Character-based barcoding, a symbiosis and potential successor of traditional taxonomy and modern DNA barcoding

Von der Naturwissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation
von
Tjard Bergmann, Dipl.-Biol.

2019

Referent: Prof. Dr. Bernd Schierwater
1. Korreferent: Prof. Rob DeSalle, Ph.D.
2. Korreferent: Prof. Dr. Felix Felmy
Tag der Promotion: 05.02.2019

„I may not have gone where I intended to go, but I think I have ended up where I needed to be.“

Douglas Adams, The Long Dark Tea-Time of the Soul

Zusammenfassung

Klassische Taxonomie ist ein wirkungsvolles Werkzeug für die Identifikation von Tieren basierend auf Ihrer Morphologie. Probleme ergeben sich jedoch bei der Identifikation ähnlich aussehender, kryptischer Arten. Eine Lösung für dieses Problem wurde im Bauplan des Lebens, der DNS, gefunden. DNS wird zum Aufbau und der Regulierung von Proteinen verwendet. Die Struktur der DNS hat hoch spezifische Bereiche, welche innerhalb einer Art konserviert sind und sich zwischen verschiedenen Arten unterscheiden. Ein bestimmter Bereich, ein 648 bp langes Fragment des mitochondrialen Cytochrome C Oxidase Untereinheit 1 (CO1) Gens, ist zu einem populären Barcode für die Artindentifikation geworden. Hier wird eine neue Barcode Technik, das sogenannte charakter-basierte Barcoden getestet, welche ähnlicher zu traditionellen Ansätzen ist.

Diese Dissertation untersucht, ob CO1 als einzelner Marker geeignet ist (a) oder mit anderen ergänzt werden sollte (b). Die Leistung von distanz- und charakter-basierten Barcodes wird evaluiert (c) und es wird getestet ob, sich charakter-basierte Barcodes für die Identifizierung kryptischer Arten eignet.

Im ersten Manuskript werden die CO1 Sequenzen von bedrohten Schildkröten Arten verglichen (a). Ein zuverlässiges Werkzeug für die Identifikation ist ein wichtiges Mittel in der Artenschutzüberwachung. Die Variabilität in der Barcode Region wird untersucht und die Eignung von distanz- und charakter-basiertem Barcoden für die Artidentifikation evaluiert (c).

Odonaten sind eine alte, artenreiche Ordnung. Da sich viele Arten in kurzer Zeit entwickelt haben, wurde beobachtet, dass sich die intra- und interspezifische Varianz in einigen Schwestergruppen überlagert. Diese Beobachtung macht Odonaten zu einem idealen Kandidaten für das Testen von CO1 (a), ND1 (b), so wie distanz- und charakter-basiertem Barcoden (c) in dem zweiten Manuskript.

Ameisen sind Paradebeispiele für einen hohen Grad an kryptischer Biodiversität, da sie eine komplexe Populationsdifferenzierung aufgrund von Hybridisierung und Artbildungsprozessen besitzen. Da die Kombination mehrerer genetischer Marker einen besseren Barcoding Ansatz darstellt, werden im dritten Manuskript drei verschiedene Marker (CO1, 28S rDNS, rhodopsin) getestet (b). Ein kombinierter, mehrschichtiger Barcode wird evaluiert und es werden einzigartige, für Regionen spezifische Merkmale identifiziert (d).

Die Ergebnisse der drei Studien zeigen, dass die Kombination mehrerer Marker den Identifikationserfolg erhöht. Charakter-basiertes Barcoden bietet in den getesteten Tiergruppen eine bessere Identifikation. Diese Methode kann genutzt werden um die Anwesenheit, Abwesenheit oder Frequenz von kryptischen Arten einzuschätzen.

Schlüsselwörter: 28S rDNS, charakter-basiertes Barcoden, CO1, distanz-basiertes Barcoden, ND1, rhodopsin

Abstract

Classic taxonomy is a powerful tool for identifying animals based on morphology but has shown to be problematic on similar looking, cryptic species. A solution to this problem has been found within the bauplan of life, the DNA (deoxyribonucleic acid). DNA is used to create and regulate proteins. The structure of DNA has highly unique sections that are conserved within species, but diverse between species. One particular section, a 648 bp long fragment of the mitochondrial cytochrome c oxidase subunit 1 (CO1) gene, has become a popular barcode for species identification. Here, a new barcoding technique, character-based barcoding more similar to traditional approaches is tested.

This thesis investigates whether CO1 is suitable as a single marker (a) or should be complemented by others (b). Performance of distance- and character-based barcoding (c) is evaluated and it is tested whether character-based barcoding can be used to identify cryptic species (d).

In the first manuscript, CO1 sequences of endangered turtle species are compared (a). Having a reliable tool for species identification is an important asset in species protection surveillance. Variability within the barcode region is assessed and the utility of both distance- and character-based methods for species identification are evaluated (c).

Odonata is an old order rich in species. As many species have evolved in a short time, it was observed that intra- and interspecific variety is overlapping in some sister groups. This observation made Odonata the ideal candidate for testing CO1 (a), ND1 (b), as well as distance- and character-based-barcoding (c) in the second manuscript.

Ants are prime examples for high degrees of cryptic biodiversity due to complex population differentiation, hybridization and speciation processes. As combinations of multiple marker regions seemed to be a better approach to barcoding, three markers (CO1, 28S rDNA, rhodopsin) are tested (b) in the third manuscript. A combined, layered approach to character-based barcoding is evaluated and unique diagnostics specific to geolocations are identified (d).

The results of all three studies show that combining multiple markers improves identification success. The character-based approach provides better identification in the tested animal groups. This method can be used to estimate presence, absence or frequency of cryptic species.

Keywords: 28S rDNA, character-based barcoding, CO1, distance-based barcoding, ND1, rhodopsin

Contents

Introduction

1

1.1 The birth of taxonomy

Among all life forms, *Homo sapiens* is neither the biggest (humongous fungus; Dodge 2000) or the fastest (falcons; Mills *et al.* 2018) nor the life form with the most expanding life span (Cnidaria are potentially immortal; Petralia *et al.* 2014). We do not possess the best hearing mechanism (moths; Nakano & Mason 2018), smell (elephants; Niimura *et al.* 2014) or eye sight (eagles; Grambo 1999 & owls; Wu *et al.* 2016) but what we have is our mind that made *Homo sapiens* a successful and expanding species. Our ability to assess our surrounding and abstract thinking allowed us to invent simple tools such as bows up to complex ones like smartphones.

Thought processes like these gave birth to taxonomy our endeavor to make sense of everything by categorizing it. The start of western scientific taxonomy can be attributed to Aristotle (384-322 BC). He was the first to classify life, e.g. subdividing vertebrate and invertebrate by animals with and without blood (Manktelow 2010). Further, he divided animals with blood into egg-bearing and live-bearing and formed within the non-blood animals the group's insects, crustacean and testacea (mollusks). These are still known today (Manktelow 2010). Only with the development of optic lenses at the end of the 16th century, taxonomic research became advanced enough to replace the ancient Greek works. Optic lenses improved investigation of morphological traits in different species. At this time, focus shifted from medical to taxonomic aspects and the collection of specimens (Manktelow 2010).

Modern taxonomy was born when Carl Linnaeus (1707-1778) published the global flora Species Plantarum in 1753 and the tenth edition of Systema Naturae in 1758 including global fauna (Manktelow 2010). For the first time, a binary form of species names called "trivial names" for both plants and animals were introduced. The simplicity of Linnaeus' trivial names revolutionized nomenclature, and soon binary nomenclature came to replace the phrase names. He transformed zoology and botany into their own sciences embraced by philosophy, order and proper systems (Manktelow 2010).

It was Jean-Baptiste de Lamarck's (1744-1829) theory of characters acquired through inheritance, named "Lamarckism" that laid the foundation for the theory of evolution presented by Charles Darwin and Alfred Russel Wallace in 1858 in London. With the shortly followed book "Origin of Species" by Charles Darwin (1859) the concept and understanding of evolution were made accessible to a broad public.

