsed in order to save space, the total number of specimens is given in brackets and the area surface of the triangle represents the amount of variation. When a terminal branch contains two species, both names are provided as well as their respective number of specimens. If a branch contains more than two species only the number of species as well as the number of specimens are given. Bootstrap values above 50 (1000 replicates) are given at the nodes.

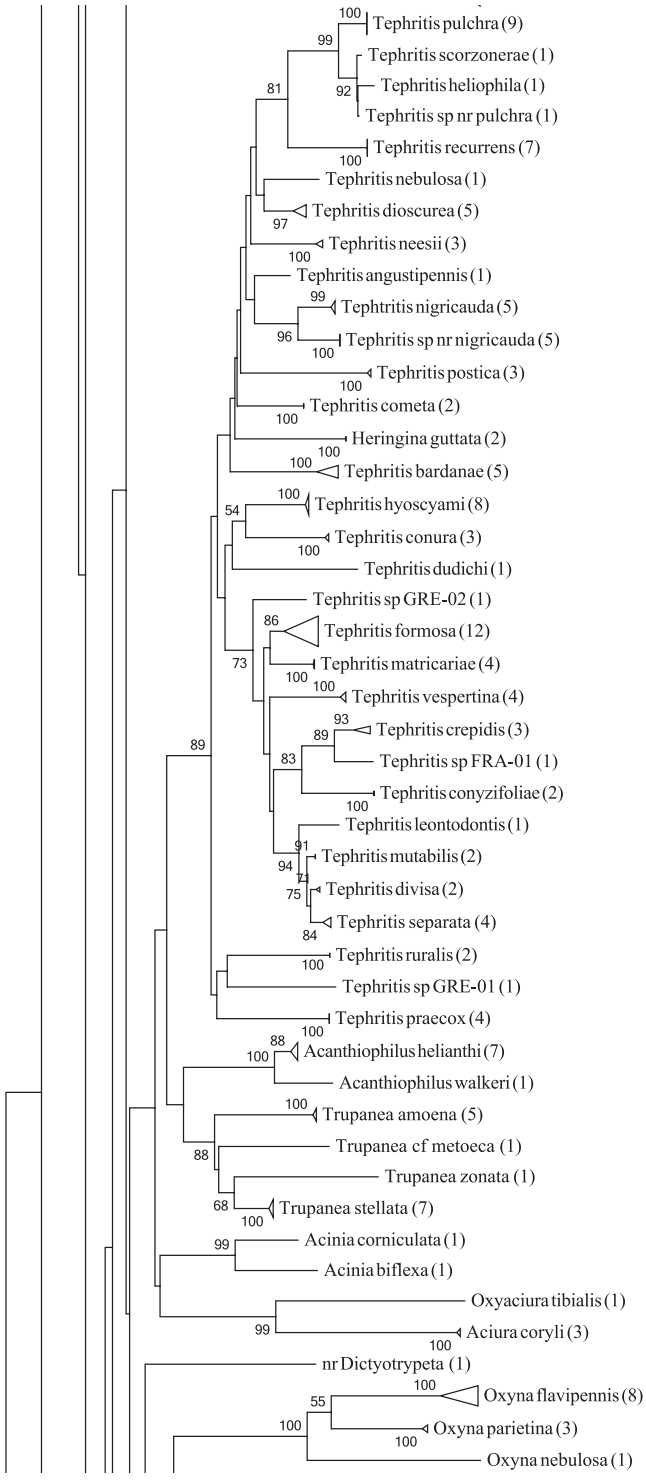


Figure 2. Continued

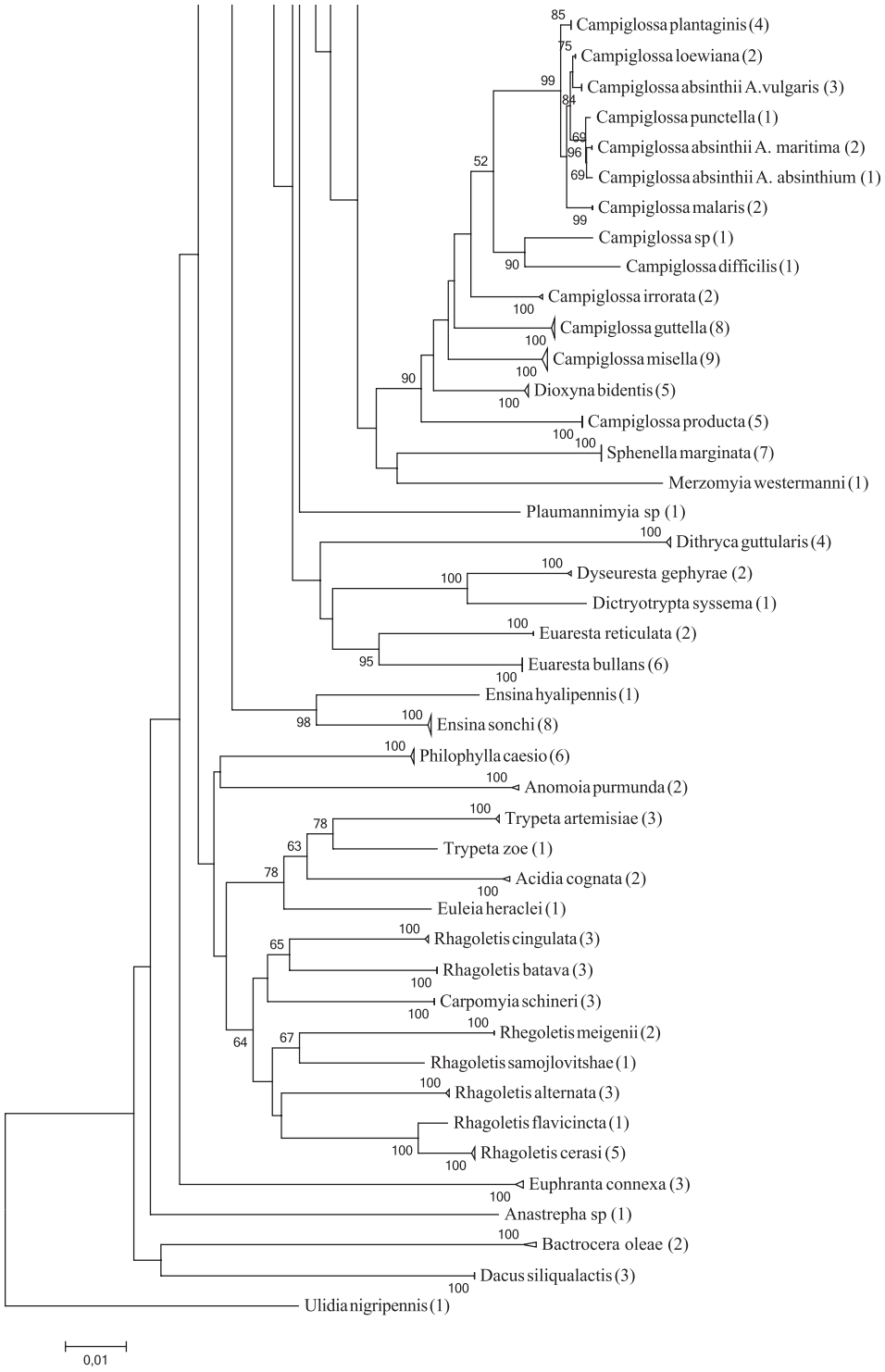


Figure 2. Continued

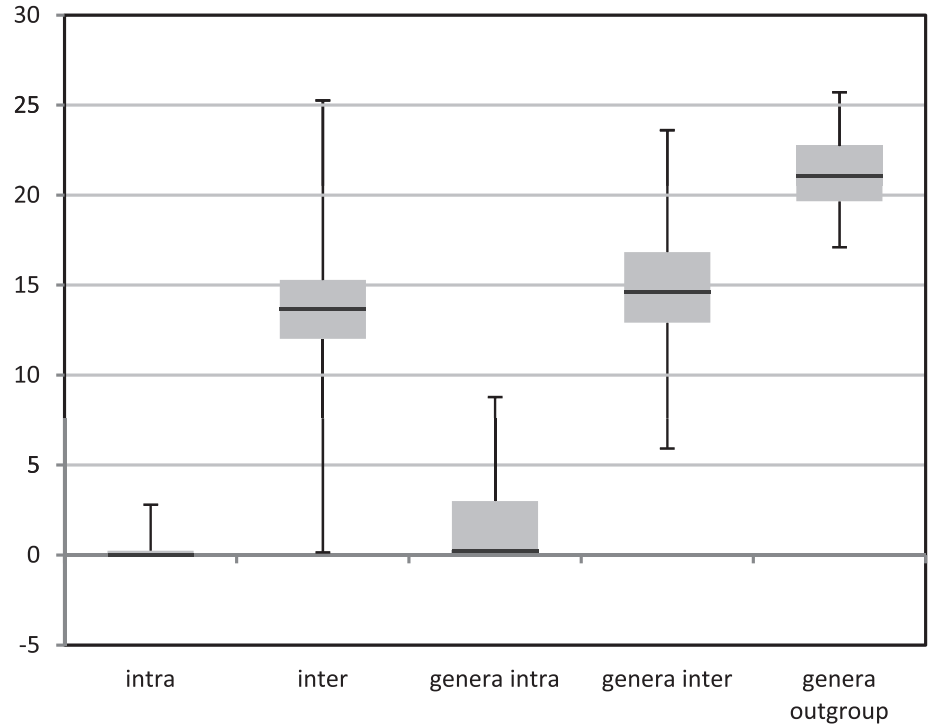


Figure 3. Box plots depicting the variation in mean distances using K2P-distance modeling of sequence divergence for intraspecific, interspecific difference among the species and genera, as well as the ingroup genera with the outgroup genus.

Table 4. Identification rates of all five criteria: Neighbour-Joining (NJT) sensu Hebert et al. (2003), revised criteria (NJT_M) according to Meier et al. (2006), and Best Match (BM), Best Close Match (BCM) and All Species Barcodes (ASB) also described by Meier et al. (2006).

Criteria	Correct ID	Ambiguous	Incorrect ID	No match
NJT	63.25%	7.38%	29.37%	-
NJT_M	61.89%	36.22%	1.80%	-
BM	78.19%	12.25%	9.54%	-
BCM (threshold 0.3%)	73.33%	10.45%	3.06%	13.15%
ASB (threshold 0.3%)	59.63%	27.02%	0.18%	13.15%

Tree-based identification

Both criteria NJT and NJT_M give comparable results with the correct identified sequences: 351 and 344 sequences respectively (Table 4). The main difference is among the number of incorrect and ambiguous sequences, for multiple placement immediately identifies the sequences as incorrect according to NJT, whereas if they still have conspecifics at the different nodes they are regarded as ambiguous according to NJT_M: 41 and 163 versus 201 and 10 sequences.

Table 5. Mean K2P-distances in percentages between the species of the *C. loewiana*-group.

C. malaris						
C. absinthii / on *A. vulgaris*	1.07					
C. loewiana	1.23	0.46				
C. punctella	1.23	0.77	0.92			
C. absinthii / on *A. absinthium*	1.23	0.77	0.92	0.30		
C. absinthii / on *A. maritima*	1.38	0.92	1.08	0.46	0.46	
C. plantaginis	1.54	0.76	0.61	0.92	0.92	1.07

The Neotropical taxa with European congeners clustered within the appropriate genus, often with a distance greater than those among the European taxa of that particular genus.

Campiglossa absinthii (Fabricius, 1805) is placed at three different branches within the NJ tree with slightly lower though similar mean distances as among the other closely related species (Table 5). All three groups originate from different *Artemisia* host-plants and might therefore represent different host-races, or perhaps even different species. Host-plant names are given in Figure 2 and are abbreviated in Figure 8.

Furthermore the NJ analysis places the genus *Dioxyna* Frey, 1945 within the genus *Campiglossa* Randani, 1876 and *Heringina* Aczél, 1940 within *Tephritis* Latreille, 1804 both of which are corroborated with the ML and MP analyses.

Similarity-based identification

Under the BM criteria 434 sequences were regarded as correctly identified, 53 incorrectly and 68 as ambiguous. The dataset contains 394 sequences with a closest match at 0%, 56 (14,21%) of them having an allospecific identical match.

The threshold for the 9^{h} percentile of the intraspecific distances has been calculated at 0.3%. Success under BCM is 73.33% (84.44% of the non-discarded queries), whereas 17 sequences were regarded as incorrectly identified, 58 ambiguous and 73 did not have a match below the threshold, the proportions of TP, FP, FN and TN were 0.733, 0.135, 0.048 and 0.082 respectively.

Under the ASB criteria 331 sequences were correctly identified, 150 were ambiguous, one was misidentified and, like BCM, 73 did not have a match below the threshold.

Discussion

Molecular identification

The discussion is confined to the success-rates of the tree-based identification criteria NJT_M and the similarity-based identification according to the BCM criteria. The

numbers are given for the other criteria as well but they are not discussed further (Figure 4). The NJT criteria gives an overrepresentation of incorrectly identified sequences, whereas BM seems to have an overoptimistic prediction of correctly identified sequences (Figure 4) (Meier et al. 2006, Virgilio et al. 2012). Like BM the ASB criteria does not take into account the possibility of multiple haplotypes for a single species and regards them, contrary to BM, as ambiguous instead of incorrect identified (Figure 4) (Meier et al. 2006).

The low success-rate is in part due to singletons and the genus *Urophora*. Of the 135 species 38 (41 when three *Urophora* singletons are included) cannot have a match simply because they lack conspecifics (7.39% of the sequences) (Meier et al. 2006, Virgilio et al. 2010, 2012). Deleting them from the dataset as to simulate a perfect world scenario with 100% taxon-coverage, for every sequences has at least one conspecific, results in a higher success-rate, increasing 5.03% and 7.72% respectively and nearly halves the discarded queries (Figure 4). *Urophora* makes up 18.56% of the entire dataset. Deleting them results in different identification-rates, for which success increases a staggering 16.21% in NJT_M and 5.43% in BCM (Figure 4). Combining the two, e.g. deleting both the singletons and *Urophora*, provides an increase correct identified queries of 23.38% and 13.77% respectively (Figure 4). Comparing these identification-rates it becomes clear that *Urophora* is largely responsible for the lack of success with molecular identification in this dataset. The ambiguity caused by the *Urophora* sequences here is due to the fact that there are not only conspecific sequences per species but also in several cases per population. These of course are identical but in most cases different from conspecific sequences from other populations, interpreted by BCM as ambiguous for they might represent different haplotypes of the same species or are in fact two different species, whereas morphologically they clearly belong to the same species. Moreover more than half of the allospecific matches are caused by the genus *Urophora*, the rest being caused by the problematic *Terellia* groups.

This stripped dataset, e.g. without singletons and without the genus *Urophora*, results in 87.1% of all queries and 93.23% of the not discarded queries as correctly identified, which is similar though slightly lower than the dataset of interceptions of Virgilio et al. (2012).

The threshold value in BCM is of strong influence on the results, as already noted by Virgilio et al. (2012). The success-rates have been calculated for a range of arbitrary threshold values between the largest observed distance and 0.00 (Figure 5). A rapid increase of accuracy can be seen to 0.84 at a threshold of 0.5%, after which it declines again to 0.78, similarly TP increases and FN decreases. Precision however never exceeds 0.86. Thus when calculating the relative ID error, linear regression shows that for a relative ID error < 0.05 the threshold value is lower than 0.00 (Figure 7a). Even when the stripped dataset is used precision only reaches 0.94 (Figure 6), therefore again producing a threshold value lower than 0.00 for a relative ID error < 0.05 (Figure 7b). This linear regression function is used by Virgilio et al. (2012) to infer the *ad hoc* threshold for the 95th percentile of the correctly identified queries and where the relative ID error does not exceed 5%. When this threshold value is lower than 0.00

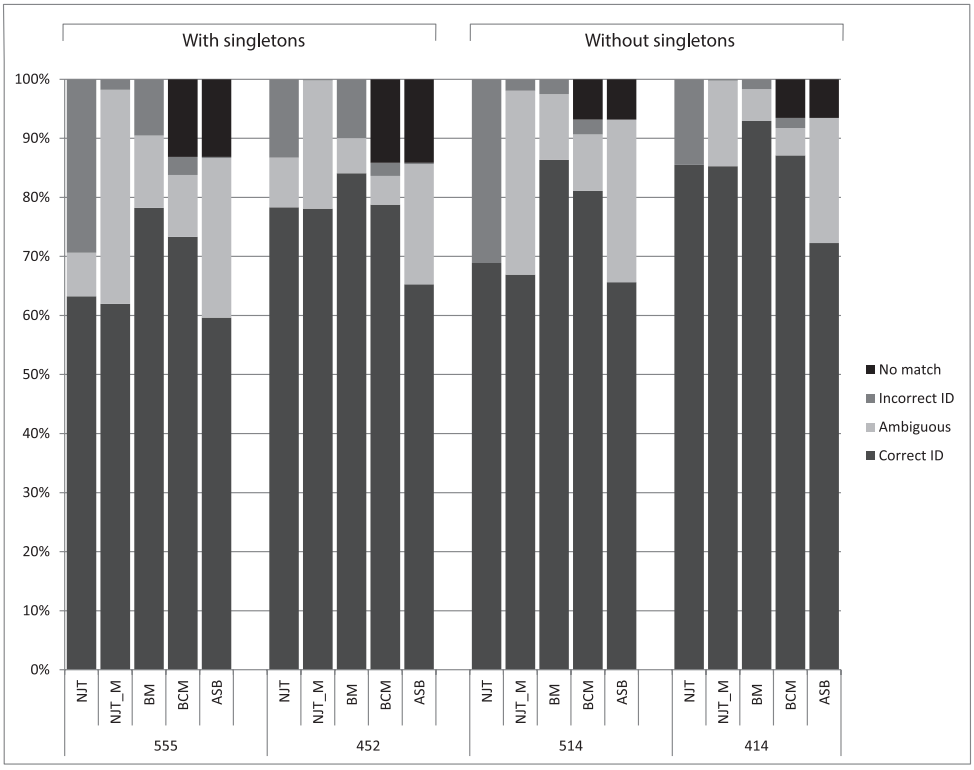


Figure 4. Identification rates of all five criteria: Neighbour-Joining (NJT) sensu Hebert et al. (2003), revised criteria (NJT_M) according to Meier et al. (2006), and Best Match (BM), Best Close Match (BCM) and All Species Barcodes (ASB) also described by Meier et al. (2006) for four different datasets, including singletons and with (n = 555) or without (n = 452) *Urophora*, and the same excluding singletons (n = 514) and (n = 414) respectively.

the dataset should be regarded as unreliable (Virgilio et al. 2012). Only when the problematic *Terellia* groups are deleted from our already stripped dataset an *ad hoc* threshold value > 0 can be inferred (Figure 7c). Therefore the dataset created here is unreliable for molecular identification. This was also clear by the number of allospecific matches as well as the ambiguity among the success-rates, resulting in an low overall success-rate. Several other groups have recently been studied in which DNA barcoding was shown to have a limited performance (Armstrong and Ball 2005, Kaila and Stahls 2006, Meier et al. 2006, Elias et al. 2007, Neigel et al. 2007, Skevington et al. 2007, Virgilio et al. 2008, Dasmahapatra et al. 2009, Jackson et al. 2011, Barr et al. 2012).

Distinguishing between true and false positives and negatives is based on morphological identification of the voucher specimens. Therefore taxonomic specialists are needed to build and check the reference database that can be used for molecular identification. Adding more morphologically correctly identified specimens will increase the understanding of the limitations of molecular identification for that particular group (Meyer and Paulay 2005, Ekrem et al. 2007, Kwong et al. 2012). Incorrectly identi-

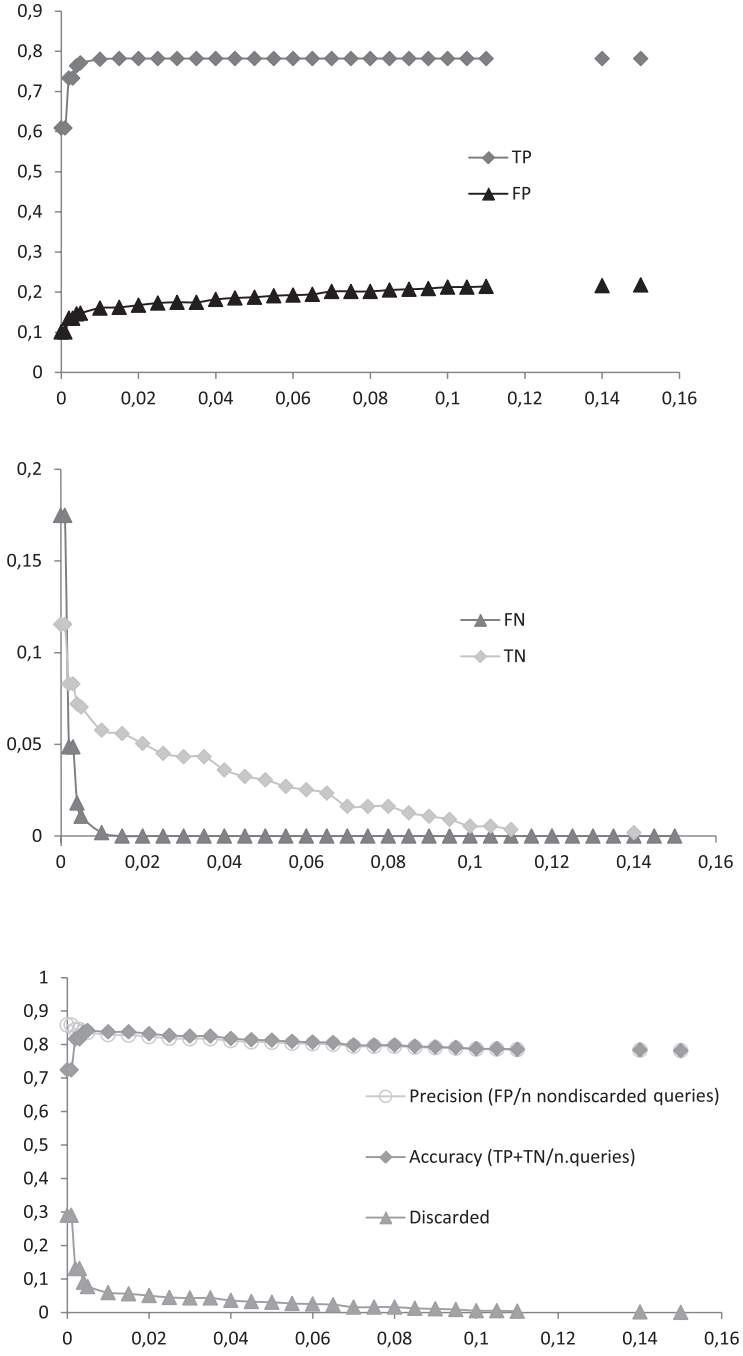


Figure 5. Best Close Match (BCM) identification of the entire dataset (n = 555). Proportions of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) are given for 30 arbitrary distance thresholds ranging from 0.15 to 0.00. For each threshold the percentages of precision, accuracy and discarded queries were calculated.

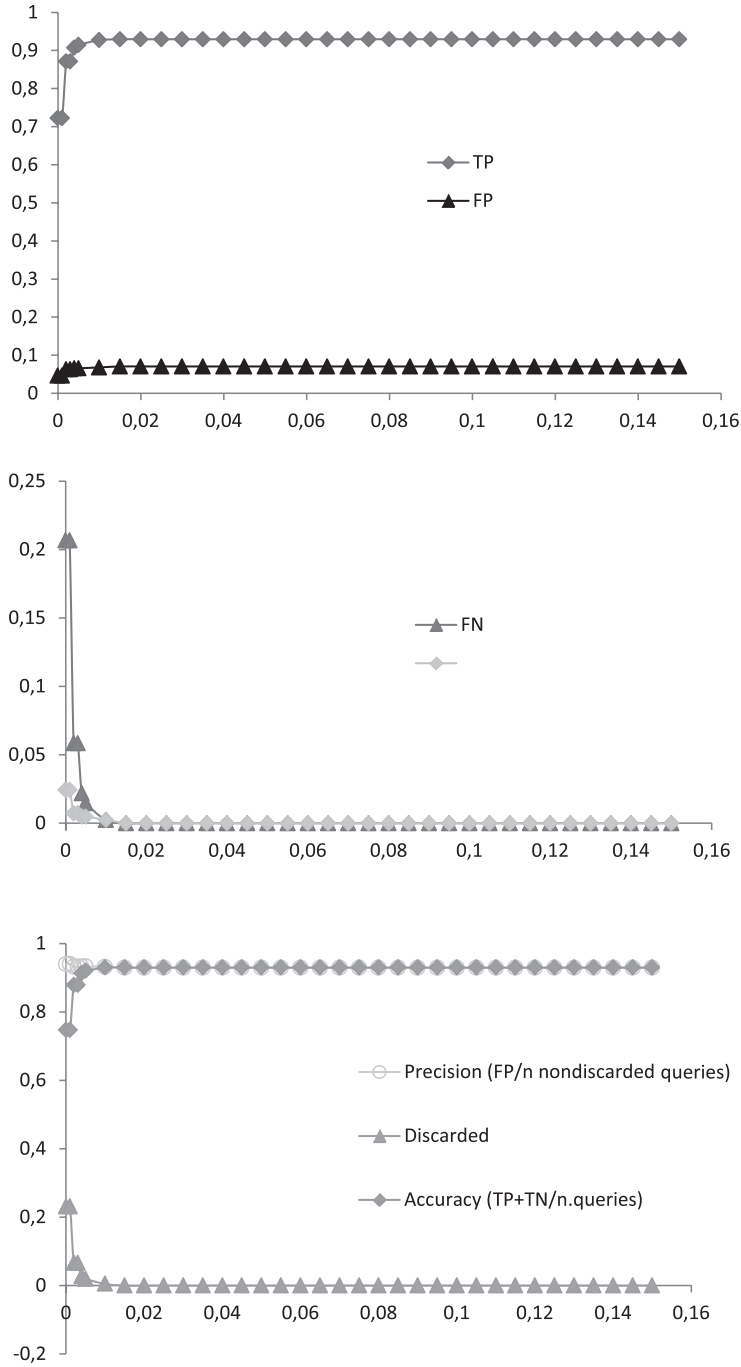


Figure 6. Best Close Match (BCM) identification of the stripped dataset, e.g. excluding singletons and *Urophora* (n = 414). Proportions of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) are given for 30 arbitrary distance thresholds ranging from 0.15 to 0.00. For each threshold the percentages of precision, accuracy and discarded queries were calculated.

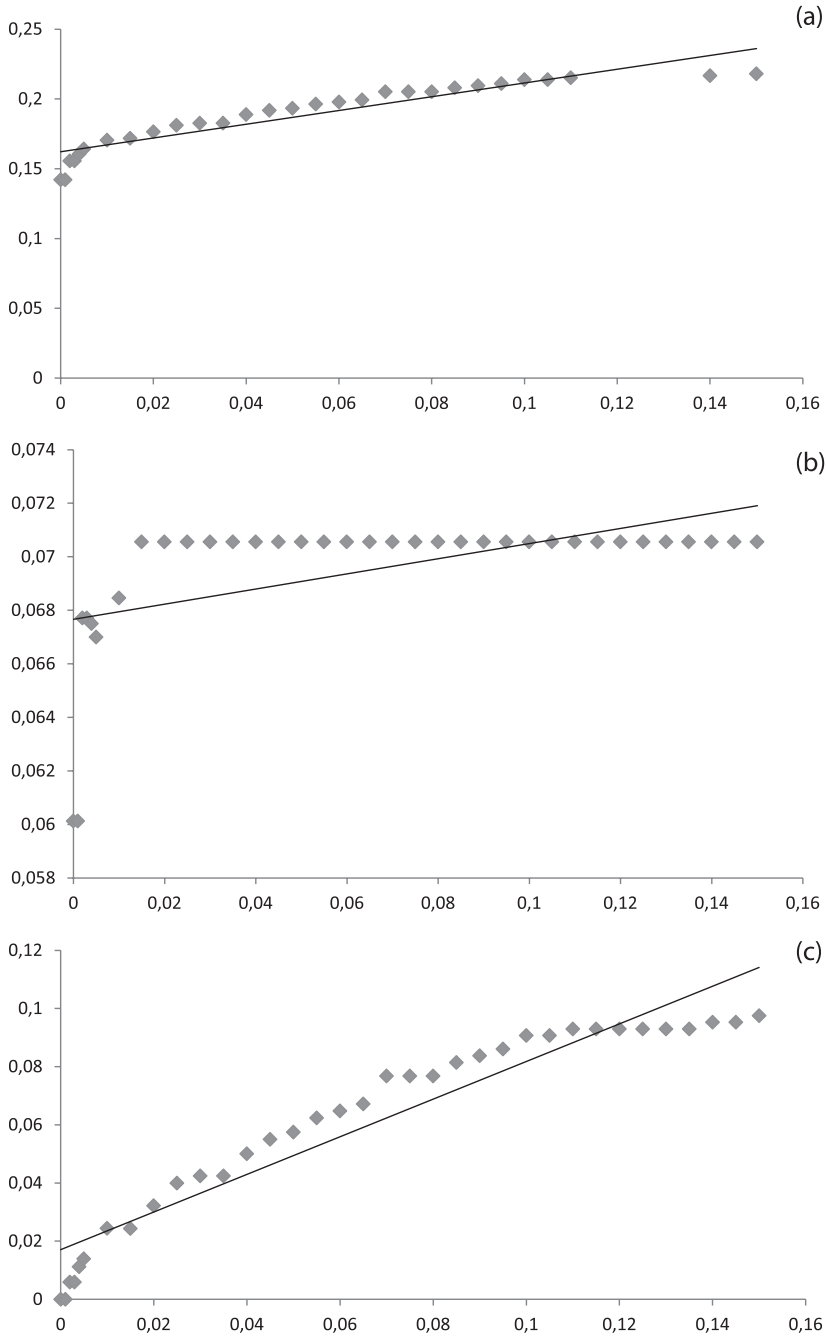


Figure 7. Relative ID errors at 30 arbitrary threshold values for a. the entire dataset (n = 555), b. the stripped dataset, e.g. excluding singletons and *Urophora* (n = 414) and c. the stripped dataset excluding the problematic *Terellia* groups. Linear regression was used to infer the *ad hoc* threshold for the 9^{h} percentile of the correctly identified queries and the relative ID error does not exceed 5%. In (a) and (b) this value is below 0.00, only in (c) this value is positive: 0.051 (R-square 0.91).

fied sequences will be added to the reference database like BOLD, for it is only human to make errors. Introducing threshold values for molecular identification will point out the obviously incorrectly identified specimens (Meier et al. 2006), but will not help with problematic groups containing for example very low interspecific distances or allospecific matches. Based on our dataset we were able to identify some problematic groups causing limitations for molecular identification of Tephritids illustrated by some examples given below.

Varying mean distances between different species groups of the same genus

The species of the genus *Campiglossa* can be identified using DNA barcodes, showing a neat mean distance of 5.2%. Looking in detail, however, shows it has a very broad range of interspecific distances, from 0.3 to 8.7%. Grouping the species into their known morphological species complexes (Merz 1992, 1994) results in a mean distances of 6.2% (4.2–8.6%), because all but one of the groups are represented by just one species (Figure 8). The five species of the *loewiana* group show a mean distance of a mere 0.9% (0.3–1.5%) (Table 5), revealing that these very closely related species are apparently difficult to separate using COI, something which has been noted before in various groups as well as Tephritids (Armstrong and Ball 2005, Kaila and Stahls 2006, Virgilio et al. 2008, Barr et al. 2012, Nieukerken et al. 2012).

Executing a BLAST on the BOLD database with one sequence of *Campiglossa malaris* Séguy, 1938 from our dataset retrieved no less than 18 sequences with a similarity of over 98%, belonging to 5 different species apart from the target species. Excluding *C. malaris* itself, the sequence with the highest similarity was one belonging to a Nearctic species, *Campiglossa farinata* (Novak, 1974) with a similarity of 99.08%. Furthermore, no less than six sequences showed a similarity of 98.93% belonging to two different species.

These differences in mean distances, especially the short ones among the *loewiana* group, indicate that it is important to include as many sequences of distinct populations per species as possible in a reference database like BOLD to preclude misidentification.

Multiple specimens

Adding specimens from geographically distinct populations is necessary in order to shed some light on the intraspecific variation caused by geography (Bergsten et al. 2012). This is clearly illustrated by adding two specimens of *Orellia falcata* (Scopoli, 1763) from Spain, which resulted in a paraphyletic placement, including the second species present in the dataset: *O. stictica* (Gmelin, 1790) (Figure 9). Both species are morphologically quite distinct and easy to recognize. Therefore either both species are so closely related that they cannot be separated based on the barcode gene and perhaps a more sensitive marker is needed, or *O. falcata* represents a complex of cryptic species.

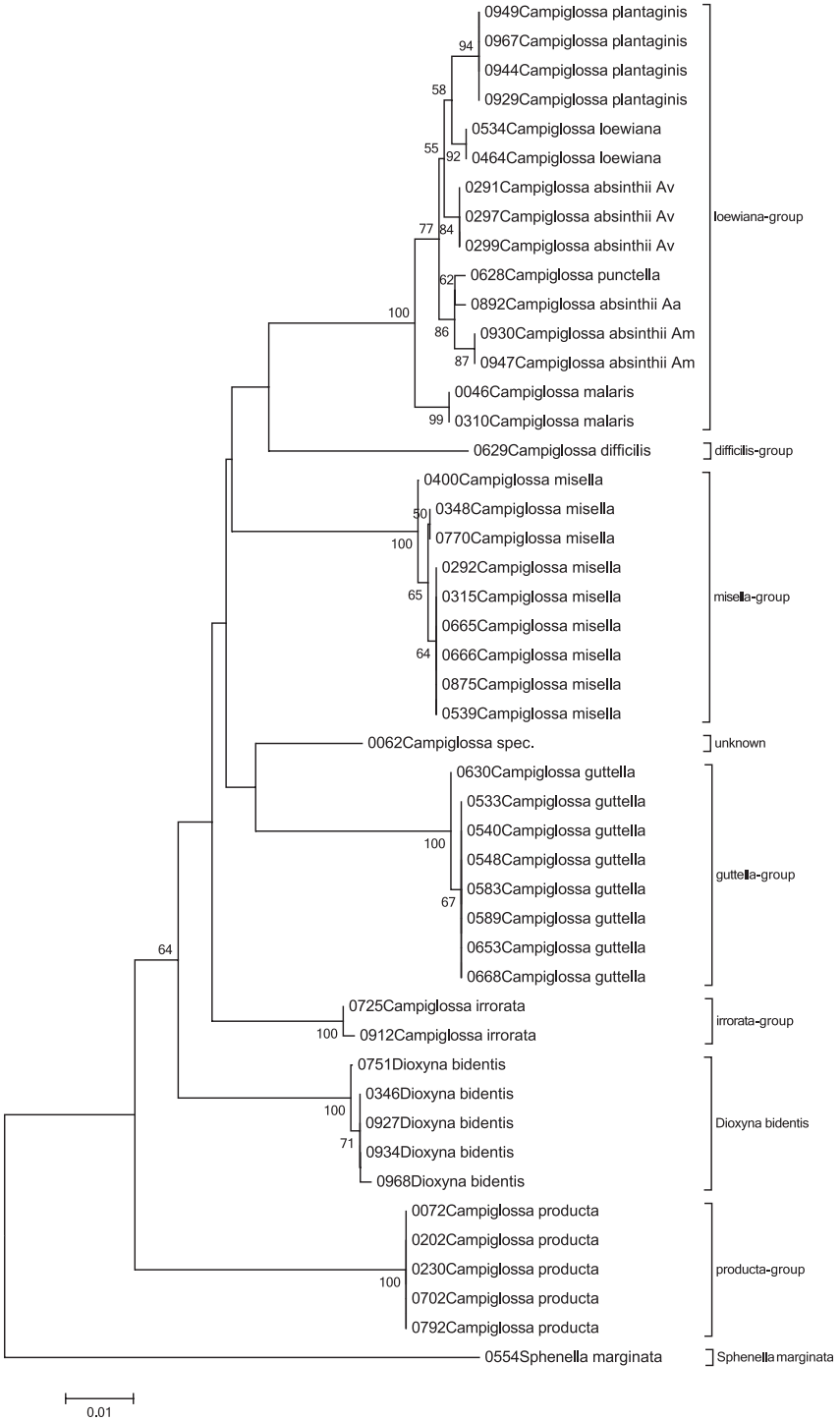


Figure 8. The Neighbour-Joining tree of the genus *Campiglossa* with *Sphenella marginata* as outgroup inferred from COI barcodes. Bootstrap values above 50 (1000 replicates) are given at the nodes.

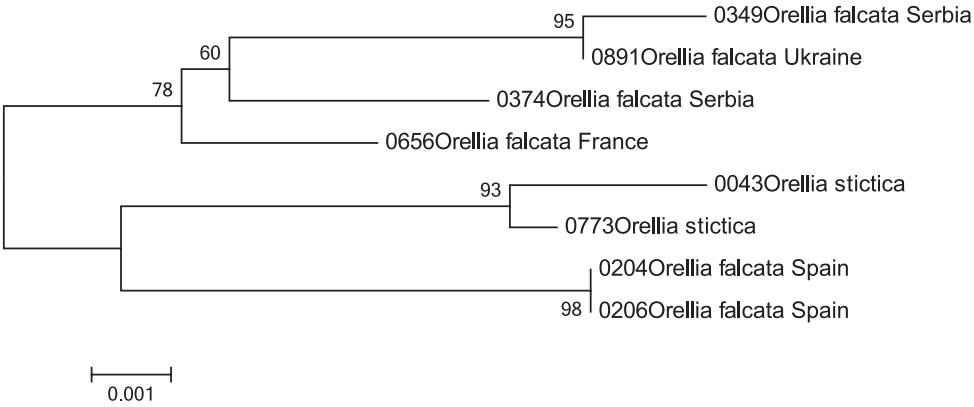


Figure 9. The Neighbour-Joining tree of the genus *Orellia* inferred from COI barcodes. Bootstrap values above 50 (1000 replicates) are given at the nodes.

Likewise it is necessary to add specimens of ecologically distinct populations as well, as is shown by the three 'host-races' of *Campiglossa absinthii* and by Smith et al. (2009) for *Chaetostomella cylindrica*.

Low interspecific variation compared to a high intraspecific variation

Looking at the NJ tree (Figure 2, 10) it is immediately obvious that the species of the genus *Urophora* cannot be separated using DNA barcodes. Jackson et al. (2011) already reported that the species of the genus *Urophora* could not be identified using DNA barcodes, having included 10 sequences belonging to three different species. In our dataset we included over 100 sequences of 16 morphologically identified species, resulting in multiple placement of several species and a mean distance of a mere 1.65% (0.3-2.45%). This limited or entire lack of performance of molecular identification is of special interest for it concerns a genus of economic importance with several species regarded as beneficiary for weed control (White and Clement 1987, White and Elson-Harris 1992). Additional genetic markers should be tested for the molecular identification of these species like Elongation Factor 1-α (EF1- α) or ribosomal Internal Transcribed Spacer 2 (ITS2) (Alvarez and Hoy 2003, Farris et al. 2010, Nieukerken et al. 2012).

The limitations of DNA barcodes for molecular identification

As is shown above, the feasibility of the use of DNA barcodes for molecular identifications relies heavily on the contents of the database used to BLAST against (Meyer and Paulay 2005, Meier et al. 2006, Ekrem et al. 2007, Virgilio et al. 2010, 2012, Kwong et al. 2012). The addition of multiple specimens per species to the database, prefer-

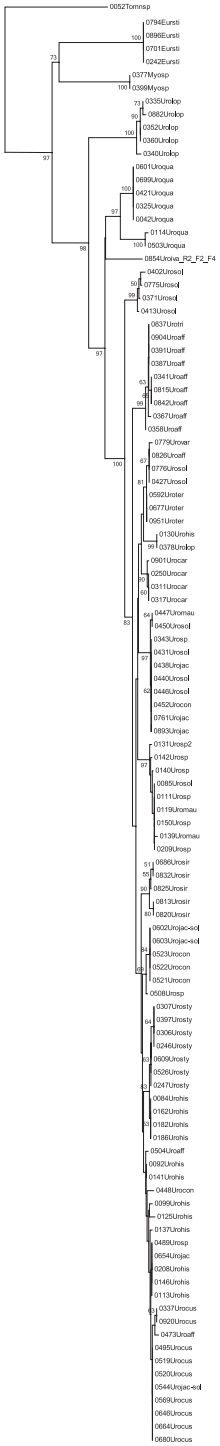


Figure 10. The Neighbour-Joining tree of the genus *Urophora* inferred from COI barcodes. Bootstrap values above 50 (1000 replicates) are given at the nodes.

ably from geographically distinct populations, as well as different ecologies, provides a much needed insight in the intraspecific versus interspecific variation of the species. Adding more species is a necessity too, because incorporating different species of the *Campiglossa loewiana*-complex clearly demonstrated that the perceived mean distance of 5.2% between the species actually represents the mean distance of the different species groups in this dataset. The mean distance of the species within the *C. loewiana*-group was a mere 0.9%. Hence threshold values like a ≥ 98% similarity as used by Li et al. (2011) or the 97% used by BOLD for a positive identification do not hold. Introducing the 9^{h} percentile threshold value increases the reliability of the identification success. Further improvement can be achieved by introducing the *ad hoc* threshold as proposed by Virgilio et al. (2012). However, as is shown by our dataset, this is not always possible. Instead of discarding the dataset as unreliable it should be used to identify the problematic groups by looking at the amount of allospecific matches, TP, FP, FN and TN. In that case these problematic groups can be flagged in the reference database so that the user can look for alternative means for identification.

Conclusion

We conclude that molecular identification of Tephritids using DNA barcoding is possible but should be treated with care due to varying performance within this group as is shown by the dataset analysed here. Even when threshold values are added groups will remain that cannot reliably be identified. We stress that a better performance is strongly dependent on an increasing input of morphologically identified specimens, containing multiple specimens of different geographical populations and different ecologies covering as much of the range of the species as possible, otherwise it remains difficult to detect cryptic species and estimate true diversity. Threshold values for both distance and relative ID error, as well as distinction between positives and negatives, both true and false, should not only be used to improve the reliability of the success for molecular identification but also to identify the problematic groups for molecular identification. These groups should be flagged in the reference database and alternative markers for molecular identification should be tested.

