

Berl Münch Tierärztl Wochenschr 122,  
446–450 (2009)  
DOI 10.2376/0005-9366-122-446

© 2009 Schlütersche  
Verlagsgesellschaft mbH & Co. KG  
ISSN 0005-9366

Korrespondenzadresse:  
tjard.bergmann@ecolevol.de

Eingegangen: 10.06.2009  
Angenommen: 25.08.2009

## Summary

## Zusammenfassung

U.S. Copyright Clearance Center  
Code Statement:  
0005-9366/2009/12211-446 \$ 15.00/0

Institut für Tierökologie und Zellbiologie der Stiftung Tierärztliche Hochschule Hannover<sup>1</sup>  
Zentrum für Experimentelle und Evolutionäre Biodiversitätsforschung der Stiftung Tierärztliche Hochschule Hannover<sup>2</sup>  
Physiologisches Institut der Stiftung Tierärztliche Hochschule Hannover<sup>3</sup>

## Character-based DNA barcoding: a superior tool for species classification

### *Charakter-basierte DNS Kodierung: ein überlegenes Werkzeug für die Klassifizierung von Arten*

Tjard Bergmann<sup>1</sup>, Heike Hadrys<sup>1</sup>, Gerhard Breves<sup>3</sup>, Bernd Schierwater<sup>1,2</sup>

In zoonosis research only correct assigned host-agent-vector associations can lead to success. If most biological species on Earth, from agent to host and from prokaryotes to vertebrates, are still undetected, the development of a reliable and universal diversity detection tool becomes a *conditio sine qua non*. In this context, in breathtaking speed, modern molecular-genetic techniques have become acknowledged tools for the classification of life forms at all taxonomic levels. While previous DNA-barcoding techniques were criticised for several reasons (Moritz and Cicero, 2004; Rubinoff et al., 2006a, b; Rubinoff, 2006; Rubinoff and Haines, 2006) a new approach, the so called CAOS-barcoding (Character Attribute Organisation System), avoids most of the weak points. Traditional DNA-barcoding approaches are based on distances, i. e. they use genetic distances and tree construction algorithms for the classification of species or lineages. The definition of limit values is enforced and prohibits a discrete or clear assignment. In comparison, the new character-based barcoding (CAOS-barcoding; DeSalle et al., 2005; DeSalle, 2006; Rach et al., 2008) works with discrete single characters and character combinations which permits a clear, unambiguous classification. In Hannover (Germany) we are optimising this system and developing a semiautomatic high-throughput procedure for hosts, agents and vectors being studied within the Zoonosis Centre of the „Stiftung Tierärztliche Hochschule Hannover“. Our primary research is concentrated on insects, the most successful and species-rich animal group on Earth (every fourth animal is a bug). One subgroup, the winged insects (Pterygota), represents the outstanding majority of all zoonosis relevant animal vectors.

**Keywords:** zoonosis, CAOS-barcoding, Pterygota, insects

In der Zoonosenforschung gilt, „nur eine korrekt bestimmte Wirt-Erreger-Vektoren Gemeinschaft erlaubt korrekte Forschung“. Wenn auf allen Ebenen, vom Vektor zum Wirt und vom Prokaryont bis Wirbeltier, die Mehrzahl der Arten, und damit der ökologisch relevanten Lebensformen, noch gar nicht beschrieben und erkannt ist, wird die Entwicklung von zuverlässigen und universell anwendbaren Diversitätsdetektoren zur *conditio sine qua non*. In diesem Kontext, wurden in nahezu atemberaubenden Tempo moderne molekulargenetische Arbeitstechniken anerkannte Hilfsmittel für die Bestimmung von Lebensformen auf allen taxonomischen Ebenen.