While Charles Darwin definitions of evolution were derived from morphological observations most of these definitions hold true on the molecular level and have become an important guideline in phylogenetic research. Although the concept of evolution was groundbreaking, it did not affect systematics in the beginning. The next important contribution to taxonomy came from Ernst Haeckel (1834-1919) and August Wilhelm Eichler (1839-1878). These two German biologists started the construction of evolutionary trees. It was Haeckel that established the term "phylogeny".

The 20th century was dominated by phenetic research, i.e. looking for differences and similarities to create systematics (Manktelow 2010). For the first time, in addition to morphology, anatomy, chromosomes, pollen, biochemistry and later proteins were investigated for meaningful characters and species definition.

In 1966, the German biologist Willig Hennig (1913-1976) founded the era of cladistics. He stated that only similarities grouping species (synapomorphies) should be used in classification, and those taxa should include all descendants from one single ancestor (rule of monophyly) (Manktelow 2010). As many other modern approaches before, cladistic was initially observed controversially. Only around 20 years later, it started to become established. In the 1980's with the invention of PCR (polymerase chain reaction), it became economically feasible to amplify DNA-sequences for use in systematics, a new tool to gather phylogenies with high resolution was born (Manktelow 2010). Simultaneously, the development of computers and software enabled the analysis and administration of large datasets. Cladistics became the most commonly used method to classify a species (Manktelow 2010).

1.2 DNA barcoding, a successor of Linnaeus taxonomy

With the development of molecular science, the study of hereditary factors in form of DNA and genes by PCR and sequencing became a new means to study and revise the knowledge about the tree of life. The understanding of the ancestry and relationships between living organisms was improved by comparing DNA sequences. The ability to better compare extinct species by the means of residual DNA was gained. When Hebert *et al.* published manuscripts describing a 648 bp long DNA fragment (Folmer region) within the CO1 (Cytochrome C oxidase subunit 1) mitochondrial marker as a tool to distinguish lepidopterans (Hebert *et al.* 2003) and the North American avifauna (Hebert *et al.* 2004), DNA barcoding was born. DNA barcoding is the concept of using a singular genetic marker, the Folmer region, to identify all animal life. Hebert declared at this time that the Folmer region is identic or at least more similar within a species and distinct to other species.

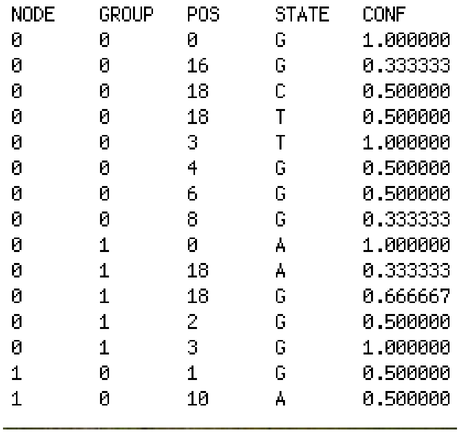
There are several advantages to barcoding compared to traditional taxonomy. For barcoding, only a small tissue sample from the specimen is needed, making this a non-invasive approach to species identification and ecosystem surveillance. As the barcode fragment is of mitochondrial origin and not part of the core DNA, multiple copies of the fragment exist in each cell. In addition, mtDNA is haploid, making it easier to extract, amplify and sequence, as only one allotype is present.

While advantageous, it is not necessary to have a taxonomic expert within the expedition when doing barcoding. The samples from the specimen can be processed in a research lab or by an independent industrial facility (today, sequencing a single sample costs around 3€) and then be classified by their unique barcode sequence. This approach makes it much easier and accurate to identify hard to distinguish species. Another advantage is that barcoding enables research on predatory species diets by collecting their feces. There is no need to perform surgery on the predators themselves or observe them closely over a long period of time.

While DNA barcoding became a success story in the last 15 years and is used by researchers all over the world through the web interface BOLD (Barcoding of Life Data System; Ratnasingham & Hebert 2007, 2013) it is not without flaws. For once, DNA barcoding is still dependent on traditional taxonomy. Reference sequences used in BOLD have to be validated by an expert through prior identification of the donor specimen. The wrong classification of reference sequences either through misidentification, cross-contamination or mislabeling of tissue samples reduces the accuracy of barcoding. Secondly, barcoding is focused on a single marker; mutations within this marker should not be set as equal to our traditional concept of species. As such, a newly discovered barcode from a specimen is not the same as a new species but rather should be used as a clue for investigation (DeSalle *et al.* 2005; DeSalle 2006). Traditional methods should proof if this specimen is a new haplotype within a prior defined group or member of a cryptic species newly discovered. Thirdly, because barcoding is focusing on a singular mitochondrial gene fragment its usability cannot be expanded to all animal groups. While it works for many phyla, such as birds or fishes (Hebert *et al.* 2004; Ward *et al.* 2005), it is problematic for other groups (Elias *et al.* 2007; Wiemers & Fiedler 2007). Especially those groups where members carry genetic markers on different strands (inner or outer strand) of the mtDNA, as has been observed in arthropods (Xu *et al.* 2006). The strands of the mostly circular mtDNA underlie different mutation rates (Rubinoff *et al.* 2006; Galtier *et al.* 2009), which highly impacts the diversity found within the Folmer region. In addition, animals with short life cycles have a higher mutation ratio than animals with long life cycles (Vassilieva & Lynch 1999; Nabholz *et al.* 2008a; Nabholz *et al.* 2008b) leading to significantly different barcoding performances. Another problem is the barcoding of groups with a history of rapid evolution such as insects. Insects were very successful in adapting to diverse ecosystems and underwent a major radiation in a very short time (Pterygotes in the Carboniferous and Endopterygota in the Permian; Smart 1963). Therefore, when different insect species are compared, the intra- and interspecific differences between these groups overlap in many instances when only the Folmer fragment is used as the identifier (Elias *et al.* 2007; Wiemers & Fiedler 2007). Lastly, rather than comparing distinct characters within barcoding, as is done with traditional taxonomy, identification is solely achieved by distance-based analysis (Hebert *et al.* 2003). In the distance-based analysis, a similarity matrix is calculated. Based on the similarity value one specimen has compared to another it is classified to the group with the best match. While this approach works very well for many groups and allows a short computational processing time, it also reduces the amount of data originally present within the dataset. Distinct data information is lost that if used could improve identification accuracy and performance.

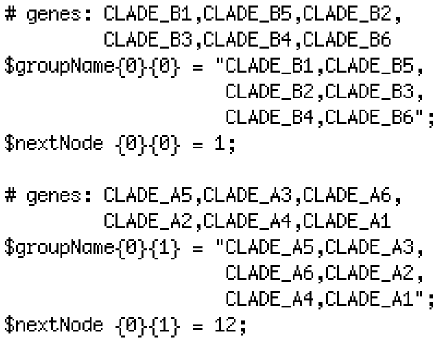
1.3 Character-based barcoding, the next step of barcoding