Acknowledgements

We thank the following persons for providing material used in this study: Kees van Achterberg (Leiden, the Netherlands), Berend Aukema (Wageningen, the Netherlands), Theodoor Heijerman (Wageningen, the Netherlands), Guido Keijl (Bakkum, the Netherlands), Roy Kleukers (Leiden, the Netherlands), Severin and Valery Korneyev (Kiev, Ukraine), Kim Meijer (Groningen, the Netherlands), Gerard Pennards (Zeist, the Netherlands), Gordon Ramel (Serron, Greece), Jeff Skevington (Ottawa, Canada), J. Smit (Duiven, the Netherlands), Wouter van Steenis (Breukelen, the Netherlands)

and Theo Zeegers (Soest, the Netherlands). Furthermore we thank Menno Reemer for the help and discussions on the analysis. We also thank Valery Korneyev and Ho-Yeon Han for valuable comments on an earlier draft of this paper. Lastly we thank Allan Norrbom and two other anonymous reviewers for their valuable comments.

References

Alvarez JM, Hoy MA (2003) Evaluation of the ribosomal ITS2 DNA sequences in separating closely related populations of the parasitoid *Ageniaspis* (Hymenoptera: Encyrtidae). Annals of the Entomological Society of America 95: 250–256. doi: 10.1603/0013-8746(2002)095[0250:EOTRID]2.0.CO;2

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. Philosophical Transactions of the Royal Society B 360: 1813–1823. doi: 10.1098/rstb.2005.1713

Barr NB, Copeland RS, De Meyer M, Masiga D, Kibogo HG, Billah MK, Osir E, Wharton RA, McPheron BA (2006) Molecular diagnostics of economically important *Ceratitis* fruit fly species (Diptera: Tephritidae) in Africa using PCR and RFLP analyses. Bulletin of Entomological Research 96: 505–521. doi: 10.1079/BER2006452

Barr NB, Islam MS, Meyer M De, McPheron BA (2012) Molecular identification of *Ceratitis capitata* (Diptera: Tephritidae) using DNA sequences of the COI barcode region. Annals of the Entomological Society of America 105: 339–350. doi: 10.1603/AN11100

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN, Ribera I, Nilsson AN, Barraclough TG, Vogler AP (2012) The effect of geographical scale of sampling on DNA sarcoding. Systematic Biology 61: 851–869. doi: 10.1093/sysbio/sys037

Boykin LM, Shatters Jr RG, Hall DG, Burns RE, Franqui RA (2006) Analysis of host preference and geographical distribution of *Anastrepha suspensa* (Diptera: Tephritidae) using phylogenetic analyses of mitochondrial cytochrome oxidase I DNA sequence data. Bulletin of Entomological Research 96: 457–469. doi: 10.1079/BER2006438

Buhay JE (2009) ''COI-like'' sequences are becoming problematic in molecular systematic and DNA barcoding studies. Journal of Crustacean Biology 29: 96–110. doi: 10.1651/08-3020.1

Cognato AI (2006) Standard percent DNA sequence difference for insects does not predict species boundaries. Journal of Economic Entomology 99: 1037–1045. doi: 10.1603/0022-0493-99.4.1037

DeSalle R, Egan MG, Siddall M (2005) The unholy Trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B 360: 1905–1916. doi: 10.1098/rstb.2005.1722

Dasmahapatra KK, Elias M, Hill RI, Hoffmans JI, Mallet J (2009) Mitochondrial barcoding detects some species that are real, and some that are not. Molecular Ecology Resources 10: 264–273. doi: 10.1111/j.1755-0998.2009.02763.x

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. doi: 10.1016/j.ympev.2006.11.021

Elias M, Hill RI, Willmott KR, Dasmahapatra KK, Brower AVZ, Mallet J, Jiggins CD (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proceedings of the Royal Society B 274: 2881–2889. doi: 10.1098/rspb.2007.1035

Farris RE, Ruiz-Arce R, Ciomperlik M, Vasquez JD, DeLeon R (2010) Development of a ribosomal DNA ITS2 marker for the identification of the thrips, *Scirtothrips dorsalis*. Journal of Insect Science 10: 1–15. doi: 10.1673/031.010.2601

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.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95–98. http://www.mbio.ncsu.edu/bioedit/page2.html

Han HY, Ro KE (2009) Molecular phylogeny of the family Tephritidae (Insecta: Diptera): new insight from combined analysis of the mitochondrial 12S, 16S en COII genes. Molecules and Cells 27: 55–66. doi: 10.1007/s10059-009-0005-3

Han HY, Ro KE, McPheron BA (2006) Molecular phylogeny of the subfamily Tephritinae (Diptera: Tephritidae) based on mitochondrial 16S rDNA sequences. Molecules and Cells 22: 78–88.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–322. doi: 10.1098/rspb.2002.2218

Jackson MD, Marshall SA, Hanner R, Norrbom AL (2011) The fruit flies (Tephritidae) of Ontario. Canadian Journal of Arthropod Identification 15: 1–251. doi: 10.3752/cjai.2011.15

Kaila L, Stahls G (2006) DNA barcodes: Evaluating the potential of COI to differentiate closely related species of *Elachista* (Lepidoptera: Gelechioidea: Elachistidae) from Australia. Zootaxa 1170: 1–26. http://www.mapress.com/zootaxa/200/zt0117026.pdf

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111–120. http://www.ncbi.nlm.nih.gov/pubmed/7463489

Knio KM, White IM, Al-Zein M (2007) Host race formation in *Chaetostomella cylindrica* (Diptera: Tephritidae): morphological and morphometric evidence. Journal of Natural History 41: 1697–1715. doi: 10.1080/00222930701494486

Kohnen A, Wisseman V, Brandl R (2009) No genetic differentiation in the rose-infesting fruit flies *Rhagoletis alternata* and *Carpomyia schineri* (Diptera: Tephritidae) across Central Europe. European Journal of Entomology 106: 315–321.

Korneyev VA, (2000) Phylogenetic relationships among higher groups of Tephritidae. In: Aluja M, Norrbom AL (eds) Fruit flies (Tephritidae): Phylogeny and evolution of behavior. CRC Press, London: 73–113.

Kwong S, Srivathsan A, Meier R (2012) An update on DNA barcoding: low species coverage and numerous unidentified sequences. Cladistics 28: 639–644. doi: 10.1111/j.1096-0031.2012.00408.x

Li Z, Li Z, Wang F, Lin W, Wu J (2011) TBIS: A web-based expert system for identification of Tephritid fruit flies in China based on DNA barcodes. Advances in Information and Communication Technology 346: 563–571. doi: 10.1007/978-3-642-18354-6_66

Maddison DR, Maddison WP (2000) MacClade 4: Analysis of phylogeny and character evolution. Version 4.0. Sinauer Associates, Sunderland, MA. http://macclade.org/macclade.html

McPheron BA, Steck GJ (1996) Fruit Fly Pests. A world assessment of their biology and management. St Lucie Press, Delray Beach, Florida, USA.

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Merz B (1992) Revision der westpalaearktischen Gattungen und Arten der *Paroxyna*-Gruppe und Revision der Fruchtfliegen der Schweiz (Diptera: Tephritidae). Dissertation ETH 9902: 342 pp.

Merz B (1994) Diptera Tephritidae. Insecta Helvetica fauna 10: 198 pp.

Meyer CP, Paulay G (2005) DNA Barcoding: error rates based on comprehensive sampling. PloS Biology 3: 2229–2238. doi: 10.1371/journal.pbio.0030422

Nakahara S, Muraj M (2008) Phylogenetic analyses of *Bactrocera* fruit flies (Diptera: Tephritidae) based on nucleotide sequences of the mitochondrial COI and COII genes. Research Bulletin of Plant Protection Japan 44: 1–12.

Neigel J, Domingo A, Stake J (2007) DNA barcoding as a tool for coral reef conservation. Coral Reefs 26: 487–499. doi: 10.1007/s00338-007-0248-4

Nieukerken EJ van, Doorenweerd C, Stokvis FR, Groenenberg DSJ (2012) DNA barcoding of the leaf-mining moth subgenus *Ectodemia* s. str. (Lepidoptera: Nepticulidae) with COI and EF1-α; two are better than one in recognising cryptic species. Contributions to Zoology 81: 1–24.

Norrbom AL, Carroll LE, Thompson FC, White IM, Freidberg A (1999) Systematic database of names. In: Thompson FC (Ed) Fruitfly Expert Identification System and Systematic Information Database. - MYIA vol. 9, Backhuys, Leiden, 65–299.

Rindal E, Brower AVZ (2011) Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data. Cladistics 27: 331–334. doi: 10.1111/j.1096-0031.2010.00342.x

Schutze MH, Yeates DK, Graham GC, Dodson G (2007) Phylogenetic relationships of antlered flies, *Phytalmia* Gerstaecker (Diptera: Tephritidae): the evolution of the antler shape and mating behaviour. Australian Journal of Entomology 46: 281–293. doi: 10.1111/j.1440-6055.2007.00614.x

Skevington JH, Kehlmaier C, Stahls G (2007) DNA barcoding: Mixed results for big-headed flies (Diptera: Pipunculidae). Zootaxa 1423: 1–26. http://www.mapress.com/zootaxa/200/zt0142026.pdf

Smit JT (2010) De Nederlandse boorvliegen (Tephritidae). Entomologische tabellen 5: 1–159.

Smith CA, Al-Zein MS, Sayar NP, Knio KM (2009) Host races in *Chaetostomella cylindrica* (Diptera: Tephritidae): genetic and behavioural evidence. Bulletin of Entomological Research 99: 425–432. doi: 10.1017/S0007485308006482

Smith-Caldas MRB, McPheron BA, Silva JG, Zucchi RA (2001) Phylogenetic relationships among species of the *fraterculus* group (*Anastrepha*: Diptera: Tephritidae) inferred

from DNA sequences of mitochondrial cytochrome oxidase I. Neotropical Entomology 30: 565–573. doi: 10.1590/S1519-566X2001000400009

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Turner CE (1996) Tephritidae in the biological control of weeds. In: McPheron BA, Steck GJ (Eds) Fruit Fly Pests: A world assessment of their biology and management. St. Lucie Press, Delray Beach, Florida, 157–164.

Virgilio M, Backeljau T, Barr N, De Meyer M (2008) Molecular evaluation of nominal species in the *Ceratitis fasciventris*, *C. anonae*, *C. rosa* complex (Diptera: Tephritidae). Molecular Phylogenetics and Evolution 48: 270–280. doi: 10.1016/j.ympev.2008.04.018

Virgilio M, Backeljau T, Nevado B, De Meyer M (2010) Comparative performances of DNA barcoding across insect orders. BMC Bioinformatics 11: 206. http://www.biomedcentral.com/1471-2105/11/206, doi: 10.1186/1471-2105-11-206

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M (2012) Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7: 1–8. doi: 10.1371/journal.pone.0031581

Wheeler QD (2008) Undisciplined thinking: morphology and Hennig's unfinished revolution. Systematic Entomology 33: 2–7. http://www.life.illinois.edu/ib/514/wheeler08.pdf, doi: 10.1111/j.1365-3113.2007.00411.x

White IM, Clement SL (1987) Systematic notes on *Urophora* (Diptera: Tephritidae) species associated with *Centaurea solstitialis* (Asteraceae: Cardueae) and other Palaearctic weeds adventives in North America. Proceedings of the Entomological Society of Washington 89: 571–580.

White IM, Elson-Harris MM (1992) Fruit flies of economic significance: their identification and bionomics. CAB International, London.

White IM, Groppe K, Sobhian R (1990) Tephritids of knapweeds, starthistles and safflower: results of a host choice experiment and the taxonomy of *Terellia luteola* (Wiedemann) (Diptera: Tephritidae). Bulletin of Entomological Research 80: 107–111. doi: 10.1017/S0007485300045983

White IM, Korneyev VA (1989) A revision of the western Palearctic species of *Urophora* Robineau-Desvoidy (Dipt., Tephritidae). Systematic Entomology 14: 327–374. doi: 10.1111/j.1365-3113.1989.tb00289.x

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

Zhang B, Liu YH, Wu WX, Wang ZL (2010) Molecular phylogeny of *Bactrocera* species (Diptera: Tephritidae: Dacini) inferred from mitochondrial sequences of 16S rDNA And COI Sequences. Florida Entomologist 93: 369–377. doi: 10.1653/024.093.0308

Appendix

Collection data of all specimens included in this study. (doi: 10.3897/zookeys.365.5819.app) File format: Adobe PDF file (pdf).

Citation: Smit J, Reijnen B, Stokvis F (2013) Half of the European fruit fly species barcoded (Diptera, Tephritidae); a feasibility test for molecular identification. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 279–305. doi: 10.3897/zookeys.365.5819
Collection data of all specimens included in this study. doi: 10.3897/zookeys.365.5819.app

ZooKeys 365: 307–328 (2013)
doi: 10.3897/zookeys.365.6027
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France

Gontran Sonet[1], Kurt Jordaens[2,3], Yves Braet[4], Luc Bourguignon[4], Eréna Dupont[4], Thierry Backeljau[1,3], Marc De Meyer[2], Stijn Desmyter[4]

1 *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (JEMU), Vautierstraat 29, 1000 Brussels, Belgium* **2** *Royal Museum for Central Africa, Department of Biology (JEMU), Leuvensesteenweg 13, 3080 Tervuren, Belgium* **3** *University of Antwerp, Evolutionary Ecology Group, Groenenborgerlaan 171, 2020 Antwerp, Belgium* **4** *National Institute of Criminalistics and Criminology, Vilvoordsesteenweg 100, 1120 Brussels, Belgium*

Corresponding author: *Gontran Sonet* (gontran.sonet@naturalsciences.be)

Academic editor: *Z. T. Nagy* | Received 1 August 2013 | Accepted 13 November 2013 | Published 30 December 2013

Citation: Sonet G, Jordaens K, Braet Y, Bourguignon L, Dupont E, Backeljau T, De Meyer M, Desmyter S (2013) Utility of GenBank and the Barcode of Life Data Systems (BOLD) for the identification of forensically important Diptera from Belgium and France. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 307–328. doi: 10.3897/zookeys.365.6027

Abstract

Fly larvae living on dead corpses can be used to estimate post-mortem intervals. The identification of these flies is decisive in forensic casework and can be facilitated by using DNA barcodes provided that a representative and comprehensive reference library of DNA barcodes is available.

We constructed a local (Belgium and France) reference library of 85 sequences of the COI DNA barcode fragment (mitochondrial cytochrome *c* oxidase subunit I gene), from 16 fly species of forensic interest (Calliphoridae, Muscidae, Fanniidae). This library was then used to evaluate the ability of two public libraries (GenBank and the Barcode of Life Data Systems – BOLD) to identify specimens from Belgian and French forensic cases. The public libraries indeed allow a correct identification of most specimens. Yet, some of the identifications remain ambiguous and some forensically important fly species are not, or insufficiently, represented in the reference libraries. Several search options offered by GenBank and BOLD can be used to further improve the identifications obtained from both libraries using DNA barcodes.

Keywords

Forensic entomology, COI, DNA barcoding, BLAST

Introduction

Insects collected on crime scenes can be used to estimate the time elapsed between death and corpse discovery, i.e. the post mortem interval or PMI (Rodriguez and Bass 1983, Joseph et al. 2011, Charabidze 2012). The correct identification of these insects is decisive in forensic casework since different species may have different developmental times under identical conditions. Erroneous identifications can therefore bias PMI estimates (Wells et al. 2001). DNA-based identification can be a valuable tool to identify immature life stages (Meiklejohn et al. 2013), fragments of insects, empty puparia (e.g. Mazzanti et al. 2010) or specimens of morphologically similar species (e.g. Meiklejohn et al. 2011, Jordaens et al. 2012). This technique relies on the comparison of a query sequence obtained from a sample collected at a crime scene with a library of reference sequences from well-identified specimens. The reference sequence showing the highest sequence similarity (= best match) with the query sequence can be used for its identification. However, the validity of this approach depends particularly on the reference library, which has to be representative, comprehensive and without misidentification or sequencing error (Wells and Stevens 2008).

In order to be of interest in court, species identifications provided by a specific reference library should be validated by assessing the likelihood of incorrect identifications using that library (Wells and Williams 2007, Wells and Stevens 2008). Sequences of a particular reference library may allow the correct identification of all species included in the library. However, if this library contains a limited set of species and ignores closely related species, then the likelihood of misidentifications is real (Wells and Stevens 2008). Moreover, the use of a reference library assembled in a different geographic area can also lead to incorrect species assignments because of geographic population structuring or eventual local hybrids (Stevens et al. 2002). Therefore, surveying local entomofaunas is a prerequisite for forensic specimen identifications (Vanin et al. 2008, Caine et al. 2009, Rolo et al. 2013). Likewise, assessing intraspecific variation and geographic substructuring is very important in forensic entomology (Wells and Williams 2007, Harvey et al. 2008, Desmyter and Gosselin 2009, Sonet et al. 2012).

The presence of pseudogene sequences and misidentified specimens in reference libraries is another problem that can constrain identification success (Wells and Stevens 2008). In order to minimise the risk of misidentifications caused by pseudogenes, an additional identification could be performed on the basis of an additional DNA fragment situated in another part of the mitochondrial genome (for example cytochrome *b*). Since most pseudogenes of mitochondrial origin are relatively short, the chance of sequencing two pseudogenes would drop substantially. Besides pseudogenes, sequences from misidentified specimens may be difficult to distinguish from haplotypes that are shared between correctly identified specimens from two different species (Whitworth et al. 2007). Increased sampling sometimes broadens the ranges of intra- and interspecific sequence divergences, even up to the point that they start overlapping so much that it becomes difficult to distinguish between the species (Wells et al. 2007).

Accurate identification of forensically important insects has been obtained using mitochondrial markers like the cytochrome *c* oxidase subunits I and II (COI and COII), cytochrome *b*, 16S rDNA, NADH dehydrogenase subunit 5, as well as nuclear markers like the ribosomal internal transcribed spacers 1 and 2, and the developmental gene bicoid (Sperling et al. 1994, Wells and Sperling 2000, Zehner et al. 2004, Guo et al. 2010, Li et al. 2010, Wang et al. 2010, Guo et al. 2011, Zaidi et al. 2011, Park et al. 2013). Among these markers, COI and COII have been predominantly used in forensic entomology (Sperling et al. 1994, Malgorn and Coquoz 1999, Vincent et al. 2000, Wallman and Donnellan 2001, Wells et al. 2001, Wells and Sperling 2001, Harvey et al. 2003, 2008, Wells and Stevens 2008, Liu et al. 2011, Boehme et al. 2012, Jordaens et al. 2012, Renaud et al. 2012). Coincidentally, a fragment of the 5' end of COI has been selected as the standard barcode marker for animal identification by the Consortium for the Barcode of Life (Hebert et al. 2003). DNA barcodes are linked to voucher specimens and are associated with additional information such as primer data and trace files. This practice allows to verify the quality of sequences and to re-examine the organism from which the DNA was extracted (Ratnasingham and Hebert 2007). Barcodes are deposited in the Barcode of Life Data Systems (BOLD) and are tagged as barcodes in GenBank. Consequently, the 5' end of COI is readily available in public reference libraries for a wide variety of dipterans of forensic interest (Wells and Stevens 2008).

In Western Europe, COI sequences from ca. 50 species of Sarcophagidae, ca. 10 species of Calliphoridae and five species of Muscidae are currently available as reference data for the identification of dipterans of forensic interest (Boehme et al. 2012, Jordaens et al. 2012). Specimens of seven species of Sarcophagidae and six species of Calliphoridae are from Belgium (Desmyter and Gosselin 2009, Jordaens et al. 2012, Marinho et al. 2012, Sonet et al. 2012). In this paper, we first extend the reference library of COI sequences with Belgian and French specimens of forensic interest belonging to two families (Calliphoridae and Muscidae) and secondly, we use these new sequences as queries to assess the validity of the identifications provided by GenBank and BOLD.

Methods

Specimens

We collected 85 adult specimens of 16 dipteran species of forensic interest from 24 localities in Belgium and three localities in France (Table 1). All Belgian specimens came from forensic cases. Three specimens from three species (*Neomyia cornicina*, *Polietes lardarius* and *Eudasyphora cyanella*) were collected on corpses but are currently not used for the calculation of the PMI. The French specimens of *Chrysomya albiceps* and *Lucilia sericata* were not collected on corpses, but were added because of their forensic interest. Morphological species identification was done by two taxonomic experts of Diptera (YB and ED), using five identification keys (D'Assis Fonseca 1968, Beĭ-Bienko 1988,

Table 1. Morphological identification and geographic origin of all specimens sampled in this study. Process ID of each sequence in the Barcode of Life Data Systems (BOLD) is given. For each haplotype (identified here by a number), the similarity of the best match (BM) retrieved from GenBank or BOLD using five search procedures is in italic if identification was ambiguous and normal font if identification was unambiguous. Ambiguous identifications where the best correct and the best incorrect matches had the same similarity with the query are indicated by an asterisk; best correct matches were more similar to the query than best incorrect matches in all other ambiguous identifications. No value is given when best matches were all of < 99% similarity and "na" when longer COI fragments were not available. Search procedures are: barcode fragment (642–658 bp) submitted to GenBank (1), to the public records of BOLD (2), to the species level records of BOLD including early releases (3), to the records of GenBank that are tagged as barcodes (4) and longer COI fragment (1412–1534 bp) submitted to GenBank (5).

Family	Species	Country	Locality	BOLD Process ID	Barcode fragment: haplotype ID	Longer COI fragment: haplotype ID	Similarity with BM in procedures 1 & 2 (%)	Similarity with BM in procedure 3 (%)	Similarity with BM in procedure 4 (%)	Similarity with BM in procedure 5 (%)
Calliphoridae	*Calliphora vicina* Robineau-Desvoidy, 1830	Belgium	Saint–Gilles/ Sint–Gillis	NICC001-13	1	1	99.72*	99.83*	99.54	99.03
		Belgium	Schaerbeek/ Schaarbeek	NICC002-13	2	na	100*	100*	99.85	na
		Belgium	Andrimont	NICC003-13	2	2	100*	100*	99.85	99.05
		Belgium	Hastière	NICC004-13	2	3	100*	100*	99.85	99.23
		Belgium	Humbeek	NICC005-13	2	4	100*	100*	99.85	99.16
		Belgium	Ixelles/Elsene	NICC006-13	3	5	99.72*	99.83*	99.54	
		Belgium	Auderghem/ Oudergem	NICC007-13	2	3	100*	100*	99.85	99.23
		Belgium	Lier	NICC008-13	4	na	100*	100*	99.85	na
		Belgium	Gent	NICC009-13	5	6	99.86*	100*	99.7	
		Belgium	Saintes	NICC010-13	6	7	99.86*	99.6*	99.7	99.1
		Belgium	Bruxelles/ Brussel	NICC011-13	2	3	100*	100*	99.85	99.23
		Belgium	Hastière	NICC012-13	7	8	99.86*	99.86*	99.7	99.03
		Belgium	Auderghem/ Oudergem	NICC013-13	8	9	100*	100*	100	99.1
		Belgium	Schaerbeek/ Schaarbeek	NICC014-13	9	10	99.72*	99.83*	99.54	99.03
	Calliphora vomitoria (Linnaeus, 1758)	Belgium	Toernich	NICC015-13	10	na	100	100	99.09	na
		Belgium	Hastière	NICC016-13	10	11	100	100	99.09	99.02
		Belgium	Andrimont	NICC017-13	11	12	99.58	100	99.38	

Family	Species	Country	Locality	BOLD Process ID	Barcode fragment: haplotype ID	Longer COI fragment: haplotype ID	Similarity with BM in procedures 1 & 2 (%)	Similarity with BM in procedure 3 (%)	Similarity with BM in procedure 4 (%)	Similarity with BM in procedure 5 (%)
		Belgium	Schaerbeek/ Schaarbeek	NICC018-13	10	11	100	100	99.09	99.02
		Belgium	Genk	NICC019-13	10	11	100	100	99.09	99.02
		Belgium	Laeken/Laken	NICC020-13	10	13	100	100	99.09	
		Belgium	Liège	NICC021-13	10	11	100	100	99.09	99.02
		Belgium	Saintes	NICC022-13	10	14	100	100	99.09	
		Belgium	Steendorp	NICC023-13	10	11	100	100	99.09	99.02
		Belgium	Gent	NICC024-13	10	11	100	100	99.09	99.02
		Belgium	Hastière	NICC025-13	10	na	100	100	99.09	na
		Belgium	Schaerbeek/ Schaarbeek	NICC026-13	10	11	100	100	99.09	99.02
		Belgium	Genk	NICC027-13	10	11	100	100	99.09	99.02
		Belgium	Sint–Laureins	NICC028-13	10	11	100	100	99.09	99.02
		Belgium	Schoonaarde	NICC029-13	12	15	99.86	100	98.94	
		Belgium	Antwerpen	NICC030-13	10	na	100	100	99.09	na
	Chrysomya albiceps (Wiedemann, 1819)	France	St Pourçain/ Sioule	NICC031-13	13	16	100	100		99.29
		France	St Pourçain/ Sioule	NICC032-13	14	na	100	100		na
		France	St Pourçain/ Sioule	NICC033-13	15	17	99.86	100		99.23
		France	St Pourçain/ Sioule	NICC034-13	13	16	100	100		99.29
		France	St Pourçain/ Sioule	NICC035-13	14	na	100	100		na
		France	Sarreguemines	NICC036-13	13	16	100	100		99.29
		Belgium	Meerdaalwoud	NICC037-13	16	na	99.86	100		na
	Cynomya mortuorum (Linnaeus, 1761)	Belgium		NICC038-13	17	18	99.73	100		
	Lucilia ampullacea Villeneuve, 1922	Belgium	Flémalle	LUCIL001-12	18	19	100	100		99.23
		Belgium	Flémalle	LUCIL002-12	19	20	99.86	99.86		99.1

Family	Species	Country	Locality	BOLD Process ID	Barcode fragment: haplotype ID	Longer COI fragment: haplotype ID	Similarity with BM in procedures 1 & 2 (%)	Similarity with BM in procedure 3 (%)	Similarity with BM in procedure 4 (%)	Similarity with BM in procedure 5 (%)
	Lucilia sericata (Meigen, 1826)	Belgium	Hastière	LUCIL061-12	20	21	99.86	100	100	99.23
		Belgium	Auderghem/ Oudergem	LUCIL062-12	21	na	100	100*	100	na
		Belgium	Gent	LUCIL063-12	22	22	100	100*	100	99.16
		Belgium	Lier	LUCIL064-12	22	23	100	100*	100	99.23
		France	Le Soler	LUCIL065-12	22	24	100	100*	100	99.29
		France	Le Soler	LUCIL066-12	23	25	99.86	100*	100	99.23
		Belgium		LUCIL067-12	22	24	100	100*	100	99.29
		Belgium	Auderghem/ Oudergem	LUCIL068-12	24	26	100	100	100	99.23
		Belgium	Schaerbeek/ Schaarbeek	LUCIL069-12	25	27	100	100*	100	99.16
		Belgium	Genk	LUCIL070-12	22	24	100	100*	100	99.29
		Belgium	Genk	LUCIL071-12	26	28	99.86	100*	100	99.23
		France		LUCIL072-12	22	29	100	100*	100	99.23
		France		LUCIL073-12	22	29	100	100*	100	99.23
		France		LUCIL074-12	22	29	100	100*	100	99.23
		France		LUCIL075-12	22	29	100	100*	100	na
		France		LUCIL076-12	22	29	100	100*	100	99.23
	Protophormia terraenovae (Robineau-Desvoidy, 1830)	Belgium	Andrimont	NICC056-13	27	na	99.86	100	99.7	na
		Belgium	Auderghem/ Oudergem	NICC057-13	27	na	99.86	100	99.7	na
		Belgium	Andrimont	NICC058-13	28	na	100	100	99.85	na
		Belgium	Andrimont	NICC059-13	27	na	99.86	100	99.7	na
		Belgium	Andrimont	NICC060-13	27	na	99.86	100	99.7	na
		Belgium	Andrimont	NICC061-13	29	na	100	100	99.85	na
		Belgium	Andrimont	NICC062-13	28	na	100	100	99.85	na
		Belgium	Andrimont	NICC063-13	30	na	100	100	100	na
		Belgium	Andrimont	NICC064-13	28	na	100	100	99.85	na
		Belgium	Andrimont	NICC065-13	31	na	99.72	100	99.85	na

Family	Species	Country	Locality	BOLD Process ID	Barcode fragment: haplotype ID	Longer COI fragment: haplotype ID	Similarity with BM in procedures 1 & 2 (%)	Similarity with BM in procedure 3 (%)	Similarity with BM in procedure 4 (%)	Similarity with BM in procedure 5 (%)
		Belgium	Andrimont	NICC066-13	28	na	100	100	99.85	na
		Belgium	Auderghem/ Oudergem	NICC067-13	27	na	99.86	100	99.7	na
Fanniidae	*Fannia* sp1	Belgium	Soignes/ Zoniën forest	NICC040-13	32	30		100		
	Fannia sp2	Belgium	Soignes/ Zoniën forest	NICC041-13	33	31		100		
	Fannia sp3	Belgium	Soignes/ Zoniën forest	NICC042-13	34	32				
Muscidae	*Eudasyphora cyanella* (Meigen, 1826)	Belgium	Soignes/ Zoniën forest	NICC039-13	35	33	100	100		
	Musca autumnalis De Geer, 1776	Belgium	Soignes/ Zoniën forest	NICC043-13	36	34	100	100		
	Muscina levida (Harris, 1780)	Belgium	Pecq	NICC044-13	37	35	100	100	100	
		Belgium	Pecq	NICC045-13	38	36	99.85	100	99.85	
		Belgium	Pecq	NICC046-13	37	37	100	100	100	
		Belgium	Pecq	NICC047-13	37	35	100	100	100	
		Belgium	Soignes/ Zoniën forest	NICC048-13	37	38	100	100	100	
	Muscina prolapsa (Harris, 1780)	Belgium	Soignes/ Zoniën forest	NICC049-13	39	39		99.85		
		Belgium	Soignes/ Zoniën forest	NICC050-13	39	na		99.85		
		Belgium	Saint–Gilles/ Sint–Gillis	NICC051-13	40	40		100		
		Belgium	Saint–Gilles/ Sint–Gillis	NICC052-13	40	40		100		
		Belgium	Soignes/ Zoniën forest	NICC053-13	40	na		100		na
	Neomyia cornicina (Fabricius, 1781)	Belgium	Soignes/ Zoniën forest	NICC054-13	41	41				
	Polietes lardarius (Fabricius, 1781)	Belgium	Soignes/ Zoniën forest	NICC055-13	42	42	99.84	99.85		

Rozkošný et al. 1997, Gregor et al. 2002, Szpila 2012). Three *Fannia* specimens (Fanniidae) could not be identified to the species level and were considered as three putative different species. Specimens were deposited as vouchers at the National Institute of Criminalistics and Criminology in Brussels, Belgium (Table 1).

Laboratory protocols

We extracted genomic DNA from one or two legs per specimen using the NucleoSpin Tissue Kit (Macherey-Nagel) and a final elution volume of 70 µl. Fragments of the COI marker were amplified using two primer pairs TY-J-1460/C1-N-2191 and C1-J-2183/TL2-N-3014 (Sperling et al. 1994, Wells and Sperling 1999). The fragment obtained with the first primer pair encompasses the barcode region of ca. 650 bp used for animals (Hebert et al. 2003). The assembly of the fragments obtained with both primer pairs generated a sequence of 1534 bp corresponding to the complete COI gene. Each 25 µl PCR reaction contained final concentrations of 0.2 mM dNTPs, 0.4 µM of each primer, 2.0 mM $MgCl_2$, 0.5 U of Taq DNA polymerase (Platinum, Invitrogen), 1 × PCR buffer and 2-4 µl DNA template. The thermal cycler program consisted of an initial denaturation step of 4 min at 94 °C, followed by 40 cycles of 30 s at 94 °C, 30 s at 45 °C and 90 s at 72 °C; with a final extension of 7 min at 72 °C. We cleaned PCR products using the NucleoFast96 PCR Kit (Macherey-Nagel) and sequenced them bidirectionally on an ABI 3130 Genetic Analyzer (Applied Biosystems) using the BigDye Terminator Cycle Sequencing Kit v3.1.

Sequence quality control and analysis

We assembled and aligned sequences in SeqScape v2.5 (Applied Biosystems) and confirmed the absence of stop codons using MEGA5 (Tamura et al. 2011). Sequences were deposited in BOLD (BOLD process ID's are given in Table 1) and GenBank. All different haplotypes were extracted from the aligned sequences using the R package PEGAS (Paradis 2010). We calculated pairwise p-distances (i.e. the proportion of sites at which two sequences differ) and searched for haplotypes that were shared among species.

Haplotypes were then used as queries to search for most similar sequences in two public databases: GenBank (NCBI, National Centre for Biotechnology Information) and BOLD (the Barcode of Life Data Systems). These most similar sequences will be called "best matches" *sensu* Meier et al. (2006) in the following. In GenBank, searches were done using MegaBLAST, the Basic Local Alignment Search Tool (BLAST) optimised for highly similar sequences (Zhang et al. 2000, Morgulis et al. 2008). In BOLD, the in-built Identification System (IDS) was applied (Ratnasingham and Hebert 2007) on two different databases: the Public Record Barcode Database (341,580 sequences; 45,368 nominal species and 11,732 interim species, or candidate species

that have not been described yet on 24 May 2013) and the Species Level Barcode Records (1,367,662 sequences; 127,679 species and 53,394 interim species on 24 May 2013). The first database comprises the same records as GenBank because both libraries regularly synchronize their published records. In BOLD, this database of public records is a collection of COI records of minimum 500 bp from the published projects of BOLD. The Species Level Barcode Records of BOLD is used by default in IDS. It contains, in addition to the published COI records, early data release of COI records with a species level identification and a minimum sequence length of 500 bp. These early releases contain all information necessary for barcodes (locality and date of sample collection, trace files and sequence information as well as voucher specimen and database identifiers), have passed computerized quality checks of BOLD but might include provisional taxonomic assignments (Hebert et al. 2010).

In total, we applied five search strategies by submitting the barcode sequences to 1) GenBank, 2) the Public Records of BOLD, 3) the Species Level Records of BOLD including early releases, as well as 4) by using the barcode sequences as queries in combination with a keyword, "barcode", in GenBank, and 5) by submitting COI sequences longer than the barcode fragment (1412–1534 bp) to GenBank. The use of the keyword "barcode" allowed us to filter the GenBank reference sequences and obtain only best matches that are tagged as barcodes, not only in the field "keyword" but also in any field of GenBank records. Longer COI sequences have not been submitted to BOLD because BOLD was developed to accept sequences from the strict barcode region only. In BOLD, IDS returns a list of maximum 99 best matches and provides a species-level identification for best close matches showing less than 1% divergence (Ratnasingham and Hebert 2007). Since BLAST searches are based on approximate alignments (regions of local similarity between sequences), species assignments are usually preferably performed on the basis of local alignments. Hence we verified that the best hits and their percentages of sequence identity obtained from the MegaBLAST searches in GenBank were identical to those (= 1 - p-distance) calculated with MEGA5 (Tamura et al. 2011) using local databases downloaded from GenBank and aligned with CLUSTAL W (Thompson et al. 1994). Identifications were made on the basis of the highly similar best matches (> 99% similarity), according to the "best close match" method of Meier et al. (2006). We qualified each best match with a similarity of > 99% as correct if it had the same species name as the query or as incorrect if it had a different species name than the query. In addition, the identification of a query was considered as unambiguous if all best matches with a similarity of > 99% had the same species name. If this was not the case, then the identification was ambiguous. For each identification, we made sure that best close matches included only records properly identified to the species level by excluding the few records with provisional identifications (a code instead of a nominal species name). We also verified whether the alignment of the query with each best match comprised at least 600 bp. When no best match of > 99% similarity was retrieved for a given query, the presence of conspecific and congeneric barcode sequences of > 500 bp was investigated in both public libraries. If present, their divergences (p-distances) with the queries were calculated using MEGA5 (Tamura et al. 2011).

Results

In total 85 sequences were obtained with more than 641 bp of the COI DNA barcode fragment, representing 42 haplotypes. The majority of them (63 sequences) involved a longer COI fragment (1412–1534 bp), representing 42 other haplotypes. Pairwise intraspecific p-distances ranged from zero to 0.5% and none of the species represented in this dataset shared haplotypes.

Search procedures 1 and 2

Using the 42 haplotypes of the barcode region as a query yielded the same results in GenBank and in the Public Record Barcode Database of BOLD. Best matches of > 99% similarity were retrieved for 36/42 haplotypes, representing 11 out of 16 species (Table 1). These best matches were either identical (17/36) or differed from the query in less than three substitutions (19/36). We obtained at least one correct best match for each query. However, species identifications were either unambiguous (18 queries, 8 species) or ambiguous (18 queries, 3 species). For two queries, best matches included species of another genus: *Musca domestica* Linnaeus, 1758 was found for *Calliphora vicina* and *Chrysomya megacephala* (Fabricius, 1794) for *Lucilia ampullacea.* In all other cases of ambiguous identification, best matches involved congenerics: *Calliphora croceipalpis* Jaennicke, 1867 was found for *Calliphora vicina*, *Lucilia cuprina* (Wiedemann, 1830) for *Lucilia sericata* and *Lucilia porphyrina* (Walker, 1856) for *Lucilia ampullacea*. Finally, the number of best matches with > 99% similarity varied from one to more than 99 per query (the number of best matches displayed by BOLD is limited to 99). For five species, less than five sequences with a similarity of > 99% were retrieved (Figure 1).

For six queries, the best matching similarities were < 93.5%. These included the haplotypes of *Fannia* sp1, sp2 and sp3, *Muscina prolapsa* and *Neomyia cornicina.* There were no COI sequences of *Muscina prolapsa* or of *Neomyia cornicina* in GenBank. For *Fannia*, fragments of the barcode region of > 500 bp were available for 14 specimens representing four species, viz. *Fannia canicularis* (Linnaeus, 1761), *Fannia scalaris* (Fabricius, 1794), *Fannia brevicauda* Chillcott, 1961 and *Fannia serena* (Fallen, 1825) but their p-distances with our three *Fannia* haplotypes ranged from 6.6% to 16.2%.

Search procedure 3

Using the Species Level Barcode Records dataset of BOLD (Table 1), highly similar best matches (> 99%) were retrieved for 40/42 queries (14/16 species). Correct best matches were retrieved for all specimens identified at the species level, but identifications were often ambiguous (25 queries, 6 species). This method yielded a higher

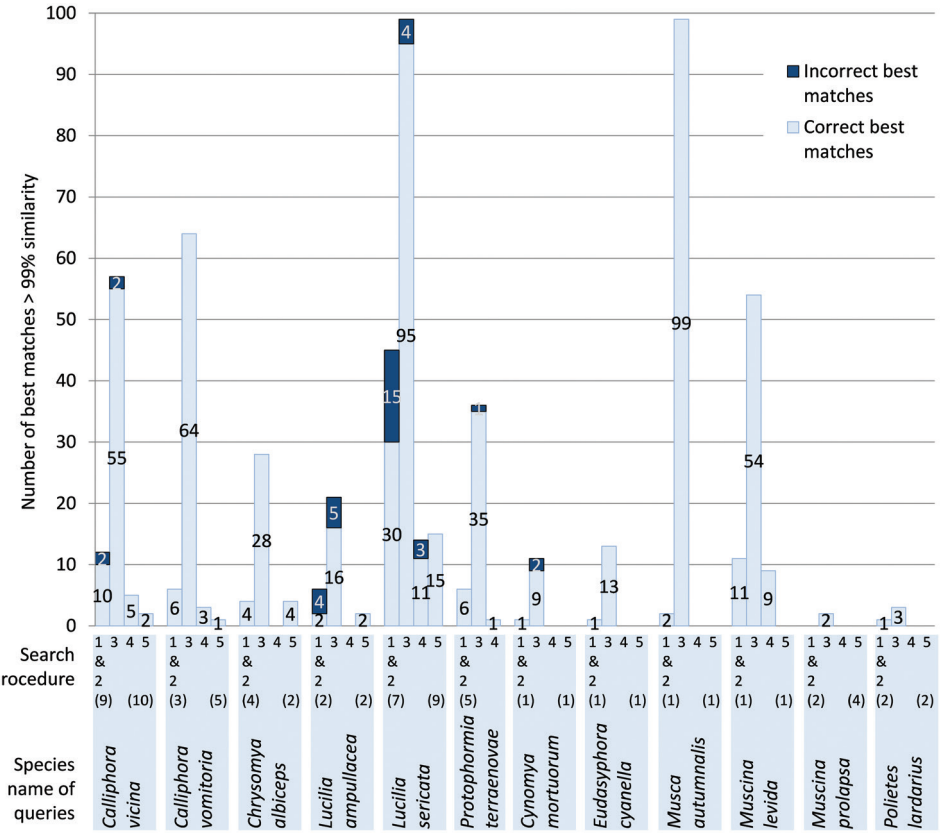


Figure 1. Best matches obtained for each species using five different search procedures: Barcode fragment (642–658 bp) submitted to GenBank (1) and the public records of BOLD (2); barcode fragment submitted to the species level records of BOLD, including early-released sequences (3); barcode fragment and keyword "barcode" submitted to GenBank (4) and longer COI fragment (1412–1534 bp) submitted to GenBank (5). Numbers of haplotypes used as queries are between parentheses. Longer COI fragments were obtained for all species except for *Protophormia terraenovae.*

proportion of best matches of > 99% similarity than when the search was restricted to public records (95% of the queries instead of 86%). However, the proportion of unambiguous identifications was smaller (38% instead of 50% of the queries; Table 2). Yet, in contrast to all the other searches, early-released sequences provided two correct matches for *Muscina prolapsa*, one match for *Fannia* sp2 and three matches for *Fannia* sp1 (correct at the genus level). The latter identification was ambiguous since two best matches showed 100% similarity with *Fannia lustrator* (Harris, 1780) and one showed 99.85% similarity with *F. pallitibia* (Rondani, 1866). The two queries for which best matches were of < 99% similarity were from *Fannia* sp3 and *Neomyia cornicina.* No barcodes were available for *Neomyia cornicina* in BOLD.

Table 2. Evaluation of the DNA-based identifications obtained in this study using five search procedures: barcode fragment submitted to GenBank (1), to the public records of BOLD (2), to the species level records of BOLD including early releases (3), to the records of GenBank that are tagged as barcodes (4) and longer COI fragment submitted to GenBank (5). Only best matches of > 99% similarity were considered. OK: correct unambiguous identification; OK +: ambiguous identification due to correct and incorrect best matches (species names associated with incorrect best matches are given with the abbreviated genus name in case of congeneric matches); *: ambiguous identifications where the best correct and the best incorrect matches had the same similarity with the query; best correct matches were more similar to the query than best incorrect matches in all other ambiguous identifications; na: longer COI fragment not available; empty cell: no best match above 99% similarity. Numbers without parentheses were obtained with the barcode fragment and numbers between parentheses were obtained with the longer COI fragment. In order to allow comparisons between the results obtained with the barcode and the longer COI datasets, values obtained with the barcode fragment of the sequences for which the longer COI fragment was available are given between brackets.

