Molecular Phylogenetics and Evolution 57 (2010) 509–517

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The genus *Drosophila* as a model for testing tree- and character-based methods of species identification using DNA barcoding

Amir Yassin [a,*,1], Therese A. Markow [b], Apurva Narechania [a], Patrick M. O'Grady [c], Rob DeSalle [a]

[a] Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West at 79th St., NY 10024, USA
[b] Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, USA
[c] Department of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 11 October 2009
Revised 2 August 2010
Accepted 19 August 2010
Available online 25 August 2010

Keywords:
Integrative taxonomy
Phylogenetic species concept
Speciation
COI
Introgression

ABSTRACT

DNA barcoding has recently been proposed as a promising tool for the (1) rapid assignment of unknown samples to described species by non-expert workers and (2) a potential method of new species discovery based on degree of DNA sequence divergence. Two broad methods have been used, one based on degree of DNA sequence variation, within and between species and another requiring the recovery of species as discrete clades (monophyly) on a phylogenetic tree. An alternative method relies on the identification of a set of specific diagnostic nucleotides for a given species (characters). The genus *Drosophila* has long served as a model system in genetics, development, ecology and evolutionary biology. As a result of this work, species boundaries within this genus are quite well delimited, with most taxa being defined by morphological characters and also conforming to a biological species concept (e.g., partial or complete reproductive isolation has used to erect and define species). In addition, some of the species in this group have also been subjected to phylogenetic analysis, yielding cases where taxa both conform and conflict with a phylogenetic species concept. Here, we analyzed 1058 *COI* sequences belonging to 68 species belonging to *Drosophila* and its allied genus *Zaprionus* and with more than a single representative to assess the performance of the three DNA barcoding methods. 26% of the species could not be defined using distance methods, i.e. had a barcoding gap of $\leqslant 0$, and 23% were not monophyletic. We focused then on four groups of closely-related species whose taxonomy is well-established on non-molecular basis (e.g., morphology, geography, reproductive isolation) and to which most of the problematic species belonged. We showed that characters performed better than other approaches in the case of paraphyletic species, but all methods failed in the case of polyphyletic species. For these polyphyletic species, other sources of evidence (e.g., morphology, geography, reproductive isolation) are more relevant than *COI* sequences, highlighting the limitation of DNA barcoding and the needs for integrative taxonomy approaches. In conclusion, DNA barcoding of *Drosophila* shows no reason to alter the 250 years old tradition of character-based taxonomy, and many reasons to shy away from the alternatives.

1. Introduction

The Barcode of Life initiative, launched in 2003 (Hebert et al., 2003) has a major goal- the rapid identification of already described species using a short stretch of the mitochondrial (mt) cytochrome oxidase I (COI) gene. This process has been called "species identification" (DeSalle, 2006, 2007; Rubinoff, 2006a,b; Rach et al., 2008). The process of species identification should be kept clear and distinct from other proposed uses of DNA sequence information in taxonomy and biodiversity studies, such as DNA taxonomy or "species discovery" using DNA sequences. Proponents have touted this approach as a solution to the so-called "taxonomic impediment" because it allows non-specialists to assign unknown samples to species using a simple PCR reaction, rather than detailed knowledge of morphology of the organism under study. Barcoding is also scalable, allowing thousands of samples to be done in parallel. The exponential growth of the number of DNA barcoding publications demonstrates the wide popularity that the initiative has gained and is continuing to gain (nearly 770 papers in Web of Science © as of July 2010).

Presently, most methods of DNA barcoding are tree-based and can fall into two broadly defined classes, distance or phylogeny-based. The first class of methods converts DNA sequences into genetic distances and then uses these distances to establish identification schemes. This approach defines a similarity threshold

* Corresponding author. Fax: +1 212 769 5277.
E-mail addresses: ayassin@amnh.org, yassin@ijm.univ-paris-diderot.fr (A. Yassin).
[1] Present address: Institut Jacques Monod, Centre National de la Recherche Scientifique, Université Paris 7 Diderot, Bât. Buffon 302B; 15 Hélène Brion St., 75205 Paris, France.

1055-7903/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2010.08.020

below which a DNA barcode is assigned to a known or a new species. While rapid taxonomic identification and species descriptions are essential in the face of the current biodiversity crisis (DeSalle and Amato, 2004; Savolainen et al., 2005; Lahaye et al., 2008), opponents of barcoding have been vocal in their criticisms of this approach, particularly as it applies to the discovery of new species. The main reason for this critique has been the somewhat arbitrary use of varying degrees of sequences divergence to assign unknowns to described or new species. Such a threshold was initially proposed to be about 3% sequence divergence (Hebert et al., 2003), and was later reduced to 1% (Ratnasingham and Hebert, 2007). Usually, reciprocal distances are visualized on a neighbor-joining (NJ) tree. However, given the diversity in mechanisms of species formation and the non-uniformity of the species rank across taxa, the application of such a universal degree of divergence is fraught with problems. Several authors (e.g., Hebert et al., 2004; Burns et al., 2007) subsequently proposed the notion of a "barcoding gap," a distance-gap between intra- and inter-specific sequences (Meyer and Paulay, 2005; Meier et al., 2006, 2008).

The second approach uses monophyly on a phylogenetic tree to assign unknown taxa to a known or new species. Statistics (bootstrap proportions) or (posterior probabilities) are often, but not always used to support barcoding conclusions (Abdo and Golding, 2007; Munch et al., 2008a,b; Lou and Golding, 2010). Two issues complicate the use of monophyly in a barcoding framework. First, the long-recognized problem of incomplete lineage sorting will yield gene genealogies that may differ in topology from locus to locus (Nielsen and Matz, 2006). Furthermore, recently divergent taxa may not be reciprocally monophyletic due to lack of time needed to coalesce (Hudson and Coyne, 2002; Knowles and Carstens, 2007). A second problem with this approach is that monophyly, while a discrete criterion, is arbitrary with respect to taxonomic level. This means that, in some taxa, clades will be evident at the subspecific level and in other lineages, clades will not appear until the subgeneric or higher levels. Therefore, monophyly may not correspond completely with a biological species – or any other species definition other than a phylogenetic one (Meier, 2008; Tan et al., 2008). In a survey of mitochondrial DNA phylogenies, 23% of 2319 nominal species were shown to be paraphyletic or polyphyletic, indicating the commonness of this phenomenon (Funk and Omland, 2003). Deciding where to draw the line in species diagnosis and delimitation is akin to the "lumping and splitting" debate that continues to frustrate traditional taxonomists.