In collaboration with the University of Columbia (Neil Indra Sarkar, Paul Planet) and the American Museum of Natural History in New York (Rob DeSalle), the Institute for Animal Ecology & Evolution developed a new approach called CAOS barcoding (CAOS = Character Attribute Organization System). Like barcoding, it uses a genetic marker (can also work with protein sequences or other data; Sarkar *et al.* 2002a; Sarkar *et al.* 2002b) as a means for classifying specimen. Unlike barcoding, it is not focused on the Folmer region. Any marker that is sufficient in identifying the phylum of interest can be applied in CAOS barcoding. While in barcoding the complete 648 bp of the Folmer region is used as data input, in CAOS barcoding only meaningful positions are compared. This means in the classification process of a query specimen, only diagnostic positions within the marker sequence are used. So instead of comparing the 648 bp between the query and reference specimen, only a subset of positions, called character attributes (CAs) are compared. Character attributes are further differentiated between "pure" and "private" characters. Pure characters are identic for members of the same group, but different for another group. Private characters are unique for one group but are not present in all members of the group. As CAOS barcoding is using CAs to distinguish one group from another and also uses these CAs to classify field samples of unknown origin like traditional taxonomy, it is dependent on distinct characters. To locate the distinct characters Neil Indra Sarkar wrote the first CAOS software based on C++ (Sarkar *et al.* 2002a; Sarkar *et al.* 2002b). In 2008, the software was integrated into a user-friendlier and DNA focused perl script called p-gnome. It was also supplemented by a classifier called p-elf (Sarkar *et al.* 2008). P-gnome needs two types of input data in order to collect the character attributes which are unique to each group within a data set. First, the raw DNA sequence data saved in nexus format and secondly a dichotomal phylogenetic tree. Neighbour joining, maximum parsimony, maximum likelihood or any other algorithm can be used to create the tree as long as each branching point is dichotomal. The tree must also be saved in nexus format. Both sequence and tree data need to be combined into a single nexus file. Either the software MacClade (Maddison & Maddison 1989) or Mesquite (Maddison & Maddison 2018) was used to achieve this goal. When this combined nexus file is entered into p-gnome, the tree data is used as a guide for the software. Starting by the root of the tree, at each branching point all sequences of the left and right branch are compared between each other. The software searches for similarities between members of the same branch and differences between members of opposing branches at each character position. If unique characters are detected, they are saved in a newly created text file (CAOS_attributesFile.txt; Fig.1.1), while the members of a branching point are saved in a separate file (CAOS_groupFile.txt; Fig.1.2). After one node has been analyzed the program proceeds to the next one and repeats the process until all nodes have been processed.



```
NODE      GROUP     POS       STATE     CONF
0         0         0         G         1.000000
0         0         16        G         0.333333
0         0         18        C         0.500000
0         0         18        T         0.500000
0         0         3         T         1.000000
0         0         4         G         0.500000
0         0         6         G         0.500000
0         0         8         G         0.333333
0         1         0         A         1.000000
0         1         18        A         0.333333
0         1         18        G         0.666667
0         1         2         G         0.500000
0         1         3         G         1.000000
1         0         1         G         0.500000
1         0         10        A         0.500000
```

Fig. 1.1.: CAOS_attributesFile.txt

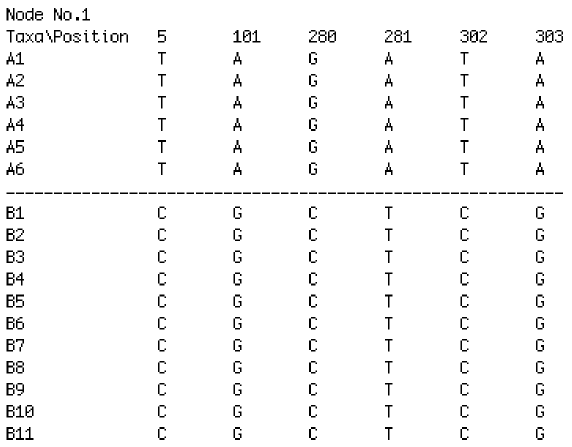


```
# genes: CLADE_B1,CLADE_B5,CLADE_B2,
         CLADE_B3,CLADE_B4,CLADE_B6
$groupName{0}{0} = "CLADE_B1,CLADE_B5,
                    CLADE_B2,CLADE_B3,
                    CLADE_B4,CLADE_B6";
$nextNode {0}{0} = 1;

# genes: CLADE_A5,CLADE_A3,CLADE_A6,
         CLADE_A2,CLADE_A4,CLADE_A1
$groupName{0}{1} = "CLADE_A5,CLADE_A3,
                    CLADE_A6,CLADE_A2,
                    CLADE_A4,CLADE_A1";
$nextNode {0}{1} = 12;
```

Fig. 1.2.: CAOS_groupFile.txt

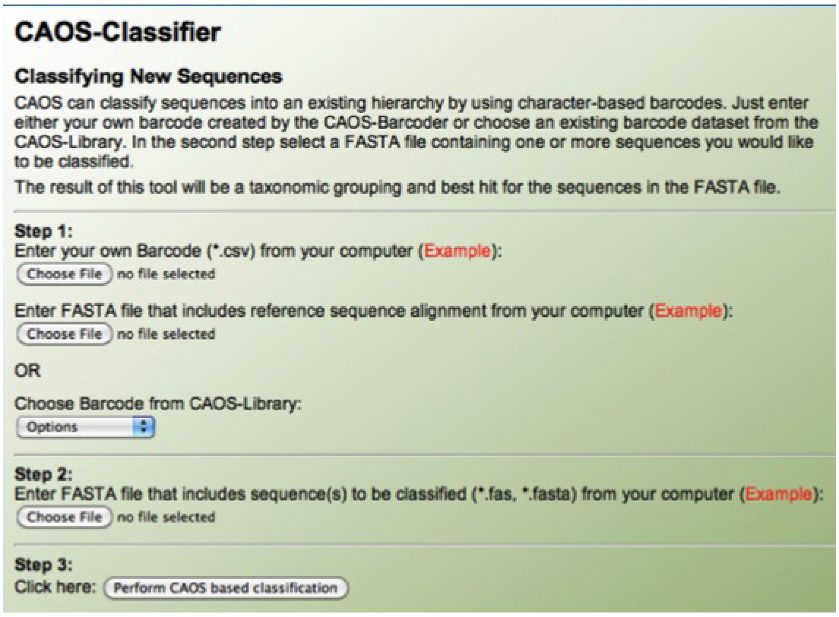
This is where the research for this thesis dealing with CAOS barcoding started: Testing p-gnome on dragonfly data showed promising results. However, a couple of issues with the program occured. The input file for p-gnome had to be saved as a nexus file. The problem with the nexus file format is that it is not uniform. Depending on the program used to create the nexus file, there are differences in the output format. At this time CAOS could only work with one of the formats. Another problem was the tree data inside the nexus format. Depending on the tree algorithm and setup it also produced different formats (e.g. numbers instead of specimen names or support values next to nodes). Using an unsupported format led to a cancellation of the analysis and an error message. In p-gnome the sequence and tree data are converted into a text file (CAOS_overviewFile.txt), which is dependent on a specific format, and is used by CAOS to extract the sequence and tree data to produce the attribute and group data. In addition, extracting the CAs and corresponding group data from the text files (see Fig.1.1 & Fig.1.2) proved to be difficult. The nodes of interest within the group data (Fig. 1.2) had to be identified and the node code representative of the group of interest had to be written down. Next, the code in the attribute file had to be found in order to extract the CAs (Fig. 1.1). This procedure was time consuming and not intuitive. The shortcomings of CAOS were discussed within the Institute of Animal Ecology and Evolution and I agreed to improve the software. The following enhancements were made: p-gnome was rewritten, the program was adapted to work with all nexus and tree formats. The program was renamed CAOS-Analyzer. In a second step, I created a program that transforms the output text files (attribute and group file) into a set of five overview table files. Each table file showing different sets of character attributes for each node within the tree (e.g. Fig.1.3).



```
Node No.1
Taxa\Position   5       101     280     281     302     303
A1              T       A       G       A       T       A
A2              T       A       G       A       T       A
A3              T       A       G       A       T       A
A4              T       A       G       A       T       A
A5              T       A       G       A       T       A
A6              T       A       G       A       T       A
-----------------------------------------------------------
B1              C       G       C       T       C       G
B2              C       G       C       T       C       G
B3              C       G       C       T       C       G
B4              C       G       C       T       C       G
B5              C       G       C       T       C       G
B6              C       G       C       T       C       G
B7              C       G       C       T       C       G
B8              C       G       C       T       C       G
B9              C       G       C       T       C       G
B10             C       G       C       T       C       G
B11             C       G       C       T       C       G
```

Fig. 1.3.: Example for one of the overview files. Here, an example for overview file 5 is illustrated, which only highlights positions where both clusters provide homogenous sPu diagnostics. In the first column, the sample names are listed, while the position and unique characters of the samples are listed in the following columns. Left and right branch data are separated by a line.