Während bisherige DNA-Barcoding-Techniken aus verschiedenen Gründen in die Kritik gerieten (Moritz and Cicero, 2004; Rubinoff et al., 2006a, b; Rubinoff, 2006; Rubinoff and Haines, 2006), umgeht eine neue Technik, das so genannte CAOS-Barcoding die Mehrzahl der bisherigen Schwachpunkte. Alle traditionellen DNA-Barcoding-Verfahren sind distanz-basiert, d. h. sie bedienen sich genetischer Distanzen und Baumbildungsverfahren zur Zuordnung und Identifikation von Arten oder Linien. Damit ist das Festlegen von „Grenzwerten“ erzwungen, und diskrete und eindeutige Zuordnungen sind nicht möglich. Das neue, so genannte charakterbasierende Barcoding-Verfahren (CAOS-Barcoding; DeSalle et al., 2005; DeSalle, 2006; Rach et al., 2008), fußt hingegen auf diskreten einzelnen Merkmalen und Merkmalskombinationen und erlaubt somit eindeutige, widerspruchsfreie Zuordnungen (Identifikationen).

In Hannover optimieren wir dieses Verfahren und entwickeln es für halbautomatische Hochdurchsatzanalysen weiter. Grundlage hierfür bilden alle Wirte, Erreger und Vektoren, die im Zoonosen-Zentrum der Tierärztlichen Hochschule Hannover bearbeitet werden. Für unsere eigene Forschung konzentrieren wir uns auf die Insekten, die erfolgreichste und artenreichste Tiergruppe der Erde (jede vierte Tierart ist ein Insekt). Eine Teilgruppe hieraus, die geflügelten Insekten (Pterygota), stellen die überragende Mehrzahl aller Zoonosen relevanter tierischer Vektoren.

**Schlüsselwörter:** Zoonose, CAOS-barcoding, Pterygota, Insekten

## Introduction

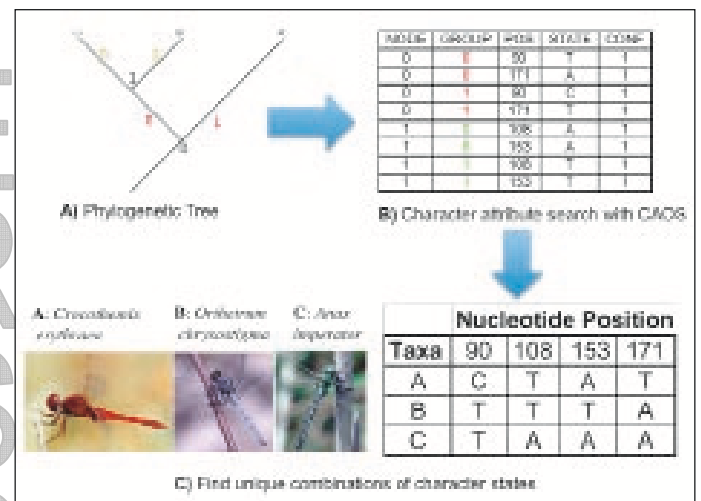
Measuring biodiversity in the field is a complex challenge, especially when the identification of the organism is restrained to a defined life history stage (e. g. the imago of many insects) or gender. Hence the announcement to establish a barcode of life based on a single DNA fragment which contains enough information to classify each possible species was for the most part welcomed with open arms and gained a lot of attention (Stoeckle, 2003; Blaxter, 2003, 2004; Savolainen et al., 2005; Frezal and Leblois, 2008). In recent years a Consortium for the Barcode of Life (CBOL, <http://barcoding.si.edu>) – an international alliance of research organisations that support the development of DNA barcoding as an international standard for species identification – was established (Marshall, 2005). Furthermore, the Barcode of Life Data Systems (<http://www.barcodinglife.org>) – a global online data management system for DNA barcodes – was developed (Ratnasingham and Hebert, 2007).