DeSalle et al. (2005) have proposed an alternative to tree-based approaches for DNA barcoding. This method identifies a set of diagnostic nucleotides in the DNA barcode sequence. The four standard nucleotides (A, T, C, G) if found in fixed states in one species can be used as simple pure diagnostics for identifying that species. In addition, sites that are polymorphic within a species can be used in combination with other sites as compound pure diagnostics to identify a single species. Adding these compound diagnostics augments the number of diagnostic character systems available to DNA barcoding. Hebert et al. (2003) have suggested that a DNA sequence of 600 nucleotides of which 15 can be assumed to be neutrally mutable permutations of the 4 DNA sequence character states can result in 4^{15} or 10^{10} diagnostics. If we expand the number of sites that can be useful in DNA barcoding and add compound pure diagnostics as possible sources of diagnostics, the number of potential barcodes massively exceeds the estimated number of extant species. Such a powerful identification potential can persist regardless to the rate of speciation or to its phylogenetic history, i.e. a maternal paraphyletic species can still be identified in light of the symplesiomorphic (ancestral) nucleotides.

In this study, we are interested in the limits of DNA barcoding as a tool in species identification. Since the goal of species identification is to use existing taxonomy as a guide and to then extract DNA sequence information that will reflect the existing taxonomy, a well known and worked out taxonomic system is needed. The family Drosophilidae offers a unique model with which to compare and contrast various DNA barcoding approaches. This family contains 4000 described species, many of which have been erected using a multitude of data types, from morphology and mating ability to genetic and genomic information (DeSalle and Grimaldi, 1991; Markow and O'Grady, 2006). While the genus-level phylogenetics of this family are not completely resolved (e.g., O'Grady et al., 2008), a number of well resolved species groups have been proposed. Thus current subgeneric classification scheme was proposed by A. H. Sturtevant (1939) to "be taken as indicating their degree of genetic relatedness." Even given the large amount of data generated since the 1930's, the species groups proposed by Sturtevant (and others) have remained well supported and are likely important units of evolutionary change. These have served as evolutionary models for the study of the mechanisms of speciation (Mallet, 2006), resulting in a wealth of genetic data for a wide range of closely-related species of a well-established taxonomy. It is under such a well resolved rubric that the various types of DNA barcoding (tree-based vs. character-based methods) can be tested and refined. The implications of this work are likely to be relevant to other areas of biology where rapid species identification is necessary but where taxonomic expertise is lacking. In addition, by examining the limits of species identification, we can also get a clearer idea of how DNA sequence information can be used in species discovery.

2. Materials and methods

2.1. Data mining

In total, nearly 1600 drosophilid *COI* sequences belonging to 301 species, 17 genera and two subfamilies were retrieved from GenBank, among which 1400 sequences could be correctly aligned, i.e. without extensive internal gaps, with CLUSTAL W (Thompson et al., 1994) using the default parameters implemented in MEGA 4 software package (Kumar et al., 2008). Analyses were limited to species with more than a single representative. This resulted in 1058 *COI* sequences belonging to two 68 species, two genera and one subfamily.

2.2. Estimation of the bracoding gap

The data matrix was not directly used in further distance and phylogenetic analyses because of the discrepancy in sequence length within and between species. A "barcoding gap" was, however, estimated as the difference between the maximal sequence divergence within species and the divergence from the closest relative species using sequence identity scores generated by BLAST (Camacho et al., 2009). In cases, where negative gaps were obtained, i.e. there was a sequence overlap between species, the neighbor-joining (NJ) distance tree inferred by BLAST under the JC-substitution model was used to examine the monophyly of the query species.

2.3. Distance and phylogenetic analyses of problematic groups

The BLAST analysis permitted to examine the extent of problematic species groups, i.e. those with overlapping sequences and non-monophyletic species. Four groups were identified and were thus subject to further analyses. For distance analyses, MEGA was used to generate pairwise intraspecific distance matrices and to reconstruct NJ trees using the K2-substitution model as recommended by Hebert et al. (2003). In addition to the MEGA-generated

NJ tree in each group, Bayesian phylogenetic trees were reconstructed for each group using the BEAST software package (Drummond and Rambaut, 2007) under coalescent and Yule speciation models separately. Two simultaneous runs of 10,000,000 generations by sampling every 1000 generations and using a burn-in period of 1000,000 generations under the GTR + I + Γ substitution model were conducted. The substitution model was suggested by the FINDMODEL software package (http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html).

For character-based analysis, the MESQUITE software package (Maddison and Maddison, 2009) was used to assign sequences to their nominal species, and the nexus file was exported to the CAOS software package (Sarkar et al., 2008) in order to define for each nominal species the set of its *COI* diagnostic nucleotides.

3. Results

3.1. COI intra- and inter-specific variations in Drosophila

Table 1 shows the width of the barcoding gap for the 68 drosophilid species. Maximal intraspecific distances ranged from 0% to 11% with a mean of 1.9 ± 0.2%. Among these scores, 28 species (41%) were above the 1% threshold, and 8 (12%) above the 3% threshold. Minimal inter-specific distances ranged from 0 to 12% with a mean of 5.1 ± 0.4%. Among these scores, 11 species (16%) were under the 1% threshold, and 41 (60%) under the 3% thresholds. The width of the barcoding gaps ranged from −5% to 11%. There were 10 species (14%) with negative gaps and 8 (12%) with a gap width of 0. For tree-based analysis, 10 species (14%) were polyphyletic, and 6 (9%) were paraphyletic. This raises the proportion of non-monophyletic species in *Drosophila* to 23%, in concordance with other phylogenetic estimates in insects. However, most of the problematic species belonged to four species groups that were investigated in more details subsequently.