In addition, two more tables were created. A) An overview file (Total_ barcode.xlsx) showing all character attribute positions and characters within the complete tree as a single table. B) A unique data file (Ref_matrix.csv) that also included all barcoding information but was formatted in a way that allows the user to use it as a means to classify new samples with a third program (CAOS-Classifier) that was invented and written by me. P-elf, a script developed together with p-gnome was intended to work as a classifier but most of the times no conclusive result was achieved with the script or the query was assigned to the wrong group. The CAOS-Classifier can identify new specimen data by a combination of character- and distance-based approaches. The program takes in query data in fasta format (Fig.1.4). Fasta has the advantage of being a simple and strict format. It is accepted by most genetic softwares. As reference-CA-database, the CAOS-Classifier uses the "Ref_matrix.csv" file created by the CAOS-Barcoder. In the first step, the CAOS-Classifier aligns the query sequences with the reference sequences (also provided as fasta file). This step is very important as the query sequences might be of varying length and it is mandatory for correct comparison of CA data between query and reference. In the second step, similar to the CAOS-Analyzer the query data is guided through a series of nodes based on the tree created for the reference dataset. Beginning at the root of the tree, for each node CAs of the left and right branch are compared with the query. If matches are detected, points are given for each match (pure CAs = 3 points; private CAs = 1 point). The branch with more points is followed and the other discarded. Once, the end is reached or both branches get the same amount of points, the query sequences are aligned with the remaining reference sequences. The best match is displayed as a hit (based on distance value; Fig.1.5) and an alignment of the best matches is created (similar to NCBI blast). In collaboration with the AMNH (Rob DeSalle) and the University of Vermont (at this time Neil Indra Sarkar was working there), I wrote a website-based interface and command line based scripts for all three programs (Analyzer, Barcoder and Classifier).



CAOS-Classifier

Classifying New Sequences

CAOS can classify sequences into an existing hierarchy by using character-based barcodes. Just enter either your own barcode created by the CAOS-Barcoder or choose an existing barcode dataset from the CAOS-Library. In the second step select a FASTA file containing one or more sequences you would like to be classified.

The result of this tool will be a taxonomic grouping and best hit for the sequences in the FASTA file.

Step 1:
Enter your own Barcode (*.csv) from your computer (Example):
Choose File no file selected

Enter FASTA file that includes reference sequence alignment from your computer (Example):
Choose File no file selected

OR

Choose Barcode from CAOS-Library:
Options

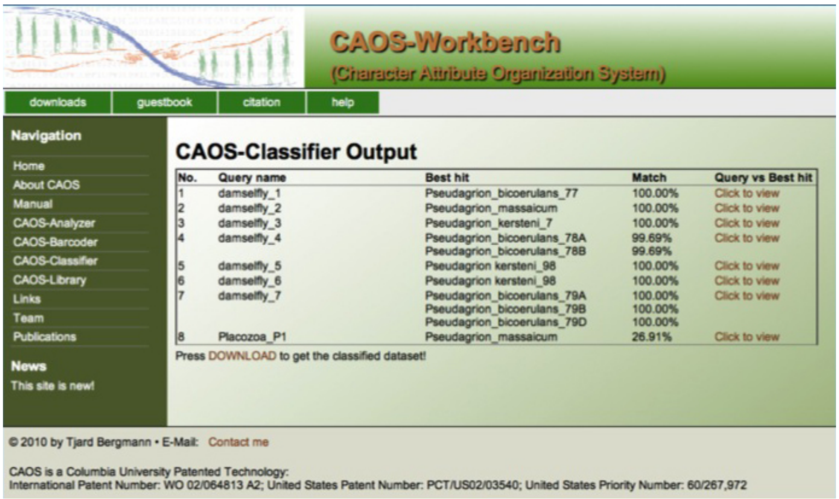
Step 2:
Enter FASTA file that includes sequence(s) to be classified (*.fas, *.fasta) from your computer (Example):
Choose File no file selected

Step 3:
Click here: Perform CAOS based classification

Fig. 1.4.: CAOS-Classifier: Data input screen taken from the CAOS-Workbench website.

1.4 The aims of this thesis

This thesis aims to better understand (1) what makes a good morphological marker, (2) what is the making of a good barcoding method and (3) how can we discover and resolve cryptic species. In order to answer these questions, we followed different approaches.



CAOS-Workbench

(Character Attribute Organization System)

This site is new!

CAOS-Classifier Output

No.	Query name	Best hit	Match	Query vs Best hit
1	damselfly_1	Pseudagrion_bicoerulans_77	100.00%	Click to view
2	damselfly_2	Pseudagrion_massaicum	100.00%	Click to view
3	damselfly_3	Pseudagrion_kersteni_7	100.00%	Click to view
4	damselfly_4	Pseudagrion_bicoerulans_78A Pseudagrion_bicoerulans_78B	99.69% 99.69%	Click to view
5	damselfly_5	Pseudagrion kersteni_98	100.00%	Click to view
6	damselfly_6	Pseudagrion kersteni_98	100.00%	Click to view
7	damselfly_7	Pseudagrion_bicoerulans_79A Pseudagrion_bicoerulans_79B Pseudagrion_bicoerulans_79D	100.00% 100.00% 100.00%	Click to view
8	Placozoa_P1	Pseudagrion_massaicum	26.91%	Click to view

Press DOWNLOAD to get the classified dataset!

© 2010 by Tjard Bergmann • E-Mail: Contact me

CAOS is a Columbia University Patented Technology:
International Patent Number: WO 02/064813 A2; United States Patent Number: PCT/US02/03540; United States Priority Number: 60/267,972

Fig. 1.5.: CAOS-Classifier: Example for data output taken from the CAOS-Workbench website.

In our first endeavor (Reid *et al.* 2011) to assess the quality of CO1 as a marker (1) and to investigate the accuracy of distance- and character-based DNA barcoding (2), we used the long living and widespread order Testudines (turtles) as a test case. Surveillance and conservation of endangered species is an important part of protecting the biodiversity of our planet. Illegal wildlife trade threatens many species, such as turtles; DNA barcoding can serve as a powerful tool in wildlife forensics. We compared the CO1 Folmer region of 174 turtle species in addition to 50 publicly available species. Combined, the data set is representative of the order Testudines (turtles). My part of this manuscript was barcoding the data and creating a CAOS barcoding website as a service platform to identify turtle specimen. The p-gnome performed character-based analysis and the corresponding table (Table 3) showing the characters was done by Brendan Reid. Within the project, I created a new character-based output using afore mentioned Analyzer, Barcoder and Classifier programs. The results were implemented in the character-based identification website as described in the manuscript.

In a second manuscript (Bergmann *et al.* 2013), we further investigated marker quality (1), barcoding method (2) and detection of cryptic species (3) by studying the taxonomically challenging order Odonata. Odonata is a species rich order (~5.800), the fast differentiation of its members over a short time span makes species identification on morphological and molecular level difficult. Odonates are an indicator for healthy ecosystems, as many members are sensitive to changes in drinking water quality. The close relationship between Odonata species and its value as an indicator for ecosystem stability makes them an intriguing case subject for evaluating distance-based DNA barcoding (BOLD) and character-based barcoding (CAOS) as well as comparing the efficiency of different markers (CO1 vs ND1). In this study, 271 odonate individuals representing 51 species were compared. Animal sampling, sequencing and distance-based data analysis was conducted by Jessica Rach, while all character-based research was my contribution.