Although Hebert and colleagues (2003a, b; 2004a, b; Smith et al., 2005; Witt et al., 2006) showed that the *cox1* (cytochrome c oxidase) c-terminal fragment is a powerful marker to assign avian and a few more animal species, other scientists report that this sequence of 650 bp fails as a barcode classifier for many other (Vences et al., 2005; Meier et al., 2006; Whitworth et al., 2007; Wiemers and Fiedler, 2007; Kane and Cronq, 2008) and consequently is not sufficient to barcode all life forms. Another complaint independent of the marker gene or set of genes chosen for barcoding involves the genetic distance approach to analyse DNA barcodes which lacks uniformity, particularly when it comes to defining species boundaries (Moritz and Cicero, 2004). Although some studies have been successful in defining DNA barcodes by means of genetic distance thresholds, for example, in butterflies and crustaceans (Hebert et al., 2004b; Lefebure et al., 2006), distance threshold boundaries seem to be ill suited as a general means for species identification (Rubinoff, 2006; Rubinoff et al., 2006b; Rubinoff and Haines, 2006). Several studies showed that the intra- and interspecific diversity between species is inconsistent and varies greatly dependent on the organisms that are observed (Vences et al., 2005) leading to broad overlaps of intra- and interspecific distances (Kipling and Rubinoff, 2004; Rubinoff et al., 2006b). One cause is the frequent translocation of genetic material into and from the nucleus and mtDNA sequence, which differs between and within species and therefore leads to different rates of evolution. This bias may hinder the accurate assignment of query sequences with distance-based methods, especially in cases of insufficient taxon sampling (Meyer and Paulay, 2005; Wiemers and Fiedler, 2007).

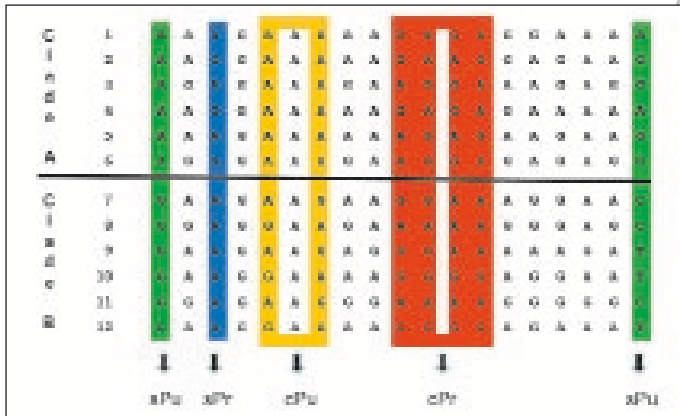
## Character-based DNA barcoding

A new method, superior to the phenetic approaches commonly used, is character-based DNA barcoding (DeSalle et al., 2005; DeSalle, 2006; Rach et al., 2008). Instead of a distance based analysis where the different sequences are compared as whole units and trees are constructed on the basis of overall similarity, the character-based model searches for diagnostic, discrete characters or combinations of characters. In the context of DNA these characters are unique SNPs (single-nucleotide polymorphisms) within DNA sequences. A character-based barcode, however, is more than the sum of several SNPs. A character-based DNA barcode is not equivalent to an SNP as a word or a sentence is not equivalent to just several letters. Single SNPs are combined through a complex algorithm to a word or a sentence to create a character-based barcode which is specific for a prescribed taxon.

These characters are not limited to DNA sequences (SNPs) alone. Amino acids (Sarkar et al., 2002b), expres-



**FIGURE 1:** Example for the application of the CAOS algorithm (Characteristic attribute organisation system): A) In the first step, after sequencing a defined barcode (here ND1) a rough guide tree is generated to identify diagnostic characters. This can be done by Maximum Likelihood, Maximum Parsimony, Bayesian or other tree building matrices. B) Starting at the root of the guide tree, CAOS searches for CAs (character attributes) at each branching node that distinguishes group 0 from group 1. This process is repeated on every node until all nodes are processed. C) The CAs discovered at each node are combined to form a character-based DNA barcode specific for each taxon, for example three insect species, like the odonates *Crocotermis erythraea*, *Orthetrum chrysostigma* and *Anax imperator*.



**FIGURE 2:** Hypothetical example of character based diagnostics (modified after Davis and Nixon, 1992). The twelve sequences represent two populations of six individuals each. The solid line through the middle of the matrix represents a barrier between the two populations. *sPu* (single pure character attributes): DNA sequence attributes in these columns are purely diagnostic characters (*sensu* Davis and Nixon, 1992). *sPr* (single private character attributes): DNA sequence attributes in this column are not purely diagnostic, but rather the G in the three individuals in the top population are 'private' to that population. *cPu* (compound pure character attributes): The DNA sequence attributes in the two columns by themselves constitute two private DNA positions. However, in combination these two columns provide a "pure" diagnostic combination (AA versus AG or GA; "compound pure" character in the terminology of Sarkar et al., 2002b). *cPr* (compound private character attributes): The four columns are neither diagnostic nor private. Yet in combination the four columns provide a diagnostic system for the top population versus the bottom. The top population is diagnosed by alternating GA or AG combinations (GAGA, AGAG, GAAG, AGGA) for the four columns, while the bottom population share combinations of GG or AA (GGGG, AAAA, GGAA).

sion patterns (Sarkar et al., 2002a) and other attributes can be used as well, so that character-based barcoding could define a taxon similar to traditional morphological identification systems and raise the number of diagnostic characters in a molecular approach to any required level of security.