3.2. The quinaria species group

The *quinaria* species group consists of 33 mostly holarctic species, four of which have been extensively used in evolutionary genetics studies: *D. falleni*, *D. innubila*, *D. subquinaria* and *D. recens*. These species are morphologically very similar, distinguished mainly by their abdominal pigmentation pattern and genitalia, with *D. innubila* more resembling *D. falleni* and *D. recens* more resembling *D. subquinaria*. The four species have been investigated for meiotic drive driven by sex-ratio distortion or the endosymbiont *Wolbachia*. Both phenomena are relevant to mitochondrial DNA variation within and between species (Shoemaker et al., 1999; Dyer and Jaenike, 2004). As shown in the *COI* NJ phenogram given in Fig. 1a, mtDNA variability in the two species *D. innubila* and *D. recens*, both infected by *Wolbachia*, is remarkably low in comparison to the two other uninfected species. *D. falleni* and *D. innubila* each form distinct clusters and can thus be identified using monophyletic criterion on the NJ tree (Fig. 1a) and the Bayesian tree (not shown). However, both species fall below the 1% divergence threshold (Ratnasingham and Hebert, 2007), and thus would be considered conspecific following phenetic barcoding methods using this cutoff. The same problem is encountered with the other species pair (*D. subquinaria* and *D. recens*) both falling below the 1% threshold. Crosses between these two species show asymmetrical hybrid inviability whose degree is strongly correlated to the *Wolbachia* infection among the cross mates (Shoemaker et al., 1999; Dyer and Jaenike, 2004). Moreover, the two species share some mitochondrial haplotypes (Fig. 1a) and neither of them forms a distinct monophyletic clade on the NJ tree even when removing the shared haplotypes (although monophyly was recovered on the Bayesian tree). Distance methods cannot distinguish between the four species because no boundary between intraspecific and inter-specific mtDNA divergences, i.e. DNA barcoding gap *sensu* Meier et al. (2008), exists (Fig. 1b). On the other hand, the application of character-based methods *sensu* DeSalle et al. (2005) can provide a combination of diagnostic nucleotides that can correctly identify the four species (Fig. 1c). Again in the character-based analysis the introgressed four haplotypes were removed as they disable DNA barcoding identification using any mitochondrial marker regardless to the identification method.

3.3. The pseudoobscura species complex

The *pseudoobscura* species complex provides another interesting case. Since the discovery of partial reproductive isolation and distinct chromosomal rearrangements between nearctic populations of the two sibling species *pseudoobscura* and *persimilis* in the 1930s, the two species have stood as an evolutionary genetic paradigm for speciation studies (Machado and Hey, 2003; Mallet, 2006). They were first considered as races and presented a taxonomic dilemma as no reliable morphological diagnoses were found except some morphometric indices (Dobzhansky and Epling, 1944). In the 1960s, an isolated population of *D. pseudoobscura* was discovered in the Andean mountains near Bogota, and also showed partial reproductive isolation from North American populations (Ayala and Dobzhansky, 1975). Because the Columbian population shared the same chromosomal configurations of *D. pseudoobscura* with no morphological difference, it was considered a subspecies and called *D. pseudoobscura bogotana*. Although such a taxonomic distinction would suggest the latter subspecies to be more closely-related to its conspecific *D. pseudoobscura pseudoobscura* than to *D. persimilis*, the NJ tree (Fig. 1d) recovered only a monophyletic *D. pseudoobscura bogotana* and reciprocally polyphyletic *D. pseudoobscura pseudoobscura* and *D. persimilis*, in concordance to geographical distribution. In spite of the partial reproductive isolation between the three taxa with hybrid males usually being sterile, no general DNA barcoding gap was found (Fig. 1e). The character-based approach also failed to distinguish between the two polyphyletic and sympatric species. Nonetheless, the monophyletic and allopatric subspecies *D. pseudoobscura bogotana* can be identified using eight fixed nucleotides (Fig. 1f).

The designation of *D. pseudoobscura bogotana* as a subspecies can then be treated as a hypothesis of species existence that can be tested using the character-based approach. The results indicate that even while lacking morphological or chromosomal characters to distinguish *D. pseudoobscura bogotana*, *COI* sequences can be used to reliably distinguish and diagnose this taxon even in the absence of an *a priori* knowledge of geographical origin of the specimens. On this basis, we suggest that this subspecies can be upgraded to the species level to become "*Drosophila bogotana*", but a thorough taxonomic revision and description have to be conducted before taking such a decision.

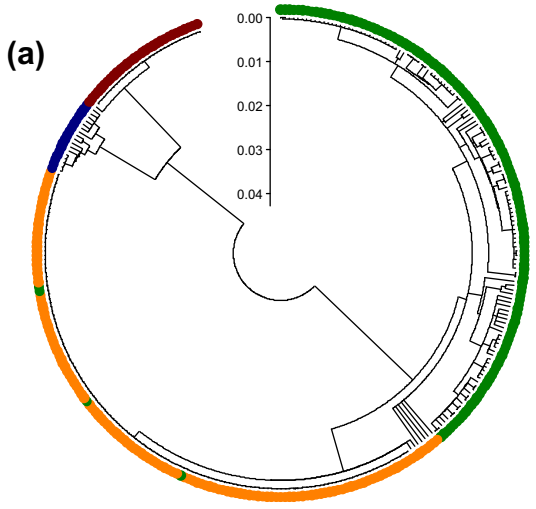
3.4. The simulans species complex

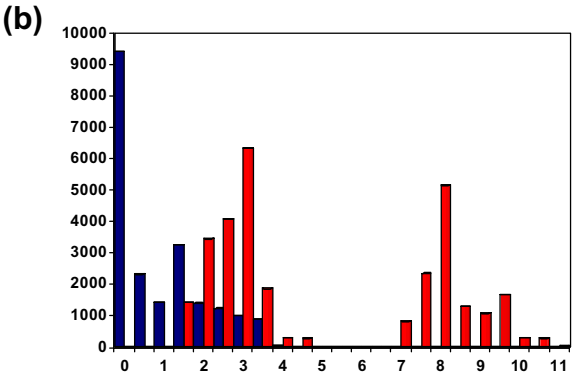
The *simulans* species complex is another classical model in evolutionary genetics and speciation research. It consists of a triad of species, one of which, *D. simulans*, is cosmopolitan, while the others, *Drosophila sechellia* and *Drosophila mauritiana*, are endemic to the Seychelles and Mauritius islands, respectively (Lachaise et al., 1988). Besides its geographical isolation, *D. sechellia* has a particular ecological niche. It breeds exclusively in fruits of *Morinda citrifolia*, which contain secondary compounds that are toxic to the other species of the *simulans* complex. The three species also show partial reproductive isolation with hybrid males being sterile, yet they can be distinguished on the bases of subtle differences of male genitalia. Early investigations of mitochondrial DNA variation

Table 1
A summary of tree-based analyses in 68 *Drosophila* species using BLAST. *N* = number of sequences, *L* = length of query sequence, min = minimal intraspecific sequence divergence (distance), max = maximal intraspecific distance, Inter = minimal inter-specific, Gap = inter − max, M = monophyletic, PA = paraphyletic, and PO = polyphyletic.