In (Paknia *et al.* 2015), the investigation is advanced on marker quality (1), barcoding method (2) and location of cryptic species (3) by focusing on ants. Ants, because of complex population differentiation, hybridization and speciation processes are prime examples for cryptic biodiversity. Here, we go one step further by testing two supplementary markers in addition to cytochrome c oxidase 1 and assessing the potential of character-based barcoding to uncover cases of potential cryptic diversity. In this manuscript data mining, tree building and ant specific topics were carried out by Omid Paknia, while I did the barcoding and analysis of the results.

1.4.1 References

Boll PK (2011) A Brief History of the Kingdoms of Life. Word Press, Earthling Nature.

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

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.

Dodge SR (2000) An even more humongous fungus. PACIFIC NORTHWEST RESEARCH STATION/USDA FOREST SERVICE, Portland, Ore.

Elias M, Hill RI, Willmott KR, et al. (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proc Biol Sci 274, 2881-2889.

Galtier N, Nabholz B, Glemin S, Hurst GD (2009) Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol Ecol 18, 4541-4550.

Grambo RL (1999) Eagles Voyageur Press, Inc., China.

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc Biol Sci 270, 313-321.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. Plos Biology 2, 1657-1663.

Jeronimo (2002) Kingdom (biology). MediaWiki, Wikipedia.

Maddison WP, Maddison DR (1989) Interactive analysis of phylogeny and character evolution using the computer program MacClade. Folia Primatol (Basel) 53, 190-202.

Maddison WP, Maddison DR (2018) Mesquite: a modular system for evolutionary analysis.

Manktelow M (2010) History of Taxonomy eds. Dept of Systematic B, Evolutionary Biology C.

Mills R, Hildenbrandt H, Taylor GK, Hemelrijk CK (2018) Physics-based simulations of aerial attacks by peregrine falcons reveal that stooping at high speed maximizes catch success against agile prey. PLoS Comput Biol 14, e1006044.

Nabholz B, Glemin S, Galtier N (2008a) Strong variations of mitochondrial mutation rate across mammals–the longevity hypothesis. Mol Biol Evol 25, 120-130.

Nabholz B, Mauffrey JF, Bazin E, Galtier N, Glemin S (2008b) Determination of mitochondrial genetic diversity in mammals. Genetics 178, 351-361.

Nakano R, Mason AC (2018) Early erratic flight response of the lucerne moth to the quiet echolocation calls of distant bats. PLoS One 13, e0202679.

Niimura Y, Matsui A, Touhara K (2014) Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res 24, 1485-1496.

Petralia RS, Mattson MP, Yao PJ (2014) Aging and longevity in the simplest animals and the quest for immortality. Ageing Res Rev 16, 66-82.

Ratnasingham S, Hebert PD (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes 7, 355-364.

Ratnasingham S, Hebert PD (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS One 8, e66213.

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J Hered 97, 581-594.

Sarkar IN, Planet PJ, Bael TE, et al. (2002a) Characteristic attributes in cancer microarrays. J Biomed Inform 35, 111-122.

Sarkar IN, Planet PJ, DeSalle R (2008) caos software for use in character-based DNA barcoding. Mol Ecol Resour 8, 1256-1259.

Sarkar IN, Thornton JW, Planet PJ, et al. (2002b) An automated phylogenetic key for classifying homeoboxes. Mol Phylogenet Evol 24, 388-399.

Smart J (1963) Explosive evolution and the phylogeny of insects. Proceedings Linnean Society London 174, 125-126.

Vassilieva LL, Lynch M (1999) The rate of spontaneous mutation for life-history traits in *Caenorhabditis elegans*. Genetics 151, 119-129.

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

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

Wu Y, Hadly EA, Teng W, et al. (2016) Retinal transcriptome sequencing sheds light on the adaptation to nocturnal and diurnal lifestyles in raptors. Sci Rep 6, 33578.

Xu W, Jameson D, Tang B, Higgs PG (2006) The Relationship Between the Rate of Molecular Evolution and the Rate of Genome Rearrangement in Animal Mitochondrial Genomes. Journal of Molecular Evolution 63, 375-392.

Experimental Studies 2

2.1 Comparing and combining distance-based and character-based approaches for barcoding turtles

MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.03032.x

Authors: B. N. Reid, M. Le, W. P. McCord, J. B. Iverson, A. Georges, T. Bergmann, G. Amato, R. DeSalle and E. Naro-Maciel

B. N. Reid: Department of Forest and Wildlife Ecology, University of Wisconsin, 1630 Linden Drive, Madison, WI 53706, USA

M. Le: Center for Natural Resources and Environmental Studies, Vietnam National University, 19 Le Thanh Tong Street, Hanoi, Vietnam; Faculty of Environmental Sciences, Hanoi University of Science, 334 Nguyen Trai Road, Hanoi, Vietnam; Department of Herpetology, American Museum of Natural History, New York, NY 10024, USA

W. P. McCord: East Fishkill Animal Hospital, 455, Route 82, Hopewell Junction, NY 12533, USA

J. B. Iverson: Department of Biology, Earlham College, Richmond, IN 47374, USA

A. Georges: Institute for Applied Ecology, University of Canberra, Canberra, ACT 2601, Australia

T. Bergmann: Institute for Animal Ecology and Evolution, Stiftung Tierärztliche Hochschule Hannover, Hannover 30559, Germany

G. Amato: Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

R. DeSalle: Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

E. Naro-Maciel: Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA; Biology Department, College of Staten Island, City University of New York, Staten Island, NY 10314, USA

Keywords: barcoding, Characteristic Attribute Organization System, species identification, turtles

2.1.1 Abstract

Molecular barcoding can serve as a powerful tool in wildlife forensics and may prove to be a vital aid in conserving organisms that are threatened by illegal wildlife trade, such as turtles (Order Testudines). We produced cytochrome oxidase subunit one (CO1) sequences (650 bp) for 174 turtle species and combined these with publicly available sequences for 50 species to produce a data set representative of the breadth of the order. Variability within the barcode region was assessed, and the utility of both distance-based and character-based methods for species identification was evaluated. For species in which genetic material from more than one individual was available (n = 69), intraspecific divergences were 1.3% on average, although divergences greater than the customary 2% barcode threshold occurred within 15 species. High intraspecific divergences could indicate species with a high degree of internal genetic structure or possibly even cryptic species, although introgression is also probable in some of these taxa. Divergences between species of the same genus were 6.4% on average; however, 49 species were <2% divergent from congeners. Low levels of interspecific divergence could be caused by recent evolutionary radiations coupled with the low rates of mtDNA evolution previously observed in turtles. Complementing distance-based barcoding with character-based methods for identifying diagnostic sets of nucleotides provided better resolution in several cases where distance-based methods failed to distinguish species. An online identification engine was created to provide character-based identifications. This study constitutes the first comprehensive barcoding effort for this seriously threatened order.

2.1.2 Introduction

Turtles (order Testudines) are highly endangered as a group, with 42% of extant species classified as threatened and 10% classified as critically endangered by the IUCN (Buhlmann *et al.* 2009). Turtles face a similar battery of threats compared with other endangered taxa, including the effects of habitat loss, invasive species, pollution, disease and climate change; however, human overexploitation represents an especially acute threat to the survival of most threatened turtle species (van Dijk *et al.* 2000; Gibbons *et al.* 2000). The turtle trade is at its most intense in China and Southeast Asia, where over 10 million individuals per year are traded as meat, pets or ingredients in traditional remedies (Turtle Conservation Fund 2002). It is important to note, however, that the Asian turtle market handles species from around the world (Cheung & Dudgeon 2006; Nijman & Shepherd 2007), with globalization of trade increasing as native Asian species become increasingly scarce.