The algorithm itself, the characteristic attribute organisation system (CAOS), is based on the fundamental concept that members of a given taxonomic group share attributes (e. g. polymorphisms) that are absent from other groups (Sarkar et al., 2002b, Rach et al., 2008). The CAOS algorithm thus identifies character-based diagnostics, here termed 'Characteristic Attributes' (CAs), for every clade at each branching node within a guide tree that is first produced from a given dataset (Fig. 1A and B). The resulting diagnostics can then be used for subsequent classification of new data into the taxonomic groupings represented by the guide tree (Fig. 1C). The guide tree is used by CAOS only as a means to identify diagnostic characters; it does not necessarily represent putative phylogenetic relationships. Thus, the guide tree can be generated using any tree-building method (Sarkar et al., 2002b).

CAs (Characteristic Attributes) are diagnostic character states (genes, amino acids, base pairs or even morphological, ecological or behavioural attributes) which

are present only in one clade but not in an alternate group that descends from the same node (Fig. 2). CAs are divided into four major groups: (i) Simple pure (*sPu*) CAs are found in all members of one clade and never within other clades, while (ii) simple private (*sPr*) CAs are shared only by some members of a clade but are absent from the other clades. Both pure and private CAs can either be simple CAs, which are confined to a single nucleotide position, or compound CAs. Compound CAs are combined states at multiple nucleotide positions. These nucleotide positions are not characteristic in and of themselves, but become diagnostic when this combination occurs only in one of the clades at a given node. (iii) Compound Pure (*cPu*) CAs are combinations that are found in all members of a clade and are never found together in samples outside that clade. (iv) Compound Private (*cPr*) CAs are combinations of gene expression values that are found in some members of a clade and never outside that clade (see DeSalle et al., 2005).

### Application of CAOS to identify diagnostic characters for insects

Character-based DNA barcoding was tested by our group in a most conservative approach by only considering simple pure (*sPu*) and simple private (*sPr*) CAs that are shared in at least 80% of all members in a given taxonomic unit.

An appropriate marker for species barcoding should show a high level of interspecific variability (to discriminate also between closely related sister species) and at the same time a lower intraspecific variability (for accurate assignment of specimens to species). Since ND1 (NADH dehydrogenase 1) sequences have been known to be highly informative at different taxonomic levels in dragonflies these are well suited as an alternative or complement to CO1 (Hadrys et al., 2006; Dijkstra et al., 2007; Groeneveld et al., 2007). The marker was approved by the Barcode of Life Consortium (CBOL, <http://barcoding.si.edu>).

Odonates were selected as test organisms because they provide an ideal platform for exploring the potential of character-based DNA barcoding. They represent a species rich, yet tractable insect order. Their different levels of habitat specificity and complex aquatic/terrestrial life cycles make them prominent surrogates for evaluating all types of freshwater ecosystems worldwide. While the imagos of the vast majority of species are readily identified by morphological, behavioural and life history traits, discrimination of the crucial larval stages still remains a major obstacle (Corbet, 1999). This is unfortunate since fast and reliable identification of larval biodiversity is instrumental in monitoring freshwater quality. Since odonates are near the base of ancient insects they represent basic model systems also for all higher zoonosis relevant insects.

Analyses of the ND1 sequence from 833 odonate specimens resulted in 22 distinguishable genera out of 25 with three or more genera specific CAs (Characteristic Attributes) from 30 SNPs selected by the CAOS algorithm. On the species level, 54 out of 64 were discriminable by three or more CAs from 23 SNPs. At least seven from 19 populations showed unique CAs. These results raised two basic questions: Why did this study fail to identify diagnostic barcodes for all samples included?