Genus (Subgenus)	Group	Species	*N*	*L*	Min	Max	Inter	Gap	Closest species	Monophyly
Drosophila	*angor*	*angor*	13	1500	0	11	11	0	*imperasitae*	M
(*Drosophila*)	*cardini*	*cardini*	3	405	1	1	7	6	*neocardini*	M
	guttifera	*guttifera*	4	440	0	2	9	7	*subquinaria*	M
	immigrans	*immigrans*	2	1304	1	1	11	10	*mercatorum*	M
	lacertosa	*lacertosa*	4	1500	1	2	7	5	*yunannensis*	M
		yunnanensis	3	1500	0	3	6	3	*lacertosa*	M
	macroptera	*macroptera*	2	413	1	1	9	8	*innubila*	M
	melanica	*euronotus*	3	439	0	0	2	2	*paramelanica*	M
		melanica	4	645	1	2	2	0	*paramelanica*	M
		micromelanica	4	1301	0	1	8	7	*tsigana*	M
		nigromelanica	2	645	1	1	7	6	*melanica*	M
		paramelanica	4	645	0	1	2	1	*euronotus*	M
		tsigana	3	1500	0	1	5	4	*longiserata*	M
	mesophragmatica	*pavani*	2	645	0	0	4	4	*gaucha*	M
	MMP	*comatifemora*	2	1263	1	1	8	7	*pectinitarsus*	M
	MT	*waddingtoni*	2	1625	0	3	2	−1	*percnosoma*	PA
	nannoptera	*nannoptera*	2	1500	1	1	12	11	*karakasa*	M
		pachea	78	661	0	2	10	8	*recens*	M
	polychaeta	*daruma*	4	1500	0	1	5	4	*latifshahi*	M
		polychaeta	2	1500	1	1	2	1	*asper*	M
	quadrisetata	*barutani*	6	1500	1	2	6	4	*spT*	M
		beppui	3	1500	1	2	6	4	*perlucida*	M
		potamophila	2	1500	1	1	1	0	*splZU*	M
	quinaria	*falleni*	18	1473	0	2	5	3	*innubila*	M
		innubila	30	1473	0	1	5	4	*falleni*	M
		limbata	2	413	1	1	3	2	*subquinaria*	M
		quinaria	3	1512	0	9	4	−5	*subquinaria*	PO
		recens	138	1432	0	1	3	2	*subquinaria*	M
		suboccidentalis	3	413	1	3	2	−1	*subquinaria*	PO
		subpalustris	2	413	0	0	2	2	*palustris*	M
		subquinaria	137	1432	0	4	3	−1	*recens*	PO
	repleta	*arizonae*	17	658	0	3	3	0	*mojavensis*	PA
		borborema	2	408	0	0	3	3	*serido*	M
		eohydei	2	408	4	4	5	1	*hydei*	PA
		hydei	5	408	0	2	5	3	*eohydei*	M
		martensis	2	408	1	1	9	8	*richardsoni*	M
		mercatorum	2	1500	1	1	10	9	*anceps*	M
		mettleri	51	591	0	1	11	10	*micromettleri*	M
		mojavensis	47	601	0	2	3	1	*arizonae*	M
		navojoa	5	1318	0	1	7	6	*tsigana*	M
		nigrospiracula	10	663	1	1	10	9	*anceps*	M
		serido	2	408	1	1	3	2	*borborema*	M
	robusta	*bai*	3	1500	5	5	10	5	*clefta*	M
		robusta	3	645	0	1	9	8	*sordidula*	M
	testacea	*neotestacea*	4	518	0	1	5	4	*orientacea*	M
		putrida	3	413	0	0	1	1	*quinaria*	PA
		testacea	2	413	1	1	5	4	*neotestacea*	M
	tripunctata	*tripunctata*	2	413	1	1	11	10	*phalerata*	M
	virilis	*kanekoi*	2	1500	1	1	9	8	*littoralis*	M
		montana	42	670	0	3	8	5	*virilis*	M
		virilis	11	670	0	2	6	4	*lummei*	M
Drosophila	*ananassae*	*bipectinata*	45	444	0	4	2	−2	*pseudoananassae*	PO
(*Sophophora*)		*malerkotliana*	20	476	0	2	0	−2	*parabipectinata*	PO
		parabipectinata	8	549	0	1	0	−1	*bipectinata*	PO
		pseudoananassae	2	476	0	0	1	1	*parabipectinata*	M
	melanogaster	*mauritiana*	2	1487	1	1	1	0	*simulans*	PO
		melanogaster	112	679	0	2	5	3	*simulans*	M
		simulans	82	599	0	3	0	−3	*mauritiana*	PA
		sechellia	2	2528	1	1	2	1	*simulans*	M
	obscura	*affinis*	3	413	0	1	8	7	*barbarae*	M
		persimilis	19	829	0	0	0	0	*pseudoobscura*	PO
		pseudoobscura	14	829	0	0	0	0	*persimilis*	PO
		pseudoobscura bogotana	14	829	0	1	2	1	*pseudoobscura*	M
	saltans	*emarginata*	4	305	0	8	11	3	*amoena*	M
		prosaltans	2	305	3	3	1	−2	*saltans*	PO
		sturtevanti	3	305	3	5	4	−1	*milleri*	M
Zaprionus	*vittiger*	*africanus*	2	670	1	1	1	0	*gabonicus*	PA
(*Zaprionus*)		*indianus*	20	670	0	2	5	3	*africanus*	M