Species	Number of haplotypes	Search procedure			
		1 & 2	3	4	5
Calliphora vicina	9 (10)	OK + *C. croceipalpis**, *Musca domestica**	OK + *C. croceipalpis**, *Musca domestica**	OK	OK
Calliphora vomitoria	3 (5)	OK	OK	OK	OK
Chrysomya albiceps	4 (2)	OK	OK		OK
Lucilia ampullacea	2 (2)	OK + *L. porphyrina*, *Chrysomya megacephala*	OK + *Chrysomya megacephala*		OK
Lucilia sericata	7 (9)	OK + *L. cuprina*	OK + *L. cuprina**	OK + *L. cuprina*	OK + *L. cuprina*
Protophormia terraenovae	5 (0)	OK	OK + *P. uralensis*	OK	na
***Fannia* sp1**	1 (1)		*F. pallitibia*, *F. lustrator*		
***Fannia* sp2**	1 (1)		*F. manicata*		
***Fannia* sp3**	1 (1)				
Cynomya mortuorum	1 (1)	OK	OK + *C. cadaverina*		
Eudasyphora cyanella	1 (1)	OK	OK		
Musca autumnalis	1 (1)	OK	OK		
Muscina levida	2 (4)	OK	OK	OK	
Muscina prolapsa	2 (2)		OK		
Neomyia cornicina	1 (1)				
Polietes lardarius	1 (1)	OK	OK		
% of species with matches > 99% similarity		**69** [67]	**88** [87]	**31** [27]	(33)
% of queries with matches > 99% similarity		**86** [86]	**95** [95]	**62** [67]	(67)
% of species with unambiguous ID		**73** [70]	**57** [62]	**80** [75]	(80)
% of queries with unambiguous ID		**50** [42]	**38** [42]	**73** [68]	(68)
% of species with ambiguous ID		**27** [30]	**43** [38]	**20** [25]	(20)
% of queries with ambiguous ID		**50** [58]	**62** [58]	**27** [32]	(32)

Search procedure 4

When both the barcode sequences and the keyword "barcode" were used as queries in GenBank, we retrieved best matches of > 99% similarity for *Calliphora vicina*, *Calliphora vomitoria*, *Lucilia sericata*, *Protophormia terraenovae* and *Muscina levida* (Tables 1 and 2). All best matches of > 99% similarity were correct and provided unambiguous identifications except for *Lucilia sericata*, which matched with both correct and incorrect species names (*Lucilia cuprina* and *Lucilia sericata*).

Search procedure 5

Haplotypes of longer COI fragments (1412–1534 bp) were also submitted to a MegaBLAST search on GenBank. Best matches of > 99% similarity were obtained for all haplotypes of *Calliphora vicina*, *Calliphora vomitoria, Chrysomya albiceps*, *Lucilia ampullacea* and *Lucilia sericata*. Like in the previous analysis, all best matches were correct and provided unambiguous identifications except for *Lucilia sericata* (best matches included *Lucilia sericata* and *Lucilia cuprina*).

Discussion

Towards a COI reference database for the forensically important dipterans in Western Europe

With this study we contributed to the establishment of a local COI reference library for fly species of forensic importance in Belgium and France. As such, we provide the first barcodes for *Muscina prolapsa* and *Neomyia cornicina*. We also extended the geographic coverage of barcodes of species which hitherto were only sampled from a limited number of localities, e.g. *Cynomya mortuorum* and *Polietes lardarius* were each represented by only one barcode sequence from the UK (Kutty et al. 2008). Similarly, barcodes of *Muscina levida* were until now only available for samples from Canada, Germany (Renaud et al. 2012) and the USA (Nakano and Honda, unpublished). Conversely, barcodes of the other species sampled here were obtained from no more than five European countries. Ideally, a reliable reference library should comprise a large sampling of sequences, not only representing the European dipteran species that are currently used in forensics (whose development times have been studied under different temperature conditions), but also those of potential forensic interest (occurring on carcasses but whose biology has been less studied) and all their close relatives. Currently, 13 species belonging to 10 genera are being used in forensic investigations (Marchenko 2001, Grassberger et al. 2002, Richards et al. 2009, Velásquez et al. 2013). Hence, the geographic coverage of GenBank and BOLD is still far from comprehensive. Yet, we did not observe intraspecific COI divergences of > 1% at COI, neither among specimens sequenced in this study nor between

them and their conspecific best matches in the public libraries (intraspecific distances among GenBank sequences were not calculated here). This indicates that geographic coverage does not always have to be complete to allow correct species identification. Nonetheless, a more comprehensive reference library may comprise more haplotypes, allowing a better assessment of the risk of incorrect identifications (Meier et al. 2006). Indeed, an increased sampling can result in a more difficult distinction between some closely related species (Bergsten et al. 2012) and this has considerable importance for courts.

Evaluation of the DNA-based identifications of forensically important flies in Belgium and France provided by GenBank and BOLD

For 86% of the barcode fragments used as queries, we retrieved highly similar conspecific sequences (> 99% similarity) from GenBank and BOLD. The more divergent best matches (< 99% similarity) obtained for the remaining 14% of the queries would have produced either incorrect (*Muscina prolapsa* and *Neomyia cornicina*) or doubtful identifications (*Fannia*) if all best matches were taken into account for identification. The better performance of the best close match method compared to the simple best match method has already been reported (e.g. Meier et al. 2006, Virgilio et al. 2010). However, even with the best close match method, our results revealed three issues that can hamper the DNA-based identification of forensically important flies in Belgium and France using GenBank or BOLD: These databases 1) do not include some fly species of forensic interest, 2) include sequences from misidentified specimens and 3) cannot always discriminate between closely related species. Below, we discuss these three issues in more detail.

1) Species not represented in the libraries

Our results showed that some fly species collected at Belgian crime scenes are not represented by COI records in GenBank and BOLD. *Muscina prolapsa*, for which no barcode sequence is present in GenBank, colonises carrion and buried remains (Gunn and Bird 2011, Prado e Castro et al. 2012). Also, the identification of *Fannia* species of forensic interest (Prado e Castro et al. 2012) is hampered by their limited representation in GenBank and BOLD. *Neomyia cornicina* is currently not used for PMI estimation but the availability of reference sequences of such species collected on crime scenes can decrease the risk of incorrect identification and help to characterize the entomofauna surrounding the crime scene (Amendt et al. 2007).

2) Sequences from misidentified specimens

Identifications based on the barcode fragment were ambiguous for 50% of the queries and for 27% of the species. Some ambiguous identifications can result from misidenti-

fied sequences in the libraries and could be corrected after re-examining the voucher specimens (Collins and Cruickshank 2013). In our study, the best matches with sequences from different genera could be the result of misidentifications: records of *Musca domestica* (GenBank accession number JQ350716) and *Chrysomya megacephala* (KC135926) matched our sequences of *Calliphora vicina* and *Lucilia ampullacea,* respectively.

3) Identification of closely related species

Still, most ambiguous identifications involved closely related species that are not necessarily incorrectly identified (Stevens et al. 2002, Sonet et al. 2012). For example, Wells et al. (2007) and Wells and Stevens (2008) showed that the barcodes of several specimens of *Lucilia cuprina* (from Hawaii and Asia) are more similar to those of *Lucilia sericata* than to those of other *Lucilia cuprina* specimens. This explains the ambiguous identification obtained here for *Lucilia sericata*. In some cases, the arbitrary similarity threshold, below which matches cannot be used for identification, is too low. Consequently, best close matches with conspecific and allospecific sequences are considered for identification, even if all conspecific best matches are closer to the query than any of the allospecific ones. To solve this problem, the similarity threshold can be adapted according to the gap between intra- and interspecific distances observed in this particular group of species (Lefébure et al. 2006, Collins and Cruickshank 2013, Puillandre et al. 2012, Virgilio et al. 2012). Here, we only used an arbitrary threshold of 99% similarity. A stricter similarity threshold (e.g. 99.5%) would resolve ambiguous identifications obtained for *Lucilia ampullacea*, for *Lucilia sericata* (but not when early releases of BOLD are used) and for *Cynomya mortuorum* (Tables 1 and 2).

Similarity values between the query and its best matches can be calculated using several methods. Here, similarities with GenBank records were determined as 1 - p-distances but no explicit information was found on the exact method used by the IDS of BOLD to determine the similarity values. Even if the IDS of BOLD applied a different method than ours, – distances are standardly corrected using the Kimura 2-parameter model (Kimura 1980) in DNA barcoding (Hebert et al. 2003) – the two searches (1 and 2) using the same queries against the same public records resulted in an identical list of highly similar best matches. Several studies have indeed observed that biases due to different distance calculation methods are less severe with similar sequences than with divergent ones (Collins et al. 2012, Fregin et al. 2012).

Expanding or restricting the search in GenBank and BOLD?

It is striking that identifications provided by GenBank and BOLD for the barcode fragment were either ambiguous or involved a rather limited number of very similar reference sequences (Figure 1). Therefore, we tested alternative search strategies to optimise the number of best matches and minimise the number of ambiguous identifications.

For this, we used different options offered by GenBank and BOLD by 1) including early releases from BOLD in the reference library, 2) adding the keyword "barcode" as a query in GenBank and 3) using longer COI sequences as queries in GenBank.

Including early releases as reference sequences in BOLD increased the number of best matches of > 99% similarity but also increased the proportion of ambiguous identifications (Table 1). Early releases might not have passed all controls that authors and reviewers make in the process of publication (e.g. Schindel et al. 2011). They are therefore more prone to errors. However, their early release allows the detection of errors and inconsistencies before publication, which is an efficient way to improve the quality of the reference libraries. In addition, they largely outnumber the published sequences and may include precious additional information such as rare haplotypes.

In order to improve the search for sequences that have been produced for DNA barcoding purposes, we added the word "barcode" to each query in GenBank. With this procedure, the number of best matches of > 99% similarity and the proportion of ambiguous matches drastically decreased. The same tendency was observed when longer COI sequences (1412–1534 bp) were used as queries. This is due to the smaller number of reference sequences that are tagged as barcodes or are longer than the standard barcode fragment. Therefore, this kind of search is currently only relevant for the identification of fly species of forensic interest that are well represented by longer COI reference sequences or that are tagged as barcodes. Moreover, longer DNA fragments are not always easy to sequence from degraded forensic samples (Mazzanti et al. 2010). Due to the limited number of best matches of > 99% similarity retrieved by these two options, it was not possible to assess their benefit when trying to minimise the proportion of ambiguous identifications.

Conclusion

Even if BOLD and GenBank contain the same public records, they offer different options for optimizing their use as reference libraries. For barcode data, we recommend using the BOLD Identification System and searching the dataset including early-released sequences (Species Level Barcode Records). This option optimises the number of best-matches and allows to verify the quality of the data (published or early-released sequence, barcode compliant or not, link with voucher specimens, etc.). When working with reference material, we encourage the early release of the data and the correction of any mistake detected at this stage (e.g. misidentification). Furthermore, entering sequences into a BOLD project gives access to a workbench with supplementary tools (tables with best matches, best close matches and construction of Neighbour-Joining trees), that are useful for quality control (Ratnasingham and Hebert 2007). If ambiguous identifications are obtained, it is possible to restrict the search to the published sequences only (BOLD or GenBank) or to the sequences that were produced in the framework of the DNA barcoding initiative. Finally, a further validation with other DNA fragments, morphological characters or ecological evidence might be necessary.

Without such a validation, identifications will remain questionable and can only be applied to more inclusive taxonomic levels (Wilson et al. 2011). Although DNA barcoding has been validated for forensic use (Dawnay et al. 2007), its applicability in forensics clearly depends on the reliability of the data and of the identification method used (Pereira et al. 2010, Linacre et al. 2011).

Acknowledgements

We wish to thank the teams of the DNA and Microtraces Analysis (NICC), especially Dr. F. Hubrecht, Dr. S. Vanpoucke, and Dr. F. Noel for their support during our work. We also thank the reviewers of the manuscript for their very constructive and pertinent comments. This research is part of the BC42W project and was carried out by the Joint Experimental Molecular Unit – JEMU, which is financed by the Belgian Science Policy Office (BELSPO).

References

Amendt J, Campobasso C, Gaudry E, Reiter C, LeBlanc H, Hall M (2007) Best practice in forensic entomology—standards and guidelines. International Journal of Legal Medicine 121: 90–104. doi: 10.1007/s00414-006-0086-x

Beĭ-Bienko GI (1988) Keys to the insects of the European part of the USSR fauna, Volume 5, Smithsonian Institution Libraries and the National Science Foundation, Washington D.C.

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN, Ribera I, Nilsson AN, Barraclough TG, Vogler AP (2012) The effect of geographical scale of sampling on DNA barcoding. Systematic Biology 61: 851–869. doi: 10.1093/sysbio/sys037

Boehme P, Amendt J, Zehner R (2012) The use of COI barcodes for molecular identification of forensically important fly species in Germany. Parasitology Research 110: 2325–2332. doi: 10.1007/s00436-011-2767-8

Caine LM, Real FC, Salona-Bordas MI, de Pancorbo MM, Lima G, Magalhaes T, Pinheiro F (2009) DNA typing of Diptera collected from human corpses in Portugal. Forensic Science International 184: e21–3. doi: 10.1016/j.forsciint.2008.10.016

Charabidze D (2012) La biologie des insectes nécrophages et leur utilisation pour dater le décès en entomologie médico-légale. Annales de la Société Entomologique de France 48: 239–252. doi: 10.1080/00379271.2012.10697773

Collins RA, Boykin LM, Cruickshank RH, Armstrong KF (2012) Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. Methods in Ecology and Evolution 3: 457–465. doi: 10.1111/j.2041-210X.2011.00176.x

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. Molecular Ecology Resources 13: 969–975. doi: 10.1111/1755-0998.12046

D'Assis Fonseca ECM (1968) Diptera Cyclorrhapha Calyptrata: Muscidae. Handbooks for the Identification of British Insects. 10 Ed. Royal Entomological Society of London, London, 119 pp.

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. doi: 10.1016/j.forsciint.2006.09.013

Desmyter S, Gosselin M (2009) COI sequence variability between Chrysomyinae of forensic interest. Forensic Science International, Genetics 3: 89–95. doi: 10.1016/j.fsigen.2008.11.002

Fregin S, Haase M, Olsson U, Alström P (2012) Pitfalls in comparisons of genetic distances: A case study of the avian family Acrocephalidae. Molecular Phylogenetics and Evolution 62: 319–328. doi: 10.1016/j.ympev.2011.10.003

Grassberger M, Reiter C (2002) Effect of temperature on development of *Liopygia* (=*Sarcophaga*) *argyrostoma* (Robineau-Desvoidy) (Diptera: Sarcophagidae) and its forensic implications. Journal of Forensic Sciences 47: 1332–6. doi: 10.1520/JFS15570J

Gregor F, Rozkošný R, Barták M, Vaňhara J (2002) The Muscidae (Diptera) of Central Europe. Folia Facultatis Scientiarum Naturalium Universitatis Masarykianae Brunensis Biologia 107: 1–280.

Gunn A, Bird J (2011) The ability of the blowflies *Calliphora vomitoria* (Linnaeus), *Calliphora vicina* (Robineau-Desvoidy) and *Lucilia sericata* (Meigen) (Diptera: Calliphoridae) and the muscid flies *Muscina stabulans* (Fallén) and *Muscina prolapsa* (Harris) (Diptera: Muscidae) to colonise buried remains. Forensic Science International 207: 198–204. doi: 10.1016/j.forsciint.2010.10.008

Guo YD, Cai JF, Li X, Xiong F, Su RN, Chen FL, Liu QL, Wang XH, Chang YF, Zhong M, Wang X, Wen JF (2010) Identification of the forensically important sarcophagid flies *Boerttcherisca peregrina*, *Parasarcophaga albiceps* and *Parasarcophaga dux* (Diptera: Sarcophagidae) based on COII gene in China. Tropical Biomedicine 27: 451–460. http://www.ncbi.nlm.nih.gov/pubmed/21399586

Guo Y, Cai J, Chang Y, Li X, Liu Q, Wang X, Wang X, Zhong M, Wen J, Wang J (2011) Identification of forensically important sarcophagid flies (Diptera: Sarcophagidae) in China, based on COI and 16S rDNA gene sequences. Journal of Forensic Sciences 56: 1534–1540. doi: 10.1111/j.1556-4029.2011.01882.x

Harvey ML, Gaudieri S, Villet MH, Dadour IR (2008) A global study of forensically significant calliphorids: implications for identification. Forensic Science International 177: 66–76. doi: 10.1016/j.forsciint.2007.10.009

Harvey ML, Dadour IR, Gaudieri S (2003) Mitochondrial DNA cytochrome oxidase I gene: potential for distinction between immature stages of some forensically important fly species (Diptera) in western Australia. Forensic Science International 131: 134–139. doi: 10.1016/S0379-0738(02)00431-0

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Hebert PDN, (2010) iBOL data & resource sharing policies, The International Barcode of Life Project (iBOL). http://ibol.org/resources/data-release-policy/ [accessed on 10 Oct 2013]

Jordaens K, Sonet G, Richet R, Dupont E, Braet Y, Desmyter S (2012) Identification of forensically important *Sarcophaga* species (Diptera: Sarcophagidae) using the mitochondrial COI gene. International Journal of Legal Medicine 127: 491–504. doi: 10.1007/s00414-012-0767-6

Joseph I, Mathew DG, Sathyan P, Vargheese G (2011) The use of insects in forensic investigations: An overview on the scope of forensic entomology. Journal of Forensic Dental Sciences 3: 89–91. doi: 10.4103/0975-1475.92154

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120. doi: 10.1007/BF01731581

Kutty SN, Pape T, Pont A, Wiegmann BM, Meier R (2008) The Muscoidea (Diptera: Calyptratae) are paraphyletic: Evidence from four mitochondrial and four nuclear genes. Molecular Phylogenetics and Evolution 49: 639–652. doi: 10.1016/j.ympev.2008.08.012

Lefébure T, Douady CJ, Gouy M, Gibert J (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: Proposal of a molecular threshold to help species delimitation. Molecular Phylogenetics and Evolution 40: 435–447. doi: 10.1016/j.ympev.2006.03.014

Li X, Cai JF, Guo YD, Wu KL, Wang JF, Liu QL, Wang XH, Chang YF, Yang L, Lan LM, Zhong M, Wang X, Song C, Liu Y, Li JB, Dai ZH (2010) The availability of 16S rRNA for the identification of forensically important flies (Diptera: Muscidae) in China. Tropical Biomedicine 27: 155–166. http://www.msptm.org/files/155_-_166_Li_X.pdf

Linacre A, Gusmão L, Hecht W, Hellmann AP, Mayr WR, Parson W, Prinz M, Schneider PM, Morling N (2011) ISFG: Recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Science International, Genetics 5: 501–505. doi: 10.1016/j.fsigen.2010.10.017

Liu Q, Cai J, Guo Y, Wang X, Gu Y, Wen J, Meng F, Yi W (2011) Identification of forensically significant calliphorids based on mitochondrial DNA cytochrome oxidase I (COI) gene in China. Forensic Science International 207: e64–5. doi: 10.1016/j.forsciint.2011.02.004

Malgorn Y, Coquoz R (1999) DNA typing for identification of some species of Calliphoridae. An interest in forensic entomology. Forensic Science International 102: 111–119. doi: 10.1016/S0379-0738(99)00039-0

Marchenko MI (2001) Medicolegal relevance of cadaver entomofauna for the determination of the time of death. Forensic Science International 120: 89–109. doi: 10.1016/S0379-0738(01)00416-9

Marinho MAT, Junqueira ACM, Paulo DF, Esposito MC, Villet MH, Azeredo-Espin AML (2012) Molecular phylogenetics of Oestroidea (Diptera: Calyptratae) with emphasis on Calliphoridae: insights into the inter-familial relationships and additional evidence for paraphyly among blowflies. Molecular Phylogenetics and Evolution 65: 840–854. doi: 10.1016/j.ympev.2012.08.007

Mazzanti M, Alessandrini F, Tagliabracci A, Wells JD, Campobasso CP (2010) DNA degradation and genetic analysis of empty puparia: genetic identification limits in forensic entomology. Forensic Science International 195: 99–102. doi: 10.1016/j.forsciint.2009.11.022

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Meiklejohn KA, Wallman JF, Dowton M (2011) DNA-based identification of forensically important Australian Sarcophagidae (Diptera). International Journal of Legal Medicine 125: 27–32. doi: 10.1080/10635150600969864

Meiklejohn KA, Wallman JF, Dowton M (2013) DNA barcoding identifies all immature life stages of a forensically important flesh fly (Diptera: Sarcophagidae). Journal of Forensic Sciences 58: 184–187. doi: 10.1111/j.1556-4029.2012.02220.x

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008) Database indexing for production MegaBLAST searches. Bioinformatics 24: 1757–1764. doi: 10.1093/bioinformatics/btn322

Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419–420. doi: 10.1093/bioinformatics/btp696

Park SH, Park CH, Zhang Y, Piao H, Chung U, Kim SY, Ko KS, Yi C-H, Jo T-H, Hwang J-J (2013) Using the developmental gene bicoid to identify species of forensically important blowflies (Diptera: Calliphoridae). BioMed Research International 2013: 538051. doi: 10.1155/2013/538051

Pereira F, Carneiro J, van Asch B (2010) A Guide for Mitochondrial DNA Analysis in Non-Human Forensic Investigations. The Open Forensic Science Journal 3: 33–44.

Prado e Castro C, Serrano A, Martins Da Silva P, García MD (2012) Carrion flies of forensic interest: a study of seasonal community composition and succession in Lisbon, Portugal. Medical and Veterinary Entomology 26: 417–431. doi: 10.1111/j.1365-2915.2012.01031.x

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology 21: 1864–1877. doi: 10.1111/j.1365-294X.2011.05239.x

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes 7: 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Renaud AK, Savage J, Adamowicz SJ (2012) DNA barcoding of Northern Nearctic Muscidae (Diptera) reveals high correspondence between morphological and molecular species limits. BMC Ecology 12: 24. doi: 10.1186/1472-6785-12-24

Richards CS, Crous KL, Villet MH (2009) Models of development for blowfly sister species *Chrysomya chloropyga* and *Chrysomya putoria*. Medical and Veterinary Entomology 23: 56–61. doi: 10.1111/j.1365-2915.2008.00767.x

Rodriguez WC, Bass WM (1983) Insect activity and its relationship to decay rates of human cadavers in east Tennessee. Journal of Forensic Sciences 28: 423–432. doi: 10.1520/JFS11524J

Rolo EA, Oliveira AR, Dourado CG, Farinha A, Rebelo MT, Dias D (2013) Identification of sarcosaprophagous Diptera species through DNA barcoding in wildlife forensics. Forensic Science International 228: 160–164. doi: 10.1016/j.forsciint.2013.02.038

Rozkošný R, Gregor F, Adrian CP (1997) The European Fanniidae (Diptera). Acta Scientiarum Naturalium Academiae Scientiarum Bohemicae Brno, Nova Series 31: 1–80.

Schindel DE, Stoeckle MY, Milensky C, Trizna M, Schmidt B, Gebhard C, Graves G (2011) Project description: DNA barcodes of bird species in the national museum of natural history, Smithsonian Institution, USA. ZooKeys 152: 87–92. doi: 10.3897/zookeys.152.2473

Sonet G, Jordaens K, Braet Y, Desmyter S (2012) Why is the molecular identification of the forensically important blowfly species *Lucilia caesar* and *L. illustris* (family Calliphoridae) so problematic? Forensic Science International 223: 153–159. doi: 10.1016/j.forsciint.2012.08.020

Sperling FA, Anderson GS, Hickey DA (1994) A DNA-based approach to the identification of insect species used for postmortem interval estimation. Journal of Forensic Sciences 39: 418–427.

Stevens JR, Wall R, Wells JD (2002) Paraphyly in Hawaiian hybrid blowfly populations and the evolutionary history of anthropophilic species. Insect Molecular Biology 11: 141–148. doi: 10.1046/j.1365-2583.2002.00318.x

Szpila K (2012) Key for identification of European and Mediterranean blowflies (Diptera, Calliphoridae) of medical and veterinary importance – adult flies. In: Gennard DE (Ed) Forensic entomology, an introduction. Wiley-Blackwell, Chichester, 77–81.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673–4680. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308517/

Vanin S, Tasinato P, Ducolin G, Terranova C, Zancaner S, Montisci M, Ferrara SD, Turchetto M (2008) Use of *Lucilia* species for forensic investigations in Southern Europe. Forensic Science International 177: 37–41. doi: 10.1016/j.forsciint.2007.10.006

Vincent S, Vian JM, Carlotti MP (2000) Partial sequencing of the cytochrome oxydase b subunit gene I: a tool for the identification of European species of blow flies for postmortem interval estimation. Journal of Forensic Sciences 45: 820–823. http://www.ncbi.nlm.nih.gov/pubmed/10914577

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M (2012) Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7: e31581. doi: 10.1371/journal.pone.0031581

Virgilio M, Backeljau T, Nevado B, De Meyer M (2010) Comparative performance of DNA barcoding across insect orders. BMC Bioinformatics 11: 206. doi: 10.1186/1471-2105-11-206

Velásquez Y, Ivorra T, Grzywacza A, Martínez-Sánchez A, Magaña C, García-Rojo A, Rojo S (2013) Larval morphology, development and forensic importance of *Synthesiomyia nudiseta* (Diptera: Muscidae) in Europe: a rare species or just overlooked? Bulletin of Entomological Research 103: 98–110. doi: 10.1017/S0007485312000491

Wallman JF, Donnellan SC (2001) The utility of mitochondrial DNA sequences for the identification of forensically important blowflies (Diptera: Calliphoridae) in southeastern Australia. Forensic Science International 120: 60–67. http://www.fsijournal.org/article/S0379-0738(01)00426-1/abstract

Wang X, Cai J, Guo Y, Chang Y, Wu K, Liu Q, Wang J, Yang L, Lan L, Zhong M, Wang X, Cheng YS, Liu Y, Chen Y, Li J, Zhang J, Peng X (2010) The availability of 16SrDNA

gene for identifying forensically important blowflies in China. Romanian Journal of Legal Medicine 18: 43–50. doi: 10.4323/rjlm.2010.43

Wells JD, Pape T, Sperling F (2001) DNA-based identification and molecular systematics of forensically important Sarcophagidae (Diptera). Journal of Forensic Sciences 46: 1098–1102. http://www.ncbi.nlm.nih.gov/pubmed/11569549

Wells JD, Sperling FA (1999) Molecular phylogeny of *Chrysomya albiceps* and *C. rufifacies* (Diptera: Calliphoridae). Journal of Medical Entomology 36: 222–226. http://www.ncbi.nlm.nih.gov/pubmed/10337087

Wells JD, Sperling FA (2000) A DNA-based approach to the identification of insect species used for postmortem interval estimation and partial sequencing of the cytochrome oxydase *b* subunit gene I: a tool for the identification of European species of blow flies for postmortem interval. Journal of Forensic Sciences 45: 1358–1359. http://www.ncbi.nlm.nih.gov/pubmed/11110202

Wells JD, Sperling FAH (2001) DNA-based identification of forensically important Chrysomyinae (Diptera : Calliphoridae). Forensic Science International 120: 110–115. doi: 10.1016/S0379-0738(01)00414-5

Wells JD, Stevens JR (2008) Application of DNA-based methods in forensic entomology. Annual Review of Entomology 53: 103–120. doi: 10.1146/annurev.ento.52.110405.091423

Wells JD, Wall R, Stevens JR (2007) Phylogenetic analysis of forensically important *Lucilia* flies based on cytochrome oxidase I sequence: a cautionary tale for forensic species determination. International Journal of Legal Medicine 121: 229–233. doi: 10.1007/s00414-006-0147-1

Wells JD, Williams DW (2007) Validation of a DNA-based method for identifying Chrysomyinae (Diptera: Calliphoridae) used in a death investigation. International Journal of Legal Medicine 121: 1–8. doi: 10.1007/s00414-005-0056-8

Whitworth TL, Dawson RD, Magalon H, Baudry E (2007) DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). Proceedings of the Royal Society B 274: 1731–1739. doi: 10.1098/rspb.2007.0062

Wilson JJ, Rougerie R, Schonfeld J, Janzen DH, Hallwachs W, Hajibabaei M, Kitching IJ, Haxaire J, Hebert PDN (2011) When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. BMC Ecology 11: 18. doi: 10.1186/1472-6785-11-18

Zaidi F, Wei S, Shi M, Chen X (2011) Utility of multi-gene loci for forensic species diagnosis of blowflies. Journal of Insect Science 11: 59. doi: 10.1673/031.011.5901

Zehner R, Amendt J, Schütt S, Sauer J, Krettek R, Povolný D (2004) Genetic identification of forensically important flesh flies (Diptera: Sarcophagidae). International Journal of Legal Medicine 118: 245–247. doi: 10.1007/s00414-004-0445-4

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 7: 203–214. doi: 10.1089/10665270050081478

ZooKeys 365: 329–336 (2013)
doi: 10.3897/zookeys.365.6034
www.zookeys.org

SHORT COMMUNICATION

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Adhoc: an R package to calculate *ad hoc* distance thresholds for DNA barcoding identification

Gontran Sonet[1], Kurt Jordaens[2,3], Zoltán T. Nagy[1], Floris C. Breman[2], Marc De Meyer[2], Thierry Backeljau[1,3], Massimiliano Virgilio[2]

1 *Royal Belgian Institute of Natural Sciences, OD Taxonomy and Phylogeny (JEMU), Vautierstraat 29, 1000 Brussels, Belgium* **2** *Royal Museum for Central Africa, Department of Biology (JEMU), Leuvensesteenweg 13, 3080 Tervuren, Belgium* **3** *University of Antwerp, Evolutionary Ecology Group, Groenenborgerlaan 171, 2020 Antwerp, Belgium*

Corresponding author: *Gontran Sonet* (gontran.sonet@naturalsciences.be)

Academic editor: *L. Penev* | Received 2 August 2013 | Accepted 2 December 2013 | Published 30 December 2013

Citation: Sonet G, Jordaens K, Nagy ZT, Breman FC, De Meyer M, Backeljau T, Virgilio M (2013) *Adhoc*: an R package to calculate *ad hoc* distance thresholds for DNA barcoding identification. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 329–336. doi: 10.3897/zookeys.365.6034

Abstract

Identification by DNA barcoding is more likely to be erroneous when it is based on a large distance between the query (the barcode sequence of the specimen to identify) and its best match in a reference barcode library. The number of such false positive identifications can be decreased by setting a distance threshold above which identification has to be rejected. To this end, we proposed recently to use an *ad hoc* distance threshold producing identifications with an estimated relative error probability that can be fixed by the user (e.g. 5%). Here we introduce two R functions that automate the calculation of *ad hoc* distance thresholds for reference libraries of DNA barcodes. The scripts of both functions, a user manual and an example file are available on the JEMU website (http://jemu.myspecies.info/computer-programs) as well as on the comprehensive R archive network (CRAN, http://cran.r-project.org).

Keywords

Species identification, accuracy, precision, relative error, reference library, COI

Introduction

The DNA barcoding initiative aims at providing a simple and standardised tool for specimen identification using a short DNA sequence from a specific region of the genome as a barcode (Hebert et al. 2003). The identification of a specimen using DNA barcoding is based on the comparison between its DNA barcode sequence (= query) and a reference library of DNA barcodes. These reference sequences satisfied a series of requirements that allow quality control (link to voucher specimen, trace files, and association with additional information such as primer and collection data). Among the approaches available for the assignment of a species name (Frézal and Leblois 2008, Austerlitz et al. 2009), methods based on sequence similarity are fast, easy and frequently applied as a first step to screen large reference libraries (Frézal and Leblois 2008). In this method, the species name of the reference sequence(s) showing the smallest genetic distance with the query (i.e. best match *sensu* Meier et al. 2006) is used for the identification (Ratnasingham and Hebert 2007). The identification provided by the best match method can be considered as true positive (TP) if a correct species name is assigned to the query or as false positive (FP) if an incorrect species name is assigned to the query (Figure 1). Yet, for many taxonomic groups, reference libraries are still incompletely representing the genetic diversity that can be found on specific and population levels. Some queries are therefore not represented by a conspecific DNA barcode in the library and will be erroneously identified according to the most similar allospecific reference barcode. Yet, the number of this sort of false positive identifications can be greatly reduced by assigning species names only when the distance between the query and its best DNA barcode match is below an arbitrary distance threshold value. With this best close match method (*sensu* Meier et al. 2006), identifications can still be TP or FP when the genetic distance between the query and its best match(es) is below the threshold. When this genetic distance is above the threshold (Figure 1), then either incorrect species name assignments can be correctly ignored (true negatives, TN) or correct species name assignments can be erroneously ignored (false negatives, FN). The determination of this distance threshold can be arbitrary (Ratnasingham and Hebert 2007) or can be based on the expected separation between intra- and interspecific distances (Meyer and Paulay 2005, Lefébure et al. 2006, Puillandre et al. 2012).

Recently, we proposed a general working strategy to deal with incomplete reference libraries of DNA barcodes (Virgilio et al. 2012). This method is based on *ad hoc* distance thresholds that are calculated for each library considering the estimated probability of relative identification errors. Indeed, by using each sequence of a reference library as a query against all other reference sequences, we can calculate (Virgilio et al. 2012, Figure 1) the relative identification error (RE) of the best close match method as FP/(TP+FP), its overall identification error (OE) as (FP+FN)/total number of queries, its accuracy as (TP+TN)/total number of sequences and its precision as TP/(TP+FP). The general procedure consists of 1) calculating the RE in a library of DNA barcodes for a number of arbitrarily chosen distance thresholds, 2) modelling the relation between

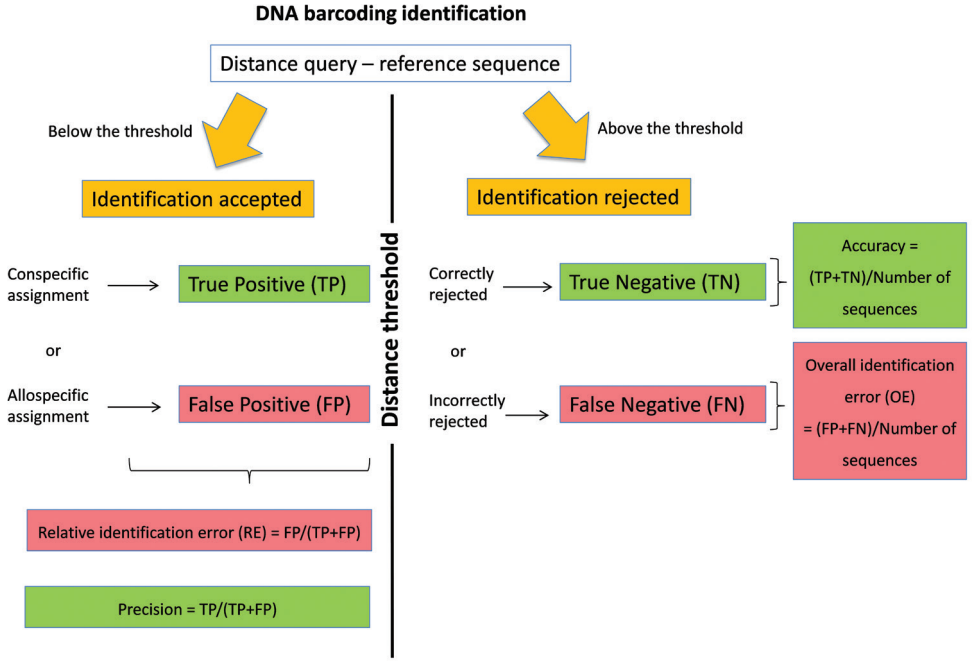


Figure 1. DNA barcoding identification using the best close match method.

distance thresholds and RE and 3) estimating the *ad hoc* threshold that would yield an estimated RE (e.g. 5%) for that particular library (Virgilio et al. 2012).

Here we introduce the R package "adhoc" including two functions, checkDNAbcd ("check DNA barcode") and adhocTHR ("*ad hoc* threshold"), which automate this procedure and calculate the *ad hoc* distance threshold.

Description of both functions

Both functions rely on the packages ape (Paradis et al. 2004), pegas (Paradis 2010) and spider (Brown et al. 2012). The first function, checkDNAbcd, imports a reference library of aligned DNA barcodes in FASTA format and provides basic descriptive statistics of the imported dataset, allowing a first quality check of the library. This function produces two tables containing species names, full sequence identifiers (as read by the function from the input file), and numbers of sequences and haplotypes for each species. CheckDNAbcd also returns the length of each reference sequence, calculates all pairwise distances and separates intra- and interspecific pairwise comparisons. The calculation of pairwise distances can be on the basis of simple uncorrected p-distances (representing the proportion of sites at which two sequences differ) or of several nucleotide substitution models such as the Kimura 2-parameter model (Kimura 1980), which is standardly used in DNA barcoding (Ratnasingham and Hebert 2007).

The second function, adhocTHR, utilises the output of the first function and performs best match and best close match identifications by taking each sequence of the reference library as a query against all other sequences of the library (Virgilio et al. 2012). For the best match identification, each query is identified as TP, FP or ambiguous false positive (FPambiguous, when both correct and incorrect species names are found as best matches). For the best close match identification, adhocTHR automatically evaluates each identification as TP, FP, FPambiguous, TN or FN and calculates the RE, OE, accuracy and precision at 30 arbitrary distance thresholds (equally distributed between zero and the largest distance observed between all pairs of query – best match). Relationships between distance thresholds and RE are then modelled through regression fitting. Regression is used to calculate the *ad hoc* distance threshold (Virgilio et al. 2012) producing an expected RE (5% by default). The function adhocTHR also produces a list of red-flagged matches (conspecific and allospecific matches responsible for the ambiguous identifications) and a table of red-flagged species names (species involved in the ambiguous identifications). The user has the possibility of modifying (1) the regression fitting (linear by default, or polynomial), (2) the number of arbitrary distance thresholds used for the fitting, (3) the estimated RE probability and (4) the treatment of ambiguous identifications. By default, the function treats ambiguous identifications as incorrect but they can optionally be ignored in the calculation or considered as correct. We recommend using this last option with caution since it will treat all red-flagged species involved in the same ambiguous identification as a single species.

As an indication, five minutes were necessary for each function to process a dataset of 5000 records (600-650 bp) on a personal computer (processor Intel Core i5 CPU M540, 2.53 GHz, 4 GB RAM with Windows 7 as operating system) using default parameters. Calculating the RE for more than 30 arbitrary distance thresholds is suggested to improve the fitting when computing time is not an issue.

When using reference libraries with particularly low levels of taxon coverage (Virgilio et al. 2010), reaching an estimated RE of 5% might not be possible, even at the most restrictive distance threshold (*viz.* distance threshold = 0.00) where only identical sequences are used for identification, all the other ones are discarded. In those cases the script will provide a warning message to inform the user that the script cannot find an *ad hoc* distance threshold for the chosen error probability.

This method has been developed for specimen identification. It is intended to optimise the identification success rate by adapting the distance threshold according to a RE estimated from a particular reference library. Hence, using this method for species delimitation requires a careful interpretation of the output (Collins and Cruickshank 2013). The estimation of the RE in DNA barcoding is an indispensable prerequisite, not only for forensic applications (Wells and Stevens 2008), but also for any further research relying on DNA barcoding identifications such as ecology or biodiversity inventories (Frézal and Leblois 2008).

The script of both functions, a user manual and an example file are available on the JEMU website (http://jemu.myspecies.info/computer-programs) and on the comprehensive R archive network (CRAN, http://cran.r-project.org). The user manual

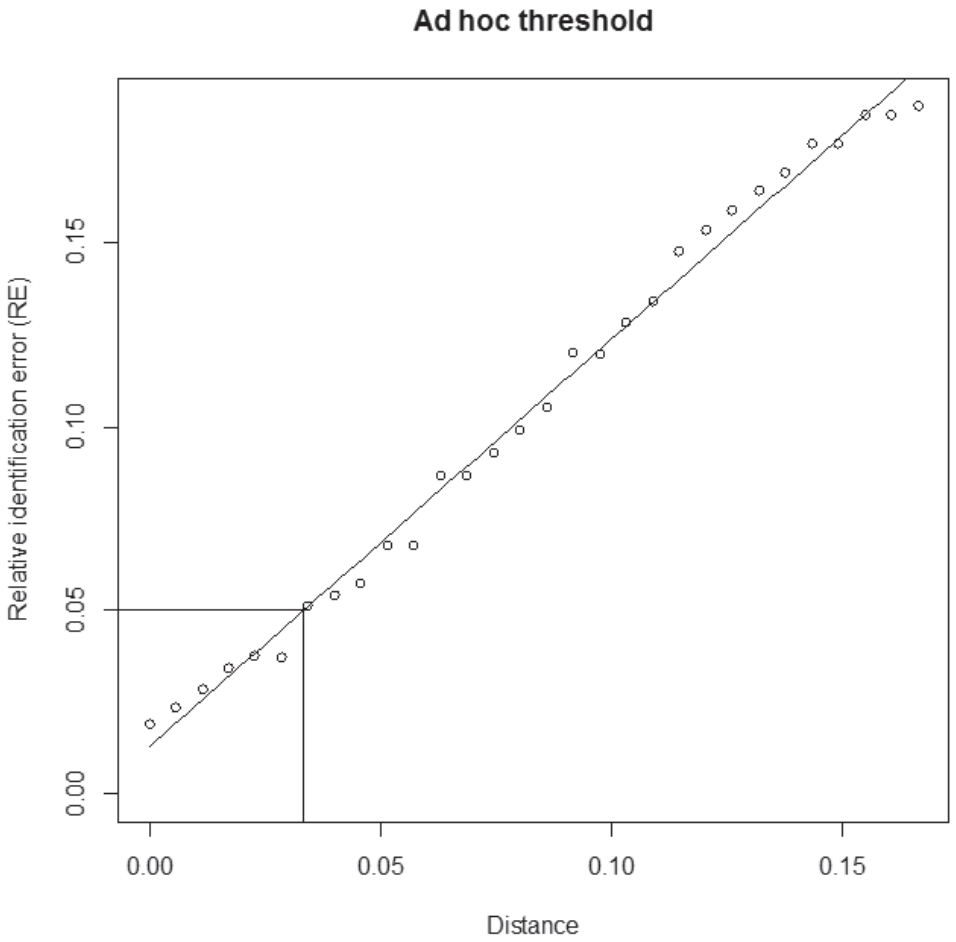


Figure 2. Estimation of the *ad hoc* distance threshold. Example of output obtained using the function adhocTHR with default settings (30 arbitrary distance thresholds, linear fit and an estimated relative identification error (RE) of 5%). The following message was given by the function: "for a RE of 0.05 use a threshold of 0.0334".

suggests a few R commands to plot (1) the distribution of sequence lengths, (2) the distribution of intra- and interspecific pairwise distances and (3) a graph representing the RE obtained with the different arbitrary distance thresholds, the linear or polynomial fitting and the distance value corresponding to the *ad hoc* threshold (Figure 2).