Is CAOS DNA barcoding an efficient and reliable technique? When considering that only simple pure (sPu) and simple private (sPr) CAs were included in our first attempt to test the potential of CAOS barcoding and that only one gene fragment (ND1) was involved in this study, the results are quite impressive. The data suggest that character-based DNA barcoding is well suited for the identification of genetic entities at different taxonomic levels. By means of the CAOS algorithm, Rach et al. (2008) were able to identify unique combinations of diagnostic characters for most of the pre-described organisms on the genus, species and population level. It is important to note again that it makes no difference for the technique, whether dragonflies or elephants or bacteria are the subject of investigation.

### Wide Impacts

The results, based on a relatively short sequence marker, showed that character-based DNA barcoding with CAOS can be an effective and reliable means for identifying diagnostics for species, sub-species and populations. The use of the barcode approach in this way can nicely complement other data in order to unravel speciation processes and identify cryptic species (DeSalle et al., 2005; Desalle, 2006). We emphasise, however, that the DNA barcodes in and of themselves do not establish that these potential units are indeed new species. Integrated taxonomic approaches (Rubinoff et al., 2006a, b) are required to accomplish a species discovery process. On the other hand, if species are already known assigning a diagnostic barcode for them is straight forward.

Because of its highly reliable classification output and easy and fast approach, a wide range of diagnostic applications can be performed with character-based DNA barcoding. It is, for example, argued that a core element of the One Health Initiative should be zoonosis research. In this matter CAOS would be an efficient tool for Monitoring Programs and Surveillance Programs which focus on the quick assignment and molecular characterisation of specimens to prevent epidemics from spreading. Once a barcode has been established, the detection of pathogens responsible for diseases, such as tuberculosis, Q-fever and others will be simple and accurate.

Another approach is the identification of cryptic species within large communities. In the last decades bats showed to be an underestimated reservoir of zoonotic diseases. While the agents of diseases are well described, only little has been researched about the factors which predestinate bats as vectors for human diseases. Molecular genetic studies indicate a high cryptic diversity within the Chiroptera and with CAOS several of the linking problems could be solved.

The practicability of CAOS barcoding for bacteria is currently being tested using the complex microfauna of the sheep rumen as a test bed. With this approach we will gain not only a better understanding of the relationship between sheep and their prokaryotes and how different forage compositions affect the microbial fauna, but we will also establish analytical routines for high-throughput identification of prokaryotes in zoonosis studies.

While character-based DNA barcoding still has to be established as a routine scientific procedure, the technology behind it is now available and ready for application

to a wide variety of questions, including those arising from zoonosis research.

### Acknowledgement

Tjard Bergmann is sponsored by the H. Wilhelm Schumann Stiftung (Grant 2009–2011).