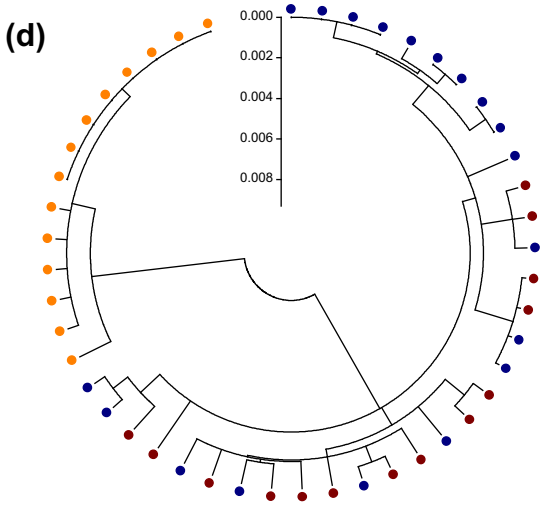
within this triad of species had revealed an interesting pattern: three mitochondrial haplotypes (called mitotypes) were found, two of which are restricted to *D. simulans* while the third is shared by all three species (Solignac and Monnerot, 1986). This is shown on the NJ tree of *COI* sequences of the *simulans* species complex (Fig. 2a) and reconfirmed on the Bayesian tree (not shown). Later investigations showed that this was due to ancient infections with three *Wolbachia* strains (Ballard, 2000). The sharing of mitotypes

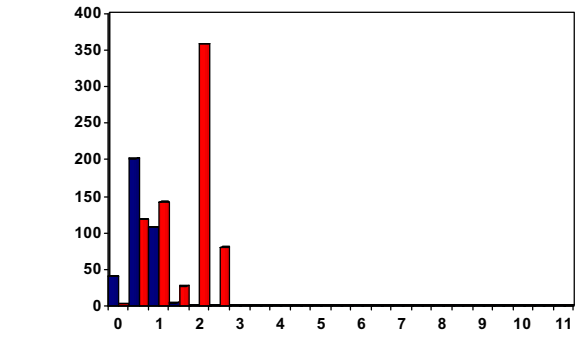




(c)

	311	313	328	364	367	379	388	448	452	457	478	502	518	529	535	545	559	676	805	943	986
falleni	T	R	C	Y	C	A	A	T	T	A	A	A	T	C	W	A	R	A	R	A	G
innubila	T	A	C	T	T	A	T	C	T	G	A	T	T	T	A	G	T	T	R	A	G
subquinaria	Y	R	T	Y	Y	T	T	Y	Y	A	T	A	A	Y	R	A	A	T	A	A	G
recens	C	A	C	C	T	T	T	T	C	A	T	A	R	C	A	A	A	R	K	T	A

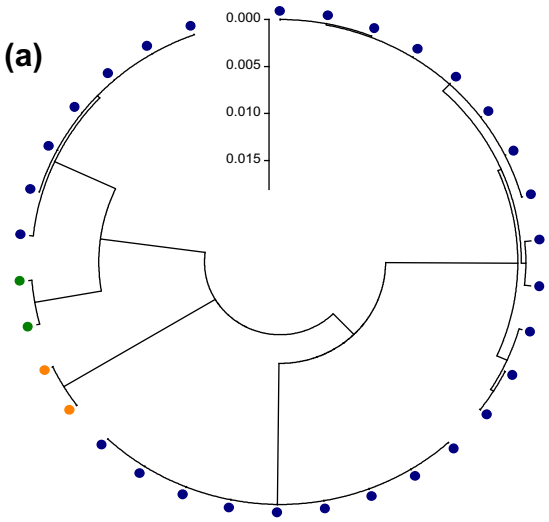


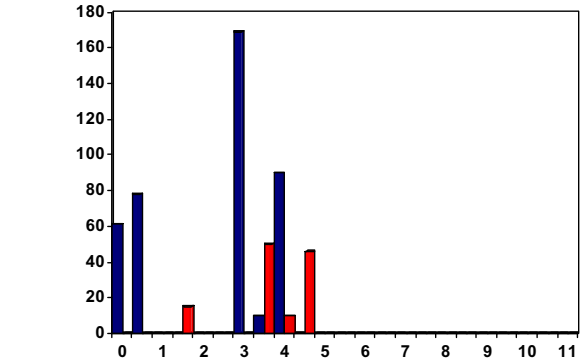


(f)

	718	775	784	818	832	868	940	946	968	1048	1090	1096	1153	1192	1195	1218	1226	1288	1309	1351	1492
persimilis	T	T	R	G	A	C	A	A	Y	T	C	C	C	G	R	A	Y	T	A	Y	T
pseudoobscura	T	Y	G	G	R	Y	A	R	C	W	Y	C	C	G	A	A	T	Y	R	C	T
bogotana	C	T	T	A	A	C	T	A	C	T	T	T	T	A	A	G	T	C	A	C	C

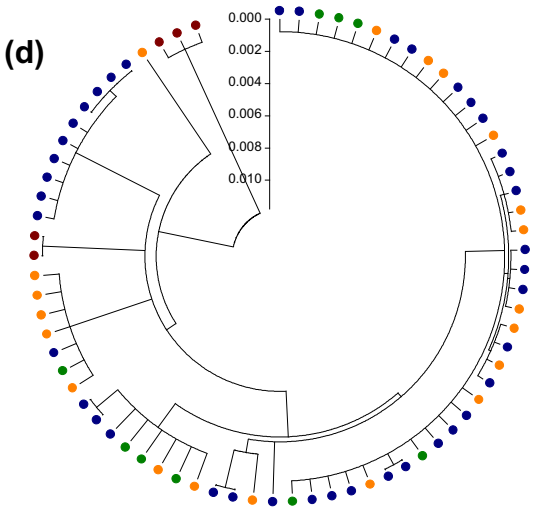
Fig. 1. Species-level DNA barcoding in the *Drosophila quinaria* species group (a–c) and the *D. pseudoobscura* species complex (d–f). (a) and (d): neighbor-joining (NJ) trees inferred from *COI* sequences. (b) and (e): histograms of intra- (in blue) and inter-specific (in red) pairwise distances between sequences. (c) and (f): combinations of diagnostic nucleotides for each species. Nucleotide numbers refer to their positions on the *D. yakuba* mitochondrial genome. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

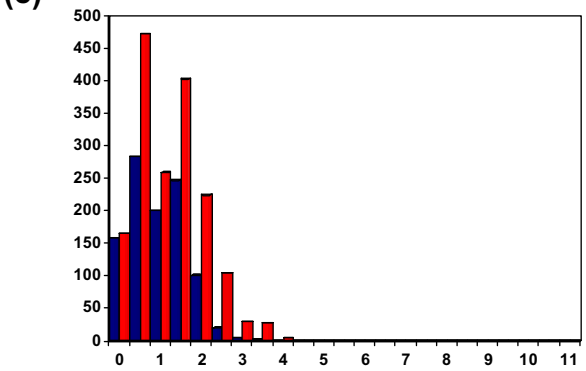




(c)