Acknowledgements

The Joint Experimental Molecular Unit (JEMU) is financed by the Belgian Federal Science Policy Office (BELSPO). The authors would like to thank Céline Poux, Bruno

Nevado and the R community for their help on the use of R, and Fabrice Clin and Grégory Canivet for help with encoding. The authors also thank the reviewers of the manuscript for their constructive suggestions.

References

Austerlitz F, David O, Schaeffer B, Bleakley K, Olteanu M, Leblois R, Veuille M, Laredo C (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10: S10. doi: 10.1186/1471-2105-10-S14-S10

Brown SDJ, Collins RA, Boyer S, Lefort M-C, Malumbres-Olarte J, Vink CJ, Cruickshank RH (2012) Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources 12: 562–565. doi: 10.1111/j.1755-0998.2011.03108.x

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. Molecular Ecology Resources 13: 969–975. doi: 10.1111/1755-0998.12046

Frézal L, Leblois R (2008) Four years of DNA barcoding: Current advances and prospects. Infection, Genetics and Evolution 8: 727–736. doi: 10.1016/j.meegid.2008.05.005

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–321. doi: 10.1098/rspb.2002.2218

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120. doi: 10.1007/BF01731581

Lefébure T, Douady CJ, Gouy M, Gibert J (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: Proposal of a molecular threshold to help species delimitation. Molecular Phylogentics and Evolution 40: 435–447. doi: 10.1016/j.ympev.2006.03.014

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–28. doi: 10.1080/10635150600969864

Meyer CP, Paulay G (2005) DNA Barcoding: error rates based on comprehensive sampling. PLoS Biology 3: e422. doi: 10.1371/journal.pbio.0030422

Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419–20. doi: 10.1093/bioinformatics/btp696

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289–290. doi: 10.1093/bioinformatics/btg412

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology 21: 1864–77. doi: 10.1111/j.1365-294X.2011.05239.x

Ratnasingham S, Hebert PDN (2007) BOLD : The Barcode of Life Data System. Molecular Ecology Notes 7: 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Virgilio M, Backeljau T, Nevado B, De Meyer M (2010) Comparative performances of DNA barcoding across insect orders. BMC Bioinformatics 11: 206. doi: 10.1186/1471-2105-11-206

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M (2012) Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. PLoS ONE 7: e31581. doi: 10.1371/journal.pone.0031581

Wells JD, Stevens JR (2008) Application of DNA-based methods in forensic entomology. Annual Review of Entomology 53: 103–120. doi: 10.1146/annurev.ento.52.110405.091423

ZooKeys 365: 337–354 (2013)
doi: 10.3897/zookeys.365.5356
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Revisiting species delimitation within the genus *Oxystele* using DNA barcoding approach

Herman Van Der Bank[1], Dai Herbert[2], Richard Greenfield[1], Kowiyou Yessoufou[3]

1 *Department of Zoology, African Centre for DNA Barcoding (ACDB), Kingsway Campus, University of Johannesburg, PO Box 524, Auckland Park 2006, South Africa* **2** *KwaZulu-Natal Museum, P. Bag 9070, Pietermaritzburg 3200, South Africa, and School of Life Sciences, University of KwaZulu-Natal, Pietermaritzburg, 3206 South Africa* **3** *Department of Botany and Plant Biotechnology, African Centre for DNA Barcoding (ACDB), Kingsway Campus, University of Johannesburg, PO Box 524, Auckland Park 2006, South Africa*

Corresponding author: *Herman Van Der Bank* (hvdbank@uj.ac.za)

Academic editor: *K. Jordaens* | Received 19 April 2013 | Accepted 13 August 2013 | Published 30 December 2013

Citation: Van Der Bank HF, Herbert D, Greenfield R, Yessoufou K (2013) Revisiting species delimitation within the genus *Oxystele* using DNA barcoding approach. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 337–354. doi: 10.3897/zookeys.365.5356

Abstract

The genus *Oxystele*, a member of the highly diverse marine gastropod superfamily Trochoidea, is endemic to southern Africa. Members of the genus include some of the most abundant molluscs on southern African shores and are important components of littoral biodiversity in rocky intertidal habitats. Species delimitation within the genus is still controversial, especially regarding the complex *O. impervia* / *O. variegata*. Here, we assessed species boundaries within the genus using DNA barcoding and phylogenetic tree reconstruction. We analysed 56 specimens using the mitochondrial gene COI. Our analysis delimits five molecular operational taxonomic units (MOTUs), and distinguishes *O. impervia* from *O. variegata*. However, we reveal important discrepancies between MOTUs and morphology-based species identification and discuss alternative hypotheses that can account for this. Finally, we indicate the need for future study that includes additional genes, and the combination of both morphology and genetic techniques (e.g. AFLP or microsatellites) to get deeper insight into species delimitation within the genus.

Keywords

Mollusca, Gastropoda, Trochidae, species delimitation, morphology

Introduction

Molluscs comprise one of the largest marine phyla, comprising more than 50,000 described species (marine species only), of which less than 10% are currently included in the global database of DNA barcodes (Radulovici et al. 2010). DNA barcoding is a genetic technique designed to standardize and accelerate species identification as an instrument facilitating conservation efforts, ecosystem monitoring, and the identification of phylogeographic and speciation patterns (Radulovici et al. 2010; but see Taylor and Harris 2012 for criticism). It has also proved valuable in population genetics and phylogenetic analyses, identification of prey in gut contents, forensic and seafood safety, invasion biology (Armstrong and Ball 2005, Bucklin et al. 2011) and in revealing cryptic species (Hebert et al. 2004, Puillandre et al. 2009, Lakra et al. 2011). One of the important uses of DNA barcoding is its ability to correctly assign several life-forms including larvae, carcass fragments and damaged specimens to species (Ward et al. 2005, Yang et al. 2012).

Although the mitochondrial cytochrome *c* oxidase subunit I gene (COI), used for barcoding purposes of animals is not efficient for all taxonomic groups (e.g. terrestrial gastropods, Davison et al. 2009; anthozoans, Huang et al. 2008), and pending the integration of the next generation sequencing into the DNA barcoding technique (Taylor and Harris 2012), the barcoding approach has proved valuable in discriminating marine biodiversity (e.g. Sun et al. 2012; see also reviews in Radulovici et al. 2010). *Oxystele* Philippi, 1847, a genus of the highly diverse marine gastropod superfamily Trochoidea (Williams et al. 2010), is endemic to southern Africa. Currently, five species are recognised (Branch et al. 2010), but delimitation within the genus is still debated (Heller and Dempster 1991, Williams et al. 2010), especially due to strong homoplasy in morphological characters traditionally used in identification keys (Hickman 1998).

In this study, our main objective was to infer species boundaries within the genus using DNA barcode. To date, attempts to resolve taxonomic issues within the genus using DNA sequence data were very limited in sample size: only one individual of each of the five recognised *Oxystele* species was generally analysed. For this purpose, we sampled 56 specimens including all five *Oxystele* species from a wide geographic distribution range. We then applied the DNA barcoding approach for taxa delimitation.

Materials and methods

Sample collections

Sampling sites were widely distributed to cover the geographical distribution range of the genus. Species identification was done using the morphological characters given in the key to *Oxystele* species provided by Heller and Dempster (1991). Collection details including GPS coordinates, altitude and photographs of specimens are available online

in the Barcode of Life Data Systems (BOLD; www.boldsystems.org) together with DNA sequences. Voucher specimens (shells) were also collected and deposited at the KwaZulu-Natal Museum (South Africa).

DNA extraction, amplification and sequencing of DNA barcodes

DNA extraction, polymerase chain reactions (PCR) and sequencing of the COI region (animal DNA barcode) were done at the Canadian Centre for DNA Barcoding (CCDB). PCR reactions followed standard CCDB protocols as described by Hajibabaei et al. (2005). This results in 51 COI DNA sequences being generated. We also included in the DNA matrix five COI sequences that we retrieved from BOLD (DQ numbers in Table 1), making the total sequences analysed to a total of 56 COI sequences. Sequence alignment was performed using Multiple Sequence Comparison by Log-Expectation (MUSCLE vs. 3.8.31, Edgar 2004). GenBank accession numbers, BOLD process identification numbers and voucher information are all available online (www.boldsystems.org). These numbers, together with authorities for the species studied are listed in Table 1.

Data analysis

We assessed the "DNA barcode gap" (Meyer and Paulay 2005) in the dataset using two approaches. First, we compared the median of interspecific distances with that of intraspecific distances (genetic distances are calculated between morphospecies). Significance of the differences between both distances was assessed using the non-parametric Wilcoxon ranked sum test. Second, we used Meier et al.'s (2008) approach, that is, we compared the smallest interspecific distance with the largest intraspecific distance. Genetic distances were measured using the Kimura 2-parameter (K2P) model (Kimura 1980). We are aware of the recent literature indicating that the K2P-model might not be the best model for DNA barcoding. However, we used this model here to allow comparison of our results with other DNA barcoding studies where K2P-model is the most frequently used model.

We also tested the discriminatory power of DNA barcoding by evaluating the proportion of correct species identification using the COI region. All sequences were labeled according to the names of the species from which the sequences were generated. The test of discriminatory power works as follows. Each sequence is considered as an unknown while the remaining sequences in the dataset are considered as the DNA barcode database used for identification. If the identification of the query is the same as the pre-considered identification (i.e. the sequence labels), the identification test is scored as "correct", and the overall proportion of correct identification corresponds to the discriminatory power of the region tested, i.e. COI. This test was done applying three approaches: the "best close match" (Meier et al. 2006), the "near neighbour" and the BOLD criteria using respectively the functions bestCloseMatch, threshID,

Table 1. Species, authority, GenBank accession numbers (*DQ*) and BOLD process ID numbers (HVDBM) of specimens studied. Specimens in bold are those for which morphological characters (weathered shell colours and patterns) failed to provide accurate identification; this is revealed in the barcoding test of species delimitation and in phylogenetic tree topology. Sample localities for *O. impervia* and *O. variegata* individuals are indicated: southern Cape[1], Robben Island[2], north-western Cape[3], Namibia[4]

Species (authority):	GenBank and process ID numbers of specimens included in this study	Composition of MOTUs based on the barcoding test of species delimitation
Oxystele sinensis (Gmelin, 1791)	DQ061089, HVDBM056-10, HVDBM083-10, HVDBM084-10, HVDBM085-10, HVDBM086-10, HVDBM087-10, HVDBM409-11, HVDBM410-11, HVDBM411-11, HVDBM412-11, HVDBM437-11	DQ061089, HVDBM056-10, HVDBM083-10, HVDBM084-10, HVDBM085-10, HVDBM086-10, HVDBM087-10, HVDBM409-11, HVDBM410-11, HVDBM411-11, HVDBM412-11, HVDBM437-11
Oxystele tabularis (Krauss, 1848)	DQ061090, HVDBM289-11, HVDBM338-11, HVDBM339-11	DQ061090, HVDBM289-11, HVDBM338-11, HVDBM339-11
Oxystele tigrina (Anton, 1838)	DQ061091, HVDBM005-10, HVDBM006-10, HVDBM013-10, HVDBM055-10, HVDBM394-11, HVDBM506-11, HVDBM507-11, HVDBM508-11, HVDBM509-11, HVDBM510-11	DQ061091, HVDBM005-10, HVDBM006-10, HVDBM013-10, HVDBM055-10, HVDBM394-11, HVDBM506-11, HVDBM507-11, HVDBM508-11, HVDBM509-11, HVDBM510-11
Oxystele variegata (Anton, 1838)	DQ061092[1], HVDBM058-10[1], HVDBM059-10[1], HVDBM070-10[1], HVDBM072-10[1], HVDBM183-10[3], HVDBM184-10[3], HVDBM185-10[3], HVDBM208-10[4], HVDBM209-10[4], HVDBM389-11[3], HVDBM393-11[1], HVDBM395-11[1], HVDBM456-11[4], HVDBM457-11[4], HVDBM511-11[2], HVDBM512-11[2], HVDBM513-11[2], HVDBM514-11[2], HVDBM515-11[2]	HVDBM072-10[1], HVDBM183-10[3], HVDBM184-10[3], HVDBM185-10[3], HVDBM208-10[4], HVDBM209-10[4], HVDBM389-11[3], HVDBM393-11[1], HVDBM395-11[1], HVDBM456-11[4], HVDBM457-11[4], HVDBM511-11[2], HVDBM512-11[2], HVDBM513-11[2], HVDBM514-11[2], HVDBM515-11[2], **HVDBM028-10[1]**
Oxystele impervia (Menke, 1843)	DQ061093[1], HVDBM022-10[1], HVDBM027-10[1], HVDBM028-10[1], HVDBM057-10[1], HVDBM071-10[1], HVDBM178-10[3], HVDBM179-10[3], HVDBM180-10[3]	**DQ061093**[1], HVDBM022-10[1], HVDBM027-10[1], HVDBM057-10[1], HVDBM071-10[1], HVDBM178-10[3], HVDBM179-10[3], HVDBM180-10[3], **DQ061092**[1], **HVDBM058-10[1]**, **HVDBM059-10[1]**, **HVDBM070-10[1]**

and nearNeighbour implemented in the program Spider 1.1-1 (Brown et al. 2012). Prior to the test, we determined the optimised genetic distance suitable as threshold for taxon identification. For this purpose, we used the function localMinima also implemented in Spider (Brown et al. 2012).

The function bestCloseMatch conducts the "best close match" analysis of Meier et al. (2006), searching for the closest individual in the dataset. If the closest individual is within a given threshold, the outcome is scored as "correct". If it is further than the given threshold, the result is "no ID" (no identification). If more than one species are tied for closest match, the outcome of the test is "ambiguous" identification. When all matches within the threshold are different species to the query, the result is scored as "incorrect".

The function threshID conducts a threshold-based analysis based on a threshold genetic distance of 1% as conducted by the "Identify Specimen" tool provided by the BOLD system (http://www.boldsystems.org/views/idrequest.php). It is more inclusive than bestCloseMatch, in that it considers all sequences within the threshold of 1%. There also four possible outcomes for threshID tests, that is, "correct", "incorrect", "ambiguous", and "no id" similar to the outcomes of the bestCloseMatch function.

The nearNeighbour function finds the closest individual and returns the score "true" (equivalent to "correct") if their names are the same, but if the names are different, the outcome is scored as "false" (equivalent to "incorrect").

Further, we performed a barcoding test of taxon delimitation. In reality, this test groups specimens into "molecular operational taxonomic units" (MOTUs; Jones et al. 2011), which are generally regarded as proxy for morpho-species (Stahlhut et al. 2013). MOTUs are defined as groups of specimens that are within the genetic threshold used for taxon delimitation. If all specimens of the same morpho-species are clustered in a single MOTU, this means that MOTUs are congruent with morpho-species, thus increasing the taxonomic value of DNA barcoding. The delimitation of MOTUs was conducted using the function tclust in the R package Spider v1.1-1. If two specimens are more distant than the threshold from each other, but both are within the threshold of a third, the function tclust identified all three individuals as a single MOTU. We also identified the composition of each MOTU using the function lapply also implemented in Spider.

Finally we complemented the test of MOTU delimitation with a phylogenetic analysis of COI sequences. We reconstructed a phylogenetic tree using Bayesian and maximum parsimony methods. The Bayesian tree was reconstructed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003). The best-fit model of DNA sequence evolution was chosen using jModelTest v0.1.1 (Posada 2008) under the Akaike information criterion (Posada and Buckley 2004). The TrN + I model was selected and used to generate the Bayesian tree. Analysis was run for nine million generations with sampling one tree every 100 generations. Two independent Bayesian analyses with four differentially heated chains were performed simultaneously. The results were visualised and checked using MEGA, and 25,000 trees were discarded as burn-in to ensure that the analysis had stabilised. Node support was assessed using posterior probability (PP) as follows: PP > 0.95: high support and PP < 0.95: no support (Alfaro and Holder 2006).

Maximum parsimony (MP) was implemented to analyse the data using PAUP* v4.10b10 (Swofford 2002). Tree searches were done using heuristic searches with 1000 random sequence additions but keeping only 10 trees. Tree bisection-reconnection was performed with all character transformations treated as equally likely i.e. Fitch parsimony (Fitch 1971). MP searches and bootstrap resampling (Felsenstein 1985) were done using PAUP* v4.10b10 (Swofford 2002).

Jujubinus exasperatus (Pennant, 1777) was used as outgroup based on Williams et al. (2010). Node support was assessed using bootstrap (BP) values: BP > 70% for strong support (Murphy et al. 2001, Wilcox et al. 2002).

Results

Our dataset includes 56 specimens: nine specimens of *O. impervia,* 12 of *O. sinensis,* four of *O. tabularis*, 11 of *O. tigrina*, and 20 specimens of *O. variegata* (Table 1). The aligned COI matrix was 654 base pairs in length, including A: 24.2%; C: 21.1%; G: 18.3% and T: 36.4%.

Interspecific distances range from 0 to 0.18 (median = 0.15) and are generally larger than intraspecific distances (range: 0-0.09; median = 0.004; Wilcoxon test, $p < 0.001$; Figure 1A). This indicates that there is a barcode gap in the dataset. Even when we compared the lowest interspecific versus the furthest intraspecific distance, we also found that barcode gap exists within the COI sequences (grey lines in Figure 1B).

We determined the optimised threshold genetic distance (d) with which we tested the discriminatory power of COI sequences and delimited MOTUs. We found d = 0.047 (Figure 2). Testing the efficacy of DNA barcoding based on this threshold,

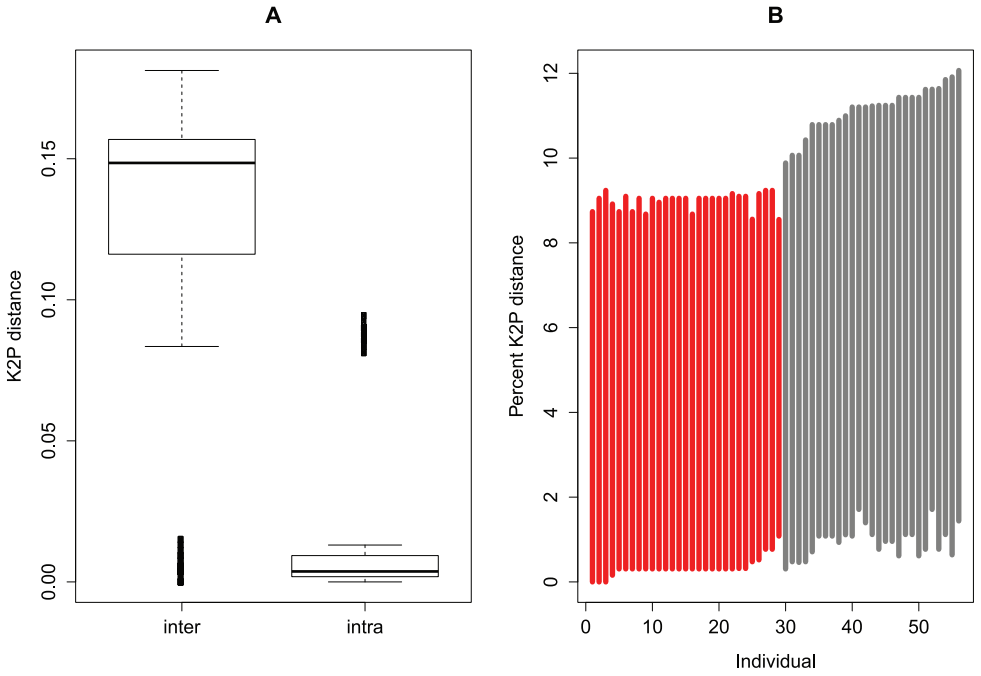


Figure 1. Evaluation of barcode gap in the dataset. **A** Boxplot of the interspecific (inter) and intraspecific genetic (intra) distances, indicating the existence of a barcode gap i.e. intraspecific distance is longer than intraspecific distance. The bottom and top of the boxes show the first and third quartiles respectively, the median is indicated by the horizontal line, the range of the data by the vertical dashed line and outliers (points outside 1.5 times the interquartile range) by Bold vertical lines **B** Lineplot of the barcode gap for the 56 *Oxsystele* specimens. For each specimen in the dataset, the grey lines indicate where the smallest interspecific distance (top of line value) is longer than the longest intraspecific distance (bottom of line value), therefore indicating existence of barcode gap; the red lines show where this pattern is reversed, and the closest non-conspecific is closer to the query than its nearest conspecific, i.e., the situation where there is no barcoding gap.

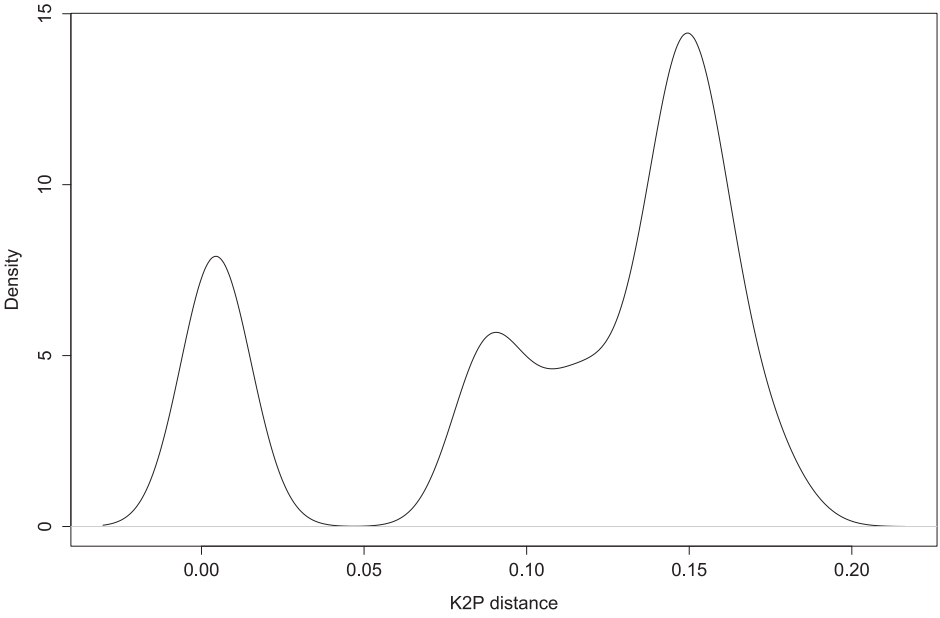


Figure 2. Determination of the threshold genetic distance for species identification. The density plot indicates transition between intra- and interspecific distances; the genetic distance corresponding to this transition (dip in the density graph, here approximately 0.05) indicates the suitable threshold to the dataset. This method does not require prior knowledge of species identity to get an indication of potential threshold values.

Table 2. Tests of barcoding identification accuracy with numbers (n) and percentages (%) of each score.

Methods	**Near neighbour**		**Best Close match**				**BOLD criteria**			
Scores	False	True	Ambiguous	Correct	Incorrect	No ID	Ambiguous	Correct	Incorrect	No ID
n (%)	7 (12.5%)	49 (87.5%)	3 (5.36%)	49 (87.5%)	4 (7.14%)	0 (0%)	27 (48.21%)	27 (48.21%)	1 (1.79%)	1 (1.79%)

we found that COI sequences performed very well in assigning DNA sequences to the correct species (Table 2). For instance, under both near neighbour and best close match methods, 87.5% of the COI sequences were correctly identified (49 specimens out of 56). However, the best close match method indicates 5.36% of ambiguity (three specimens), i.e. both correct and incorrect species are within the given threshold; and 7.14% of incorrect identification (four specimens). Also, for 12.5% of sequences (seven specimens) the near neighbour method results in "incorrect". Using the BOLD method (threshold = 1%), we obtained poor barcoding performance, that is, we have as many correct as ambiguous results (48.21% respectively; i.e. 27 specimens). The BOLD method also indicates one "incorrect" and one "no id" (Table 2).

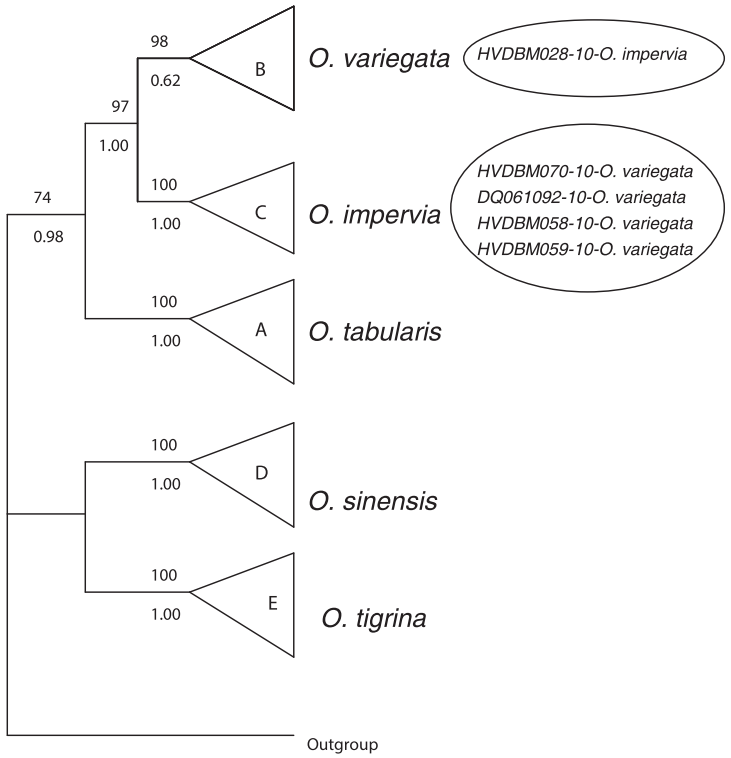


Figure 3. Summary of both Bayesian and parsimonious trees. Values above branches indicate bootstrap supports; values under branches indicate posterior probability. All distinguished species are indicated at the tip of the tree. Branches without values indicate non-supported nodes; the small circle indicates a specimen of *O. impervia* (HVDBM028-10) that was misidentified based on morphology; large circle indicates four specimens morphologically indistinguishable from *O. variegata* (HVDBM070-10; DQ061092; HVDBM058-10; HVDBM059-10), but that are, based on both barcoding analysis of species delimitation (see Table 1) and phylogenetic tree analysis identified as *O. impervia* (see also Appendices 1 and 2).

Further, all the 56 specimens included in this study were grouped into five MOTUs based on our threshold (Table 1). Using tree-based analysis, we also found five strongly supported groupings (PP = 1.00; BP = 100%), identified as A–E (Figure 3), except that the grouping B corresponding to *O. variegata* is only well supported in the MP analysis (BP = 98%). The composition of these five groupings matches that of MOTUs and comprises *O. tabularis* (A), *O. variegata* (B), *O. impervia* (C), *O. sinensis* (D), and *O. tigrina* (E) (Figures 3, Appendix 1 and 2).

Discussion

The concept of DNA barcoding was first proposed as a technique to accelerate species identification within micro-organisms (Nanney 1982). However, it has now been

generalised as a potential method that can help characterise and discover new species in broader taxonomic groups (Hebert et al. 2004, Van der Bank et al. 2012). In the animal kingdom, the COI region has proved valuable as a DNA barcode for many taxonomic groups, but it can also be problematic for others (Moritz 2004, Ebach and Holdrege 2005, Schindel and Miller 2005, Köhler 2007, Huang et al. 2008).

We first tested COI's potential as a good barcode for the genus *Oxystele*. A good barcode candidate is expected to exhibit a barcode gap (Meyer and Paulay 2005), i.e. higher genetic variation between than within species (Hebert et al. 2003). Various options are currently available to evaluate the barcode gap. We used two approaches. We compared the median of interspecific versus intraspecific distances. We found that interspecific distance is significantly greater than intraspecific distance, suggesting that there is a barcode gap in COI data. We also applied the approach of Meier et al. (2008); i.e. compared the smallest interspecific versus the greatest intraspecific distances), rather than comparing just the median distances. This approach also reveals existence of a barcode gap, thus confirming COI as a potential DNA region for taxon identification within *Oxystele*. This DNA region has also proved successful for barcoding identification in other mollusc taxonomic groups (Davison et al. 2009, Köhler and Glaubrecht 2009, Feng et al. 2011a,b, Sun et al. 2012; but see Sauer and Hausdorf 2012 for limitation of single-locus DNA sequences).

In addition, we found that COI has a strong discriminatory power (85%) within the genus *Oxystele* especially using the best close match and near neighbour methods. This gives support to the efficacy of COI for identification purposes within the genus. However, the application of BOLD identification criteria yields a poor identification success i.e. < 50% and similar proportion of ambiguity (Table 2). The poor performance of COI using BOLD criteria should not be seen as a result of barcoding inefficiency, but should rather be linked to the untested 1% threshold used in BOLD identification (see Meyer and Paulay 2005).

Our analysis of barcoding-based taxon delimitation results in five MOTUs, of which three correspond to morphology-delimited species: *O. sinensis, O. tabularis* and *O. tigrina* (Table 1). These results are also supported by phylogeny-based analysis of species delimitation. However, four specimens identified morphologically as *O. variegata* are included by the barcoding taxon delimitation test within the MOTU of *O. impervia*. Similarly, one specimen identified morphologically as *O. impervia* is grouped within the MOTU of *O. variegata* (Figure 3). These mismatches between morpho-species identification and barcoding-based taxon delimitation (MOTUs) reflect the controversy surrounding species boundaries and/or the identification key (e.g. Heller and Dempster's (1991) key) currently used to distinguish the *impervia/variegata* complex.

Why the mismatch between MOTU and morpho-species? Potential explanations include unsuitable morphology-based taxon delimitation, species paraphyly (– including but not restricted to ancestral polymorphism), and on-going gene flow (i.e., the two taxa are not distinct species or they hybridize; see Funk and Omland 2003). Specifically, Funk and Omland (2003) demonstrated that about 25% of animal species are para- or even polyphyletic, suggesting that the non-monophyly of *O. variegata* and *O.*

impervia in the examined gene tree is not necessarily an argument against their species status. This provides further evidence of the limitations of DNA barcoding in general. It is also possible that the rate of speciation events is slower or greater than that of morphological differentiation; e.g. rapid morphological changes can occur with little or no evolutionary changes (Adams et al. 2002); and this could be driven for example by habitat specialisation (Collar et al. 2010).

In our attempt to resolve the taxonomic uncertainty, we also used the phylogenetic tree reconstruction. The results are similar to those of MOTUs, that is, one specimen morphologically identified as *O. impervia,* grouped on the phylogeny with *O. variegata* (grouping B, Figure 3, Appendix 1 and 2), but this grouping B has strong support only in MP analysis.

The controversy regarding the complex has been reported in previous studies (Heller and Dempster 1991, Williams et al. 2010), likely reflecting the limitations in morphological characters (Hickman 1998) on which the current identification key is based. Heller and Dempster (1991) reported that *O. impervia* and *O. variegata* should be considered as two different species based on shell colour, radula cusp indentation, ecological (*O. impervia* occurs higher up the shore than *O. variegata)*, and fixed allozyme differences at one enzyme-coding locus (out of 22). However, the overlaps in ecological zones and interspecific overlap of up to 66% in radula cusp indentation (Heller and Dempster 1991) indicate that these criteria (ecology and radula indentation) might be unreliable for taxon identification.

In addition, Heller and Dempster (1991) described 24 different photos of shell colours and patterns of typical *O. impervia* and *O. variegata* (12 photos for each species), but the differentiation they proposed is still unclear and could lead to multiple interpretations as indicated in the words such as “very infrequently”, “off-white”, or “greenish-grey” and “almost never” that they used to distinguish between both species. Also, overlaps in colours and weathered shells make Heller and Dempster’s (1991) keys unreliable to identify some individuals (e.g. see Figure 4). Specimens of both *O. impervia* and *O. variegata* are commonly weathered to some extent, resulting in shell colour being indistinct or scarcely discernible. Some specimens (e.g. as shown in Figure 4) can only be tentatively identified because they exhibit unusual colour patterns, not clearly consistent with published photos in Heller and Dempster (1991).

Williams et al. (2010) however suggested that *O. impervia* and *O. variegata* should be regarded as one species based on analysis from a single individual from each species. DH inspected the morphology of the samples (available on MorphoBank) used in the study by Williams et al. (2010) and confirmed that the shell of specimen DQ061092-10 is very typical of that of *O. variegata*, but that DQ061093-10 has a more intermediate form with a finer colour pattern. He concluded that the latter is not obviously referable to any one of *O. impervia* and/or *O. variegata,* more than to the other. In this study, the fact that both specimens come out not only on the phylogeny in the grouping of *O. impervia* (grouping C on the phylogeny; with strong support from PP and BP; Figures 3, Appendix 1 and 2), but also in the MOTU delimitation (Table 1), is surprising (particularly DQ061092-10, which is morphologically typical of *O. variegata*).

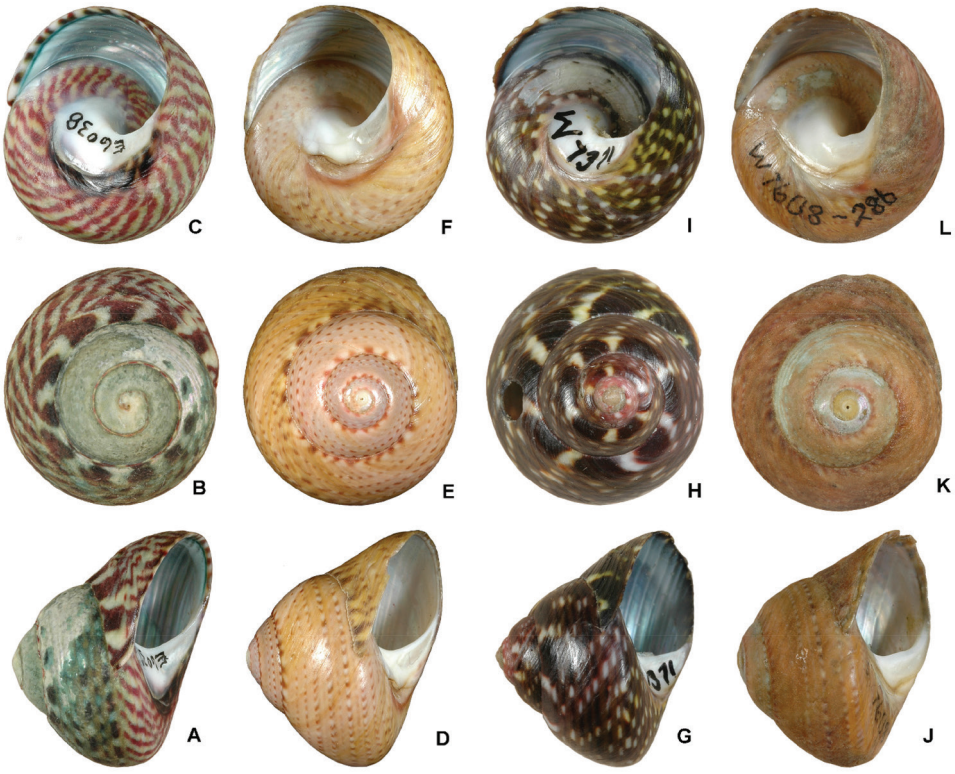


Figure 4. Patterns of shell colour within the genus *Oxystele*. **A–C** *Oxystele variegata* from Namibia, 5 km north of Swakopmund, diameter 22.2 mm (NMSA E6038) **D–F** *Oxystele impervia* from the Western Cape, Groen Rivier, diameter 22.3 mm (NMSA E7353) **G–I** *Oxystele* sp. from the Eastern Cape, Tsitsikamma National Park, diameter 16.5 mm (HVDBM058-10, NMSA W7371); the colour pattern of these specimens suggests *O. variegata*, but these specimens group within the unit of *O. impervia* **J–L** *Oxystele* sp. from the Northern Cape, Noup, diameter 18.0 mm (HVDBM185-10, NMSA W7608); the colour pattern suggests *O. impervia*, but they group with *O. variegata* (see Figures 4 and Appendix 2 for the phylogenetic groupings of these specimens and node supports; these groupings contradict their morphological identification).

One of six polymorphic loci (glycyl-leucine peptidase or peptidase A; Van der Bank 2002) indicated fixed allele differences between *O. impervia* and *O. variegata*, and this was the most convincing characteristic to differentiate between both species (Heller and Dempster 1991). Williams et al. (2010) argue that differences in allele frequency could result from selection pressures (e.g. peptidase in *Mytilus*; Hilbish 1985). They further indicate that differences in habitat preferences, as reported for the *impervia/variegata* complex, could subject them to variation in salinity or temperature, which could lead to variation not only in diets but also in allozymes and morphology.

Indeed morphological differentiation between both species can be difficult. Some of the shell colours and patterns are similar, and radula morphology could be altered

as a result of differences in diet, age and other factors. For example, Padilla (1998) demonstrated that two species of Gastropoda "produce differently shaped teeth when fed different foods, displaying intraspecific variability as extreme as would usually be considered to define different species". Such variation in morphological characters has also been reported to be misleading in other groups such as spiders where the description of almost 50% of the known species was mistakenly based on the same species (Coddington and Levi 1991). Indeed molluscs are well-known to exhibit considerable intraspecific variation in shell morphology (Colgan et al. 2007; Figure 4), and high adaptive capacity to various environmental conditions, leading to striking ecological, morphological and behavioural disparity among specimens within the same species (Ponder et al. 2008).

In this study, most of the specimens that group within unexpected MOTUs were collected from different localities, suggesting possible shell colour variation due to variation in environmental conditions. For example, specimens of *O. variegata* from Namibia and Robben Island clustered on the phylogeny, but those from north-western and southern Africa (Cape) did not. The Cape is renowned for its bad weather as indicated in its common name of "The Cape of Storms", resulting in weathering of individuals (i.e. see "Ships in trouble in Cape waters"; http://www.e-gnu.com/shipwreck_update.html).

Conclusion

The split we found on the phylogeny and species delimitation analyses between *O. impervia* and *O. variegata* does not correspond with the nominal, morphologically-based identifications, indicating the need for the combination of morphological features and genetic data for further analysis. It is also possible that the COI gene alone is insufficient to discriminate species within the genus. We therefore suggest that future analysis should use a multi-gene approach. However, Donald et al. (2005) have studied three genes including two mitochondrial (16S + COI) and one nuclear (actin), and Williams et al. (2010) used one nuclear and three mitochondrial genes; but neither study was successful in teasing apart both species. We would therefore suggest that additional techniques such as AFLP or microsatellites should be applied in an attempt to reveal the status of *O. impervia* and *O. variegata*. Nevertheless, our analyses using barcoding confirm the existence of five MOTUs (probably suggestive of five species), with *O. variegata* being a distinct species from *O. impervia*.

Acknowledgements

We would like to thank the Government of Canada through Genome Canada and the Ontario Genomics Institute (2008-OGI-ICI-03) for the DNA sequencing. The research was supported by the ACDB and partially by the Toyota Enviro Outreach program

2010. The Tsitsikamma National Parks Board gave permission for sample collection and the Kwazulu-Natal Museum processed the voucher specimens. We thank Stephanus Voges and Bronwen Curry (Ministry of Fisheries and Marine Resources, Namibia), Estelle Esterhuizen (Robben Island Nature Conservation) and Gerhard Groenewald (Klipbokkop Nature Reserve) for assistance with sample collections. This work is based on research supported in part by the National Research Foundation of South Africa.

References

Adams PA, Pandey N, Rezzi S, Casanova J (2002) Geographic variation in the Random Amplified Polymorphic DNAs (RAPDs) of *Juniperus phoenicea*, *J. p.* var. *canariensis*, *J. p.* subsp. *eumediterranea*, and *J. p.* var. *turbinata*. Biochemical Systematic Ecology 30: 223–229. doi: 10.1016/S0305-1978(01)00083-7

Alfaro ME, Holder MT (2006) The posterior and the prior in Bayesian Phylogenetics. Annual Review of Ecology Evolution and Systematics 37: 19–42. doi: 10.1146/annurev.ecolsys.37.091305.110021

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. Philosophical Transactions of the Royal Society B 360: 1813–1823. doi: 10.1098/rstb.2005.1713

Branch GM, Griffiths CL, Branch ML, Beckley LE (2010) Two oceans: a guide to the marine life of southern Africa. Struik Nature, Cape Town.

Brown SDJ, Collins RA, Boyer S, Lefort M-C, Malumbres-Olarte J, Vink CJ, Cruickshank RH (2012) Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Molecular Ecology Resources. http://cran.r-project.org; http://spider.r-forge.r-project.org/, doi: 10.1111/j.1755-0998.2011.03108.x

Bucklin A, Steinke D, Blanco-Bercial L (2011) DNA barcoding of marine metazoa. Annual Review of Marine Science 3: 471–508. doi: 10.1146/annurev-marine-120308-080950

Coddington JA, Levi HW (1991) Systematics and evolution of spiders (Araneae). Annual Reviews of Ecology and Systematics 22: 565–592. doi: 10.1146/annurev.es.22.110191.003025

Colgan DJ, Ponder WF, Beacham E, Macaranas J (2007) Molecular phylogenetics of Caenogastropoda (Gastropoda: Mollusca). Molecular Phylogenetics and Evolution 42: 717–737. doi: 10.1016/j.ympev.2006.10.009

Collar DC, Schulte JA, O'meara BC, Losos JB (2010) Habitat use affects morphological diversification in dragon lizards. Journal of Evolutionary Biology 23: 1033–1049. doi: 10.1111/j.1420-9101.2010.01971.x

Davison A, Blackie RLE, Scothern GP (2009) DNA barcoding of stylommatophoran land snails: a test of existing sequences. Molecular Ecology Resources 9: 1092–1101. doi: 10.1111/j.1755-0998.2009.02559.x

Donald KM, Kennedy M, Spencer HG (2005) The phylogeny and taxonomy of austral monodontine topshells (Mollusca: Gastropoda: Trochidae), inferred from DNA sequences. Molecular Phylogenetics and Evolution 37: 474–483. doi: 10.1016/j.ympev.2005.04.011

Ebach MC, Holdrege C (2005) DNA barcoding is no substitute for taxonomy. Nature 434: 697. doi: 10.1038/43469

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797. doi: 10.1093/nar/gkh340

Feng Y, Li Q, Kong L, Zheng X (2011a) COI-based DNA barcoding of Arcoida species (Bivalvia: Pteriomorphia) along the coast of China. Molecular Ecology Resources 11**:** 435–441.