### References

- Blaxter M (2003):** Molecular systematics: Counting angels with DNA. *Nature* 421: 122–124.
- Blaxter ML (2004):** The promise of a DNA taxonomy. *Philos Trans R Soc Lond B Biol Sci* 359:669–679.
- Corbet PS (1999):** Dragonflies: behaviour and ecology of Odonata. Ithaca, NY: Cornell University Press.
- Davis JI, Nixon KC (1992):** Populations, genetic variation, and the delimitation of phylogenetic species. *Syst Biol* 41, 421–435.
- DeSalle R (2006):** What's in a character? *J Biomed Inform* 39: 6–17.
- DeSalle R, Egan MG, Siddall M (2005):** The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philos Trans R Soc Lond B Biol Sci* 360: 1905–1916.
- Dijkstra KDB, Groeneveld LF, Clausnitzer V, Hadrys H (2007):** The Pseudagrion split: molecular phylogeny confirms the morphological and ecological dichotomy of Africa's most diverse genus of Odonata (Coenagrionidae). *Int J Odonatol* 10, 31–41.
- Frezal L, Leblois R (2008):** Four years of DNA barcoding: current advances and prospects. *Infect Genet Evol* 8: 727–736.
- Groeneveld LF, Clausnitzer V, Hadrys H (2007):** Convergent evolution of gigantism in damselflies of Africa and South America? Evidence from nuclear and mitochondrial sequence data. *Mol Phylogenet Evol* 42: 339–346.
- Hadrys H, Clausnitzer V, Groeneveld LV (2006):** The present role and future promise of conservation genetics for forest Odonates. In: Rivera A, Forests and dragonflies, Sofia, Bulgaria; Moscow, Russia: Pensoft Publishers. 279–299.
- Hebert PD, Ratnasingham S, deWaard JR (2003a):** Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* 270 Suppl 1: S96–99.
- Hebert PD, Cywinska A, Ball SL, deWaard JR (2003b):** Biological identifications through DNA barcodes. *Proc Biol Sci* 270: 313–321.
- Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004a):** Identification of Birds through DNA Barcodes. *PLoS Biol* 2: e312.
- Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004b):** Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc Natl Acad Sci U S A* 101: 14812–14817.
- Kane NC, Cronk Q (2008):** Botany without borders: barcoding in focus. *Mol Ecol* 17: 5175–5176.
- Kipling WW, Rubinoff D, (2004):** Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20, 47–55.
- Lefebure 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. *Mol Phylogenet Evol* 40: 435–447.
- Marshall, E (2005):** Taxonomy. Will DNA bar codes breathe life into classification? *Science* 307: 1037.
- 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. *Syst Biol* 55: 715–728.
- Meyer CP, Paulay G (2005):** DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol* 3: e422.
- Moritz C, Cicero C (2004):** DNA barcoding: promise and pitfalls. *PLoS Biol* 2: e354.
- Rach J, Desalle R, Sarkar IN, Schierwater B, Hadrys H (2008):** Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc Biol Sci* 275: 237–247.
- Ratnasingham S, Hebert PD (2007):** bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol Ecol Notes* 7: 355–364.
- Rubinoff D (2006):** Utility of mitochondrial DNA barcodes in species conservation. *Conserv Biol* 20: 1026–1033.
- Rubinoff D, Haines WP (2006):** Hyposmocoma molluscivora description. *Science* 311: 1377.
- Rubinoff D, Cameron S, Will K (2006a):** Are plant DNA barcodes a search for the Holy Grail? *Trends Ecol Evol* 21: 1–2.
- Rubinoff D, Cameron S, Will K (2006b):** A genomic perspective on the shortcomings of mitochondrial DNA for „barcoding“ identification. *J Hered* 97: 581–594.
- Sarkar IN, Thornton JW, Planet PJ, Figurski DH, Schierwater B, DeSalle R (2002a):** An automated phylogenetic key for classifying homeoboxes. *Mol Phylogenet Evol* 24: 388–399.
- Sarkar IN, Planet PJ, Bael TE, Stanley SE, Siddall M, DeSalle R, Figurski DH (2002b):** Characteristic attributes in cancer microarrays. *J Biomed Inform* 35: 111–122.
- Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005):** Towards writing the encyclopedia of life: an introduction to DNA barcoding. *Philos Trans R Soc Lond B Biol Sci* 360: 1805–1811.
- Smith MA, Fisher BL, Hebert PD (2005):** DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philos Trans R Soc Lond B Biol Sci* 360: 1825–1834.
- Stoeckle M (2003):** Taxonomy, DNA, and the bar code of life. *Bio-science* 53: 796–797.
- Vences M, Thomas M, van der Meijden A, Chiari Y, Vieites DR (2005):** Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Front Zool* 2: 5.
- Whitworth TL, Dawson RD, Magalon H, Baudry E. (2007):** DNA barcoding cannot reliably identify species of the blowfly genus *Protophormia* (Diptera: Calliphoridae) *Proc Biol Sci* 274(1619): 1731–1739.
- Wiemers M, Fiedler K (2007):** Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Front Zool* 4: 8.
- Witt JD, Threlloff DL, Hebert PD (2006):** DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Mol Ecol* 15: 3073–3082.

**Address for correspondence:**

Tjard Bergmann (Doktorand)  
 Institut für Tierökologie und Zellbiologie  
 Stiftung Tierärztliche Hochschule Hannover  
 Bünteweg 17d  
 30559 Hannover  
 Germany  
 tjard.bergmann@ecolevol.de