	16	70	88	94	95	127	244	304	308	329	412	473	517	574	595	616	679	784	883	904	913
simulans	A	T	A	A	T	T	A	T	T	T	C	C	A	T	A	T	C	A	Y	A	C
sechellia	G	C	A	A	T	C	A	T	T	T	T	C	A	C	G	T	C	G	A	A	C
mauritiana	A	T	T	C	C	T	G	C	C	C	T	T	T	T	A	C	T	A	G	G	T





(f)

	842	847	868	871	892	904	928	940	961	974	978	1021	1050	1051	1055	1060	1096	1114	1135	1144	1147
malerkotliana	T	R	T	T	C	A	A	A	Y	A	T	T	G	R	T	T	T	C	T	T	A
bipectinata	T	A	Y	Y	Y	A	A	A	Y	M	Y	T	K	R	Y	Y	Y	C	T	T	R
pseudoananas.	Y	R	T	Y	Y	R	R	R	C	A	T	Y	G	A	T	T	T	Y	Y	W	A
parabipectinata	T	A	T	T	C	A	A	A	Y	A	T	T	G	R	T	T	T	C	T	T	A

Fig. 2. Species-level DNA barcoding in the *Drosophila simulans* (a–c) and the *D. pseudoobscura* (d–f) species complexes. (a) and (d): neighbor-joining (NJ) trees inferred from *COI* sequences. (b) and (e): histograms of intra- (in blue) and inter-specific (in red) pairwise distances between sequences. (c) and (f): combinations of diagnostic nucleotides for each species. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the phylogenetic patterns make *D. simulans* a paraphyletic species, *i.e.* an ancestral species where two of its related but isolated populations (*D. mauritiana* and *D. sechelia*) have evolved into two distinct species (McDermott and Kliman, 2008). In this case, distance methods cannot distinguish *D. simulans*, as some of its populations are closer to other species than to their conspecific populations bearing different mitotypes (Fig. 2b), although a pure diagnostic nucleotide change can be found at site 412 for *D. simulans* (Fig. 2c) once again illustrating the efficacy of the character-based approach in identifying paraphyletic species.

3.5. The bipectinata species complex

The *bipectinata* species complex consists of four oriental species of the *ananassae* species group. This complex is interesting because in spite of its low genetic variation, partial reproductive isolation and overlapping geographical ranges, its species show a high degree of morphological divergence (Kopp and Barmina, 2005). Tree-based methods fail to discriminate the four species using the established cutoff (Fig. 2d and e), and no pure diagnostics were found using character-based barcoding (Fig. 2f). While both the character and distance-based methods fail on this data set, this example merely illustrates a failure of molecular characters to discriminate or diagnose the species where morphological characters do, a situation that should clearly demonstrate the need for a variety of character types to be used in species identification (the so-called integrative taxonomy approach; Dayrat, 2005). We point out that the molecular characters we used to examine these species are most likely neutral markers. On the other hand, there is some evidence that the morphological characters used in discriminating the four species, e.g., abdominal pigmentation and sex comb morphology, are under sexual selection and thus may follow the speciation history more closely than neutral maternally-inherited lineages. Such a conflict between molecular and sexually-selected morphological traits have also been observed in other recently diverged species that usually relies on visual cues in courtship (e.g., East African cichlids, Verheyen et al., 2003).

4. Discussion

Our examination of the limits of DNA barcoding approaches for species identification purposes have resulted in four major conclusions. Two of these conclusions concern the efficacy of DNA barcoding in species identification of drosophilids and two of the conclusions allow us to make specific statements about species discovery and taxonomic approaches using DNA barcoding approaches. First, no single distance threshold can be applied in species identification at least for drosophilids. A 1% sequence divergence threshold as proposed by Ratnasingham and Hebert (2007) will result in splitting 41% of *bona fide* species and lumping of 16% of distinct species. A 3% sequence divergence threshold as initially proposed by Hebert et al. (2003) will result in splitting 12% of *bona fide* species and lumping 76% of distinct species. Although this was previously demonstrated and suggested the use of a "barcoding gap," i.e. the difference between maximal intraspecific and minimal inter-specific distances (Meyer and Paulay, 2005; Meier et al., 2008), a gap >1 was only found in 74% of species.

Second, when species boundaries are well-defined, i.e. barcoding gap >1, all methods, whether tree-based (distance or phylogenetic) or character-based perform well in identifying species. However, in problematic cases (26% of species), distance approaches consistently fail to discriminate closely-related species in the four *Drosophila* model groups. This is in concordance with Meier et al. (2006) finding of relatively low identification success (<70%) of tree-based DNA barcoding in 449 Dipteran species. Because the major rationale for DNA barcoding as articulated by the consortium (Hebert et al., 2003) is the rapid identification of nominal species, with the eventual consequence of possible discovery of cryptic ones, this failure can be viewed as a major shortcoming of either tree-based approaches specifically or of DNA barcoding in general. Monophyly-based methods fail to correctly identify species in several *Drosophila* species groups where no DNA barcoding gap can be defined. While the character-based method also fails in polyphyletic species, paraphyletic species were still can be identified by sets of pure and combined nucleotides using the character-based approach (*sensu* DeSalle et al., 2005). Character-based approaches thus increase the identification success by nearly 9% over tree-based methods. Indeed, several studies have already shown the failure of tree-based methods to identify paraphyletic species (e.g., Trewick, 2008; Robinson et al., 2009; Lukhtanov et al., 2009; Fazekas et al., 2009; Wild, 2009).

Third, character-based DNA barcoding also allows the erection of hypotheses to assist in the discovery of new species or proposing taxonomic decisions such as in the case of the subspecies *D. pseudoobscura bogotana*. This occurs when a DNA barcode of a query specimen does not bear any of the diagnostic nucleotides of identified and barcoded species. No *a priori* assumptions of genetic similarity or cladogenic rate and pattern are thus imposed on the discovery of a species in concordance to the heterogeneity of speciation mechanisms as well as with the character-based taxonomic tradition. This is an important advantage for the character-based method because formulating nomenclatural hypotheses form a major objective of the taxonomic practice. In the Preface of the 4th edition of the International Code of Zoological Nomenclature (ICZN, 1999) it has been highlighted that the traditional nomenclature is "too permissive, in so far as it may be equally applied to paraphyletic as to monophyletic groups." Tree-based methods are, however, too prescriptive in that they only recognize monophyletic species. For example, applying a tree-based method on the *COI* sequences of *D. simulans* would split this species into two species in spite of the lack of any other morphological, ecological, geographical or reproductive isolation criterion supporting such a split.