Feng Y, Li Q, Kong L, Zheng X (2011b) DNA barcoding and phylogenetic analysis of Pectinidae (Mollusca: Bivalvia) based on mitochondrial COI and 16S rRNA genes. Molecular Biology Report 38: 291–299. doi: 10.1007/s11033-010-0107-1

Felsenstein J (1985) Confidence levels on phylogenies: an approach using the bootstrap. Evolution 39: 783−791. DOI: 10.2307/2408678

Fitch WM (1971) Towards defining the course of evolution: minimum change for a specific tree topology. Systematic Zoology 20: 406–416. doi: 10.2307/2412116

Funk DJ, Omland KE (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. doi: 10.1146/annurev.ecolsys.34.011802.132421

Hajibabaei M, De Waard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, Hebert PDN (2005) Critical factors for assembling a high volume of DNA barcodes. Philosophical Transactions of the Royal Society B 360: 1959–1967. doi: 10.1098/rstb.2005.1727

Hebert PDN, Ratnasingham S, de Waard JR (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London B (Supplement) 270: S96–S99. doi: 10.1098/rsbl.2003.0025

Hebert PDN, Penton EH, Burns J, Janzen DJ, Hallwachs W (2004) 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 USA 101: 14812–14817. doi: 10.1073/pnas.0406166101

Heller J, Dempster Y (1991) Detection of two coexisting species of *Oxystele* (Gastropoda, Trochidae) by morphological and electrophoretic analysis. Journal of Zoology 223: 395–418. doi: 10.1111/j.1469-7998.1991.tb04773.x

Hickman CJ (1998) A field guide to sea stars and other echinoderms of Galápagos. Sugar Spring Press, Lexington, VA, USA, 83 pp.

Hilbish TJ (1985) Demographic and temporal structure of an allele frequency cline in the mussel *Mytilus edulis*. Marine Biology 86: 163–171. doi: 10.1007/BF00399023

Huang DW, Meier R, Todd PA, Chou LM (2008) Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. Journal of Molecular Evolution 66: 167–174. doi: 10.1007/s00239-008-9069-5

Jones M, Ghoorah A, Blaxter M (2011) jMOTU and Taxonerator: Turning DNA barcode sequences into annotated operational taxonomic units. PLoS ONE 6: e19259. doi: 10.1371/journal.pone.0019259

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120. doi: 10.1007/BF01731581

Köhler F (2007) From DNA taxonomy to barcoding - how a vague idea evolved into a biosystematic tool. Mitteilungen aus dem Museum für Naturkunde in Berlin Zoologische Reihe 83: 44–51. doi: 10.1002/mmnz.200600025

Köhler F, Glaubrecht M (2009) Uncovering an overlooked radiation: molecular phylogeny and biogeography of Madagascar's endemic river snails (Caenogastropoda: Pachychilidae: *Madagasikara* gen. nov.). Biological Journal of the Linnean Society 99: 867–894. doi: 10.1111/j.1095-8312.2009.01390.x

Lakra WS, Verma MS, Goswami M, Lal KK, Mohindra V, Punia P, Gopalakrishnan A, Ward RD, Hebert P (2011) DNA barcoding Indian marine fishes. Molecular Ecology Resources 11: 60–71. doi: 10.1111/j.1755-0998.2010.02894.x

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Meier R, Zhang G, Ali F (2008) The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Systematic Biology 57: 809–813. doi: 10.1080/10635150802406343

Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS Biology 3: 2229–2238. doi: 10.1371/journal.pbio.0030422

Moritz C (2004) DNA barcoding: promise and pitfalls. PLoS Biology 2: e354. doi: 10.1371/journal.pbio.0020354

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E et al. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348–2351. doi: 10.1126/science.1067179

Nanney DL (1982) Genes and phenes in *Tetrahymena*. BioScience 32: 783–788. doi: 10.2307/1308971

Padilla DK (1998) Inducible phenotypic plasticity of the radula in *Lacuna* (Gastropoda: Littorinidae). The Veliger 41: 201–204.

Ponder WF, Colgan DJ, Healy JM, Hützel A, Simone LRL, Strong EE (2008) Caenogastropoda. In: Ponder WF, Lindberg DR (Eds) Phylogeny and Evolution of the Mollusca. University of California Press, Berkeley, 331–383. doi: 10.1525/california/9780520250925.003.0013

Posada D (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25: 1253–1256. doi: 10.1093/molbev/msn083

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology 53: 793–808. doi: 10.1080/10635150490522304

Puillandre N, Cruaud C, Kantor YI (2009) Cryptic species in *Gemmuloborsonia* (Gastropoda: Conoidea). Journal of Molluscan Studies 76: 11–23. doi: 10.1093/mollus/eyp042

Radulovici AE, Archambault P, Dufresne F (2010) DNA barcodes for marine biodiversity: moving fast forward? Diversity 2: 450–472. doi: 10.3390/d2040450

Ronquist F, Huelsenbeck JP (2003) MrBayes 3.1.2: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574. doi: 10.1093/bioinformatics/btg180

Sauer J, Hausdorf B (2012) A comparison of DNA-based methods for delimiting species in a Cretan land snail radiation reveals shortcomings of exclusively molecular taxonomy. Cladistics 28: 300–316. doi: 10.1111/j.1096-0031.2011.00382.x

Schindel DE, Miller SE (2005) DNA barcoding, a useful tool for taxonomists. Nature 435: 17. doi: 10.1038/43501

Stahlhut JK, Fernández-Triana J, Adamowicz SJ, Buck M, Goulet H, Hebert PDN, Huber JT, Merilo MT, Sheffield CS, Woodcock T, Smith MA (2013) DNA barcoding reveals diversity of Hymenoptera and the dominance of parasitoids in a sub-arctic environment. BMC Ecology 13: 2. doi: 10.1186/1472-6785-13-2

Sun Y, Li Q, Kong L, Zheng X (2012) DNA barcoding of Caenogastropoda along coast of China based on the COI gene. Molecular Ecology Resources 12: 209–218. doi: 10.1111/j.1755-0998.2011.03085.x

Swofford DL (2002) PAUP*: phylogenetic analysis using parsimony (* and other methods), version 4.10. Sinauer, Sunderland, Massachusetts.

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12: 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Van der Bank FH (2002) A review of gene nomenclature for enzyme-coding loci generally used in allozyme studies. Trends in Comparative Biochemistry and Physiology 9: 197–203. doi: 10.3750/AIP2012.42.4.04

Van der Bank HF, Greenfield R, Daru BH, Yessoufou K (2012) DNA barcoding reveals micro-evolutionary changes and river system-level phylogeographic resolution of African silver catfish, *Schilbe intermedius* (Actinopterygii: Siluriformes: Schilbeidae) from seven populations across different African river systems. Acta Ichthyologica Et Piscatoria 42: 307–320.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B 360: 1847–1857. doi: 10.1098/rstb.2005.1716

Wilcox T, Zwick D, Heath T, Hillis D (2002) Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Molecular Phylogenetics and Evolution 25: 361–371. doi: 10.1016/S1055-7903(02)00244-0

Williams ST, Donald KM, Spencer HG, Nakano T (2010) Molecular systematics of the marine gastropod families Trochidae and Calliostomatidae (Mollusca: Superfamily Trochoidea). Molecular Phylogenetics and Evolution 54: 783–809. doi: 10.1016/j.ympev.2009.11.008

Yang JB, Wang YP, Moller M, Gao LM, Wu D (2012) Applying plant DNA barcodes to identify species of *Parnassia* (Parnassiaceae). Molecular Ecology Resources 12: 267–275. doi: 10.1111/j.1755-0998.2011.03095.x

Appendix 1

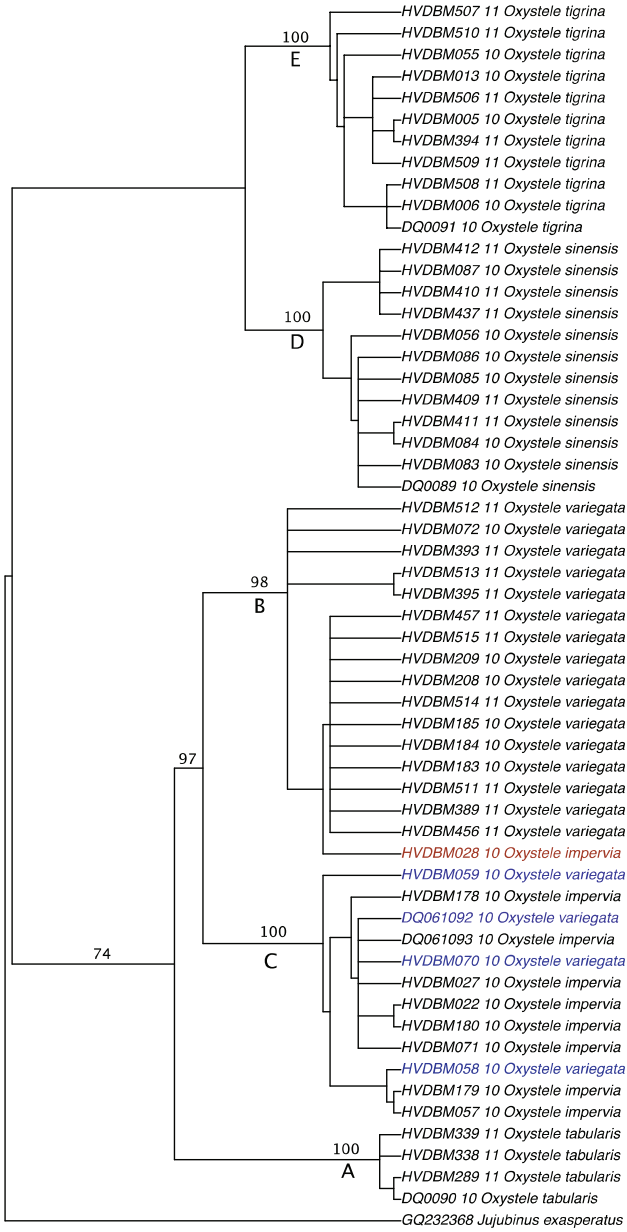


Figure S1. The only parsimonious tree obtained from the maximum parsimony (MP) analysis. Topology of species groupings is similar to that of the Bayesian tree (see Figure 3). Node supports are reported on the branches; the first value is bootstrap support from MP analysis; the second value in bracket indicates the posterior probability obtained from Bayesian analysis; only moderate to high node support values are indicated; *Jujubinus exasperatus* is used as outgroup; A-E indicates different possible species-units in the dataset: A (*O. tabularis*), B (*O. variegata*), C (*O. impervia*), D (*O. sinensis*), E (*O. tigrina*), as in Figure 3.

Appendix 2

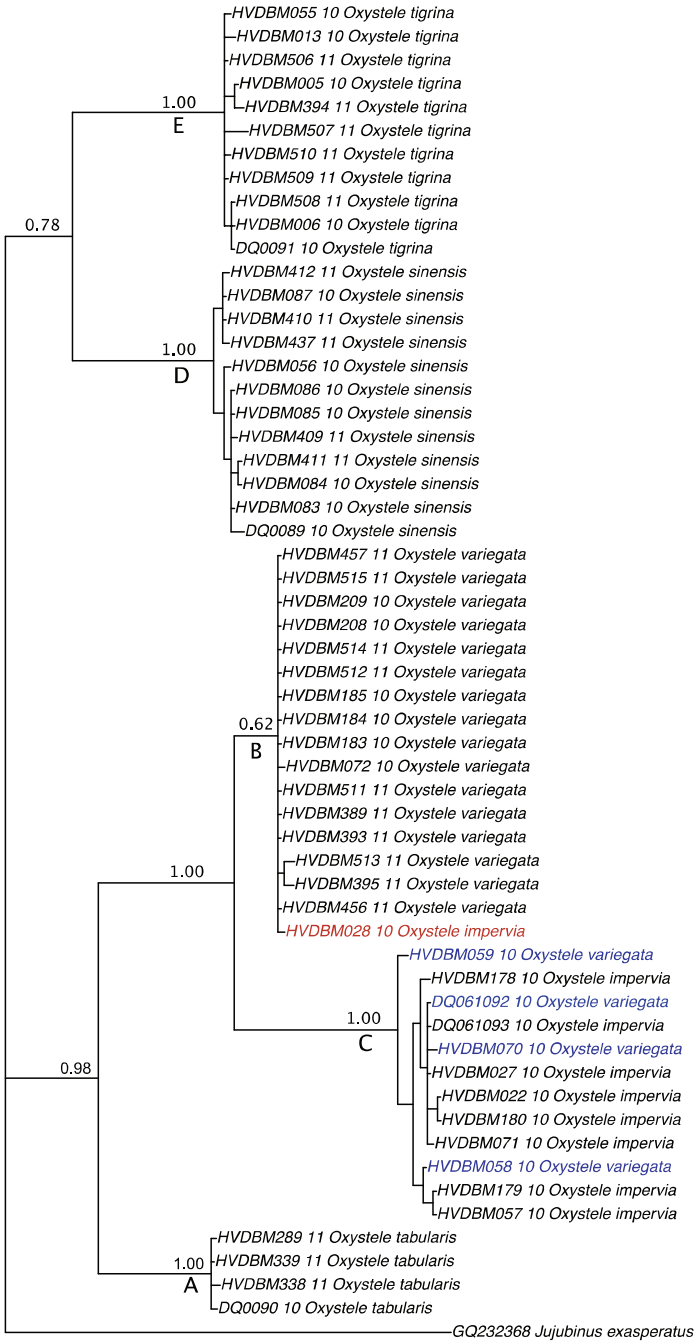


Figure S2. Bayesian tree assembled using MrBayes indicating the groupings of specimens and the posterior probability of the nodes.

ZooKeys 365: 355–379 (2013)
doi: 10.3897/zookeys.365.5776
www.zookeys.org

SHORT COMMUNICATION

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes)

Maarten P. M. Vanhove[1,2,*], Bart Tessens[3,*], Charlotte Schoelinck[4], Ulf Jondelius[5], D. Tim J. Littlewood[6], Tom Artois[3], Tine Huyse[1,7]

1 *Laboratory of Biodiversity and Evolutionary Genomics, Department of Biology, University of Leuven, Leuven, Belgium* **2** *Present address: Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic* **3** *Research Group Zoology: Biodiversity & Toxicology, Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium* **4** *Aquatic animal health, Fisheries and Oceans Canada, Moncton, NB, Canada* **5** *Department of Invertebrate Zoology, Swedish Museum of Natural History, Stockholm, Sweden* **6** *Division of Parasites & Vectors, Department of Life Sciences, Natural History Museum, London, United Kingdom* **7** *Department of Biology, Royal Museum for Central Africa, Tervuren, Belgium*

Corresponding author: *Maarten P.M. Vanhove* (maarten.vanhove@bio.kuleuven.be)

Academic editor: *K. Jordaens* | Received 11 June 2013 | Accepted 2 December 2013 | Published 30 December 2013

Citation: Vanhove MPM, Tessens B, Schoelinck C, Jondelius U, Littlewood DTJ, Artois T, Huyse T (2013) Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 355–379. doi: 10.3897/zookeys.365.5776

Abstract

Some taxonomic groups are less amenable to mitochondrial DNA barcoding than others. Due to the paucity of molecular information of understudied groups and the huge molecular diversity within flatworms, primer design has been hampered. Indeed, all attempts to develop universal flatworm-specific COI markers have failed so far. We demonstrate how high molecular variability and contamination problems limit the possibilities for barcoding using standard COI-based protocols in flatworms. As a consequence, molecular identification methods often rely on other widely applicable markers. In the case of Monogenea, a very diverse group of platyhelminth parasites, and Rhabdocoela, representing one-fourth of all free-living flatworm taxa, this has led to a relatively high availability of nuclear ITS and 18S/28S rDNA sequences on GenBank. In a comparison of the effectiveness in species assignment we conclude that mitochondrial and nuclear ribosomal markers perform equally well. In case intraspecific information is needed, rDNA sequences can guide the selection of the appropriate (i.e. taxon-specific) COI primers if available.

* These authors contributed equally to this work.

Keywords
mitochondrial DNA, Monogenea, primer design, ribosomal DNA, Rhabdocoela, turbellarians

Introduction

Many biodiversity studies tend to focus on conspicuous fauna, ignoring the vast species diversity and ecological importance of less sizeable animals such as parasitic or meiofaunal taxa, including flatworms (Windsor 1998, Marcogliese 2004, Fonseca et al. 2010). To deal with the huge task of assessing their biodiversity and systematics, a variety of molecular-based methods have been proposed. These include a (phylo)genetic approach (Brooks and Hoberg 2001), DNA barcoding (Besansky et al. 2003) and amplicon-based next generation sequencing of environmental samples (Fonseca et al. 2010). DNA barcoding aims to use the sequence diversity of one or more uniform target genes to identify species (e.g. Stoeckle 2003, Hebert et al. 2004, Meusnier et al. 2008). Such a standardized approach is particularly promising in understudied taxa and for organisms where morphological identification is complicated, in case of heteromorphic generations, sexual dimorphism, a lack of suitable characters or, for example in parasites, the existence of larval stages that have not been characterized yet (Leung et al. 2009, Radulovici et al. 2010 and references therein). There are indications that the commonly used barcoding gene, cytochrome *c* oxidase subunit I (COI), is also suitable for tree reconstruction and molecular dating (e.g. for insects: Gaunt and Miles 2002, for digeneans: Brant and Loker 2009). Phylogenetic inference (or, for that matter, phylogeography or population assignment) is certainly a potential added value of COI barcoding. It does not, however, lie at its core (Moritz and Cicero 2004). While we do not intend to provide a review here on the pros and cons of barcoding, we completely agree with Besansky et al. (2003) that barcoding is not an end in itself. Rather, it is a sequence-based tool that may facilitate and accelerate identification of previously characterized species, e.g. from environmental samples, or that may assist in the detection of cryptic species (Vilas et al. 2005, Locke et al. 2010b, Nadler and Pérez-Ponce de Léon 2011, Jörger and Schrödl 2013).

It is important to consider the characteristics of the COI gene, warranting its common use as a barcoding gene. Being a mitochondrial gene, it has a maternal inheritance, lacks introns, undergoes no recombination, and primers are available for potentially much of the animal kingdom (Folmer et al. 1994, Hebert et al. 2003a). This supposed availability of universal primers is a core advantage, although truly universal applicability is questionable (Stoeckle 2003, Radulovici et al. 2010, Taylor and Harris 2012). Another asset is that intraspecific genetic distances are usually much lower than interspecific distances (Hebert et al. 2003b). Hebert et al. (2003a, b) demonstrated the wide applicability of COI, both at various taxonomic levels and across a range of taxa. What is the state-of-affairs, then, in barcoding abundant but inconspicuous animals such as those belonging to the parasitic realm or the meiobenthos? With the example

of two species-rich and understudied groups of flatworms, rhabdocoels and monogeneans, we aim to evaluate the potential of COI for barcoding, and assess the potential of alternative ribosomal DNA markers.

COI barcoding in monogenean and rhabdocoel flatworms: not a one-stop shop

The acquisition of COI markers for flatworms has opened up many new research avenues. COI data have proven useful in parasitic, meiofaunal or other flatworms (e.g. Elsasser et al. (2009) for guinea worms, Lázaro et al. (2009) for triclads, Sanna et al. (2009) for proseriates, Locke et al. (2010a) for diplostomid digeneans). However, the amplified COI fragment may lie outside the barcoding region (Moszczynska et al. 2009). The use of COI in flatworms can also entail amplification or sequencing problems (e.g. Larsson et al. (2008) for catenulids) or can simply be insufficiently explored (Casu et al. (2009) for proseriates). Moreover, COI amplification in flatworms may require the development of taxon-specific primers (Moszczynska et al. 2009, Sanna et al. 2009). Indeed, truly "universal" barcoding primers for flatworms are either lacking to date or underperform for certain groups (Littlewood 2008, Moszczynska et al. 2009). Within flatworms, there is considerable amino acid sequence variability in the region where Folmer et al. (1994) designed the "universal" COI primers, and flatworms seem radically different from other metazoans in amino acid content over the COI gene (Figure 1). Hence, it is easy to understand why it is difficult to find a set of primers that perform well for a wide range of flatworms. Indeed, despite their diversity, neither monogeneans nor rhabdocoels were well covered in papers central to the development of the barcoding idea, although these included flatworms. For example, Folmer et al. (1994) mentioned that the COI primers proved successful in a polyclad and a digenean flatworm, while Hebert et al. (2003b) scrutinized COI sequences from several families of cestodes, digeneans and triclads, but only included one monogenean family (Polystomatidae) and no rhabdocoels. The unavailability of truly ubiquitous PCR primers and conditions is suboptimal and undermines the use of COI as a barcoding marker universal to flatworms.

Monogenea is a species-rich group within the parasitic flatworms, a lot of the diversity of which remains unexplored. Indeed, only an estimated 2200 – 5000 species have been described (Hoberg 1997, Whittington 1998 and references therein), with a remaining 20,000 presently undescribed species (Whittington 1998). Monogenea mostly includes ectoparasites of cold-blooded amphibious or aquatic vertebrates, while some are endoparasites or infect aquatic invertebrates (Pugachev et al. 2009). Though clearly not as widespread in use as nuclear rDNA (Littlewood 2008) (see below), COI markers may offer high resolution for monogenean barcoding (Hansen et al. 2007). COI sequences are available from an increasing range of monogeneans, including representatives of Ancyrocephalidae, Capsalidae, Chauhaneidae, Chimaericolidae, Diclidophoridae, Discocotylidae, Diplectanidae, Diplozoidae, Gastrocotylidae, Gotocotylidae, Mazocraeidae, Microcotylidae, Plectanocotylidae, Polystomatidae

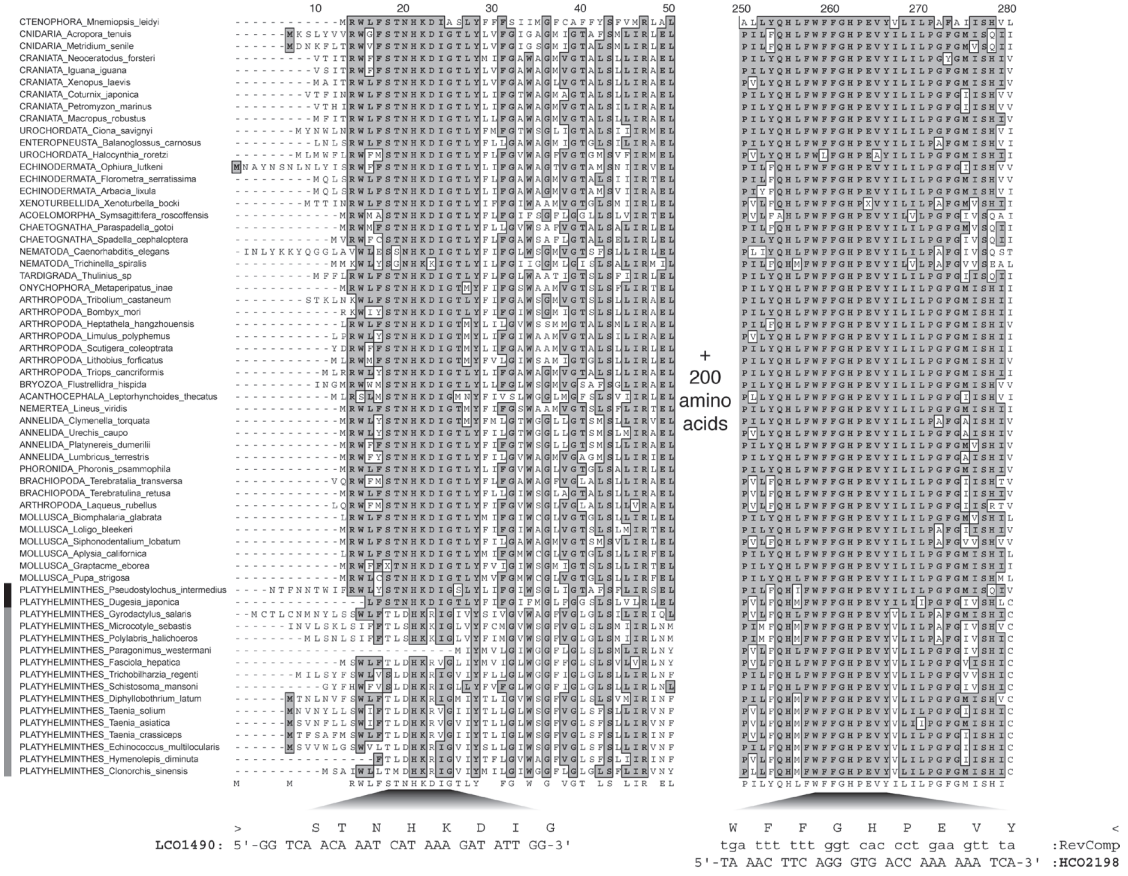


Figure 1. Fragments of an alignment of complete mitochondrial cytochrome *c* oxidase subunit I genes, as amino acids, for a diversity of metazoan taxa indicating the positions of the Folmer et al. (1994) primers. Shading indicates sequence identity to a consensus sequence calculated where > 50% taxa share amino acid identity at any position. All data available from MitoZoa database (Lupi et al. 2010). Vertical bar indicates platyhelminth taxa; black indicates 'turbellarians', grey indicates neodermatans.

and Pyragraphoridae (Telford et al. 2000, Jovelin and Justine 2001, Plaisance et al. 2008, Mladineo et al. 2009, 2013, Perkins et al. 2010, Li et al. 2011, Poisot et al. 2011, Schoelinck et al. 2012, Stefani et al. 2012, Zhang et al. 2012). COI sequences published for monogeneans are regularly positioned in the region amplified by the widely used ASmit primers (Littlewood et al. 1997). The COI gene region in question does not match the commonly used "barcoding fragment" *sensu* Folmer et al. (1994). Indeed, the 3' primer (LCO1490) of the Folmer et al. (1994) set overlaps with the 5' primer of ASmit (Asmit1) so that these fragments only overlap at primer-binding sites; the Folmer et al. (1994) fragment is ~ 459 bp and the adjoining ASmit fragment ~ 445 bp. Each of these fragments is less than a third of the complete COI gene. However, short barcoding fragments in general need not be problematic (Meusnier et al. 2008). Moreover, their length can of course be extended, e.g. in combination with a schistosomatid primer from Lockyer et al. (2003b) (e.g. up to *ca.* 580 bp in Vanhove (2012)).

Rhabdocoela is one of the most species-rich clades of free-living "turbellarian" flatworms with over 1 500 described species (Van Steenkiste et al. 2013). The suitability of the COI gene for DNA barcoding of rhabdocoels has not yet been explored. Since there are only four COI sequences from three species published in GenBank on 25 November 2013, a first goal of this paper is to obtain COI barcode data from different rhabdocoels by means of cloning. This approach is obviously not suited for large-scale applications, but can be used to identify possible contaminating factors and to establish a dataset of COI sequences that allows the development of new taxon-specific primers.

The ribosomal DNA region and its use in species recognition in flatworms

Various fragments of the nuclear ribosomal DNA, like the genes for 18S, 5.8S and 28S rRNA, and the internal transcribed spacers ITS-1 and ITS-2, evolve at different rates, making them suitable for assessing genetic divergence at various levels (Hillis and Dixon 1991). The ribosomal RNA genes are rather conserved, allowing the design of primers for a wide range of taxa. Additional methodological advantages include the multicopy structure of rDNA (allowing amplification of little DNA template, e.g. in minute animals or museum specimens) and its concerted evolution, leading to low intraspecific variation (Hillis and Dixon 1991, Nieto Feliner and Rosselló 2007). The phylogenetic or taxonomic application of nuclear ITS rDNA is especially established in plants and fungi (Nieto Felliner and Rosselló 2007) but also popular in a wide range of animal taxa (e.g. Odorico and Miller 1997), including flatworms (e.g. Nolan and Cribb 2005, Brant and Loker 2009).

In monogeneans, various portions of the rDNA, and most often the spacer regions ITS-1 and ITS-2, are considered to adequately mirror differences between morphologically recognized species (Cunningham 1997, Matějusová et al. 2001, Meinilä et al. 2002, Ziętara and Lumme 2002). They are also useful in identifying cryptic species (e.g. Pouyaud et al. 2006). As a consequence, these sequence fragments are often included in species descriptions (e.g. Huyse and Malmberg 2004, García-Vásquez et

al. 2007, 2011, Paetow et al. 2009, Paladini et al. 2009, 2010, 2011a, b, Přikrylová et al. 2009a, b, 2012a, b, Rokicka et al. 2009, Vaughan et al. 2010, Schelkle et al. 2011, Vanhove et al. 2011b, Ziętara et al. 2012, Řehulková et al. 2013). This goes especially for representatives of *Gyrodactylus* von Nordmann, 1832. As is often the case, ITS sequences in monogeneans display little (or no) intraspecific variation (Meinilä et al. 2002, Huyse et al. 2006, Přikrylová et al. 2012a, but see Vanhove et al. 2011b), precluding comparisons of interspecific *versus* intraspecific genetic diversity which is an important part of COI barcoding (Stoeckle 2003, Hebert et al. 2004). It is, however, unfortunate that many studies do not address potential intraspecific ITS diversity at all. By often sufficing with one or a few sequenced individuals per species, in general, ITS rDNA has been used for phylogenetic positioning in species descriptions rather than as a barcoding fragment.

In rhabdocoels, the 18S and 28S rDNA has been used extensively for phylogenetic analysis (Willems et al. 2006, Van Steenkiste et al. 2013). These gene fragments can be obtained very easily in rhabdocoels using universal primers. Most rhabdocoel morphospecies have unique 18S and 28S rDNA sequences, except for a few species of the genera *Microdalyellia* Gieysztor, 1938 and *Castrada* Schmidt, 1862. No data from the spacer regions are currently available.

From these examples, it is clear that the various portions of the nuclear rDNA region render it a versatile region for genetic approaches to systematics of both monogeneans and rhabdocoels. An additional advantage is the availability of primers that seem to be flatworm-universal (Lockyer et al. 2003a, Telford et al. 2003) or that are even applicable to a much wider range of organisms ranging from fungi to schistosome flatworms (White et al. 1990, Barber et al. 2000, Sonnenberg et al. 2008, Moszczynska et al. 2009). However, the use of rDNA markers for barcoding has rarely been formally tested in monogeneans and rhabdocoels. As a second goal of this paper, we will therefore formally test the usefulness of some candidate rDNA barcoding markers in selected cases in monogeneans and rhabdocoels. Whenever possible, we directly compare the performance of these rDNA markers to that of the traditional COI mitochondrial marker.

Materials and methods

Amplification success of COI in rhabdocoel flatworms

A total of 27 species of rhabdocoels (from 21 genera covering 15 out of the 35 rhabdocoel families) were collected from freshwater, marine or brackish water sites. Specimens were collected as described in Schockaert (1996) and stored in ethanol for subsequent molecular work. All specimens were studied alive and documented through drawings, pictures and videos. Specimen collection and sequence data are provided in Appendix 1.

DNA was extracted from whole or partial specimens using the QIAamp DNA micro kit (QIAGEN) according to the manufacturer's instructions. Extracts were stored in duplicates (40 and 20 µl) for each specimen. The Folmer et al. (1994) region of the COI gene was amplified with the primers LCO1490 (5'-GGTCAACAAATCATAAAGTTGG-3') and an adapted version of the HCO2198 primer (5'-TCATAGTAGCCSYTGTAAAATAAGCTCG-3') using a touchdown PCR protocol [95 °C for 4 min, 2 × (94 °C for 30 s, 58 °C for 30 s, 72 °C for 30 s), 2 × (94 °C for 30 s, 56 °C for 30 s, 72 °C for 30 s), 5 × (92 °C for 40 s, 45 °C for 40 s, 72 °C for in 15 s), 35 × (94 °C for 30 s, 51 °C for 40 s, 72 °C for 1 min 15 s), 72 °C for 10 min]. Illustra puReTaq Ready-To-Go PCR beads (GE Healthcare) were used to prepare reactions containing 3 µl DNA-extract, 0.2 µM of each primer and water for a final volume of 25 µl. PCR products were checked on 1.4% agarose gels stained with Gelred (Biotum Inc.), then were cleaned in Nucleofast 96 PCR plates (Macherey-Nagel, Düren). PCR products were then cloned using the TOPO TA for Sequencing Cloning Kit (Invitrogen) according to the manufacturer's instructions. From each PCR product, eight colonies were picked and bidirectionally sequenced on an ABI3130XL Automated DNA sequencer (Applied Biosystems, Hitachi). Sequences were visually inspected and assembled in Geneious Pro v5.7.5 (Biomatters Ltd).

To check for possible contamination we first submitted all sequences of each clone to BLAST search on the NCBI website (http://www.ncbi.nlm.nih.gov). To further identify sequences that did not have a strong match in GenBank we aligned them to a reference dataset of the Folmer et al. (1994) region COI sequences of possible contaminants such as known food items. This reference dataset was constructed with sequences collected from GenBank (see Appendix 2) and included the following taxa: Platyhelminthes, Arthropoda, Gastropoda, Bivalvia, Nematoda, Cnidaria. Sequences were aligned in ClustalX v2 (Larkin et al. 2007). A Neighbour-Joining (NJ) tree based on Kimura 2-Parameter (K2P) (Kimura 1980) distances was calculated in MEGA5 (Tamura et al. 2011).

Test cases for barcoding with ribosomal and mitochondrial markers

Three test cases were analyzed to demonstrate the potential of different markers for DNA barcoding in Monogenea and Rhabdocoela. The first consisted of 33 species from four genera from the monogenean family Diplectanidae infecting groupers from the Indo-Pacific (from Schoelinck 2012). This dataset contained 117 sequences of the COI gene and the nuclear 28S rDNA region (see Appendix 3). A second test case consisted of eight species from the monogenean genus *Gyrodactylus* (from Vanhove 2012). The species included are parasites of Balkan freshwater gobies (Vanhove et al. 2012, 2013). This dataset contained 35 sequences of the ITS-1 – 5.8S rDNA – ITS-2 rDNA region, 17 sequences of the COI gene and 38 sequences of the cytochrome *c* oxidase subunit II gene (COII) gene (see Appendix 3). The latter, of which over 600 bp was amplified by the primers developed by Bueno Silva (2011) is a promising additional marker

for *Gyrodactylus* (Vanhove 2012). These degenerate primers can be optimized for other monogenean families (W.A. Boeger and M. Bueno Silva, personal communication), though there does not seem to be a single COII protocol which is generally suitable for non-gyrodactylid monogeneans. As a third case, we reanalyzed the nuclear 18S and 28S data from the analysis of Van Steenkiste et al. (2013) for the rhabdocoel genus *Gieysztoria* Ruebush and Hayes, 1939 (see Appendix 4). Additionally, we sequenced the ITS-1 – 5.8S rDNA – ITS-2 rDNA region from the same specimens (sequences deposited in GenBank under accession numbers KF953866–KF953883; see Appendix 4).

The K2P-distance model (Kimura 1980) was used to calculate sequence divergences between and within species. Histograms of intra- and interspecific distance frequencies were reconstructed in R v2.15.2 (R Core Team 2012) using scripts made available by G. Sonet (RBINS – Brussels, Belgium). For the monogenean test cases (test case 1 and 2), the proportion of correctly identified specimens was estimated with the program Species Identifier using the best match (BM) and best close-match (BCM) criteria of Meier et al. (2006). The threshold used in the BCM analysis was the "best compromise threshold" (BCTh) based on cumulative distribution curves of intra- and interspecific K2P-distances calculated in R (Lefébure et al. 2006). Species represented by a single sequence in the dataset were removed as they will generate incorrect identifications under the BM and BCM criterion because there are no other conspecifics in the dataset. For this reason, the BM and BCM criterion was not used on the rhabdocoel dataset (test case 3) where there are many species represented by a single sequence.

Results

Amplification success of COI in rhabdocoel flatworms

A BLAST search of the 169 clones that could successfully be sequenced showed that contamination originated both from external DNA sources (*Homo sapiens, Bos taurus* – the latter possibly stemming from liver fed to flatworm cultures, or from bovine serum albumin used in the laboratory) and from food items eaten by the worms (Arthropoda, Annelida, Rotifera, Cnidaria, Ciliophora). Most rhabdocoels are so small that DNA has to be extracted from whole animals, which potentially results in the amplification of food items present in the animal. Only two sequences could be identified by BLAST as belonging to flatworms. This is, however, not very surprising given that there is currently only one rhabdocoel COI sequence overlapping with the Folmer et al. (1994) region available in GenBank. Only by aligning the sequences that did not have a significant BLAST hit (accounting for 100 of the 169 clones) to a reference dataset containing published COI sequences from different flatworm species and some possible food items (Appendix 2), were we able to identify a clade of flatworm sequences (Figure 2). From the original 169 clones that were sequenced, only 19 sequences, belonging to 13 (out of a total of 27 investigated) species were identified as belonging to Platyhelminthes. Genetic diversity among these newly identified COI sequences from 13 rhabdocoel species is high (average pairwise K2P-distance = 0.284).

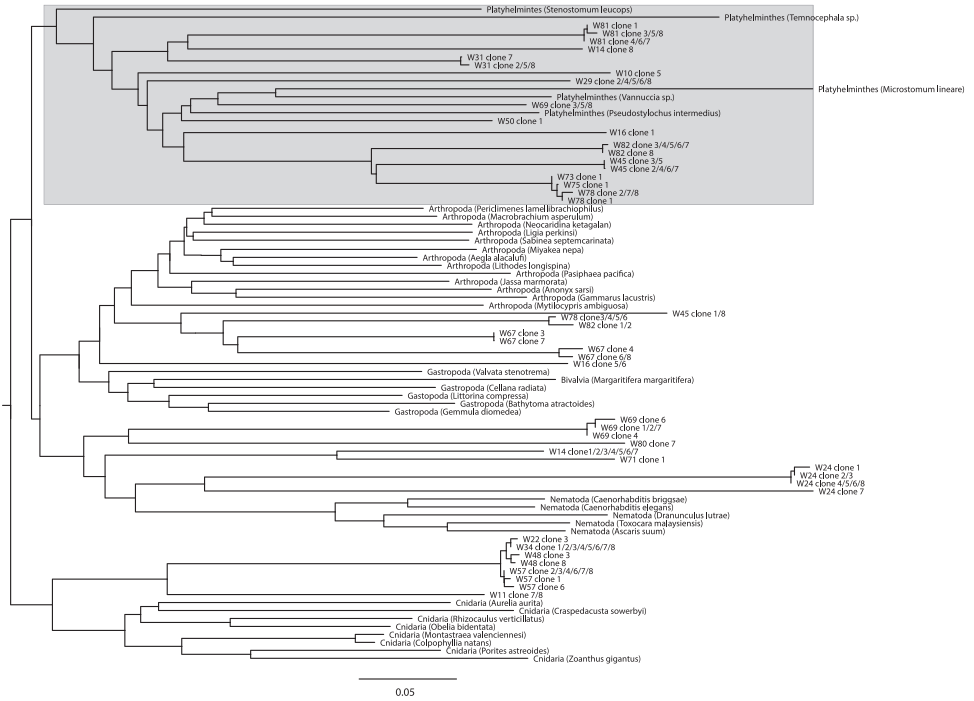


Figure 2. Neighbour-Joining tree based on Kimura 2-Parameter (Kimura 1980) distances for COI DNA sequences for 100 clones from 27 rhabdocoel species, five flatworm COI sequences available from GenBank and 31 reference COI sequences from taxa that are potential food sources for rhabdocoels. The clade with platyhelminth sequences is indicated in gray.

Test cases for barcoding with ribosomal and mitochondrial markers

Histograms of intra- and interspecific K2P-distances are given in Figure 3. Only for the COI gene of *Gyrodactylus* there was a clear barcoding gap (3-11%). In all other cases there was overlap between the distribution of intra- and interspecific K2P-distances. In the Diplectanidae dataset (test case 1) the BCTh values were 14.5% for COI and 0.74% for 28S (Figure 4). In *Gyrodactylus* (test case 2) the BCTh was 5.3% for COII, 6.5% for COI and 1.39% for the entire ITS-1 – 5.8S – ITS-2 fragment (Figure 4). Alignment of ITS fragments needs to take into account many indels, even in this dataset with closely related species.

In Diplectanidae, the identification success for the 33 species was high for both COI and 28S (Table 1). In the COI dataset there was only a single incorrect identification. In the 28S dataset there were no misidentifications, but nine identifications were ambiguous because *Diplectanum nanus* Justine, 2007 and *D. parvum* Justine, 2008 share an 28S sequence despite an average COI divergence of 1.9%. In *Gyrodactylus*, the identification success of the eight species was 100% with all three markers.

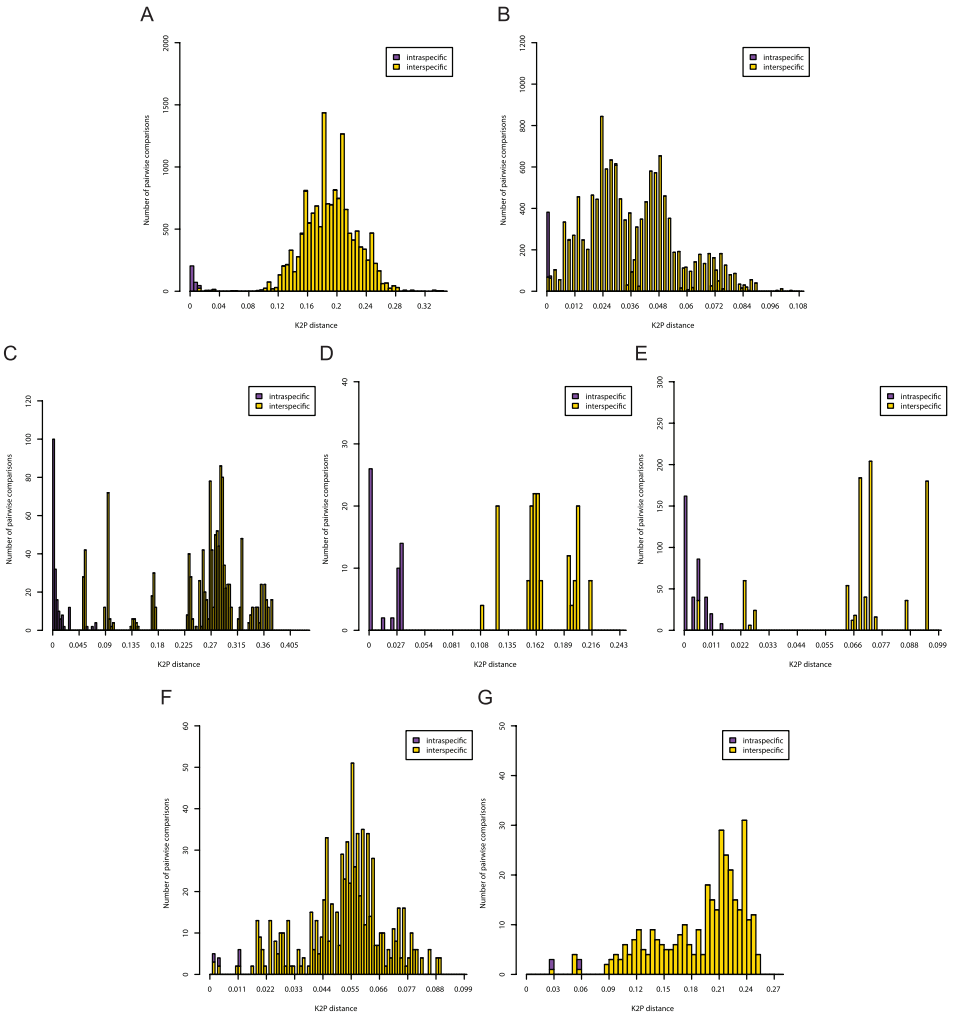


Figure 3. Pairwise distance (K2P) distributions of intra- and interspecific sequence divergences for the COI gene in Diplectanidae (**A**), 28S rDNA region in Diplectanidae (**B**), the COII gene in *Gyrodactylus* (**C**), the COI gene in *Gyrodactylus* (**D**), the ITS rDNA region in *Gyrodactylus* (**E**), the 28S rDNA region in *Gieysztoria* (**F**) and the ITS – 5.8S – ITS2 rDNA region in *Gieysztoria* (**G**).