The forensic applications of DNA barcoding have great potential as a means for quantifying and regulating trade in endangered turtle species (Ogden *et al.* 2009; Alacs *et al.* 2010). Previous studies have shown that, given a comprehensive sequence database, CO1 can serve as a reliable forensic marker for identifying unknown zoological material to the species level (Dawnay *et al.* 2007). The forensic applications proposed for barcoding run the gamut from identifying fish species in commercial markets (Costa & Carvalho 2007) to investigating bird airplane collisions (Dove *et al.* 2008). Recently, barcoding has been shown to be a reliable means of identifying material in the bushmeat trade (Eaton *et al.* 2010). Despite the promise of utilizing DNA barcoding as a tool for their conservation, turtles have been underrepresented in the global barcoding effort. Prior to the initiation of this research, sequences from only 52 species had been deposited in the Barcode of Life Datasystems database (BOLD, accessed 26 February 2009), and the species barcoded were also heavily skewed towards Asian pond turtles (family Geoemydidae) and tortoises (family Testudinidae). Turtles therefore represented a significant gap in the barcode catalogue that we intended to fill.

This report provides novel CO1 barcode sequences for 174 turtle species. The species barcoded here were chosen because they either appear on the IUCN Red List, indicating that they are species of conservation concern which would probably benefit from the forensic applications of barcoding, or because they belong to clades that are underrepresented within the Testudines with regard to previous barcoding efforts. Publicly available sequences as well as sequences for sea turtles produced in a previous study (Naro-Maciel *et al.* 2010) were added to these novel sequences to better evaluate variability and identification success across the entire order. Distance-based (Hebert *et al.* 2003, 2004) and character-based approaches to barcoding (DeSalle *et al.* 2005; Kelly *et al.* 2007) were both evaluated to determine the effectiveness in distinguishing turtle species. While application of the barcode information gleaned here to quantifying or controlling the wildlife trade is beyond the scope of this report, this information represents a potentially powerful tool for combating the anthropogenic challenges currently faced by turtles on the global scale.

2.1.3 Material & Methods

Taxonomy, sample selection and acquisition

A list of all turtle species on the IUCN Red List (in every category except for ’Extinct’) was compiled (IUCN 2009) and cross-referenced against a list of turtle species already present in the BOLD database to produce a master list of red-listed species without barcodes. The IUCN’s taxonomic designations were checked against the most widely accepted account of turtle taxonomy (Turtle Taxonomic Working Group 2007) at the time of compilation and revised accordingly. The taxonomy used in this work does not account for several very recent changes in nomenclature (such as the reorganization of several chelid species into the new genus *Myuchelys*; Georges & Thomson 2009). When several alternate genera were listed for a species, the species was assigned to a genus in a way that minimized the total number of genera under consideration. Non-IUCN-listed species from two turtle families (Chelidae and Pelomedusidae) that were underrepresented in the BOLD database were also added to the master list.

Species on this master list that were already represented in the American Museum of Natural History (AMNH)’s collection, either as extracted DNA or frozen tissue, were obtained directly from the museum. Availability of the remaining species was determined by querying the Association of Zoos and Aquariums (AZA)’s zoo holdings database, ISIS (http://www.isis.org) and the museum herpetological collections database Herp-NET (http://herpnet.org). Once sources were identified, blood or tissue samples were obtained from a collaborating zoo, museum, university or from the authors’ (Georges, Iverson, McCord) collections. In cases where species were protected by national law or listed under one of the appendices of the Convention on International Trade in Endangered Species, care was taken to obtain all relevant permits and observe applicable regulations for the collection of samples and transfer of specimens between institutions. When possible, aliquots of blood or tissue samples obtained from private collections have been deposited into the Ambrose Monell Cryo Collection (AMCC) at the AMNH for future reference. Owing to the nature of the sampling, original collection locality information was unavailable for many samples, including samples obtained from zoo animals and specimens obtained from the pet trade. Where available, voucher numbers and locality information have been uploaded as annotation to the Genbank and BOLD records for the novel sequences presented in this study.

DNA extraction and sequencing

DNA was extracted from blood or tissue using a DNeasy Tissue kit (QIAGEN Inc., Valencia, CA, USA). The CO1 barcode region was amplified from most species using either turtle-specific or universal primers from previous studies or primers designed in the course of this study (Table 2.1). PCR conditions for all primer sets except the universal CO1-3 primer cocktail were as follows: 95 °C for 5 m; 35 cycles of 95 °C for 45 s, 54 °C for 45 s, 72 °C for 45 s; 72 °C for 6 m; 4 °C indefinitely. PCR for the CO1-3 primer cocktail (utilizing primers VF2_t1, FishF2_t1, FishR2_t1 and FR1d_t1) was run according to Ivanova *et al.* 2007 (94 °C for 2 m; 35 cycles of 94 °C for 30 s, 52 °C for 40 s and 72 °C for 1 m; 72 °C for 10 m; 4 °C indefinitely). PCR products were cleaned on a BIOMEK automated apparatus using the Ampure system. Cycle sequencing was performed using BigDye reagents (Perkin Elmer, Waltham, MA, USA). Both strands of all PCR products were sequenced with the same primers and used to amplify the products except in the case of CO1-3 primer cocktail products, which were sequenced using the M13F and M13R primers. Cycle sequencing PCR was run as follows: 96 °C for 5 m; 35 cycles of 94 °C for 15 s, 50 °C for 15 s, 60 °C for 4 m; 4 °C indefinitely. Cycle sequencing products were ethanol precipitated and run on an ABI3770 automated sequencer (Applied Biosystems, Foster City, CA, USA).

Tab. 2.1.: Primers used in this study. 5' positions are relative to the published mitochondrial sequence for *Chrysemys picta*

Primer name	Sequence	Reference	5' position
L-turtCOI	5'-ACTCAGCCATCTTACCTGTGATT-3'	Stuart and Parham 2004	5384
L-turtCOIc	5'-TACCTGTGATTTTAACCCGTTGAT-3'	Stuart and Parham 2004	5396
H-turtCOIb	5'-GTTGCAGATGTAAAATAGGCTCG-3'	Stuart and Parham 2004	6327
H-turtCOIc	5'-TGGTGGGCTCATACAATAAAGC-3'	Stuart and Parham 2004	6273
LCO1490	5'-GGTCAACAAATCATAAAGATATTGG-3'	Folmer *et al.* 1994	5423
HCO2198	5'-TAAACTTCAGGGTGACCAAAAAATCA-3'	Folmer *et al.* 1994	6132
VF2_t1	5'-TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC-3'	Ward *et al.* 2005	5426*
FishF2_t1	5'-TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC-3'	Ward *et al.* 2005	5426*
FishR2_t1	5'-CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA-3'	Ward *et al.* 2005	6129*
FR1d_t1	5'-CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAATCARAA-5'	Ivanova *et al.* 2007	6129*
M13F	5'-TGTAAAACGACGGCCCAGT-3'	Messing 1983	n/a
M13R	5'-CAGGAAACAGCTATGAC-3'	Messing 1983	n/a
HturtCOIk^{a}	5'-GGTGGGCTCATACAATAAAACC-3'	This study	6272
LturtCOIk^{a}	5'-CTACTAACCATAAAGACATCGGTACCC-3'	This study	5426
HturtCOIa^{b}	5'-CATACAATGAATCCCAGGAATCCGAT-3'	This study	6264
LturtCOIa^{b}	5'-CGCTGACTATTTTCTACTAATC-3'	This study	5413
Fbat2^{b}	5'-CTACTAATCATAAAGACATTGG-3'	This study	5426
Rbat1^{b}	5'-TAGGCAACTACGTGTGAGATTAT-3'	This study	6180
Fpodo1^{c}	5'-CAAACCATAAAGATATTGGCACCC-3'	This study	5429
Rpodo1^{c}	5'-GATATTATTGCTCATACTATTCC-3'	This study	6237
Fpelu1^{d}	5'-CCCGTTGATTATTCTCCACTAACC-3'	This study	5411
Rpelu1^{d}	5'-GATGCTATGGCTCAAACTATTCC-3'	This study	6237
Fpyx1^{e}	5'-CTCTACTAACCATAAAGATAT-3'	This study	5424

⋆Excluding engineered 5' M13 sequence.
Novel primers with superscript annotations were used for amplifying several species from these specific families:
(a) Kinosternidae. (b)Chelidae. (c) Podocnemididae. (d) Pelomedusidae. (e) Testudinidae.