The fourth conclusion is that no single source of data can be applied universally to identify a species (DeSalle et al., 2005). In identifying species of *Drosophila*, morphological characters are better for the *quinaria* subgroup and the *bipectinata* complex; morphological and geographical characters, for the *simulans* complex; and geographical, karyological and molecular characters, for the *pseudoobscura* subgroup. Dayrat (2005) has previously called to the use of different sources of evidence in the taxonomic practice and to not only rely on morphology, an approach he called 'integrative taxonomy.' Our study in *Drosophila* shows also that DNA sequences cannot be used by themselves in identifying species, and need always to be compared to other sources. In other words, primacy should be given to morphology or DNA sequences in identifying species according to which type of characters corroborates better with other sources of evidence (i.e. geography, ecology or reproductive isolation). This is a strong argument against DNA taxonomy using coalescent-based approaches in species delineation, which usually delineates species as monophyletic clades, such as by identifying transitions from coalescent to speciation branching patterns on a phylogenetic tree (the so-called MYC model by Pons et al., 2006). Indeed, levels of gene flow (e.g., introgressive hybridization) which can maintain the biological cohesion of polyphyletic species (Bull et al., 2006) were shown to convolute the estimated number of species using MYC (Papadopoulou et al., 2008). Moreover, demographic increase following geographical range expansion was also shown to alter the coalescence branching pattern and thus affects the MYC thresholds (Yassin et al., 2009). We have applied the MYC algorithm on the Bayesian tree reconstructed for

each of the four model species groups analyzed here. In addition to having failed to delineate paraphyletic and polyphetic species, the MYC thresholds were highly affected by the *Wolbachia*-driven selective sweeps in the *quinaria* subgroup and were unable to delineate the uninfected monophyletic species (*i.e. D. falleni* and *D. recens*). A good example of integrative taxonomy using DNA barcoding is shown in the case of skipper butterflies, wherein molecular, morphological and ecological data were used to identify species whose DNA barcodes can differ by only one to three nucleotides (Hebert et al., 2004; Burns et al., 2007). Another example from the drosophilid family is that of the invasive species *Zaprionus indianus* where reproductive isolation experiments coupled with DNA barcodes and ecological information identified two morphologically cryptic species lacking invasive capacities (Yassin et al., 2008).

In our view, species identification using DNA barcoding is a simple process whereby known species boundaries are used in conjunction with DNA sequences to establish DNA sequence diagnostics for such species. Species discovery on the other hand is an integrated decision-making process, requiring the corroboration of different sources of data that include morphology, geography, ecology, behavior and molecules. Since the dawn of modern Linnean thought, taxonomists have taken advantage of every technical advance relevant to their field (e.g., microscopy, karyology, protein mobility and DNA sequence data), as well as of species concept theory (e.g., biological, ecological and phylogenetic). However, the single taxonomic practice that sets this science apart from others has never been altered during the last 250 years. This practice is the use of sets of reliable diagnostic characters to define taxa (infraspecific, specific and ultraspecific). We see no reason to alter this approach now with the addition of DNA barcodes to the taxonomist's toolbox, and many reasons to shy away from the alternatives.

Acknowledgments

The authors are grateful for two anonymous reviewers for their constructive comments on the manuscript. This work was supported by grants from the Fondation Bettencourt-Schueller and the Richard Lounsbery Foundation to Amir Yassin. Rob DeSalle thanks the Korein Foundation at the American Museum of Natural History and the Lewis and Dorothy Cullman Program in Molecular Systematics at the AMNH.

References

Abdo, Z., Golding, G.B., 2007. A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. Syst. Biol. 56, 44–56.

Ayala, F.J., Dobzhansky, T., 1975. A new subspecies of *Drosophila pseudoobscura* (Diptera: Drosophilidae). Pan-Pac. Entomol. 50, 211–219.

Ballard, J.W.O., 2000. Comparative genomics of mitochondrial DNA in *Drosophila simulans*. J. Mol. Evol. 51, 64–75.

Bull, V., Beltrán, M., Jiggins, C.D., McMillan, W.O., Bermingham, E., Mallet, J., 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. BMC Biol. 4, 11.

Burns, J.M., Janzen, D.H., Hajibabaei, M., Hallwachs, W., Hebert, P.D.N., 2007. DNA barcodes of closely related (but morphologically and ecologically distinct species of butterflies (Hesperiidae) can differ by only one to three nucleotides. J. Lepidopt. Soc. 61, 138–153.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulous, J., Bealer, K., Madden, T.L., 2009. BLAST plus: architecture and applications. BMC Bioinf. 10, 421.

Dayrat, B., 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85, 407–415.

DeSalle, R., 2006. Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. Conserv. Biol. 20, 1545–1547.

DeSalle, R., 2007. Phenetic and DNA taxonomy; a comment on Waugh. BioEssays 29, 1289–1290.

DeSalle, R., Amato, G., 2004. The expansion of conservation genetics. Nat. Rev. Genet. 5, 702–712.

DeSalle, R., Grimaldi, D.A., 1991. Morphological and molecular systematics of the Drosophilidae. Ann. Rev. Ecol. Syst. 22, 447–475.

DeSalle, R., Egan, M.G., Siddall, M., 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Phil. Trans. R. Soc. Lond. B 360, 1905–1916.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Dobzhansky, T., Epling, C., 1944. Taxonomy, geographic distribution, and ecology of *Drosophila pseudoobscura* and its relatives. Carnegie Inst. Washington Publ. 554, 1–46.

Dyer, K.A., Jaenike, J., 2004. Evolutionary stable infection by a male-killing endosymbiont in *Drosophila innubila*: molecular evidence from the host and parasite genomes. Genetics 168, 1443–1455.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C.H., Newmaster, S.G., Hajibabaei, M., Husband, B.C., 2009. Are plant species inherently arder than animal species using DNA barcoding markers? Mol. Ecol. Resour 9, 130–139.