Discussion

In order for COI to function as a widely used barcoding marker, ideally primers should be available allowing amplification of the gene under standard conditions for a wide range of taxa. For rhabdocoels, a taxon where the acquisition of COI data is clearly lagging behind, our results show that using universal COI barcoding primers is problematic. Universal primers seem to amplify non-rhabdocoel DNA much more efficient. This leads to contamination problems where several sequences are present in the PCR product and the resulting chromatogram becomes difficult to interpret.

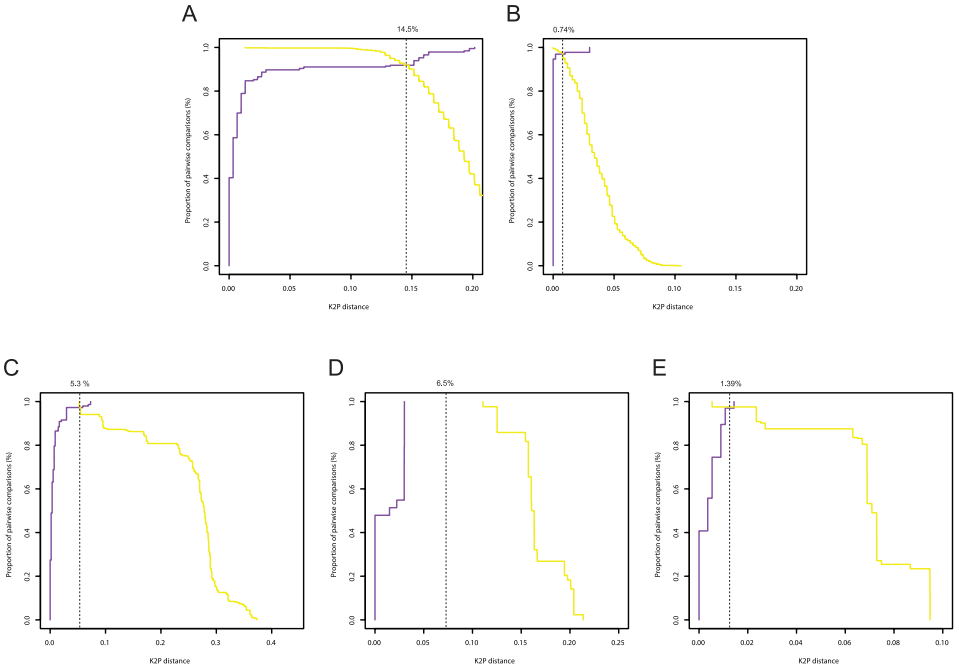


Figure 4. Optimum threshold defined by the intersection between the cumulative frequency distribution curves of the intraspecific (purple) and the interspecific (yellow) pairwise distances for the COI gene in Diplectanidae (**A**), 28S rDNA region in Diplectanidae (**B**), the COII gene in *Gyrodactylus* (**C**), the COI gene in *Gyrodactylus* (**D**), the ITS rDNA region in *Gyrodactylus* (**E**).

Table 1. Identification success, with best compromise threshold (BCTh) values used, as determined via the best match (BM) and best close-match (BMC) criteria.

Dataset		Threshold (%)	Correct	Ambiguous	Incorrect	No match closer than threshold
Diplectanidae COI	BM	-	116 (99,15%)	0	1 (0,85%)	-
	BCM	14,50%	116 (99,15%)	0	1 (0,85%)	0
Diplectanidae 28S	BM	-	108(92,3%)	9 (7,69%)	0	-
	BCM	0,74%	107(91.45%)	9 (7,69%)	0	1 (0,85%)
Gyrodactylus COII	BM	-	38 (100%)	0	0	-
	BCM	5,30%	38 (100%)	0	0	0
Gyrodactylus COI	BM	-	15 (100%)	0	0	-
	BCM	6,50%	15 (100%)	0	0	0
Gyrodactylus ITS	BM	-	35 (100%)	0	0	-
	BCM	1,39%	35 (100%)	0	0	0

Problems with limited success of universal barcoding primers and with contamination by associated fauna are known from other animals as well, e.g. marine free-living nematodes (Derycke et al. 2010). Because of the high variation within the obtained rhabdocoel COI sequences it was also not possible to use this dataset to develop in-

ternal rhabdocoel-specific primers. Efforts to establish COI barcoding protocols in rhabdocoels should therefore probably focus on smaller taxonomic entities within this taxon.

What can alternative markers offer?

Though less acute than in rhabdocoels, amplification success in our view is the biggest limitation to a wider use of COI barcoding in monogeneans as well. Despite the recent increase in published monogenean mitogenomes (e.g. Huyse et al. 2007, 2008, Park et al. 2007, Plaisance et al. 2007, Perkins et al. 2010, Kang et al. 2012, Zhang et al. 2011, 2012), universal COI barcoding primers have not yet been developed for monogeneans, let alone for flatworms in general. While advances in mitogenomics will hopefully facilitate the development of primer combinations for additional molecular markers, the mitochondrial genomes seem variable to such an extent that such primers will often forcibly be taxon-specific (as exemplified for the Folmer et al. (1994) region of COI in Figure 1). Hence, we agree with McManus et al. (2004) that the "post-genomic era" has clearly not dawned yet for parasitic flatworms, or, we suggest, flatworms in general. Moreover, as barcoding should be a user-friendly technique, ideally suitable also to the non-molecularly trained, relying on a set of taxon- or marker-specific protocols does not seem an ideal way forward. The nuclear rDNA region, including the ITS, is a better candidate in terms of widely suitable and versatile molecular markers. Their continued (and increased) use would of course exacerbate the existing "bandwagon effect" (e.g. Nieto Feliner and Rosselló 2007), but this need not be a problem. Indeed, any barcoding approach is only as good as the resulting available datasets. This limitation is evident even in better-studied taxa like fishes, for which barcoding efforts are considerable (Ward et al. 2009, Taylor and Harris 2012). While widely used COI barcoding primers are available (Ward et al. 2005), Vanhove et al. (2011a) demonstrated that for gobies, even within Europe, mitochondrial 12S and 16S rDNA yielded a bigger reference dataset and were hence better suited for phylogenetic assignment of unidentified species. Needless to say, similar problems exist for helminths (e.g. Palesse et al. 2011: philometrid nematodes). In contrast to mitochondrial markers, rDNA sequences can very easily be retrieved from both Monogenea and Rhabdocoela. The number of monogenean or rhabdocoel flatworms covered by rDNA sequences presently far outnumbers those for which COI data are available. This is clearly illustrated by the number of sequences available in GenBank (on 29 November 2013): a) Rhabdocoela: 233 18S, 144 28S, 0 ITS-1, 0 ITS-2 and 4 COI and b) Monogenea: 2298 rDNA and 1250 COI sequences, of which one-third from only three species: *Gyrodactylus salaris* Malmberg, 1957, *G. arcuatus* Bychowsky, 1933 and *Gotocotyla sawara* Ishii, 1936. Despite the importance of reference datasets, this in itself may not be an argument to favour rDNA over COI as a barcoding marker. The information content of the respective markers should be compared.

Our analysis of the distributions of intra- and interspecific K2P-sequence divergence shows that, in most cases, there is no clear DNA barcode gap in either COI

or rDNA. However, since coalescent depths are known to vary among species, such overlap is to be expected and has indeed been reported in many other taxa (see, for example, Wiemers and Fiedler 2007, Virgilio et al. 2010, Breman et al. 2013). As Collins and Cruickshank (2013) recently argued, this lack of a barcode gap does not necessarily mean that these markers are not suited for species level identifications because there might still exist a "local barcode gap".

Our analyses of Diplectanidae and *Gyrodactylus* show that both rDNA and mitochondrial markers can be highly effective for species identification. It is clear that the slower evolutionary rate of the rDNA markers does not necessarily make them less suited for DNA barcoding. We therefore suggest, also for monogeneans, to continue using rDNA markers. Both the 28S and ITS region could potentially be used as barcode marker. Our analysis of *Gieysztoria* shows that the faster evolving ITS region does not necessarily show a more pronounced DNA barcode gap (Figure 3). The choice between both markers should therefore be based on the species that need to be identified. The 28S region can be aligned more easily between distantly related species than the ITS region. Indeed, alignment problems have been reported for ITS in several monogeneans (Desdevises et al. 2000, Poisot et al. 2011). This limits the applicability of this marker to phylogeny reconstruction and genetic distance calculation, but does not preclude its use in species recognition. Indeed, while different rates of concerted evolution cause difficulties in phylogeographic analyses (Harris and Crandall 2000), various homogenization mechanisms most often lead to clear distinctions at the species level (Odorico and Miller 1997). Likewise, while the non-coding nature of ITS allows substantial length differences possibly precluding reliable alignment, this is of less concern when working with closely related species (Nieto Felliner and Rosselló 2007).

Yet, the slower evolving rDNA genes might not be suited to discriminate between very recently diverged species. More conservative than ITS-1 and ITS-2, they are more suitable for deeper phylogeny reconstruction than for example the detection of cryptic species. This was evident in our analysis of Diplectanidae where *D. nanus* and *D. parvum* shared a 28S rDNA sequence while their difference amounted to a maximum of 3.2% in COI. However, in most cases, the 18S and 28S rRNA genes can also differentiate among closely related monogenean and rhabdocoel species (e.g. Gilmore et al. 2012, Van Steenkiste et al. 2013). There are exceptions to this rule (e.g. Přikrylová et al. 2013 for identical 18S sequences in recently diverged *Gyrodactylus* species), which is not surprising given the extensive divergence rate variation throughout Monogenea (Olson and Littlewood 2002).

Unfortunately, because rDNA has exclusively been used in a phylogenetic setting in Rhabdocoela, there is too little information about intraspecific distances to formally test its use as a barcoding marker for rhabdocoels. We suggest that further efforts to establish a DNA barcoding protocol focus on the 28S rDNA region instead of the ITS region because the overlap between intra- and interspecific distances is not smaller in the faster evolving ITS, and because the ITS region is very difficult to align, even between closely related sequences.

The way forward

Given the different applicability of the various markers, we suggest the approach offered by Moszczynska et al. (2009) for digeneans would be a suitable way forward in our target organisms as well. Widely applicable rDNA primers could be used in an initial, prospective step. Once the organisms in question have been assigned to a lower taxonomic rank, appropriate COI primers for the taxon can be selected, when sequences from a faster-evolving and mitochondrial marker are desired, for example to assess for recently diverged or cryptic species. This is, of course, highly dependent on the availability of such COI primers, which we showed to be problematic in certain taxa. Although this differs from the "classical" approach of barcoding with a standard marker and protocol, a combined use of COI with portions of the nuclear rDNA region fulfills most promises of DNA barcoding in monogeneans and rhabdocoels.

Acknowledgements

Walter A. Boeger (Universidade Federal do Paraná, Brazil), Thierry Backeljau, Marc De Meyer and Kurt Jordaens (Joint Experimental Molecular Unit, Royal Belgian Institute of Natural Sciences/Royal Museum for Central Africa, Belgium), Filip A.M. Volckaert (University of Leuven, Belgium) and Niels Van Steenkiste (Hasselt University, Belgium/Fisheries and Oceans Canada) are gratefully acknowledged for their input into this research. We thank Gontran Sonet (Royal Belgian Institute of Natural Sciences, Belgium) for providing some of the R-scripts and the anonymous reviewers who commented on this manuscript. T.H. was, at the time of writing, a post-doctoral fellow of the Research Foundation – Flanders (FWO-Vlaanderen). M.P.M.V. was supported by KU Leuven – VES/12/005 and by Research Programme G.0553.10 of the Research Foundation – Flanders, and is currently funded by Czech Science Foundation project no. P505/12/G112 (ECIP - Centre of excellence). This research received support from the SYNTHESYS Project (http://www.synthesys.info/) which is financed by European Community Research Infrastructure Action under the FP7 Integrating Activities Programme. Diplectanid molecular analyses were supported by the "Service de Systématique Moléculaire" of the Muséum national d'histoire naturelle (CNRS UMS 2700) and the network "Bibliothèque du Vivant" funded by the CNRS, the Muséum national d'histoire naturelle, the INRA and the CEA (Genoscope).

References

Barber KE, Mkoji GM, Loker ES (2000) PCR-RFLP analysis of the ITS2 region to identify *Schistosoma haematobium* and *S. bovis* from Kenya. American Journal of Tropical Medicine and Hygiene 62: 434–440. www.ajtmh.org/content/62/4/434

Besansky NJ, Severson DW, Ferdig MT (2003) DNA barcoding of parasites and invertebrate disease vectors: what you don't know can hurt you. Trends in Parasitology 19: 545–546. doi: 10.1016/j.pt.2003.09.015

Brant SV, Loker ES (2009) Molecular systematic of the avian schistosome genus *Trichobilharzia* (Trematoda: Schistosomatidae) in North America. Journal of Parasitology 95: 941–63. doi: 10.1645/GE-1870.1

Breman FC, Jordaens K, Sonet G, Nagy ZT, Van Houdt J, Louette M (2013) DNA barcoding and evolutionary relationships in *Accipiter* Brisso, 1760 (Aves, Falconiformes: Accipitridae) with a focus on African and Eurasian representatives. Journal of Ornithology 154: 265–287. doi: 10.1007/s10336-012-0892-5

Brooks DR, Hoberg EP (2001) Parasite systematics in the 2t century: opportunities and obstacles. Trends in Parasitology 17: 273–275. doi: 10.1016/S1471-4922(01)01894-3

Bueno Silva M (2011) Cofilogeografia: estruturação geográfica, demografia histórica e associação entre species de *Gyrodactylus* (Monogenoidea: Gyrodactylidae) e hospedeiros *Scleromystax* (Siluriformes: Callichthyidae). PhD thesis, Universidade Federal do Paraná, Curitiba, Brazil.

Casu M, Lai T, Sanna D, Cossu P, Curini-Galletti M (2009) An integrative approach to the taxonomy of the pigmented European *Pseudomonocelis* Meixner, 1943 (Platyhelminthes: Proseriata). Biological Journal of the Linnean Society 98: 907–922. doi: 10.1111/j.1095-8312.2009.01316.x

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. Molecular Ecology Resources 13: 969–975. doi: 10.1111/1755-0998.12046

Cunningham CO (1997) Species variation within the internal transcribed spacer (ITS) region of *Gyrodactylus* (Monogenea: Gyrodactylidae) ribosomal RNA genes. Journal of Parasitology 83: 215–219. http://www.jstor.org/stable/3284442, doi: 10.2307/3284442

Derycke S, Vanaverbeke J, Rigaux A, Backeljau T, Moens T (2010) Exploring the use of cytochrome oxidase c subunit I (COI) for DNA barcoding of free-living marine nematodes. PLoS ONE 5: e13716. doi: 10.1371/journal.pone.0013716

Desdevises Y, Jovelin R, Jousson O, Morand S (2000) Comparison of ribosomal DNA sequences of *Lamellodiscus* spp. (Monogenea, Diplectanidae) parasitizing *Pagellus* (Sparidae, Teleostei) in the North Mediterranean Sea: species divergence and coevolutionary interactions. International Journal for Parasitology 30: 741–746. doi: 10.1016/S0020-7519(00)00051-5

Elsasser SC, Floyd R, Hebert PDN, Schulte-Hostedde AI (2009) Species identification of North American guinea worms (Nematoda: *Dracunculus*) with DNA barcoding. Molecular Ecology Resources 9: 707–712. doi: 10.1111/j.1755-0998.2008.02393.x

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3: 294–299.

Fonseca VG, Carvalho GR, Sung W, Johnson HF, Power DM, Neill SP, Packer M, Blaxter ML, Lambshead PJD, Thomas WK, Creer S (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nature Communications 1: 98. doi: 10.1038/ncomms1095

García-Vásquez A, Hansen H, Shinn AP (2007) A revised description of *Gyrodactylus cichlidarum* Paperna, 1968 (Gyrodactylidae) from the Nile tilapia, *Oreochromis niloticus niloticus* (Cichlidae), and its synonymy with *G. niloticus* Cone, Arthur et Bondad-Reantaso, 1995. Folia Parasitologica 54: 129–140. http://hdl.handle.net/1893/10236, http://folia.paru.cas.cz/detail.php?id=20841

García-Vásquez A, Hansen H, Christison KW, Bron JE, Shinn AP (2011) Description of three new species of *Gyrodactylus* von Nordmann, 1832 (Monogenea) parasitizing *Oreochromis niloticus niloticus* (L.) and *O. mossambicus* (Peters) (Cichlidae). Acta Parasitologica 56: 20–33. doi: 10.2478/s11686-011-0005-2

Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Molecular Biology and Evolution 19: 748–761. http://mbe.oxfordjournals.org/content/19/5/748

Gilmore SR, Cone DK, Lowe G, King SF, Jones SRM, Abbott CL (2012) Molecular phylogeny of *Gyrodactylus* (Monogenea) parasitizing fishes in fresh water, estuarine, and marine habitats in Canada. Canadian Journal of Zoology 90: 776–786. doi: 10.1139/z2012-040

Hansen H, Bakke TA, Bachmann L (2007) DNA taxonomy and barcoding of monogenean parasites: lessons from *Gyrodactylus*. Trends in Parasitology 23: 363–367. doi: 10.1016/j.pt.2007.06.007

Harris DJ, Crandall KA (2000) Intragenomic variation within ITS1 and ITS2 of freshwater crayfishes (Decapoda: Cambaridae): implications for phylogenetic and microsatellite studies. Molecular Biology and Evolution 17: 284–291. http://mbe.oxfordjournals.org/content/17/2/284, doi: 10.1093/oxfordjournals.molbev.a026308

Hebert PDN, Cywinska A, Ball SL, de Waard JR (2003a) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B 270: 313–322. doi: 10.1098/rspb.2002.2218

Hebert PDN, Ratnasingham S, de Waard JR (2003b) Barcoding animal life: cytochrome *c* oxidase subunit I divergences among closely related species. Proceedings of the Royal Society of London B (Supplement) 270: S96-S99. doi: 10.1098/rsbl.2003.0025

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. PLoS Biology 2: e312. doi: 10.1371/journal.pbio.0020312

Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Quarterly Review of Biology 66: 411–453. http://www.jstor.org/stable/2831326, doi: 10.1086/417338

Hoberg EP (1997) Phylogeny and historical reconstruction: host-parasite systems as keystones in biogeography and ecology. In: Reaka-Kudla ML, Wilson DE, Wilson EO (Eds) Biodiversity II: understanding and protecting our biological resources. Joseph Henry Press, Washington, D.C., 243–262.

Huyse T, Malmberg G (2004) Molecular and morphological comparisons between *Gyrodactylus ostendicus* sp. nov. (Monogenea: Gyrodactylidae) on *Pomatoschistus microps* (Krøyer) and *G. harengi* Malmberg, 1957 on *Clupea harengus membras* L. Systematic Parasitology 58: 105–113. doi: 10.1023/B:SYPA.0000029423.68703.43

Huyse T, Pampoulie C, Audenaert V, Volckaert FAM (2006) First report of *Gyrodactylus* spp. (Platyhelminthes: Monogenea) in the western Mediterranean sea: molecular and morphological descriptions. Journal of Parasitology 92: 682–690. doi: 10.1645/GE-690R.1

Huyse T, Plaisance L, Webster BL, Mo TA, Bakke TA, Bachmann L, Littlewood DTJ (2007) The mitochondrial genome of *Gyrodactylus salaris* (Platyhelminthes: Monogenea), a pathogen of Atlantic salmon (*Salmo salar*). Parasitology 134: 739–747. doi: 10.1017/S0031182006002010

Huyse T, Buchmann K, Littlewood DTJ (2008) The mitochondrial genome of *Gyrodactylus derjavinoides* (Platyhelminthes: Monogenea) – a mitogenomic approach for *Gyrodactylus* species and strain identification. Gene 417: 27–34. doi: 10.1016/j.gene.2008.03.008

Jörger KM, Schrödl M (2013) How to describe a cryptic species? Practical challenges of molecular taxonomy. Frontiers in Zoology 10: 59. doi: 10.1186/1742-9994-10-59

Jovelin R, Justine J-L (2001) Phylogenetic relationships within the polyopisthocotylean monogeneans (Platyhelminthes) inferred from partial 28S rDNA sequences. International Journal for Parasitology 31: 393–401. doi: 10.1016/S0020-7519(01)00114-X

Kang S, Kim J, Lee J, Kim S, Min G-S, Park J-K (2012) The complete mitochondrial genome of an ectoparasitic monopisthocotylean fluke *Benedenia hoshinai* (Monogenea: Platyhelminthes). Mitochondrial DNA 23: 176–178. doi: 10.3109/19401736.2011.588223

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120. doi: 10.1007/BF01731581

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948. doi: 10.1093/bioinformatics/btm404

Larsson K, Ahmadzadeh A, Jondelius U (2008) DNA taxonomy of Swedish Catenulida (Platyhelminthes) and a phylogenetic framework for catenulid classification. Organisms, Diversity & Evolution 8: 399–412. doi: 10.1016/j.ode.2008.09.003

Lázaro EM, Sluys R, Pala M, Angela Stocchino G, Baguñà J, Riutort M (2009) Molecular barcoding and phylogeography of sexual and asexual freshwater planarians of the genus *Dugesia* in the Western Mediterranean (Platyhelminthes, Tricladida, Dugesiidae). Molecular Phylogenetics and Evolution 52: 835–845. doi: 10.1016/j.ympev.2009.04.022

Lefébure T, Douady CJ, Gouy M, Gilbert J (2006) Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. Molecular Phylogenetics and Evolution 40: 435–447. doi: 10.1016/j.ympev.2006.03.014

Leung TLF, Donald KM, Keeney DB, Koehler AV, Peoples RC, Poulin R (2009) Trematode parasites of Otago Harbour (New Zealand) soft-sediment intertidal ecosystems: life cycles, ecological roles and DNA barcodes. New Zealand Journal of Marine and Freshwater Research 43: 857–865. doi: 10.1080/00288330909510044

Li M, Shi S-F, Brown CL, Yang T-B (2011) Phylogeographical pattern of *Mazocraeoides gonialosae* (Monogenea, Mazocraeidae) on the dotted gizzard shad, *Konosirus punctatus*, along the coast of China. International Journal for Parasitology 41: 1263–1272. doi: 10.1016/j.ijpara.2011.07.012

Littlewood DTJ (2008) Platyhelminth systematics and the emergence of new characters. Parasite 15: 333–341. doi: 10.1051/parasite/2008153333

Littlewood DTJ, Rohde K, Clough KA (1997) Parasite speciation within or between host species? Phylogenetic evidence from site-specific polystome monogeneans. International Journal for Parasitology 27: 1289–1297. doi: 10.1016/S0020-7519(97)00086-6

Locke SA, McLaughlin JD, Dayanandan S, Marcogliese DJ (2010a) Diversity and specificity in *Diplostomum* spp. metacercariae in freshwater fishes revealed by cytochrome *c* oxidase I and internal transcribed spacer sequences. International Journal for Parasitology 40: 333–343. doi: 10.1016/j.ijpara.2009.08.012

Locke SA, McLaughlin JD, Marcogliese DJ (2010b) DNA barcodes show cryptic diversity and a potential physiological basis for host specificity among Diplostomoidea (Platyhelminthes: Digenea) parasitizing freshwater fishes in the St. Lawrence River, Canada. Molecular Ecology 19: 2813–2827. doi: 10.1111/j.1365-294X.2010.04713.x

Lockyer AE, Olson PD, Littlewood DTJ (2003a) Utility of complete large and small subunit rRNA genes in resolving the phylogeny of the Neodermata (Platyhelminthes): implications and a review of the cercomer theory. Biological Journal of the Linnean Society 78: 155–171. doi: 10.1046/j.1095-8312.2003.00141.x

Lockyer AE, Olson PD, Østergaard P, Rollinson D, Johnston DA, Attwood SW, Southgate VR, Horak P, Snyder SD, Le TH, Agatsuma T, McManus DP, Carmichael AC, Naem S, Littlewood DTJ (2003b) The phylogeny of the Schistosomatidae based on three genes with emphasis on the interrelationships of *Schistosoma* Weinland, 1858. Parasitology 126: 203–224. doi: 10.1017/S0031182002002792

Lupi R, D'Onorio de Meo P, Picardi E, D'Antonio M, Paoletti D, Castrignanò T, Pesole G, Gissi C (2010) MitoZoa: A curated mitochondrial genome database of metazoans for comparative genomics studies. Mitochondrion 10: 192–199. doi: 10.1016/j.mito.2010.01.004

Marcogliese DJ (2004) Parasites: small players with crucial roles in the ecological theater. EcoHealth 1: 151–164. doi: 10.1007/s10393-004-0028-3

Matějusová I, Gelnar M, McBeath AJA, Collins CM, Cunningham CO (2001) Molecular markers for gyrodactylids (Gyrodactylidae: Monogenea) from five fish families (Teleostei). International Journal for Parasitology 31: 738–745. doi: 10.1016/S0020-7519(01)00176-X

McManus DP, Le TH, Blair D (2004) Genomics of parasitic flatworms. International Journal for Parasitology 34: 153–158. doi: 10.1016/j.ijpara.2003.11.003

Meier R, Shiyang K, Vaidya G, Ng PKL (2006) DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology 55: 715–728. doi: 10.1080/10635150600969864

Meinilä M, Kuusela J, Ziętara M, Lumme J (2002) Primers for amplifying ~820 bp of highly polymorphic mitochondrial COI gene of *Gyrodactylus salaris*. Hereditas 137: 72–74. doi: 10.1034/j.1601-5223.2002.1370110.x

Meusnier I, Singer GAC, Landry J-F, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 9: 214. doi: 10.1186/1471-2164-9-214

Mladineo I, Šegvić T, Grubišić L (2009) Molecular evidence for the lack of transmission of the monogenean *Sparicotyle chrysophrii* (Monogenea, Polyopisthocotylea) and isopod *Ceratothoa oestroides* (Crustacea, Cymothoidae) between wild bogue (*Boops boops*) and cage-

reared sea bream (*Sparus aurata*) and sea bass (*Dicentrarchus labrax*). Aquaculture 295: 160–167. doi: 10.1016/j.aquaculture.2009.07.017

Mladineo I, Šegvić-Bubić T, Stanić R, Desdevises Y (2013) Morphological plasticity and phylogeny in a monogenean parasite transferring between wild and reared fish populations. PLoS ONE 8: e62011. doi: 10.1371/journal.pone.0062011

Moritz C, Cicero C (2004) DNA barcoding: promise and pitfalls. PLoS Biology 2: e354. doi: 10.1371/journal.pbio.0020354

Moszczynska A, Locke SA, McLaughlin JD, Marcogliese DJ, Crease TJ (2009) Development of primers for the mitochondrial cytochrome *c* oxidase I gene in digenetic trematodes (Platyhelminthes) illustrates the challenge of barcoding parasitic helminths. Molecular Ecology Resources 9: 75–82. doi: 10.1111/j.1755-0998.2009.02634.x

Nadler SA, Pérez-Ponce de Léon G (2011) Integrating molecular and morphological approaches for characterizing parasite cryptic species: implications for parasitology. Parasitology 138: 1688–1709. doi: 10.1017/S003118201000168X

Nieto Feliner G, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Molecular Phylogenetics and Evolution 44: 911–919. doi: 10.1016/j.ympev.2007.01.013

Nolan MJ, Cribb TH (2005) The use and implications of ribosomal DNA sequencing for the discrimination of digenean species. Advances in Parasitology 60: 101–163. doi: 10.1016/S0065-308X(05)60002-4

Odorico DM, Miller DJ (1997) Variation in the ribosomal internal transcribed spacers and 5. rDNA among five species of *Acropora* (Cnidaria; Scleractinia): patterns of variation consistent with reticulate evolution. Molecular Biology and Evolution 14: 465–473. http://mbe.oxfordjournals.org/content/14/5/465

Olson PD, Littlewood DTJ (2002) Phylogenetics of the Monogenea – evidence from a medley of molecules. International Journal for Parasitology 32: 233–244. doi: 10.1016/S0020-7519(01)00328-9

Paetow L, Cone D, Huyse T, McLaughlin J, Marcogliese D (2009) Morphology and molecular taxonomy of *Gyrodactylus jennyae* n. sp (Monogenea) from tadpoles of captive *Rana catesbeiana* Shaw (Anura), with a review of the species of *Gyrodactylus* Nordmann, 1832 parasitising amphibians. Systematic Parasitology 73: 219–227. doi: 10.1007/s11230-009-9183-9

Paladini G, Cable J, Fioravanti ML, Faria PJ, Di Cave D, Shinn AP (2009) *Gyrodactylus orecchiae* sp. n. (Monogenea: Gyrodactylidae) from farmed populations of gilthead seabream (*Sparus aurata*) in the Adriatic Sea. Folia Parasitologica 56: 21–28. http://hdl.handle.net/1893/1419, http://folia.paru.cas.cz/detail.php?id=21138

Paladini G, Cable J, Fioravanti ML, Faria PJ, Shinn AP (2010) The description of *Gyrodactylus corleonis* sp. n. and *G. neretum* sp. n. (Platyhelminthes: Monogenea) with comments on other gyrodactylids parasitising pipefish (Pisces: Syngnathidae). Folia Parasitologica 57: 17–30. http://hdl.handle.net/1893/9998, http://folia.paru.cas.cz/detail.php?id=21368

Paladini G, Hansen H, Fioravanti ML, Shinn AP (2011a) *Gyrodactylus longipes* n. sp. (Monogenea: Gyrodactylidae) from farmed gilthead seabream (*Sparus aurata* L.) from the Mediterranean. Parasitology International 60: 410–418. doi: 10.1016/j.parint.2011.06.022

Paladini G, Huyse T, Shinn AP (2011b) *Gyrodactylus salinae* n. sp. (Platyhelminthes: Monogenea) infecting the south European toothcarp *Aphanius fasciatus* (Valenciennes) (Teleostei, Cyprinodontidae) from a hypersaline environment in Italy. Parasites & Vectors 4: 100. doi: 10.1186/1756-3305-4-100

Palesse S, Meadors WA, de Buron I, Roumillat WA, Strand AE (2011) Use of molecular tools in identification of philometrid larvae in fishes: technical limitations parallel our poor assessment of their biodiversity. Parasitology Research 109: 1725–1730. doi: 10.1007/s00436-011-2481-6

Park J-K, Kim K-H, Kang S, Kim W, Eom KS, Littlewood DTJ (2007) A common origin of complex life cycles in parasitic flatworms: evidence from the complete mitochondrial genome of *Microcotyle sebastis* (Monogenea: Platyhelminthes). BMC Evolutionary Biology 7: 11. doi: 10.1186/1471-2148-7-11

Perkins EM, Donnellan SC, Bertozzi T, Whittington ID (2010) Closing the mitochondrial circle on paraphyly of the Monogenea (Platyhelminthes) infers evolution in the diet of parasitic flatworms. International Journal for Parasitology 40: 1237–1245. doi: 10.1016/j.ijpara.2010.02.017

Plaisance L, Huyse T, Littlewood DTJ, Bakke TA, Bachmann L (2007) The complete mitochondrial DNA sequence of the monogenean *Gyrodactylus thymalli* (Platyhelminthes: Monogenea), a parasite of grayling (*Thymallus thymallus*). Molecular and Biochemical Parasitology 154: 190–194. doi: 10.1016/j.molbiopara.2007.04.012

Plaisance L, Rousset V, Morand S, Littlewood DTJ (2008) Colonization of pacific islands by parasites of low dispersal abilities: phylogeography of two monogenean species parasitizing butterflyfishes in the Indo-West Pacific Ocean. Journal of Biogeography 35: 76–87. doi: 10.1111/j.1365-2699.2007.01794.x

Poisot T, Verneau O, Desdevises Y (2011) Morphological and molecular evolution are not linked in *Lamellodiscus* (Plathyhelminthes, Monogenea). PLoS ONE 6: e26252. doi: 10.1371/journal.pone.0026252

Pouyaud L, Desmarais E, Deveney M, Pariselle A (2006) Phylogenetic relationships among monogenean gill parasites (Dactylogyridea, Ancyrocephalidae) infesting tilapiine hosts (Cichlidae): systematic and evolutionary implications. Molecular Phylogenetics and Evolution 38: 241–249. doi: 10.1016/j.ympev.2005.08.013

Přikrylová I, Matějusová I, Musilová N, Gelnar M (2009a) *Gyrodactylus* species (Monogenea: Gyrodactylidae) on the cichlid fishes of Senegal, with the description of *Gyrodactylus ergensi* sp. nov. from Mango tilapia, *Sarotherodon galilaeus* L. (Teleostei: Cichlidae). Parasitology Research 106: 1–6. doi: 10.1007/s00436-009-1600-0

Přikrylová I, Matějusová I, Musilová N, Gelnar M, Harris PD (2009b) A new gyrodactylid (Monogenea) genus on gray bichir, *Polypterus senegalus* (Polypteridae) from Senegal (West Africa). Journal of Parasitology 95: 555–560. doi: 10.1645/GE-1652.1

Přikrylová I, Blažek R, Vanhove MPM (2012a) An overview of the *Gyrodactylus* (Monogenea: Gyrodactylidae) species parasitizing African catfishes, and their morphological and molecular diversity. Parasitology Research 110: 1185–1200. doi: 10.1007/s00436-011-2612-0

Přikrylová I, Blažek R, Gelnar M (2012b) *Gyrodactylus malalai* sp. nov. (Monogenea: Gyrodactylidae) from Nile tilapia, *Oreochromis niloticus* (L.) and Redbelly tilapia, *Tilapia zillii* (Gervais)

(Teleostei: Cichlidae) in the Lake Turkana, Kenya. Acta Parasitologica 57: 122–130. doi: 10.2478/s11686-012-0017-6

Přikrylová I, Vanhove MPM, Janssens SB, Billeter PA, Huyse T (2013) Tiny worms from a mighty continent: high diversity and new phylogenetic lineages of African monogeneans. Molecular Phylogenetics and Evolution 67: 43–52. doi: 10.1016/j.ympev.2012.12.017

Pugachev ON, Gerasev PI, Gussev AV, Ergens R, Khotenowsky I (2009) Guide to Monogenoidea of freshwater fish of Palaearctic and Amur Regions. Ledizione-Ledi Publishing, Milan, 564 pp.

R Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Radulovici AE, Archambault P, Dufresne F (2010) DNA barcodes for marine biodiversity: moving fast forward? Diversity 2: 450–472. doi: 10.3390/d2040450

Řehulková E, Mendlová M, Šimková A (2013) Two new species of *Cichlidogyrus* (Monogenea: Dactylogyridae) parasitizing the gills of African cichlid fishes (Perciformes) from Senegal: morphometric and molecular characterization. Parasitology Research 112: 1399–1410. doi: 10.1007/s00436-013-3291-9

Rokicka M, Lumme J, Ziętara MS (2009) Two new Antarctic *Gyrodactylus* species (Monogenoidea): description and phylogenetic characterization. Journal of Parasitology 95: 1112–1119. doi: 10.1645/GE-2002.1

Sanna D, Lai T, Francalacci P, Curini-Galletti M, Casu M (2009) Population structure of the *Monocelis lineata* (Proseriata, Monocelididae) species complex assessed by phylogenetic analysis of the mitochondrial cytochrome *c* oxidase subunit I (COI) gene. Genetics and Molecular Biology 32: 864–867. doi: 10.1590/S1415-47572009005000076

Schelkle B, Paladini G, Shinn AP, King S, Johnson M, van Oosterhout C, Mohammed RS, Cable J (2011) *Ieredactylus rivuli* gen. et sp. nov. (Monogenea, Gyrodactylidae) from *Rivulus hartii* (Cyprinodontiformes, Rivulidae) in Trinidad. Acta Parasitologica 56: 360–370. doi: 10.2478/s11686-011-0081-3

Schockaert ER (1996) Turbellarians. In: Hall GS (Ed) Methods for the examination of organismal diversity in soils and sediments. CAB International, Wallingford, 221–226.

Schoelinck C (2012) Systématique évolutive des Diplectanidae (Plathelminthes, Monogenea) parasites des Mérous des récifs coralliens (Perciformes, Serranidae). PhD thesis, Université Pierre et Marie Curie, Paris, France.

Schoelinck C, Cruaud C, Justine J-L (2012) Are all species of *Pseudorhabdosynochus* strictly host specific? A molecular study. Parasitology International 61: 356–359. doi: 10.1016/j.parint.2012.01.009

Sonnenberg R, Nolte AW, Tautz D (2007) An evaluation of LSU D1-D2 sequences for their use in species identification. Frontiers in Zoology 4: 6. doi: 10.1186/1742-9994-4-6

Stefani F, Aquaro G, Azzurro E, Colorni A, Galli P (2012) Patterns of genetic variation of a Lessepsian parasite. Biological Invasions 14: 1725–1736. doi: 10.1007/s10530-012-0183-3

Stoeckle M (2003) Taxonomy, DNA, and the bar code of life. BioScience 53: 796–797. http://www.jstor.org/stable/10.1641/0006-3568%282003%29053%5B0796%3ATDATBC%5D2.0.CO%3B2

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28: 2731–2739. doi: 10.1093/molbev/msr121

Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12: 377–388. doi: 10.1111/j.1755-0998.2012.03119.x

Telford MJ, Herniou EA, Russell RB, Littlewood DTJ (2000) Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. Proceedings of the National Academy of Sciences of the USA 97: 11359–11364. doi: 10.1073/pnas.97.21.11359

Telford MJ, Lockyer AE, Cartwright-Finch C, Littlewood DTJ (2003) Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. Proceedings of the Royal Society of London B 270: 1077–1083. doi: 10.1098/rspb.2003.2342

Vanhove MPM (2012) Species flocks and parasite evolution. Towards a co-phylogenetic analysis of monogenean flatworms of cichlids and gobies. PhD thesis, KU Leuven, Leuven, Belgium.

Vanhove MPM, Kovačić M, Koutsikos NE, Zogaris S, Vardakas LE, Huyse T, Economou AN (2011a) First record of a landlocked population of marine *Millerigobius macrocephalus* (Perciformes: Gobiidae): observations from a unique spring-fed karstic lake (Lake Vouliagmeni, Greece) and phylogenetic positioning. Zoologischer Anzeiger 250: 195–204. doi: 10.1016/j.jcz.2011.03.002

Vanhove MPM, Snoeks J, Volckaert FAM, Huyse T (2011b) First description of monogenean parasites in Lake Tanganyika: the cichlid *Simochromis diagramma* (Teleostei, Cichlidae) harbours a high diversity of *Gyrodactylus* species (Platyhelminthes, Monogenea). Parasitology 138: 364–380 (erratum in 138: 403). doi: 10.1017/S0031182010001356

Vanhove MPM, Economou AN, Zogaris S, Larmuseau MHD, Giakoumi S, Kalogianni E, Volckaert FAM, Huyse T (2012) Phylogenetics and biogeography of the Balkan "sand gobies" (Teleostei, Gobiidae): vulnerable species in need of taxonomic revision. Biological Journal of the Linnean Society 105: 73–91. doi: 10.1111/j.1095-8312.2011.01781.x

Vanhove MPM, Economou AN, Zogaris S, Giakoumi S, Zanella D, Volckaert FAM, Huyse T (2013) The *Gyrodactylus* (Monogenea, Gyrodactylidae) parasite fauna of freshwater sand gobies (Teleostei, Gobioidei) in their centre of endemism, with description of seven new species. Parasitology Research. doi: 10.1007/s00436-013-3693-8

Van Steenkiste N, Tessens B, Willems W, Backeljau T, Jondelius U, Artois T (2013) A comprehensive molecular phylogeny of Dalytyphloplanida (Platyhelminthes: Rhabdocoela) reveals multiple escapes from the marine environment and origins of symbiotic relationships. PLoS ONE 8: e59917. doi: 10.1371/journal.pone.0059917

Vaughan DB, Christison KW, Hansen H, Shinn AP (2010) *Gyrodactylus eyipayipi* sp. n. (Monogenea: Gyrodactylidae) from *Syngnathus acus* (Syngnathidae) from South Africa. Folia Parasitologica 57: 11–15. http://hdl.handle.net/1893/9958, http://folia.paru.cas.cz/detail.php?id=21366

Vilas R, Criscione CD, Blouin MS (2005) A comparison between mitochondrial DNA and the ribosomal internal transcribed regions in prospecting for cryptic species of platyhelminth parasites. Parasitology 131: 839–846. doi: 10.1017/S0031182005008437

Virgilio M, Backeljau T, Nevado B, De Meyer M (2010) Comparative performance of DNA barcoding across insect orders. BMC Bioinformatics 11: 206. doi: 10.1186/1471-2105-11-206

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) Barcoding Australia's fish species. Philosophical Transactions of the Royal Society B 360: 1847–1857. doi: 10.1098/rstb.2005.1716

Ward RD, Hanner R, Hebert PDN (2009) The campaign to DNA barcode all fishes, FISH-BOL. Journal of Fish Biology 74: 329–356. doi: 10.1111/j.1095-8649.2008.02080.x

White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: White TJ (Ed) PCR protocols: a guide to methods and applications. Academic Press, San Diego, 315–322.