Sequence variability and distance-based species identification

Novel sequences were assembled and edited in Sequencher (Gene Codes Corporation) and added to a set of publicly available sequences downloaded from BOLD. As nuclear paralogues (numts) have already been detected in several turtle species (Stuart & Parham 2004; Spinks & Shaffer 2007), all sequences were systematically screened to identify numts. Multiple primer pairs were used in most cases to increase the chance of amplifying the true mitochondrial sequence, and all suspected numts (sequences with premature stop codons or frameshift mutations) were expunged from the data set. Sequences were aligned in MEGA 4 (Tamura *et al.* 2007) and trimmed to a region 650 nucleotides in length. The fragment used here begins at base pair 62 of the complete CO1 sequence (base pair 5453 of the complete *Chrysemys picta* mitochondrial genome), with codon 22 in the translated CO1 amino sequence being the first complete codon in the fragment. These sites are designated as the first nucleotide and amino acid positions, respectively, in our data set.

Sequence composition and substitution pattern for the entire data set, the number of variable nucleotide and amino acid sites in the data set, and pairwise

Kimura 2-parameter (K2P) sequence divergences within groups at multiple taxonomic levels (intraspecific, between species of the same genus and between species of different genera in the same family) were calculated in MEGA 4. The K2P substitution model rather than a more realistic model was used to calculate distances to allow for repeatability of analyses through the BOLD engine and comparison with canonical distance-based barcoding studies (Hebert *et al.* 2003, 2004). The distribution of pairwise K2P values at each taxonomic level was visualized using a density plot in R (R Foundation for Statistical Computing, Vienna, Austria). Pearson product-moment correlations and Spearman rank correlations between sample size and mean intraspecific distance were also calculated in R to determine whether the number of available samples affected estimates of intraspecific distance.

Two neighbour-joining trees, one for pleurodiran species (side-necked turtles) and one for cryptodiran species (all other turtles), were constructed in MEGA 4 strictly to allow for the visualization of K2P distances for all novel sequences produced in this study. Trees were displayed using the Interactive Tree of Life web service (http://itol.embl.de; Letunic & Bork 2006). Previously published sequences were excluded from these trees because of space considerations. Species were organized into one of four categories (after Hebert *et al.* 2004) based on pairwise K2P distances. The categories used were as follows: Category I (maximum intraspecific distance <2%, minimum interspecific distance >2%), Category II (maximum intraspecific distance ≥2%, minimum interspecific distance >2%), Category III (maximum intraspecific distance <2%, minimum interspecific distance ≤2%) and Category IV (maximum intraspecific distance ≥2%, minimum interspecific distance ≤2%). In species where only one individual was sampled, categories I and II and categories III and IV were conflated as only interspecific distances could be measured.

Character-based analysis and online identification engine

Pure unique identifying characters, defined here as single-nucleotide states that distinguish a species from others in its family, were determined for each family using the Characteristic Attribute Organization System (CAOS; Sarkar *et al.* 2002, 2008; Bergmann *et al.* 2009). When all members of a species share these characters, they are termed 'simple pure characters' (sensu Sarkar *et al.* 2002). Characters were identified at the family level to correspond with the previous studies (Kelly *et al.* 2007; Rach *et al.* 2008; Damm *et al.* 2010; Naro-Maciel *et al.* 2010; Yassin *et al.* 2010). A guide tree was first produced using the maximum parsimony module in Phylip (v3.67; Felsenstein 1989) and modified to group individual samples according to current species designations (Turtle Taxonomic Working Group 2007). This guide tree was then incorporated into a nexus file containing CO1 sequence data in MacClade (v4.06; Maddison & Maddison 2000), and the p-gnome script (Rach *et al.* 2008; Sarkar *et al.* 2008) was used to identify characters. The proportion of all species exhibiting within-family identifying characters, as well as the proportion in each family, was calculated. Finally, the number of species exhibiting within-family characters for each of the distance-based categories was evaluated.

An online identification engine ('Project Turtle' in the Ruby-CAOS website, http://boli.uvm.edu/CAOS-workbench/htdocs/CAOS.php) was designed to allow for the implementation of the character-based identification method in a manner similar to the user-friendly BOLD interface for distance data. Sequences supplied to the website are first assigned to a family, after which the CAOS-Classifier script in RubyCAOS is employed to establish species identity using the family-level characters described here. If a positive identification is made, the site provides a link to the species description in the Turtles of the World database (http://nlbif.eti.uva.nl/bis/turtles.php); if no identification is possible, a list of possible species is provided.

2.1.4 Results

Taxonomic range and Red List coverage

Information for the taxa included in this study is given in Table S1 (Supporting information). Overall, 220 species from all 14 chelonian families (four of which had no representation in the barcode database before) are represented in the final data set. Of the 204 valid, extant turtle species on the Red List, 35 (17%) had been previously barcoded and another 149 (73%) were barcoded in this study. Owing to the rarity of many of these turtles, multiple samples were not available for all species; however, two or more sequences were available from 69 of the species included in this study.

Barcode fragment variability and distance-based species identification

Approximately half of the nucleotide positions (51.8%) were variable across the data set. Nucleotide composition showed a bias against G consistent with that observed previously in turtles (Spinks *et al.* 2004), and transitions were more frequent than transversions. Approximately two-fifths (40.7%) of amino acid positions were variable (Table 2.2).

Tab. 2.2.: Nucleotide substitution pattern, nucleotide frequencies, and nucleotide and amino acid variability as estimated in MEGA 4. Transitions rates are in bold, while transversion rates are italicized.

Maximum composite likelihood estimate of substitution pattern				
	A	T	C	G
A	-	*4.58*	*4.37*	**7.58**
T	*4.58*	-	**23**	*2.74*
C	*4.58*	**24.16**	-	*2.74*
G	**12.74**	*4.57*	*4.36*	-
Nucleotide frequencies				
A				0.281
T				0.282
C				0.268
G				0.168
Proportion of sites variable				
		Variable	Total	% Variable
Nucleotide		337	650	52
Amino acid		88	216	41

Mean intraspecies K2P divergence across 1403 possible pairwise combinations was 1.3% (Fig. 2.1). Variance was high, however [standard deviation (SD) = 2.2%], and pairwise intraspecific distances >2% were observed in 15 of the 69 species with n > 2. The Pearson and Spearman tests for correlation between sample size and intraspecific divergence gave conflicting results (Pearson's r = 0.01, P = 0.91; Spearman's rho = 0.26, P = 0.029). This indicates a positive relationship between relative (but not absolute) sample size and intraspecific divergence, meaning that although intraspecific distances may be somewhat underestimated in undersampled species there is no linear relationship between sample size and divergence. Mean pairwise divergence between congeneric individuals was 6.4% (SD = 2.6%, Fig.

2.1). Pairwise K2P differences of <2% were observed between 49 species. Mean intrafamily divergence was 13.6% (SD = 4.3%, Fig. 2.1). All sequences were uploaded to BOLD and analysed using the BOLD interface, yielding similar results in all cases. Genus and species groupings for novel sequences on the distance-based trees (Fig. 2.2) were broadly congruent with the accepted taxonomy (although some accepted genera and species were not monophyletic on the tree). Very low levels of divergence (<1%) were apparent between certain species in some genera (*Elseya, Pseudemys, Graptemys, Trachemys, Kinosternon, Mesoclemmys*), while very high levels of intraspecies divergence (>4%) were observed in five species (*Kinosternon integrum*, *Elseya novaeguineae*, *Emydura subglobosa*, *Acanthochelys radiolata* and *Amyda cartilaginea*). For species with multiple samples, 43 (62%) were placed in Category I, 9 (13%) were placed in Category II, 11 (16%) were placed in Category III and 6 (9%) were placed in Category IV. For species with one sample, 119 (79%) were placed in Category I/II and 32 (21%) were placed in Category III/IV (Fig. 2.3).