Funk, D.J., Omland, K.E., 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Ann. Rev. Ecol. Syst. 34, 397–423.

Hebert, P.D.N., Cywinska, A., Ball, S.L., deWaard, J.R., 2003. Biological identification through DNA barcodes. Proc. R. Soc. Lond. B 270, S96–S99.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Nat. Acad. Sci. USA 101, 14812–14817.

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

Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895.

Kopp, A., Barmina, O., 2005. Evolutionary history of the *Drosophila bipectinata* species complex. Genet. Res. 85, 23–46.

Kumar, S., Dudley, J., Nei, M., Tamura, K., 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. 9, 299–306.

Lachaise, D., Cariou, M.-L., David, J.R., Lemeunier, F., Tsacas, L., Ashburner, M., 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. 22, 159–225.

Lahaye, R., Van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G., Savolainen, V., 2008. DNA barcoding the floras of biodiversity hotspots. Proc. Nat. Acad. Sci. USA 105, 2923–2928.

Lou, M., Golding, G.B., 2010. Assigning sequences to species in the absence of large interspecific differences. Mol. Phyl. Evol. 56, 187–194.

Lukhtanov, V.A., Sourakov, A., Zakharov, E.V., Hebert, P.D.N., 2009. DNA barcoding Central Asian butterflies: increasing geographical dimension does not successfully reduce the success of species identification. Mol. Ecol. Resour. 9, 1302–1310.

Machado, C.A., Hey, J., 2003. The causes of phylogenetic conflict in a classic *Drosophila* species group. Proc. R. Soc. Lond. B. 270, 1193–1202.

Maddison, W. P., Maddison, D. R., 2009. MESQUITE: a modular system for evolutionary analysis (http://mesquiteproject.org).

Mallet, J., 2006. What has *Drosophila* genetics revealed about speciation? TREE 21, 386–393.

Markow, T.A., O'Grady, P.M., 2006. Drosophila: a Guide to Species Identification and Use. Elsevier, Amsterdam.

McDermott, S.R., Kliman, R.M., 2008. Estimation of isolation times in the *Drosophila simulans* complex. PLoS ONE 3, e2442.

Meier, R. 2008. DNA Sequences in Taxonomy: Opportunities and Challenges. Pages 95-128 in The New Taxonomy Systematics Association Special Volume Q. Wheeler, ed.

Meier, R., Kwong, S., Vaidya, G., Ng, P.K.L., 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. 55, 715–728.

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. Syst. Biol. 57, 809–813.

Meyer, C.P., Paulay, G., 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3, 2229–2238.

Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E., Nielsen, R., 2008a. Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst. Biol. 57, 750–757.

Munch, K., Boomsma, W., Willerslev, E., Nielsen, R., 2008b. Fast phylogenetic DNA barcoding. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 3997–4002.

Nielsen, R., Matz, M., 2006. Statistical approaches for DNA barcoding. Syst. Biol. 55, 162–169.

O'Grady, P.M., Lapoint, R.T., Bennett, G.M., 2008. The potential and peril of the supertree approach: a response to van der Linde and Houle. Insect Syst. Evol. 39, 269–280.

Papadopoulou, A., Bergsten, J., Fujisawa, T., Monaghan, M.T., Barraclough, T.G., Vogler, A.P., 2008. Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete clusters. Phil. Trans. R. Soc. B 363, 2987–2996.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D., Vogler, A.P., 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55, 595–609.

Rach, J., DeSalle, R., Sarkar, I.N., Schierwater, B., Hadrys, H., 2008. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proc. R. Soc. Lond. B 275, 237–247.

Ratnasingham, S., Hebert, P.D.N., 2007. BOLD: the Barcode of Life Data system (www.barcodinglife.org). Mol. Ecol. Notes 7, 355–364.

Robinson, E.A., Blagoev, G.A., Hebert, P.D.N., Adamowicz, S.J., 2009. Prospects for using DNA barcoding to identify spiders in species-rich genera. Zookeys 16, 27–46.

Rubinoff, D., 2006a. DNA barcoding evolves into the familiar. Conserv. Biol. 20, 1548–1549.

Rubinoff, D., 2006b. Utility of mitochondrial DNA barcodes in species conservation. Conserv. Biol. 20, 1026–1033.

Sarkar, I.N., Planet, P.J., DeSalle, R., 2008. CAOS software for use in character-based DNA barcoding. Mol. Ecol. Res. 8, 1256–1259.

Savolainen, V., Cowan, R.S., Vogler, A.P., Roderick, G.K., Lane, R., 2005. Towards writing the encyclopedia of life: an introduction to DNA barcoding. Phil. Trans. R. Soc. Lond. B 360, 1805–1811.

Shoemaker, D.D., Katju, V., Jaenike, J., 1999. *Wolbachia* and the evolution of reproductive isolation between *Drosophila recens* and *Drosophila subquinaria*. Evolution 53, 1157–1164.

Solignac, M., Monnerot, M., 1986. Race formation, speciation, and introgression within *Drosophila simulans*, *D. Mauritiana,* and *D. sechellia* inferred from mitochondrial DNA analysis. Evolution 40, 531–539.

Sturtevant, A.H., 1939. On the subdivision of the genus *Drosophila*. Proc. Nat. Acad. Sci. USA 25, 137–141.

Tan, S.H.D., Ali, F., Kutty, S.N., Meier, R., 2008. The need for specifying species concepts: how many species of silvered langurs (*Trachypithecus cristatus* group) should be recognized? Mol. Phylogen. Evol. 49, 688–689.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Res. 22, 4673–4680.

Trewick, S.A., 2008. DNA barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera: Acrididae). Cladistics 24, 240–254.

Verheyen, E., Salzburger, W., Snoeks, J., Meyer, A., 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science 300, 325–329.

Wild, A.L., 2009. Evolution of the Neotropical ant genus *Linepithema*. Syst. Ent. 34, 49–62.

Yassin, A., Capy, P., Madi-Ravazzi, L., Ogereau, D., David, J.R., 2008. DNA barcode discovers two cryptic species and two geographical radiations in the invasive drosophilid *Zaprionus indianus*. Mol. Ecol. Res. 8, 491–501.

Yassin, A., Amédégnato, C., Cruaud, C., Veuille, M., 2009. Molecular taxonomy and species delimitation in Andean *Schistocerca* (Orthoptera: Acrididae). Mol. Phyl. Evol. 53, 404–411.