Whittington ID (1998) Diversity "down under": monogeneans in the Antipodes (Australia) with a prediction of monogenean biodiversity worldwide. International Journal for Parasitology 28: 1481–1493. doi: 10.1016/S0020-7519(98)00064-2

Wiemers M, Fiedler K (2007) Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). Frontiers in Zoology 4: 8. doi: 10.1186/1742-9994-4-8

Willems WR, Wallberg A, Jondelius U, Littlewood DTJ, Backeljau T, Schockaert ER, Artois TJ (2006) Filling a gap in the phylogeny of flatworms: relationships within the Rhabdocoela (Platyhelminthes), inferred from 18S ribosomal DNA sequences. Zoologica Scripta 35: 1–17. doi: 10.1111/j.1463-6409.2005.00216.x

Windsor DA (1998) Most of the species on Earth are parasites. International Journal for Parasitology 28: 1939–1941. doi: 10.1016/S0020-7519(98)00153-2

Zhang J, Wu X, Xie M, Xu X, Li A (2011) The mitochondrial genome of *Polylabris halichoeres* (Monogenea: Microcotylidae). Mitochondrial DNA 22: 3–5. doi: 10.3109/19401736.2011.588223

Zhang J, Wu X, Xie M, Li A (2012) The complete mitochondrial genome of *Pseudochauhanea macrorchis* (Monogenea: Chauhaneidae) revealed a highly repetitive region and a gene rearrangement hot spot in Polyopisthocotylea. Molecular Biology Reports 39: 8115–8125. doi: 10.1007/s11033-012-1659-z

Ziętara MS, Lumme J (2002) Speciation by host-switching and adaptive radiation in a fish parasite genus *Gyrodactylus* (Monogenea, Gyrodactylidae). Evolution 56: 2445–2458. doi: 10.1111/j.0014-3820.2002.tb00170.x

Ziętara MS, Lebedeva D, Muñoz G, Lumme J (2012) A monogenean fish parasite, *Gyrodactylus chileani* n. sp., belonging to a novel marine species lineage found in the South-Eastern Pacific and the Mediterranean and North Seas. Systematic Parasitology 83: 159–167. doi: 10.1007/s11230-012-9379-2

Appendix 1

Supplementary table 1. (doi: 10.3897/zookeys.365.5776.app1) File format: Microsoft Excel file (xls).

Explanation note: List of clones sequenced in this study with species on which PCR was performed.

Citation: Vanhove MPM, Tessens B, Schoelinck C, Jondelius U, Littlewood DTJ, Artois T, Huyse T (2013) Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 355–379. doi: 10.3897/zookeys.365.5776 Supplementary table 1. doi: 10.3897/zookeys.365.5776.app1

Appendix 2

Supplementary table 2. (doi: 10.3897/zookeys.365.5776.app2) File format: Microsoft Excel file (xls).

Explanation note: Reference sequences downloaded from GenBank with accession numbers.

Citation: Vanhove MPM, Tessens B, Schoelinck C, Jondelius U, Littlewood DTJ, Artois T, Huyse T (2013) Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 355–379. doi: 10.3897/zookeys.365.5776 Supplementary table 2. doi: 10.3897/zookeys.365.5776.app2

Appendix 3

Supplementary table 3. (doi: 10.3897/zookeys.365.5776.app3) File format: Microsoft Excel file (xls).

Explanation note: List of species and number of sequences from each marker used in the monogenean test cases.

Citation: Vanhove MPM, Tessens B, Schoelinck C, Jondelius U, Littlewood DTJ, Artois T, Huyse T (2013) Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 355–379. doi: 10.3897/zookeys.365.5776 Supplementary table 3. doi: 10.3897/zookeys.365.5776.app3

Appendix 4

Supplementary table 4. (doi: 10.3897/zookeys.365.5776.app4) File format: Microsoft Excel file (xls).

Explanation note: List of species and GenBank accession numbers from the genus *Gieysztoria*.

Citation: Vanhove MPM, Tessens B, Schoelinck C, Jondelius U, Littlewood DTJ, Artois T, Huyse T (2013) Problematic barcoding in flatworms: A case-study on monogeneans and rhabdocoels (Platyhelminthes). In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 355–379. doi: 10.3897/zookeys.365.5776 Supplementary table 4. doi: 10.3897/zookeys.365.5776.app4

ZooKeys 365: 381–411 (2013)
doi: 10.3897/zookeys.365.5859
www.zookeys.org

REVIEW ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Reviewing population studies for forensic purposes: Dog mitochondrial DNA

Sophie Verscheure[1,2], Thierry Backeljau[2,3], Stijn Desmyter[1]

1 *National Institute of Criminalistics and Criminology, Vilvoordsesteenweg 100, B-1120, Brussels, Belgium*
2 *University of Antwerp (Evolutionary Ecology Group), Groenenborgerlaan 171, B-2020, Antwerp, Belgium*
3 *Royal Belgian Institute of Natural Sciences (OD "Taxonomy and Phylogeny" and JEMU), Vautierstraat 29, B-1000, Brussels, Belgium*

Corresponding author: *Sophie Verscheure* (sophie.verscheure@just.fgov.be)

Academic editor: *M. De Meyer* | Received 25 June 2013 | Accepted 14 December 2013 | Published 30 December 2013

Citation: Verscheure S, Backeljau T, Desmyter S (2013) Reviewing population studies for forensic purposes: Dog mitochondrial DNA. In: Nagy ZT, Backeljau T, De Meyer M, Jordaens K (Eds) DNA barcoding: a practical tool for fundamental and applied biodiversity research. ZooKeys 365: 381–411. doi: 10.3897/zookeys.365.5859

Abstract

The identification of dog hair through mtDNA analysis has become increasingly important in the last 15 years, as it can provide associative evidence connecting victims and suspects. The evidential value of an mtDNA match between dog hair and its potential donor is determined by the random match probability of the haplotype. This probability is based on the haplotype's population frequency estimate. Consequently, implementing a population study representative of the population relevant to the forensic case is vital to the correct evaluation of the evidence. This paper reviews numerous published dog mtDNA studies and shows that many of these studies vary widely in sampling strategies and data quality. Therefore, several features influencing the representativeness of a population sample are discussed. Moreover, recommendations are provided on how to set up a dog mtDNA population study and how to decide whether or not to include published data. This review emphasizes the need for improved dog mtDNA population data for forensic purposes, including targeting the entire mitochondrial genome. In particular, the creation of a publicly available database of qualitative dog mtDNA population studies would improve the genetic analysis of dog traces in forensic casework.

Keywords

Forensics, Mitochondrial DNA, Dog, Random match probability, Population study, Sampling strategy

Introduction

Dogs (*Canis lupus familiaris*) are common and widespread in human society and hence, dog trace material is frequently encountered in forensic casework. Usually, this trace material involves hair, which is easily dispersed either through immediate contact with a dog or indirectly via an intermediate carrier, thus leaving a signature of the dog. Consequently, determining whether a particular dog could have donated the hair found at a crime scene may provide associative evidence (dis)connecting victims and suspects. For example, dog hairs could have been transferred from a victim's clothes to the trunk of a perpetrator's car during transportation of a body. Linking these hairs to the victim's dog could connect the suspect to the crime.

Most dog hairs collected at crime scenes are naturally shed and are in the telogen phase. As such, because they contain only limited amounts of, usually degraded, nuclear DNA (nDNA), they are ill suited for nDNA analysis. Conversely, mainly as a result of its high copy number and much smaller size (Nass 1969, Bogenhagen and Clayton 1974), mitochondrial DNA (mtDNA) is quantitatively and qualitatively better preserved than nDNA in telogenic hairs and hence is far more suitable for analysis, as e.g. demonstrated in Gagneux et al. (1997) and Allen et al. (1998). To identify the mammal taxon that shed the hair, DNA barcoding can be applied through analysis of an mtDNA marker with little variation within and sufficient variation among taxa, often a part of cytochrome *b* or cytochrome *c* oxidase subunit I in forensics (Linacre and Tobe 2011). On the other hand, in order to individualize dog hairs as accurately as possible, it is necessary to analyze mtDNA regions that show high variability among dogs and low intra-individual variation (heteroplasmy). As for human traces, this type of analysis focuses on the non-coding control region or D-loop (Wilson et al. 1993, Holland and Parsons 1999), which in dog mtDNA comprises about 1200 bp consisting of two hypervariable regions (HV-I and HV-II) separated by a Variable Number of Tandem Repeats (VNTR) region (Figure 1). This VNTR is a 10 bp tandem repeat with variable repeat numbers, both between and within individuals (length heteroplasmy). Because of its high level of length heteroplasmy, this repeat region is mostly not considered in forensics (Fridez et al. 1999). Several publications illustrate forensic casework involving control region analysis of dog traces, such as Savolainen and Lundeberg (1999), Schneider et al. (1999), Branicki et al. (2002), Aaspõllu and Kelve (2003), Halverson and Basten (2005) and Scharnhorst and Kanthaswamy (2011).

In general, mtDNA is maternally inherited (Sato and Sato 2013). In theory, this means that all dogs sharing a maternal line have the same mtDNA haplotype barring mutations. Hence, a match between the mtDNA of the dog hair found at a crime scene and that of a dog suspected of donating the trace, may be due to either of three possibilities: (1) the dog hair from the crime scene is from the suspected donor, (2) the hair from the crime scene is from a dog of the same maternal lineage as the suspected donor, (3) the mtDNA from the crime scene is by coincidence identical to that of the suspected donor. In order to assess the evidential weight of a match under the last scenario, one must calculate the haplotype's random match probability, the probability

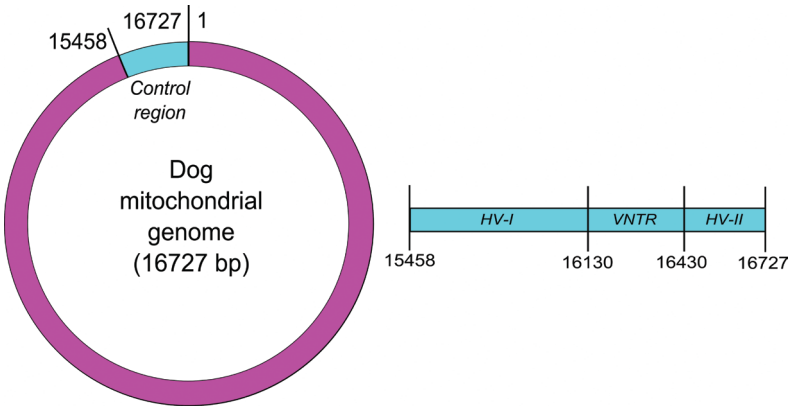


Figure 1. Position of the control region and its subregions within the Kim et al. (1998) reference dog mitochondrial genome.

that within a given population two randomly selected dogs will share the same haplotype by chance (Holland and Parsons 1999).

The random match probability is determined by the frequency estimate of the haplotype in the population of interest. The more common a haplotype, the higher is the probability that two dogs share this haplotype by chance, thus decreasing the evidential value of a match with this mtDNA type. Consequently, this sort of forensic applications requires the accurate estimation of haplotype frequencies in a population relevant to the criminal case.

The goal of this publication is to draw people's attention to the importance of implementing a dog mtDNA population study representative of the population of interest in a forensic case. It will provide an overview of the most important issues to keep in mind both when performing a population study of your own, as well as when considering to use published mtDNA data. First of all, sampling strategy characteristics are discussed such as sample size, maternal relatedness, breed status of the sampled dogs, and their geographic origin. Next, the importance of the quality of the sequence data is emphasized. In addition, the need to expand the sequenced DNA fragment in dog mtDNA studies is illustrated. Finally, the advantages of, and the criteria for, the assembly of an international, publicly available dog mtDNA population database of the highest quality, are pinpointed.

Estimating population frequencies of dog mtDNA haplotypes for forensic purposes

Background

The accuracy of haplotype frequency estimates almost entirely depends on the characteristics of the population sample that is used to represent the relevant population, i.e.

the population to which the donor of the trace is supposed to belong. Hence, biased population samples may lead to haplotype frequency estimates that diverge from the true population values.

To explore the impact of biased reference population samples in dog studies, we relied on current practices in human mtDNA population analyses and data derived from a selection of papers on haplotype variation in the dog mtDNA control region or the entire mitochondrial genome (mtGenome). Table 1 summarizes the main characteristics of the 58 dog studies used in this review. It includes studies with forensic aims, but also phylogenetic population and breed studies.

Dog mtDNA studies quite often do not meet the standards required for generating and publishing forensic human mtDNA population data. Briefly, these standards include: (1) providing a good documentation of the sampling strategy and a detailed description of the sampled individuals and the population, (2) avoiding sampling bias due to population substructure, (3) applying high quality mtDNA sequencing protocols and describing them clearly, (4) avoiding errors by handling and transferring data electronically, (5) performing quality checks of the generated data by e.g. haplogrouping or quasi-median network analysis and (6) making the full sequences publicly and electronically available preferably through either GenBank (Benson et al. 2013) or a forensic database such as EMPOP (Parson and Bandelt 2007, Parson and Dür 2007, Carracedo et al. 2010, Parson and Roewer 2010, Carracedo et al. 2013). These standards will be discussed here in relation to dog mtDNA studies.

Sampling strategy and its reporting

Strategies to sample mtDNA from dog populations are rarely well documented. Hence, it is often not clear to what extent the population samples adequately represent the populations from which they were drawn. Not seldom, sampling efforts are indeed limited to "sampling by convenience", i.e. relying on opportunistic sampling from locations as veterinary clinics and laboratories, dog shows, training schools and animal shelters. Obviously, it can be doubted whether these sampling locations are representative random samples of the "free" living relevant dog community (Parson and Bandelt 2007). Moreover, the types of sampling locations are often not even specified. In addition, basic information on the sampled dogs is often rudimentary, as many studies do not mention (1) how many dogs are mixed-breed or purebred, (2) to which breeds the dogs belong and/or (3) whether the geographic information provided refers to the region of origin of the breeds or to the actual region where the dogs were sampled (Table 1).

Several publications have provided recommendations on population sampling strategies for both dog and human mtDNA in forensics (Parson and Bandelt 2007, Pereira et al. 2010, Webb and Allard 2010, Linacre et al. 2011, Scharnhorst and Kanthaswamy 2011). Although more detailed guidelines are lacking, the main issue with

Table 1. Overview of the characteristics of sampling and sequence analysis in 58 canine mtDNA studies. Number of dogs sampled and, when specified in the publication, the number of dog breeds and mixed-breed, feral or village dogs in the sample; Origin of sample: new or extracted from previous studies as a comparison or to supplement the population sample (see reference numbers, except unpublished data by van Ash et al. (59); Koop et al. (60) and Shahid et al. (61)); Sampling region (or the geographic region of origin of included dog breeds if unclear from publication); Intention to avoid the inclusion of maternal relatives; GenBank accession numbers of new data are stated when applicable; un, unknown; s, skeletal remains of various age; ±, when variable, all sequences extracted from GenBank or the publication have this region in common; <, selected from this number of dogs from the same publication; * Larger region is mentioned in the publication, but only this part is available; Characteristics can differ from what is stated in publication if potential clerical errors were adapted, e.g. **publication states 246 instead of 233 as the sequences from reference 22 were included twice.

Publication			Sample						Sequence analysis	
	Reference	Aim of study	# Dogs	# Breeds	# Mixed-breed	Origin of sample	Sampling region	Avoid relatives	mtDNA region	Availability of new sequence data
1	Rothuizen et al. (1995)	mtDNA variability study	11	11	0	new	The Netherlands	un	Repeat region*	Publication
2	Okumura et al. (1996)	Phylogenetic breed study	94	24	0	new	Japan	un	15458-16129 16420-16727	D83599-D83613 D83616-D83638
3	Savolainen et al. (1997)	Forensic, Phylogenetic Population study	102	52	0	new	Sweden	YES	15431-15687	Publication
4	Tsuda et al. (1997)	Phylogenetic population study	34	24	0	new	Japan, Korea, Mongolia, Indonesia	un	15458-16130*	AB007380-AB007403
5	Vilà et al. (1997)	Phylogenetic population study	140 15	67 un	5 un	new < 140 new	un	un	15431-15687 ± 15393-16076 ± 16508-16549	AF005280-AF005295 AF008143-AF008157 AF008168-AF008182
6	Kim et al. (1998)	First complete dog mtGenome	1	1	0	new	Korea	YES	1-16727	U96639
7	Okumura et al. (1999)	Phylogenetic breed study	7 84 94	un un 24	un un 0	new new 2	Japan	un	15483-15679 15458-16129 16420-16727	AB031089-AB031107
8	Vilà et al. (1999)	Phylogenetic breed study	19 140	1 67	0 5	new 5	US, Mexico un	YES	15431-15687* 15431-15687	Publication
9	Randi et al. (2000)	Phylogenetic population study	41 9	30 0	0 9	new new	Switzerland Italy	un	15458-16000	AF115704-AF115718
10	Kim et al. (2001)	Phylogenetic breed study	25	11	0	new	Korea	NO	15622-16030	AF064569-AF064579 AF064581-AF064585
11	Branicki et al. (2002)	Forensic population study	12	11	1	new	Poland	un	15431-15687	AF345977-AF345982

	Publication		Sample						Sequence analysis	
	Reference	Aim of study	# Dogs	# Breeds	# Mixed-breed	Origin of sample	Sampling region	Avoid relatives	mtDNA region	Availability of new sequence data
12	Savolainen et al. (2002)	Phylogenetic population study	526 128	un	un	new 2, 4	Europe, Asia, Africa, Arctic America	un	± 15458-16039	AF531654-AF531741
13	Takahasi et al. (2002)	Inheritable disorder study	365	49	un	new	Japan	un	15458-16055	AB055010-AB055055
14	Valière et al. (2003)	Phylogenetic population study	50 15	un un	un un	new new	France, Switzerland France, Portugal	un	± 15519-15746	AF487730-AF487735 (excl. AF487732) AF487747-AF487751 AF338772-AF338788
15	Wetton et al. (2003)	Forensic population study	105 246	un	un	new 2, 3, 9	UK Japan, Switzerland, Italy, Sweden	un	15431-16030 ± 15458-15687	AY928903-AY928932
16	Pereira et al. (2004)	Catalogue of published datasets	58 1089	1 un	0 un	59 2, 4, 10, 12, 13, 15	Portugal Europe, Asia, Africa, Arctic America	un	15458-16039 ± 15622-16030	Publication
17	Savolainen et al. (2004)	Phylogenetic population study	22 1 654	un un un	un un un	new new 2, 4, 12	SE-Asia, India Polynesia Europe, Asia, Africa, Arctic America	un	15458-16039 15458-15720 ± 15458-16039	AY660647-AY660650 Publication
18	Sharma et al. (2004)	Phylogenetic population study	24	0	24	new	India	un	15443-15783	AY333727-AY333737
19	Angleby and Savolainen (2005)	Forensic & Phylogenetic Population study	35 74 758	19 52 un	9 2 un	new new 2, 4, 12, 15	Germany Europe Europe, Asia, Africa, Arctic America	YES	15458-16039 ± 15458-16030	AY656703-AY656710
20	Halverson and Basten (2005)	Forensic population study	348	88	45	new	US	un	15431-16085	Not published
21	van Asch et al. (2005)	Phylogenetic breed study	143 144	4 9	0 0	new 2, 4, 12, 13	Portugal Europe, Asia, Africa, Arctic America	YES	15372-16083 ± 15458-16030	Publication
22	Björnerfeldt et al. (2006)	Phylogenetic population study	88 14	53 13	0 0	new < 88 new	Sweden	un	part of HV-I 1-16727	Not published DQ480489-DQ480502

Publication			Sample						Sequence analysis	
	Reference	Aim of study	# Dogs	# Breeds	# Mixed-breed	Origin of sample	Sampling region	Avoid relatives	mtDNA region	Availability of new sequence data
23	Pires et al. (2006)	Phylogenetic breed study	143 21	11 0	0 21	new new	Portugal, Spain, Morocco Portugal, Azores, Tunisia	YES	15211-16096	AY706476-AY706524
24	Ryabinina (2006)	Phylogenetic breed study	84 20	3 2	0 0	new 12	Russia Turkey	un	15458-15778 ± 15458-16039	DQ403817-DQ403837
25	Sundqvist et al. (2006)	Phylogenetic breed study	100	20	0	new	Sweden	un	15431-15687	Publication
26	Eichmann and Parson (2007)	Forensic population study	133	46	38	new	Austria	un	15458-16727	Publication
27	Gundry et al. (2007)	Forensic population study Forensic breed study	61 64	41 2	0 0	new new	US	un	15455-16727	AY240030-AY240157 (excluding AY240073 AY240094, AY240155)
28	Baute et al. (2008)	Forensic population study	83 159	30	0	new 27, 30	US	un	15595-15654	Publication
29	Hassell et al. (2008)	Forensic population study Forensic breed study	96 15	79 1	0 0	new new	UK	un	15458-16039 15458-16131 16428-16727	Not published
30	Himmelberger et al. (2008)	Forensic population study	36 22 179	11 un	20 un	new 60 2, 4, 5, 6, 10, 27	US (California) un Europe, Asia, North-America	un	15456-16063 15433-16139 ± 15622-16030	EF122413-EF122428 AF098126-AF098147
31	Parra et al. (2008)	Phylogenetic breed study	52	5	0	new	Spain	un	15458-16105	EF380216-EF380225
32	Baranowska et al. (2009)	Inheritable disorder study	7	1	0	new	Sweden	NO	1-16727	FJ817358-FJ817364
33	Boyko et al. (2009)	Phylogenetic population study	309 17 un	0 0 un	309 17 un	new new 12, 23	Egypt, Uganda, Namibia US (mostly Puerto Rico) East-Asia, Africa	YES	± 15454-16075 ± 15458-16039	GQ375164-GQ375213
34	Desmyter and Comblez (2009)	Forensic population study	117	60	24	new	Belgium	YES	15458-16130 16431-16727	Not published
35	Koban et al. (2009)	Phylogenetic breed study	114 un	2 un	0 un	new 12	Turkey Europe, Asia, Africa	YES	15458-16039	EF660078-EF660191

Publication			Sample						Sequence analysis	
	Reference	Aim of study	# Dogs	# Breeds	# Mixed-breed	Origin of sample	Sampling region	Avoid relatives	mtDNA region	Availability of new sequence data
36	Pang et al. (2009)	Phylogenetic population study	907	un	un	new, 61	Old World, Arctic America	un	± 15458-16039	EU816456-EU816557
			669	un	un	2, 4, 6, 12, 22			1-15511	
			135	un	un	< 907 + 669			15535-16039 16551-16727	EU789638-EU789786
			34	un	un	6, 22, 61			1-16727	AY656737-AY656755
37	Webb and Allard 2009a	Forensic population study	427	139	118	new	US	YES	± 15458-16114 ± 16484-16727	EU223385-EU223811
			125			27			15455-16727	
38	Webb and Allard 2009b	Forensic population study	64	43	11	37	US	YES	± 1-16129 ± 16434-16727	EU408245-EU408308
			15	14	0	6, 22	Korea, Sweden		1-16727	
39	Muñoz-Fuentes et al. (2010)	Phylogenetic population study	29	un	un	new	Canada	un	15361-15785	FN298190-FN298218
40	Smalling et al. (2010)	Forensic population study	220	0	220	new	US	YES	15456-16063	FJ501174-FJ501203
			429			30, 37			± 15458-16063	
41	Ardalan et al. (2011)	Phylogenetic population study	325	un	un	new	Europe, SW-Asia	un	15458-16039	HQ261489 HQ452418- HQ452423
			1576	un	un	2, 4, 6, 12, 22, 36, 61	Old World, Arctic America		± 15458-16039	HQ452432- HQ452433 HQ452466- HQ452477
42	Brown et al. (2011)	Phylogenetic population study	200	0	200	new	Middle East/SW-Asia	un	15482-15867	HQ287728- HQ287744
			231	0	231	new	SE-Asia			
			1576	un	un	2, 4, 6, 12, 22, 36, 61	Old World, Arctic America		± 15458-16039	
43	Castroviejo-Fisher et al. (2011)	Phylogenetic population study	371	0	371	new	the Americas	un	± 15491-15755	HQ126702- HQ127072
			29	un	un	39				
44	Klütsch et al. 2011a	Phylogenetic population study	280	33	0	new	Europe, Arctic America, East-Asia	YES	15458-16039 ± 15458-16039	GQ896338-GQ896345
			234			36				

	Publication		Sample						Sequence analysis	
	Reference	Aim of study	# Dogs	# Breeds	# Mixed-breed	Origin of sample	Sampling region	Avoid relatives	mtDNA region	Availability of new sequence data
45	Klütsch et al. 2011b	Point heteroplasmy pedigree study	180 131	18 2	0 0	new new	Europe, Arctic America, East-Asia	NO	15458-16039	Publication
46	Kropatsch et al. (2011)	Phylogenetic breed study	77 34	26 1	0 0	new new	Germany	NO	15458-16124	Publication
47	Li et al. (2011)	Phylogenetic breed study	1 33	1 un	0 un	new 22, 32, 38, 61	China Sweden, US	YES	1-16727	HM048871
48	Sindičić et al. (2011)	Forensic species ID & Phylogenetic population study	20	0	20	new	Croatia	un	15465-15744	GU324475-GU324486
49	Bekaert et al. (2012)	Validation of forensic analysis method	41 550	29	3	new 27, 37	Belgium US	un	± 15458-16092 ± 16474-16703 ± 15458-16114 ± 16484-16727	HM561524-HM561546 HQ845266-HQ845282
50	Chakirou et al. (2012)	Phylogenetic breed study	78	3	0	new	Romania	YES	± 15251-16068	HE687017-HE687019
51	Desmyter and Gijsbers (2012)	Forensic population study Forensic breed study	208 778 107 337	60 6 6	68 0 0	new, 34 15, 26, 27, 37 new < 208 new, 13, 19, 26, 27, 37	Belgium UK, Austria, US Belgium Worldwide	YES	15458-16129 16430-16727 ± 15458-16030 15458-16129 16430-16727 ± 15458-16039	HM560872-HM560932
52 53	Głażewska et al. (2012) & Głażewska and Prusak (2012)	Phylogenetic breed study	34 un	2 un	0 un	new GenBank	Poland Worldwide	NO	15426-16085	HM007196-HM007200
54	Imes et al. (2012)	Forensic population study	100 233**	98 un	0 un	new 6, 22, 36, 38, 61	US, Australia, Canada, Columbia, Uruguay Worldwide	YES	± 1-16129 ± 16430-16727 ± 1-15511 ± 15535-16039 ± 16551-16727	JF342807-JF342906
55	Li and Zhang (2012)	Phylogenetic breed study	47 439	1 un	0 un	new GenBank	Tibet, surrounding areas Worldwide	YES	± 582 bp of control region	Not published

Publication			Sample						Sequence analysis	
	Reference	Aim of study	# Dogs	# Breeds	# Mixed-breed	Origin of sample	Sampling region	Avoid relatives	mtDNA region	Availability of new sequence data
56	Oskarsson et al. (2012)	Phylogenetic population study	305	un	un	new	SE-Asia, E-Asia	YES	15458-16039	HQ452439-HQ452465
			350	un	un	4, 12, 36				
			1224	un	un	2, 4, 6, 12, 22, 36, 61	Old World, Arctic America		± 15458-16039	
			1	un	un	17	Polynesia		15458-15720	
57	Brown et al. (2013)	Phylogenetic population study	2	un	un	new	Alaska, Greenland	un	± 367 bp of HV-I	JX185397
			51	1	0	new	Arctic America		± 15580-16016	
			78	2	0	2, 12, 36, 44			± 15458-16039	
58	Suárez et al. (2013)	Phylogenetic breed study	324	5	0	new	Canary Islands	YES	15361-16086	Publication
			986	un	un	15, 26, 27, 34, 37, 51	UK, Austria, Belgium, US		± 15458-16030	

sampling a population in a representative manner is to avoid over- and underestimating haplotype frequencies. Sampling bias causes regarding the number and features of sampled individuals will be discussed in relation to dog mtDNA.

Sample size

Using a random subsampling method (Pereira et al. 2004a), Webb and Allard (2010) assessed the influence of increased sample size on the distribution of haplotype frequency estimates in dog mtDNA population samples. They predicted that adding another 100 dogs to sample sizes of less than 650 dogs for HV-I and 750 dogs for HV-I and -II increases estimates of e.g. haplotype number and exclusion probability (i.e. the probability that two randomly chosen dogs from a sample have different haplotypes) with ≥ 5%. Table 1 shows that unless data are pooled, the majority of forensic dog control region population studies have rather small sample sizes of about 100 or fewer dogs.

Generally, the number of observed haplotypes increases with sample size (Table 2), while the proportion of rare haplotypes (i.e. encountered only once or twice) goes down. Consequently, exclusion probability largely remains the same with sample size expansion (Webb and Allard 2010) (Table 2). Under-sampling the population particularly affects the frequencies of haplotypes that remain rare while increasing sample size (Holland and Parsons 1999). This overestimation of rare haplotypes is illustrated when comparing nine forensic dog mtDNA studies. Many of the haplotypes with the highest frequencies in population samples of ≤ 100 dogs, have lower frequencies in larger sized studies (Table 2). For example, haplotype C5 occurs in 4.9% of the 61 dogs in the Gundry et al. (2007) study, while its frequency estimate is maximum 0.5% in other US studies in Table 2. Limited sample size thus tends to overestimate haplotype frequencies, which decreases the evidential value of an mtDNA match. Since overestimations do not inflate the risk of incriminating a false suspect, under-sampling can be deemed a conservative error (Salas et al. 2007).

Maternal relationships

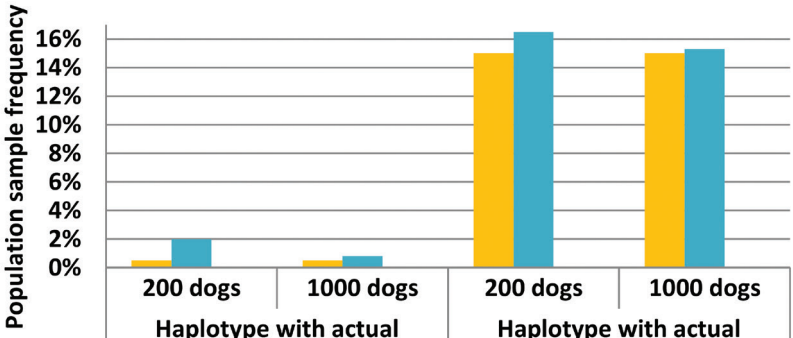
A randomized population sample for forensics should be allowed to include relatives if it is supposed to be unbiased (Brenner 2010). However, many population samples are assembled by convenience and could therefore contain more maternal relatives than expected from a randomized sample (Bodner et al. 2011).

The impact of a biased inclusion of maternal relatives in a forensic population study is rarely addressed, but generally decreases the genetic diversity of the population sample (Webb and Allard 2009a, Webb and Allard 2010). In small population samples, it particularly affects the risk of over-representing rare haplotypes (Bodner et al. 2011). By way of example, Figure 2 demonstrates that the impact of including 4 maternally related dogs is 5 times higher in a sample of 200 compared to 1000 dogs. In the smaller sample, the biased inclusion of 4 maternal relatives sharing a rare haplotype

Table 2. Comparison of haplotype number, P_E and haplotypes with the 10 highest frequencies in selected dog mtDNA studies. Exclusion probability (P_E) is based on the part of the control region studied in the publication (further details on exact region in Table 1) excluding the repeat region; characteristics can differ from publication if potential clerical errors were adapted; the 3 universally most frequent haplotypes are in bold (**A11, B1** and **A17**); (x) US and a minority from Australia, Canada, Uruguay and Columbia; (xx) Haplotype names are analogous to Savolainen et al. 2002, Savolainen et al. 2004, Angleby and Savolainen 2005, Pang et al. 2009, Klütsch et al. 201, Ardalan et al. 2011 and Oskarsson et al. 2012, or are in italic when unavailable and publication name was used; (xxx) Haplotypes are based on 15458-16039 except for Wetton et al. 2003 (15458-16030). Only the Kim et al. (1998) reference nucleotide was considered in case of a heteroplasmic site.

	Population studies for forensic purposes																		**Breed studies for forensic or phylogenetic purposes**					
	Europe								**US**															
	Wetton et al. 2003		**Angleby and Savolainen 2005**		**Eichmann and Parson 2007**		**Desmyter and Gijsbers 2012**		**Gundry et al. 2007**		**Himmelberger et al. 2008**		**Webb and Allard 2009a**		**Smalling et al. 2010**		**Imes et al. 2012**		**Okumura et al. 1996**		**Gundry et al. 2007**		**Suárez et al. 2013**	
Sampling region	UK		Germany		Austria		Belgium		US		US		US		US		US (x)		Japan		US		Canary Islands	
Studied part of control region (CR)	HV-I		HV-I		entire CR		HV-I+II		entire CR		HV-I		HV-I+II		HV-I		HV-I		HV-I+II		entire CR		HV-I	
# Dogs (# breeds/# mixed-breed)	105(un/un)		35(19/9)		133 (46/38)		208 (60/68)		61(41/0)		36(11/20)		552 (139/118)		649		100(98/0)		94(24/0)		64(2/0)		324(5/0)	
# Haplotypes	31		13		40		58		32		16		104		71		34		38		13		16	
Exclusion probability	0.93		0.86		0.93		0.92		0.93		0.89		0.96		0.92		0.91		0.93		0.8		0.86	
Haplotypes with 10 highest frequency estimates (%) (xx) (xxx)	**B1**	14.2	**A11**	28.6	**A11**	18.0	**B1**	16.8	**A11**	14.8	**B1**	16.7	**B1**	15.6	**B1**	18.0	**B1**	17.0	A18	13.8	A33	29.7	**B1**	25.9
	A18	11.4	**A17**	14.3	**A17**	12.0	**A11**	15.4	**B1**	13.1	**A11**	13.9	**A17**	10.9	**A11**	12.6	**A11**	15.0	A68	10.6	A16	28.1	**A17**	14.5
	A17	10.5	**B1**	14.3	**B1**	12.0	**A17**	15.4	A18	9.8	A16	13.9	**A11**	10.3	**A17**	11.7	A18	14.0	C3	10.6	**B1**	26.6	A20	14.2
	A2	8.6	C3	8.6	A19	8.3	A19	6.7	**A17**	6.6	**A17**	13.9	A18	9.2	A18	11.1	**A17**	8.0	**A17**	9.6	A5	6.3	B6	14.2
	A11	8.6	A2	5.7	A2	6.8	A18	5.8	C3	4.9	A18	11.1	A16	6.7	A16	6.2	A2	7.0	B14	8.5	*Gundry_24*	6.3	A19	8.3

	Population studies for forensic purposes																				Breed studies for forensic or phylogenetic purposes					
	Europe								US																	
	Wetton et al. 2003		Angleby and Savolainen 2005		Eichmann and Parson 2007		Desmyter and Gijsbers 2012		Gundry et al. 2007		Himmelberger et al. 2008		Webb and Allard 2009a		Smalling et al. 2010		Imes et al. 2012		Okumura et al. 1996		Gundry et al. 2007		Suárez et al. 2013	
Haplotypes with 10 highest frequency estimates (%) (xx) (xxx)	A19	5.7	A19	5.7	A18	6.8	A16	4.8	C5	4.9	A1	2.8	A33	3.4	A19	3.1	A22	4.0	A19	4.3	**A11**	1.6	**A11**	6.5
	A16	4.8	B6	5.7	A16	3.8	A22	3.8	A1	3.3	A15	2.8	C3	3.1	C3	3.1	A5	3.0	A2	3.2	**A17**	1.6	A22	4.9
	A20	3.8	A5	2.9	A1	3.0	A2	2.4	A2	3.3	A22	2.8	A2	2.5	A2	2.5	A16	3.0	**A11**	3.2			*A17+*	2.8
	A26	2.9	A20	2.9	A153	3.0	C1	2.4	A64	3.3	A26	2.8	A19	2.4	A5	1.7	*A167**	3.0	**B1**	3.2			A18	2.8
	A1	2.9	A33	2.9	A22	2.3	C2	2.4	C2	3.3	A28	2.8	A5	2.0	A22	1.7	A19	2.0	A29	2.1			A33	2.5
	C1	2.9	A44	2.9	A26	2.3			*Gundry_31*	3.3	A29	2.8			C2	1.7	A24	2.0	A70	2.1				
			A70	2.9	A33	2.3					A64	2.8							A72	2.1				
			A82	2.9	C1	2.3					A140	2.8							B12	2.1				
											A156	2.8							C1	2.1				
											B3	2.8												
											C3	2.8												



no maternal relatives	0.50%	0.50%	15.00%	15.00%
4 maternal relatives	2.00%	0.80%	16.50%	15.30%
Frequency increase	1.50%	0.30%	1.50%	0.30%

Figure 2. Impact of including maternally related dogs in population samples of 200 versus 1000 dogs on the estimation of the frequencies of rare haplotypes.

even causes its population sample frequency to be quadrupled. Moreover, even when observed only once in a population study, a haplotype that is rare in a population is already typically overrepresented in that sample (Holland and Parsons 1999). Therefore, oversampling maternal relatives should be avoided.

Although not specifically mentioned as a criterion for human mtDNA data in the international EMPOP database (Parson and Dür 2007), published population studies submitted to EMPOP do state that, as far as could be ascertained, the sampled individuals are unrelated. Examples are Brandstätter et al. (2007), Irwin et al. (2007), Saunier et al. (2009) and Prieto et al. (2011). Moreover, maternal relatives are removed in database updates. For human mtDNA population studies, it has been recommended to assess familial relationships by screening both available donor information and nDNA variation using microsatellites (Bodner et al. 2011).

For dog studies there is no consistent practice in dealing with maternal relationships in the population samples. Only about half of the 58 dog mtDNA studies in Table 1 mention whether or not they had the intention to avoid maternal relatives. Obviously, the usefulness of studies that do not provide this information may be doubtful in forensics. How maternal relationships were assessed is often not specified either, but it usually involves collecting information about the dogs from their owners. These background records can be used to verify whether dogs sharing a haplotype are e.g. from the same breed or whether their places of residence or those of their parents coincide (Webb and Allard 2009a, Desmyter and Gijsbers 2012). However, since dogs can have lots of offspring, there could be many maternal relatives, and these may be hard to track. Also, information provided by owners is not necessarily reliable (Webb and Allard 2009a) or available, and even registered pedigree records can be erroneous or incomplete (Kropatsch et al. 2011).

Purebred versus mixed-breed dogs

Another characteristic that can affect the haplotype frequency distribution in a population sample is potential population substructure due to the existence of dog breeds. Indeed, although generally mtDNA does not allow dogs to be grouped into their respective breeds (Okumura et al. 1996, Savolainen et al. 1997, Tsuda et al. 1997, Vilà et al. 1997, Kim et al. 2001, Wetton et al. 2003, Angleby and Savolainen 2005, van Asch et al. 2005, Pires et al. 2006, Sundqvist et al. 2006, Eichmann and Parson 2007, Gundry et al. 2007, Himmelberger et al. 2008, Parra et al. 2008, Desmyter and Comblez 2009, Kropatsch et al. 2011, Bekaert et al. 2012, Desmyter and Gijsbers 2012, Suárez et al. 2013), haplotype frequencies can differ between breeds, as well as between specific breeds and the entire dog mtDNA gene pool (Savolainen et al. 1997, Vilà et al. 1999, Angleby and Savolainen 2005, van Asch et al. 2005, Pires et al. 2006, Ryabinina 2006, Eichmann and Parson 2007, Gundry et al. 2007, Hassell et al. 2008, Himmelberger et al. 2008, Parra et al. 2008, Koban et al. 2009, Webb and Allard 2009a, Kropatsch et al. 2011, Desmyter and Gijsbers 2012, Brown et al. 2013, Suárez et al. 2013). Because of the overrepresentation of certain haplotypes in specific breeds in comparison to the general dog population, a dog trace mtDNA type might provide an indication about the breed(s) to which it may belong. However, such information should be used with caution and police investigations should not only focus on the more likely breed (Angleby and Savolainen 2005, Hassell et al. 2008, Desmyter and Gijsbers 2012).

Obviously, the over- and underrepresentation of particular breeds in a population sample compared to the population from which the sample is drawn, may bias haplotype frequency estimates (Desmyter and Gijsbers 2012). Therefore, in theory, dogs should be randomly sampled in order to correctly represent the breed composition of the population of interest (Scharnhorst and Kanthaswamy 2011). In addition, it is recommended that population samples reflect the actual proportions of mixed-breed versus purebred dogs in the population. However, mixed-breed dogs are often underrepresented in population studies (Smalling et al. 2010) and many population studies even only include purebred dogs (Table 1). Moreover, since most samples are collected by convenience, deviations from the actual breed composition of the population and overrepresentation of rare dog breeds are to be expected. This should be taken into account when using this sort of data in forensic casework (Eichmann and Parson 2007).

Against this background, Savolainen et al. (1997) attempted to adjust the number of dogs per breed in their population sample, so as to more accurately reflect the countrywide breed composition in Sweden. Later, Angleby and Savolainen (2005) stated that the exclusion probability of population samples containing dogs of the 20 most common breeds in Sweden represent the Swedish population more accurately than population samples containing the 100 most common breeds. Still, the number of dogs per breed and the overall number of breeds in a sample can be overestimated, since these data largely depend on the owner's subjective opinion (Himmelberger et al. 2008, Webb and Allard 2009a).

Some authors have indicated that population studies specific for single breeds may be forensically relevant in the rare event that the breed of the dog that donated the crime scene trace is known, for example by eye-witness reports (Savolainen et al. 1997, Wetton et al. 2003, Desmyter and Gijsbers 2012). Obviously, the evidential value of an mtDNA match can be quite different when based on a general rather than a breed-specific population study. Studies focused on specific breeds have been published, mostly aiming at verifying the accuracy of pedigree records and tracing its population genetic features (e.g. demographic history, region of origin, hybridization events, etc.). Examples are Vilà et al. (1999), van Asch et al. (2005), Kropatsch et al. (2011) and Suárez et al. (2013).

Including pedigree data can improve intra-breed mtDNA diversity studies. In theory, an appropriate selection of representative individuals from existing maternal lines from pedigrees allows to capture all mtDNA haplotypes of a breed within a population while minimizing the amount of laboratory work. The frequencies of these haplotypes can be estimated from the numbers of offspring in each maternal line in the breed population (Głażewska et al. 2013). Of course, to this end pedigree records need to be accurate and complete (Głażewska et al. 2013). Unfortunately, this is not always the case, as has been shown in e.g. Weimaraner dogs (Kropatsch et al. 2011).

Analyzing the haplotype frequency distribution within breeds can also give insight into differences between published population studies. An example of the impact of breed associated sample bias was given by Desmyter and Gijsbers (2012). These authors noted that the US population sample of Webb and Allard (2009a) included 64 dogs of two Retriever breeds from Gundry et al. (2007). This could have biased the frequency estimates of haplotypes A16 and A33 in the US sample, since these haplotypes are very common in Retrievers (Desmyter and Gijsbers 2012). Additionally, mtDNA studies focusing on specific breeds rather than on entire populations, clearly show lower amounts of variation (expressed in terms of exclusion probability) than population studies from similar geographical regions (Table 2). Also, the sets of haplotypes with the ten highest frequencies can be quite different (Table 2). In order to compensate for purebred related biases, Himmelberger et al. (2008) increased the number of mixed-breed dogs in their US population sample and claimed that in this way their sample was more representative than previous US dog mtDNA population samples. However, their sample still showed an unusually high frequency of haplotype A16 (Table 2), most probably because their sample was limited to only 36 dogs, 13 of which were either purebred or mixed-breed Retrievers.

Geographic origin

To evaluate the significance of a haplotype match between a dog trace and its suspected donor, a population sample should reliably reflect the population to which the donor of the trace is supposed to belong. As such, one might wonder about the importance of the geographic origin of the sampled dogs in a sampling strategy.

Probably the most important macrogeographic issue to consider in dog studies, is the fact that dog populations in Southeast Asia show almost the entire dog mtDNA diversity, while elsewhere in the world only parts of this diversity is present (Savolainen et al. 2002, Pang et al. 2009, Ardalan et al. 2011, Brown et al. 2011). This suggests that SE Asia is the region where dogs were first domesticated and from where domesticated dogs were spread throughout the rest of the world (Savolainen et al. 2002, Pang et al. 2009). Another noticeable macrogeographic structuring in dog mtDNA is that haplotype group d1 is almost exclusively found in Scandinavian and Finnish breeds, in which sometimes over 50% of the dogs have a d1 haplotype (Klütsch et al. 2011a). Obviously, this sort of macrogeographic mtDNA differentiation should be considered in population sampling, since oversampling dog breeds of SE Asian or Scandinavian/Finnish origin in local population samples elsewhere in the world can bias haplotype frequency estimates. For example, Angleby and Savolainen (2005) demonstrated that dogs of East Asian origin in Europe carried a number of haplotypes that are absent in native European breeds. Moreover, the frequencies of globally common haplotypes differ between Asian and European samples (Angleby and Savolainen 2005). This is also illustrated by the composition and frequency distribution of the most common dog haplotypes in the breed study from Japan by Okumura et al. (1996) and those in the forensic population studies from Europe and the US (Table 2).

Quality of nucleotide sequence data

The description of haplotypes is a source of error and confusion when comparing population studies. Typically, haplotypes are aligned to a reference sequence using software supplemented with annotation rules in order to record them unambiguously as an alpha-numeric code. This code is a shortened annotation of the sequence string, consisting of differences to the reference sequence. For example, the HV-I alpha-numeric code of haplotype A11 is 15639A, 15814T and 16025C (Angleby and Savolainen 2005). Analogous to human mtDNA analyses, Pereira et al. (2004b) recommended to set the L-strand of the first published complete dog mtGenome (Kim et al. 1998) as the reference standard. In order to identify different haplotypes and enable their comparison, Pereira et al. (2004b) listed a number of rules to align sequences to the Kim et al. reference (1998) and to unambiguously record polymorphisms. These rules are based on those for human mtDNA (Carracedo et al. 2000, Wilson et al. 2002b, Wilson et al. 2002b). Length heteroplasmy in the VNTR region of the dog's mtDNA control region complicates the numbering system of the nucleotide positions. To simplify this, Pereira et al. (2004b) decided that numbering the nucleotide positions after this repeat region should start at position 16430 regardless of the number of repeats (Figure 1). Nevertheless, even with a standard reference haplotype, a numbering system and annotation rules, variation can still be miscoded, such as for the polyC-polyT-polyC region from position 16661 to 16674 in HV-II (Table 3).

Table 3. Illustration of different annotations for the HV-II polyC-polyT-polyC haplotype with 6 C's, 8 T's and 2 C's. Annotation (1) was used by Gundry et al. (2007), while Eichmann and Parson (2007) and Desmyter and Gijsbers (2012) applied annotation (2) because of different alignments to the Kim et al. (1998) reference sequence of 3C8T3C.

#C#T#C	16661	16662	16663	16663.1	16663.2	16663.3	16664	16665	16666	16667	16668	16669	16670	16671	16672	16673	16674
3C8T3C	**C**	**C**	**C**	-	-	-	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**C**	**C**	**C**
6C8T2C (1)	C	C	C	C	C	-	C	T	T	T	T	T	T	T	T	C	C
6C8T2C (2)	C	C	C	C	C	C	T	T	T	T	T	T	T	T	C	C	-

Haplotypes can also be denoted by names. However, it is not good practice to provide only haplotype names in publications, like e.g. Sundqvist et al. (2006) did. This introduces ambiguities if the same names are used elsewhere for other haplotypes. For the same reason, it is ill advised to use haplotype names that differ from GenBank entries, as was done by e.g. Smalling et al. (2010). The haplotype names established by Savolainen et al. (2002), Savolainen et al. (2004) and Angleby and Savolainen (2005), were expanded by Pang et al. (2009) and Webb and Allard (2009a), such that they both used the same names for different new haplotypes. Since then, names of new haplotypes often overlap between publications that are building further onto the names of both of these publications, e.g. Smalling et al. (2010), Ardalan et al. (2011), Klütsch et al. (2011a) and Imes et al. (2012). In addition, applying previously published haplotype names can be difficult because the analyzed mtDNA region may differ (Pereira et al. 2004b).

Mistakes occur relatively often while copying and editing sequence data. Therefore, guidelines have been published to minimize making these clerical errors and to detect them more easily (Bandelt et al. 2001, Bandelt et al. 2004, Yao et al. 2004, Salas et al. 2005). For example, alpha-numeric codes presented in the form of a matrix-based dot table are particularly error-prone and difficult to read (Parson and Bandelt 2007, Parson and Roewer 2010). In practice, several clerical errors have been observed in dog mtDNA studies. For example, alignment with the Kim et al. (1998) reference sequence of the GenBank entries corresponding to the HV-I codes in Table 2 of Imes et al. (2012), revealed several inconsistencies. A deletion at position 15932 in many haplotypes in Table 2 of Imes et al. (2012) cannot be observed in most of the GenBank entries. As such, this deletion defined two artificial haplotypes. Furthermore, Table 2 of Imes et al. (2012) did not include two haplotypes deposited in GenBank, while variant base 15665C was not recorded for haplotype A170*.

As more mtGenome data are generated, coding regions SNPs are encountered that appear to be characteristic for particular control region haplotypes and haplogroups (Verscheure, unpublished data). Such SNPs can help to indicate potential sequence or clerical errors. For example, the control region sequence of mtGenome haplotype A169* (A11 after removal of the deletion at 15932) belongs to haplogroup A (Imes et al. 2012), but the SNPs in the rest of its mtGenome are typical for haplogroup B. This might be due to artificial recombination, caused by mixing up amplicons from dif-

ferent individuals, either during laboratory work or data editing (Bandelt et al. 2001, Bandelt et al. 2004). Similarly, the entire mtGenome sequence of Imes et al. (2012) haplotype A167* is more typical of haplogroup C than of haplogroup A.

As shown above, deposition of sequence data in GenBank provides an opportunity to verify sequence data quality. Unfortunately, in contrast to good practice, 15 of the 58 studies reviewed here did not submit any sequence to GenBank, but only provided alpha-numeric codes or haplotype names. Moreover, several papers did not even disclose the haplotype sequences or their estimated population frequencies (Table 1). When studies did deposit sequences in GenBank, they did so either only for new haplotypes, for all observed haplotypes, or for all sampled dogs. These various practices may confound subsequent analyses. For example, Imes et al. (2012) extracted the mtGenomes of the same 14 dogs twice from GenBank, because these sequences were uploaded in GenBank both by Björnerfeldt et al. (2006) and Pang et al. (2009). Obviously, these duplicated datasets introduce bias in the estimation of mtGenome haplotype frequencies and mtDNA diversity in the Imes et al. (2012) study.

Dog mtDNA studies show a large variety of analysis methods as well. Consequently, the quality of these analyses might vary. Next to annotation issues, several sequence quality issues have been observed while reviewing dog mtDNA studies. For example, Webb and Allard (2009a) reported sequence reading difficulties in the HV-II region because of length heteroplasmy in the VNTR and the polyC-polyT-polyC region. Nevertheless, in about 190 sequences Webb and Allard (2009a) observed that positions 16430, 16431, 16432 and/or 16433 directly adjacent to the VNTR at the start of HV-II are deleted in comparison to the Kim et al. (1998) reference sequence. Since deletions have not been reported at these sites in any of the other reviewed studies, we suggest these should be considered missing data due to reading difficulties. Therefore, it is recommended to verify this issue using additional primers, other alignment software and visual inspection of the alignments. Webb and Allard (2009a) interpreted these sites as highly informative and counted sequences differing only at these positions (e.g. A1 and d) as different haplotypes. If these deletions indeed resulted from reading difficulties, then they artificially increased the haplotype number and exclusion probability. A second example is that about 65% of the mtGenome sequences deposited in GenBank by Imes et al. (2012) contain ambiguities outside the VNTR with up to 130 N's per sequence and stretches of up to 110 adjacent N's, i.e. ambiguous bases due to the presence of dye blobs (Imes et al. 2012). If such ambiguities occur at informative sites, then these sequence quality issues can affect the frequency estimates of mtGenome haplotypes and the SNPs that define them.

Thus, caution and proofreading is necessary for both new sequences and those extracted from papers and databases. Therefore, Berger et al. (2012) published a detailed workflow for generating high quality HV-I and -II data from dogs based on experience from human mtDNA analysis. For forensic mtDNA analysis, these and other authors recommend to sequence each position at least twice, preferably on both mtDNA strands, so as to minimize sequencing errors (Wilson et al. 1993, Carracedo et al. 2000, Tully et al. 2001, Parson and Bandelt 2007, Berger et al. 2012). Finally, when extracting

sequences from GenBank, it is important to realize that quality control of a database entry relies on the submitting scientist. Hence, it is not surprising that the reliability of GenBank data has been questioned, such as by Harris (2003) and Yao et al. (2009).

Exploring the entire mtGenome to improve discriminatory power

The majority of dogs have haplotypes that are frequent in most dog populations worldwide. As a result, even if there are many rare haplotypes, the discriminatory power of the dog mtDNA control region is limited (Savolainen et al. 1997, Wetton et al. 2003, Angleby and Savolainen 2005, Halverson and Basten 2005, Eichmann and Parson 2007, Gundry et al. 2007, Baute et al. 2008, Hassell et al. 2008, Himmelberger et al. 2008, Desmyter and Comblez 2009, Webb and Allard 2009a, Smalling et al. 2010, Desmyter and Gijsbers 2012, Imes et al. 2012). This is well illustrated by comparing the mtDNA characteristics of nine forensic population studies which all consider at least positions 15458 to 16030 in HV-I. Almost half of the sampled dogs have haplotypes B1, A11 or A17 with average population frequency estimates of 15.3%, 15.2% and 11.5%. In addition, many other frequent haplotypes are shared between samples (Table 2). Hence, dog mtDNA matches will often have limited forensic value.

Evidently, expanding the length of the surveyed sequence will increase the number of polymorphic sites and thus may improve the discriminatory power of the mtDNA control region in dogs. However, most population studies did not include HV-II and as such missed important variation that often allows splitting up HV-I haplotypes. Hence, sequencing at least both HV-I and HV-II is recommended for forensic population studies (Eichmann and Parson 2007, Gundry et al. 2007, Desmyter and Comblez 2009, Webb and Allard 2009a, Webb and Allard 2010, Desmyter and Gijsbers 2012, Imes et al. 2012).

A number of complete control region haplotypes still show high population frequencies. Therefore, it is advised to further increase the discriminatory power of dog mtDNA by surveying population samples for entire mtGenomes (Webb and Allard 2009a). This is indeed a trend in the last years with very promising results (Webb and Allard 2009b, Imes et al. 2012). However, the use of SNPs in the coding region in forensics will require many more mtGenome studies (Irwin et al. 2011).

Population study versus database

Not every forensic laboratory has the resources to conduct large-scale population studies. As such, supplementing smaller, local samples with published data allows capturing more mtDNA variability. However, this practice may bias the haplotype frequency distribution in the pooled sample compared to the population of interest, because of (1) sample heterogeneity, (2) inconsistent sequence quality, (3) clerical errors and (4)

the difficulty of sequence comparisons due to variation in sequence lengths, alignment procedures, and sequence annotation. Relying on a public dog mtDNA database instead of, or in addition to, published local population data may be a trustworthy alternative, provided that the sequences are carefully reviewed before inclusion in the database. As such, submitting population sample data to the database could be an obligatory quality check with which studies have to comply before they are published. This is often demanded for human mtDNA population data (Carracedo et al. 2010, Parson and Roewer 2010, Carracedo et al. 2013).

To establish a reliable dog mtDNA database, inspiration can be found in the European DNA profiling group (EDNAP) mtDNA population database (EMPOP) for human mtDNA haplotypes useful in forensic casework. EMPOP stresses the need for generating mtDNA sequence data of the highest quality (Parson et al. 2004, Parson and Dür 2007) and established guidelines to achieve this. Briefly, these guidelines recommend: (1) application of a high quality mtDNA determination method that covers the entire sequenced region at least twice, (2) electronic transfer and transcription of sequence results, (3) compliance to generally accepted alignment and annotation guidelines and (4) data verification through haplogrouping and quasi-median network analysis (Brandstätter et al. 2007, Parson and Dür 2007). Against this background, the interlaboratory study by van Asch et al. (2009) emphasized a similar need for such guidelines for dog mtDNA analyses.

Next to the need for high quality mtDNA population data from all around the world, three other important requirements for building a dog mtDNA database are discussed hereafter. Firstly, management by a central laboratory is indispensable to perform the quality assessment of submitted population samples, to maintain and update the database software and web portal, and to communicate about it to the users. After submission to EMPOP, this laboratory reviews the population sample data for errors by e.g. examining the raw sequence data and using quasi-median network analysis (Bandelt and Dür 2007, Parson and Dür 2007, Zimmermann et al. 2011). Indeed, allocating mtDNA sequences to specific haplogroups may indicate which mutations are expected and may help to detect potential artificial recombination (Bandelt et al. 2001, Bandelt et al. 2004, Bandelt et al. 2012). Additionally, EMPOP provides its users with software for network analysis that may point out potential errors within the data based on phylogenetic background information. Thorough phylogenetic knowledge of dog mtDNA haplotypes could allow the adaptation of such software for dog mtDNA population samples.

Secondly, the database should be searchable and provide tools for comparison of various mtDNA sequence ranges. EMPOP uses the SAM search engine, which translates the queried haplotype and all database entries into sequence strings that are more easily comparable than alpha-numeric codes. In this way, it avoids generating biased haplotype frequency estimates caused by alignment and annotation inconsistencies making that database entries remain undetected in a database search even if they are identical to the queried haplotype (Röck et al. 2011).

Finally, the database should sufficiently document background information on the specimens. This enables the selection of subsets of samples in the database relevant to a specific case, such as dogs from specific geographic regions, of particular breeds, etc. In casework, selection of a suitable dataset is vital to a correct evaluation of evidence. Weighing the evidence against several database subdivisions is recommended to consider which one provides the most appropriate and conservative estimate of a haplotype's random match probability (Salas et al. 2007).

FidoSearch™, a canine mtDNA database with search software, was developed for use in casework by the Institute of Pathology and Molecular Immunology in Porto, Portugal in collaboration with Mitotyping Technologies in Pennsylvania, USA (Melton et al. 2011). However, it is not publicly available and its data entries were assembled from GenBank. Hence, FidoSearch™ is not an appropriate alternative for the creation of a publicly available, high quality and comprehensive dog mtDNA database.

Conclusions

In order to meet forensic quality standards, a dog mtDNA population sample needs to be representative of the population of interest to the case. To this end, several recommendations can be made for performing and publishing a dog mtDNA population study for forensic purposes: (1) provide sufficiently detailed information on the population of interest, the sampling strategy and the sampled dogs, (2) include at least several hundred dogs in the population sample, (3) intend to avoid biased inclusion of maternal relatives, (4) use a population sample reflecting the dog population where the crime occurred, (5) the composition of the population sample in terms of purebred and mixed-breed dogs, groups of breeds of a particular geographic origin, and dogs belonging to specific breeds, should be proportional to the studied population, (6) apply a high quality and validated analytical methodology and run quality control steps to minimize the risk of errors during either laboratory work or data processing, (7) submit the haplotype sequence strings to a publicly available database such as GenBank and (8) follow the Pereira et al. (2004b) rules when converting haplotype sequences into alpha-numeric codes denoting differences in relation to the Kim et al. (1998) reference sequence. These recommendations also apply when supplementing your own data with published data. In addition, keep in mind that sequence files in a database such as GenBank do not provide raw sequence data and can hide ambiguous results.

All things considered, this review emphasizes the need for more forensically relevant, high quality dog mtDNA population studies. In addition, it stresses the need for a publicly available dog mtDNA population database that assembles easily comparable and thoroughly checked population data from all around the world. Finally, expanding mtDNA studies from the control region to the entire mtGenome is recommended to enhance the discriminatory power of forensic dog mtDNA analysis.

Acknowledgements

S. Verscheure is a PhD student at the University of Antwerp supported by a grant from the Belgian Federal Public Planning Service Science Policy. This work was conducted within the framework of FWO Research Community W0.009.11N "Belgian Network for DNA Barcoding".

References

Aaspóllu A, Kelve M (2003) The first criminal case in Estonia with dog's DNA data admitted as evidence. International Congress Series 1239: 847–851.

Allen M, Engström AS, Meyers S, Handt O, Saldeen T, von Haeseler A, Pääbo S, Gyllensten U (1998) Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. Journal of Forensic Sciences 43: 453–464.

Angleby H, Savolainen P (2005) Forensic informativity of domestic dog mtDNA control region sequences. Forensic Science International 154: 99–110. doi: 10.1016/j.forsciint.2004.09.132

Ardalan A, Klütsch CF, Zhang AB, Erdogan M, Uhlén M, Houshmand M, Tepeli C, Ashtiani SR, Savolainen P (2011) Comprehensive study of mtDNA among Southwest Asian dogs contradicts independent domestication of wolf, but implies dog-wolf hybridization. Ecology and Evolution 1: 373–385. doi: 10.1002/ece3.35

Bandelt HJ, Dür A (2007) Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. Molecular Phylogenetics and Evolution 42: 256–271. doi: 10.1016/j.ympev.2006.07.013

Bandelt HJ, Lahermo P, Richards M, Macaulay V (2001) Detecting errors in mtDNA data by phylogenetic analysis. International Journal of Legal Medicine 115: 64–69. doi: 10.1007/s004140100228

Bandelt HJ, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. International Journal of Legal Medicine 118: 267–273. doi: 10.1007/s00414-004-0455-2

Bandelt HJ, van Oven M, Salas A (2012) Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. International Journal of Legal Medicine 126: 901–916. doi: 10.1007/s00414-012-0762-y

Baranowska I, Jäderlund KH, Nennesmo I, Holmqvist E, Heidrich N, Larsson NG, Andersson G, Wagner EG, Hedhammar Å, Wibom R, Andersson L (2009) Sensory ataxic neuropathy in golden retriever dogs is caused by a deletion in the mitochondrial tRNATyr gene. PLoS Genetics 5: e1000499. doi: 10.1371/journal.pgen.1000499

Baute DT, Satkoski JA, Spear TF, Smith DG, Dayton MR, Malladi VS, Goyal V, Kou A, Kinaga JL, Kanthaswamy S (2008) Analysis of forensic SNPs in the canine mtDNA HV1 mutational hotspot region. Journal of Forensic Sciences 53: 1325–1333. doi: 10.1111/j.1556-4029.2008.00880.x

Bekaert B, Larmuseau MH, Vanhove MP, Opdekamp A, Decorte R (2012) Automated DNA extraction of single dog hairs without roots for mitochondrial DNA analysis. Forensic Science International, Genetics 6: 277–281. doi: 10.1016/j.fsigen.2011.04.009

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Research 41: D36–42. doi: 10.1093/nar/gks1195

Berger C, Berger B, Parson W (2012) Sequence analysis of the canine mitochondrial DNA control region from shed hair samples in criminal investigations. Methods in Molecular Biology 830: 331–348. doi: 10.1007/978-1-61779-461-2_23

Björnerfeldt S, Webster MT, Vilà C (2006) Relaxation of selective constraint on dog mitochondrial DNA following domestication. Genome Research 16: 990–994. doi: 10.1101/gr.5117706

Bodner M, Irwin JA, Coble MD, Parson W (2011) Inspecting close maternal relatedness: Towards better mtDNA population samples in forensic databases. Forensic Science International, Genetics 5: 138–141. doi: 10.1016/j.fsigen.2010.10.001

Bogenhagen D, Clayton DA (1974) The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. Journal of Biological Chemistry 249: 7991–7995.

Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, Corey L, Degenhardt JD, Auton A, Hedimbi M, Kityo R, Ostrander EA, Schoenebeck J, Todhunter RJ, Jones P, Bustamante CD (2009) Complex population structure in African village dogs and its implications for inferring dog domestication history. Proceedings of the National Academy of Sciences of the USA 106: 13903–13908. doi: 10.1073/pnas.0902129106

Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W (2007) Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. Forensic Science International 166: 164–175. doi: 10.1016/j.forsciint.2006.05.006

Branicki W, Kupiec T, Pawłowski R (2002) Analysis of dog mitochondrial DNA for forensic identification purposes. Problems of Forensic Sciences (Z Zagadnien Nauk Sadowych) 50 (L): 91–98.

Brenner CH (2010) Fundamental problem of forensic mathematics--the evidential value of a rare haplotype. Forensic Science International, Genetics 4: 281–291. doi: 10.1016/j.fsigen.2009.10.013

Brown SK, Darwent CM, Sacks BN (2013) Ancient DNA evidence for genetic continuity in arctic dogs. Journal of Archaeological Science 40: 1279–1288. doi: 10.1016/j.jas.2012.09.010

Brown SK, Pedersen NC, Jafarishorijeh S, Bannasch DL, Ahrens KD, Wu JT, Okon M, Sacks BN (2011) Phylogenetic distinctiveness of Middle Eastern and Southeast Asian village dog Y chromosomes illuminates dog origins. PLoS ONE 6: e28496. doi: 10.1371/journal.pone.0028496

Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M (2000) DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. Forensic Science International 110: 79–85. doi: 10.1016/S0379-0738(00)00161-4

Carracedo A, Butler JM, Gusmão L, Linacre A, Parson W, Roewer L, Schneider PM (2013) New guidelines for the publication of genetic population data. Forensic Science International, Genetics 7: 217–220. doi: 10.1016/j.fsigen.2013.01.001

Carracedo A, Butler JM, Gusmão L, Parson W, Roewer L, Schneider PM (2010) Publication of population data for forensic purposes. Forensic Science International, Genetics 4: 145–147. doi: 10.1016/j.fsigen.2010.02.001

Castroviejo-Fisher S, Skoglund P, Valadez R, Vilà C, Leonard JA (2011) Vanishing native American dog lineages. BMC Evolutionary Biology 11: 73. doi: 10.1186/1471-2148-11-73

Chakirou O, Vlaic A, Carşai TC, Bâlteanu VA, Coşier V (2012) Aspects regarding the molecular characterisation of the Romanian Shepherd dog breeds. Animal Biology & Animal Husbandry 4: 28–31.

Desmyter S, Comblez S (2009) Belgian dog mitochondrial DNA database for forensics. Forensic Science International, Genetics Supplement Series 2: 286–287. doi: 10.1016/j.fsigss.2009.08.110

Desmyter S, Gijsbers L (2012) Belgian canine population and purebred study for forensics by improved mitochondrial DNA sequencing. Forensic Science International, Genetics 6: 113–120. doi: 10.1016/j.fsigen.2011.03.011

Eichmann C, Parson W (2007) Molecular characterization of the canine mitochondrial DNA control region for forensic applications. International Journal of Legal Medicine 121: 411–416. doi: 10.1007/s00414-006-0143-5

Fridez F, Rochat S, Coquoz R (1999) Individual identification of cats and dogs using mitochondrial DNA tandem repeats? Science & Justice 39: 167–171. doi: 10.1016/S1355-0306(99)72042-3

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6: 861–868. doi: 10.1111/j.1365-294X.1997.tb00140.x

Głażewska I, Prusak B (2012) Evaluation of the effectiveness of introducing new alleles into the gene pool of a rare dog breed: Polish Hound as the example. Czech Journal of Animal Science 57: 248–254.

Głażewska I, Prusak B, Gralak B (2013) Pedigrees as a source of information in mtDNA studies of dogs and horses. Animal Genetics 44: 227–230. doi: 10.1111/j.1365-2052.2012.02388.x

Głażewska I, Zielińska S, Prusak B (2012) Formation of a new dog population observed by pedigree and mtDNA analyses of the Polish Hovawart. Archiv Tierzucht 55: 391–401.

Gundry RL, Allard MW, Moretti TR, Honeycutt RL, Wilson MR, Monson KL, Foran DR (2007) Mitochondrial DNA analysis of the domestic dog: control region variation within and among breeds. Journal of Forensic Sciences 52: 562–572. doi: 10.1111/j.1556-4029.2007.00425.x

Halverson JL, Basten C (2005) Forensic DNA identification of animal-derived trace evidence: tools for linking victims and suspects. Croatian Medical Journal 46: 598–605.

Harris DJ (2003) Can you bank of GenBank. Trends in Ecology & Evolution 18: 317–319. doi: 10.1016/S0169-5347(03)00150-2

Hassell R, Heath P, Musgrave-Brown E, Ballard D, Harrison C, Thacker C, Catchpole B, Syndercombe Court D (2008) Mitochondrial DNA analysis of domestic dogs in the UK.

Forensic Science International, Genetics Supplement Series 1: 598–599. doi: 10.1016/j.fsigss.2007.10.187

Himmelberger AL, Spear TF, Satkoski JA, George DA, Garnica WT, Malladi VS, Smith DG, Webb KM, Allard MW, Kanthaswamy S (2008) Forensic utility of the mitochondrial hypervariable region 1 of domestic dogs, in conjunction with breed and geographic information. Journal of Forensic Sciences 53: 81–89. doi: 10.1111/j.1556-4029.2007.00615.x

Holland MM, Parsons TJ (1999) Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Science Review 11: 21–50.

Imes DL, Wictum EJ, Allard MW, Sacks BN (2012) Identification of single nucleotide polymorphisms within the mtDNA genome of the domestic dog to discriminate individuals with common HVI haplotypes. Forensic Science International, Genetics 6: 630–639. doi: 10.1016/j.fsigen.2012.02.004

Irwin J, Egyed B, Saunier J, Szamosi G, O'Callaghan J, Padar Z, Parsons TJ (2007) Hungarian mtDNA population databases from Budapest and the Baranya county Roma. International Journal of Legal Medicine 121: 377–383. doi: 10.1007/s00414-006-0128-4

Irwin JA, Parson W, Coble MD, Just RS (2011) mtGenome reference population databases and the future of forensic mtDNA analysis. Forensic Science International, Genetics 5: 222–225. doi: 10.1016/j.fsigen.2010.02.008

Kim KS, Jeong HW, Park CK, Ha JH (2001) Suitability of AFLP markers for the study of genetic relationships among Korean native dogs. Genes & Genetic Systems 76: 243–250. doi: 10.1266/ggs.76.243

Kim KS, Lee SE, Jeong HW, Ha JH (1998) The complete nucleotide sequence of the domestic dog (*Canis familiaris*) mitochondrial genome. Molecular Phylogenetics and Evolution 10: 210–220. doi: 10.1006/mpev.1998.0513

Klütsch CF, Seppälä EH, Fall T, Uhlén M, Hedhammar Å, Lohi H, Savolainen P (2011a) Regional occurrence, high frequency but low diversity of mitochondrial DNA haplogroup d1 suggests a recent dog-wolf hybridization in Scandinavia. Animal Genetics 42: 100–103. doi: 10.1111/j.1365-2052.2010.02069.x

Klütsch CF, Seppälä EH, Uhlén M, Lohi H, Savolainen P (2011b) Segregation of point mutation heteroplasmy in the control region of dog mtDNA studied systematically in deep generation pedigrees. International Journal of Legal Medicine 125: 527–535. doi: 10.1007/s00414-010-0524-7

Koban E, Saraç ÇG, Açan SC, Savolainen P, Togan İ (2009) Genetic relationship between Kangal, Akbash and other dog populations. Discrete Applied Mathematics 157: 2335–2340. doi: 10.1016/j.dam.2008.06.040

Kropatsch R, Streitberger K, Schulte-Middelmann T, Dekomien G, Epplen JT (2011) On ancestors of dog breeds with focus on Weimaraner hunting dogs. Journal of Animal Breeding and Genetics 128: 64–72. doi: 10.1111/j.1439-0388.2010.00874.x

Li Y, Li Q, Zhao X, Xie Z, Xu Y (2011) Complete sequence of the Tibetan Mastiff mitochondrial genome and its phylogenetic relationship with other Canids (Canis, Canidae). Animal 5: 18–25. doi: 10.1017/S1751731110001370

Li Y, Zhang YP (2012) High genetic diversity of Tibetan Mastiffs revealed by mtDNA sequences. Chinese Science Bulletin 57: 1483–1487. doi: 10.1007/s11434-012-4995-4

Linacre A, Gusmão L, Hecht W, Hellmann AP, Mayr WR, Parson W, Prinz M, Schneider PM, Morling N (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Science International, Genetics 5: 501–505. doi: 10.1016/j.fsigen.2010.10.017

Linacre A, Tobe SS (2011) An overview to the investigative approach to species testing in wildlife forensic science. Investigative Genetics 2: 2. doi: 10.1186/2041-2223-2-2

Melton T, Sikora J, Fernandes V, Pereira L (2011) FidoTyping™ and FidoSearch™: Validation of a forensic canine mitochondrial DNA protocol and a new on-line canid mitochondrial hypervariable region database. Mitotyping Technologies, State College, PA, USA, and Institute of Pathology and Molecular Immunology, Porto, Portugal, http://www.mitotyping.com/59859292811581/lib/59859292811581/_files/FidoTyping.ppt

Muñoz-Fuentes V, Darimont CT, Paquet PC, Leonard JA (2010) The genetic legacy of extirpation and re-colonization in Vancouver Island wolves. Conservation Genetics 11: 547–556. doi: 10.1007/s10592-009-9974-1

Nass MM (1969) Mitochondrial DNA. I. Intramitochondrial distribution and structural relations of single- and double-length circular DNA. Journal of Molecular Biology 42: 521–528. doi: 10.1016/0022-2836(69)90240-X

Okumura N, Ishiguro N, Nakano M, Matsui A, Sahara M (1996) Intra- and interbreed genetic variations of mitochondrial DNA major non-coding regions in Japanese native dog breeds (*Canis familiaris*). Animal Genetics 27: 397–405. doi: 10.1111/j.1365-2052.1996.tb00506.x

Okumura N, Ishiguro N, Nakano M, Matsui A, Shigehara N, Nishimoto T, Sahara M (1999) Variations in mitochondrial DNA of dogs isolated from archaeological sites in Japan and neighbouring islands. Anthropological Science 107: 213–228. doi: 10.1537/ase.107.213

Oskarsson MC, Klütsch CF, Boonyaprakob U, Wilton A, Tanabe Y, Savolainen P (2012) Mitochondrial DNA data indicate an introduction through Mainland Southeast Asia for Australian dingoes and Polynesian domestic dogs. Proceedings of the Royal Society B 279: 967–974. doi: 10.1098/rspb.2011.1395

Pang JF, Klütsch C, Zou XJ, Zhang AB, Luo LY, Angleby H, Ardalan A, Ekström C, Sköllermo A, Lundeberg J, Matsumura S, Leitner T, Zhang YP, Savolainen P (2009) mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. Molecular Biology and Evolution 26: 2849–2864. doi: 10.1093/molbev/msp195

Parra D, Méndez S, Cañón J, Dunner S (2008) Genetic differentiation in pointing dog breeds inferred from microsatellites and mitochondrial DNA sequence. Animal Genetics 39: 1–7. doi: 10.1111/j.1365-2052.2007.01658.x

Parson W, Bandelt HJ (2007) Extended guidelines for mtDNA typing of population data in forensic science. Forensic Science International, Genetics 1: 13–19. doi: 10.1016/j.fsigen.2006.11.003

Parson W, Brandstätter A, Alonso A, Brandt N, Brinkmann B, Carracedo A, Corach D, Froment O, Furac I, Grzybowski T, Hedberg K, Keyser-Tracqui C, Kupiec T, Lutz-Bonengel S, Mevag B, Ploski R, Schmitter H, Schneider P, Syndercombe-Court D, Sørensen E, Thew H, Tully G, Scheithauer R (2004) The EDNAP mitochondrial DNA population

database (EMPOP) collaborative exercises: organisation, results and perspectives. Forensic Science International 139: 215–226. doi: 10.1016/j.forsciint.2003.11.008

Parson W, Dür A (2007) EMPOP–a forensic mtDNA database. Forensic Science International, Genetics 1: 88–92. doi: 10.1016/j.fsigen.2007.01.018

Parson W, Roewer L (2010) Publication of population data of linearly inherited DNA markers in the International Journal of Legal Medicine. International Journal of Legal Medicine 124: 505–509. doi: 10.1007/s00414-010-0492-y

Pereira F, Carneiro J, van Asch B (2010) A Guide for Mitochondrial DNA Analysis in Non-Human Forensic Investigations. The Open Forensic Science Journal 3: 33–44.

Pereira L, Cunha C, Amorim A (2004a) Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. International Journal of Legal Medicine 118: 132–136. doi: 10.1007/s00414-003-0424-1

Pereira L, van Asch B, Amorim A (2004b) Standardisation of nomenclature for dog mtDNA D-loop: a prerequisite for launching a *Canis familiaris* database. Forensic Science International 141: 99–108. doi: 10.1016/j.forsciint.2003.12.014

Pires AE, Ouragh L, Kalboussi M, Matos J, Petrucci-Fonseca F, Bruford MW (2006) Mitochondrial DNA sequence variation in Portuguese native dog breeds: diversity and phylogenetic affinities. Journal of Heredity 97: 318–330. doi: 10.1093/jhered/esl006

Prieto L, Zimmermann B, Goios A, Rodriguez-Monge A, Paneto GG, Alves C, Alonso A, Fridman C, Cardoso S, Lima G, Anjos MJ, Whittle MR, Montesino M, Cicarelli RM, Rocha AM, Albarrán C, de Pancorbo MM, Pinheiro MF, Carvalho M, Sumita DR, Parson W (2011) The GHEP-EMPOP collaboration on mtDNA population data–A new resource for forensic casework. Forensic Science International, Genetics 5: 146–151. doi: 10.1016/j.fsigen.2010.10.013

Randi E, Lucchini V, Christensen MF, Mucci N, Funk SM, Dolf G, Loeschcke V (2000) Mitochondrial DNA Variability in Italian and East European Wolves: Detecting the Consequences of Small Population Size and Hybridization. Conservation Biology 14: 464–473. doi: 10.1046/j.1523-1739.2000.98280.x

Röck A, Irwin J, Dür A, Parsons T, Parson W (2011) SAM: String-based sequence search algorithm for mitochondrial DNA database queries. Forensic Science International, Genetics 5: 126–132. doi: 10.1016/j.fsigen.2010.10.006

Rothuizen J, de Gouw H, Hellebrekers LJ, Lenstra HA (1995) Variable structures of mitochondrial DNA in dogs. Veterinary Quarterly 17: 22–23. doi: 10.1080/01652176.1995.9694575

Ryabinina OM (2006) Mitochondrial DNA Variation in Asian Shepherd Dogs. Russian Journal of Genetics 42: 748–751. doi: 10.1134/S1022795406070088

Salas A, Bandelt HJ, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. Forensic Science International 168: 1–13. doi: 10.1016/j.forsciint.2006.05.037

Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochemical and biophysical research communications 335: 891–899. doi: 10.1016/j.bbrc.2005.07.161

Sato M, Sato K (2013) Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. Biochimica et biophysica acta 1833: 1979–1984. doi: 10.1016/j.bbamcr.2013.03.010

Saunier JL, Irwin JA, Strouss KM, Ragab H, Sturk KA, Parsons TJ (2009) Mitochondrial control region sequences from an Egyptian population sample. Forensic Science International, Genetics 3: e97–103. doi: 10.1016/j.fsigen.2008.09.004

Savolainen P, Leitner T, Wilton AN, Matisoo-Smith E, Lundeberg J (2004) A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. Proceedings of the National Academy of Sciences of the USA 101: 12387–12390. doi: 10.1073/pnas.0401814101

Savolainen P, Lundeberg J (1999) Forensic evidence based on mtDNA from dog and wolf hairs. Journal of Forensic Sciences 44: 77–81.

Savolainen P, Rosén B, Holmberg A, Leitner T, Uhlén M, Lundeberg J (1997) Sequence analysis of domestic dog mitochondrial DNA for forensic use. Journal of Forensic Sciences 42: 593–600.

Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T (2002) Genetic evidence for an East Asian origin of domestic dogs. Science 298: 1610–1613. doi: 10.1126/science.1073906

Scharnhorst G, Kanthaswamy S (2011) An assessment of scientific and technical aspects of closed investigations of canine forensics DNA–case series from the University of California, Davis, USA. Croatian Medical Journal 52: 280–292. doi: 10.3325/cmj.2011.52.280

Schneider PM, Seo Y, Rittner C (1999) Forensic mtDNA hair analysis excludes a dog from having caused a traffic accident. International Journal of Legal Medicine 112: 315–316. doi: 10.1007/s004140050257

Sharma DK, Maldonado JE, Jhala YV, Fleischer RC (2004) Ancient wolf lineages in India. Proceedings of the Royal Society of London B (Supplement) 271: S1–S4. doi: 10.1098/rsbl.2003.0071

Sindičić M, Gomerčić T, Galov A, Arbanasić H, Kusak J, Slavica A, Huber Đ (2011) Mitochondrial DNA control region as a tool for species identification and distinction between wolves and dogs from Croatia. Veterinarski Archiv 81: 249–258.

Smalling BB, Satkoski JA, Tom BK, Szeto WY, Erickson BJ-A, Spear TF, Smith DG, Budowle B, Webb KM, Allard M, Kanthaswamy S (2010) Geographic Differences in Mitochondrial DNA (mtDNA) Distribution Among United States (US) Domestic Dog Populations. The Open Forensic Science Journal 3: 22–32.

Suárez NM, Betancor E, Fregel R, Pestano J (2013) Genetic characterization, at the mitochondrial and nuclear DNA levels, of five Canary Island dog breeds. Animal Genetics. doi: 10.1111/age.12024

Sundqvist AK, Björnerfeldt S, Leonard JA, Hailer F, Hedhammar Å, Ellegren H, Vilà C (2006) Unequal contribution of sexes in the origin of dog breeds. Genetics 172: 1121–1128. doi: 10.1534/genetics.105.042358

Takahasi S, Miyahara K, Ishikawa H, Ishiguro N, Suzuki M (2002) Lineage classification of canine inheritable disorders using mitochondrial DNA haplotypes. The Journal of Veterinary Medical Science 64: 255–259. doi: 10.1292/jvms.64.255

Tsuda K, Kikkawa Y, Yonekawa H, Tanabe Y (1997) Extensive interbreeding occurred among multiple matriarchal ancestors during the domestication of dogs: evidence from inter- and intraspecies polymorphisms in the D-loop region of mitochondrial DNA between dogs and wolves. Genes & Genetic Systems 72: 229–238. doi: 10.1266/ggs.72.229

Tully G, Bär W, Brinkmann B, Carracedo A, Gill P, Morling N, Parson W, Schneider P (2001) Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. Forensic Science International 124: 83–91. doi: 10.1016/S0379-0738(01)00573-4

Valière N, Fumagalli L, Gielly L, Miquel C, Lequette B, Poulle M-L, Weber J-M, Arlettaz R, Taberlet P (2003) Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. Animal Conservation 6: 83–92. doi: 10.1017/S1367943003003111

van Asch B, Albarran C, Alonso A, Angulo R, Alves C, Betancor E, Catanesi CI, Corach D, Crespillo M, Doutremepuich C, Estonba A, Fernandes AT, Fernandez E, Garcia AM, Garcia MA, Gilardi P, Gonçalves R, Hernández A, Lima G, Nascimento E, de Pancorbo MM, Parra D, Pinheiro Mde F, Prat E, Puente J, Ramírez JL, Rendo F, Rey I, Di Rocco F, Rodríguez A, Sala A, Salla J, Sanchez JJ, Solá D, Silva S, Pestano Brito JJ, Amorim A (2009) Forensic analysis of dog (*Canis lupus familiaris*) mitochondrial DNA sequences: an inter-laboratory study of the GEP-ISFG working group. Forensic Science International, Genetics 4: 49–54. doi: 10.1016/j.fsigen.2009.04.008

van Asch B, Pereira L, Pereira F, Santa-Rita P, Lima M, Amorim A (2005) MtDNA diversity among four Portuguese autochthonous dog breeds: a fine-scale characterisation. BMC Genetics 6: 37. doi: 10.1186/1471-2156-6-37

Vilà C, Maldonado JE, Wayne RK (1999) Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. Journal of Heredity 90: 71–77. doi: 10.1093/jhered/90.1.71

Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK (1997) Multiple and ancient origins of the domestic dog. Science 276: 1687–1689. doi: 10.1126/science.276.5319.1687

Webb K, Allard M (2010) Assessment of minimum sample sizes required to adequately represent diversity reveals inadequacies in datasets of domestic dog mitochondrial DNA. Mitochondrial DNA 21: 19–31. doi: 10.3109/19401730903532044

Webb KM, Allard MW (2009a) Identification of forensically informative SNPs in the domestic dog mitochondrial control region. Journal of Forensic Sciences 54: 289–304. doi: 10.1111/j.1556-4029.2008.00953.x

Webb KM, Allard MW (2009b) Mitochondrial genome DNA analysis of the domestic dog: identifying informative SNPs outside of the control region. Journal of Forensic Sciences 54: 275–288. doi: 10.1111/j.1556-4029.2008.00952.x

Wetton JH, Higgs JE, Spriggs AC, Roney CA, Tsang CS, Foster AP (2003) Mitochondrial profiling of dog hairs. Forensic Science International 133: 235–241. doi: 10.1016/S0379-0738(03)00076-8

Wilson MR, Allard MW, Monson K, Miller KW, Budowle B (2002a) Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Science International 129: 35–42. doi: 10.1016/S0379-0738(02)00206-2

Wilson MR, Allard MW, Monson KL, Miller KWP, Budowle B (2002b) Further Discussion of the Consistent Treatment of Length Variants in the Human Mitochondrial DNA Control Region. Forensic Science Communications 4.

Wilson MR, Stoneking M, Holland MM, DiZinno JA, Budowle B (1993) Guidelines for the Use of Mitochondrial DNA Sequencing in Forensic Science. Crime Laboratory Digest 20: 68–77.

Yao YG, Bravi CM, Bandelt HJ (2004) A call for mtDNA data quality control in forensic science. Forensic Science International 141: 1–6. doi: 10.1016/j.forsciint.2003.12.004

Yao YG, Salas A, Logan I, Bandelt HJ (2009) mtDNA data mining in GenBank needs surveying. The American Journal of Human Genetics 85: 929–933; author reply 933. doi: 10.1016/j.ajhg.2009.10.023

Zimmermann B, Röck A, Huber G, Krämer T, Schneider PM, Parson W (2011) Application of a west Eurasian-specific filter for quasi-median network analysis: Sharpening the blade for mtDNA error detection. Forensic Science International, Genetics 5: 133–137. doi: 10.1016/j.fsigen.2010.10.003