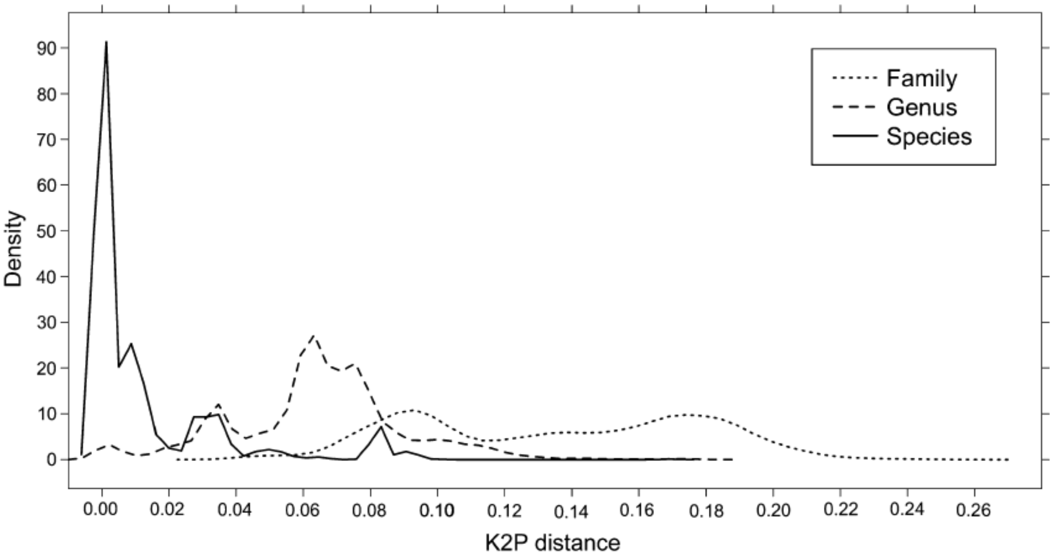


Fig. 2.1.: Density plot of Kimura 2-parameter (K2P) divergences within each taxonomic level.

Character-based identification

Characteristic Attribute Organization System analysis produced sets of simple identifying characters capable of distinguishing species from all others in their respective

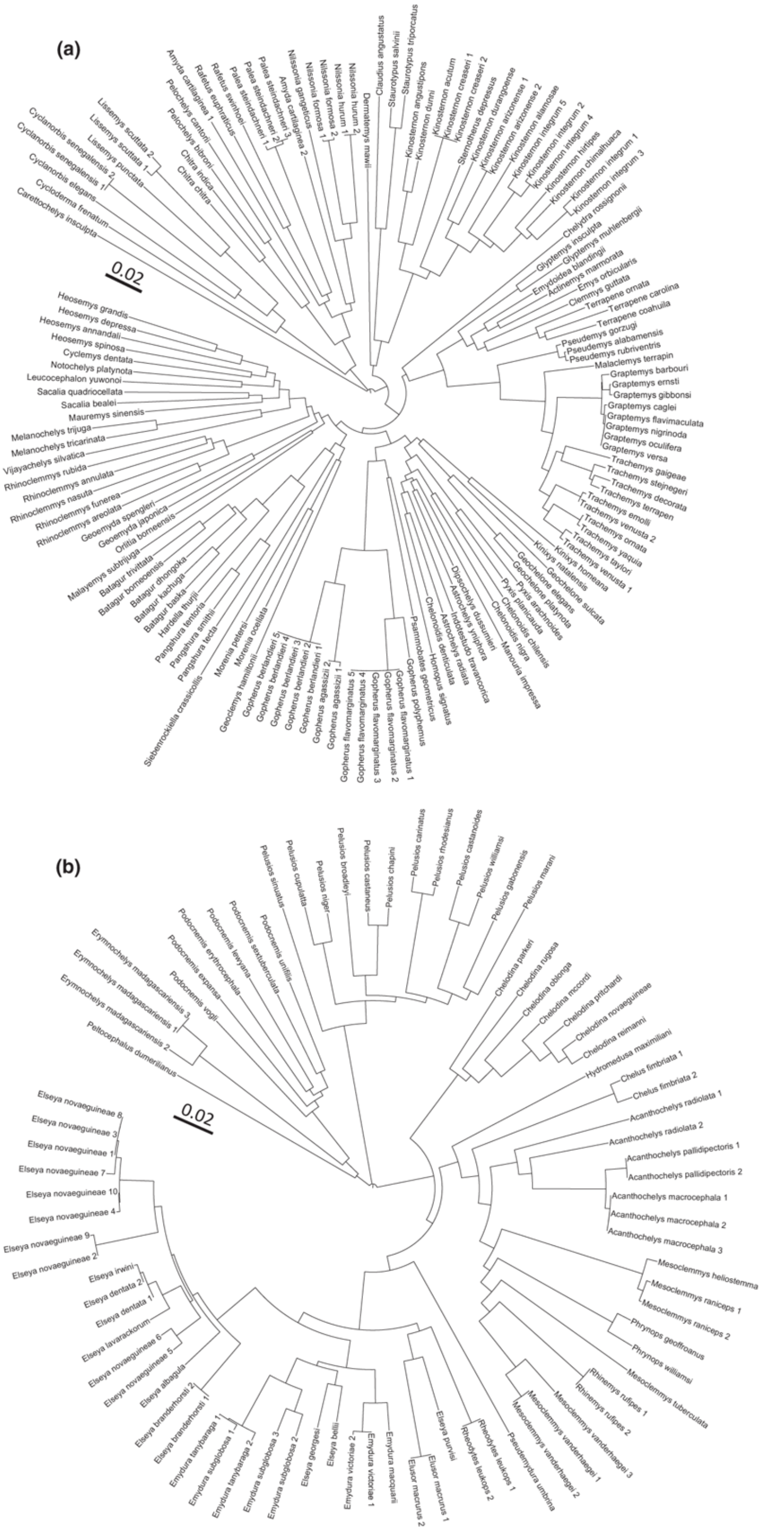


Fig. 2.2.: Neighbour-joining trees of CO1 sequences produced in this study, organized by suborder. (a) Pleurodires. (b) Cryptodires.

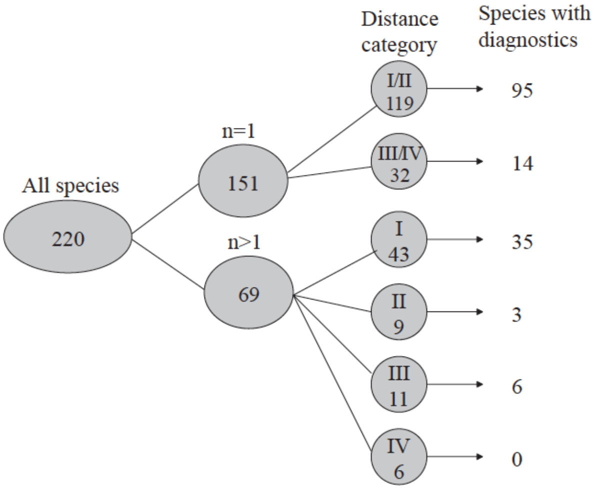


Fig. 2.3.: Number of species in each distance category that exhibit identifying characters at the family level.

families for 155 of the 218 species (71%) in nonmonotypic families. The proportion of species in a given family possessing simple diagnostic traits (Fig. 2.4) varied from 100% (Cheloniidae, Chelydridae, Pelomedusidae, Podocnemididae) to lower than 60% (Emydidae, Geoemydidae). Example sets of simple identifying characters (in which some characters identified by CAOS are excluded for reasons of space) are shown for the families Podocnemididae (Table 2.3a) and Trionychidae (Table 2.3b). Identifying characters could be found in 130 of the 162 species (80%) successfully distinguished by a distance-based threshold (i.e. species in categories I or I/II). Identifying characters were found for 23 of 58 species (40%) in which classification by a distance threshold failed (i.e. species in Categories II, III, III/IV or IV) (Fig. 2.3).

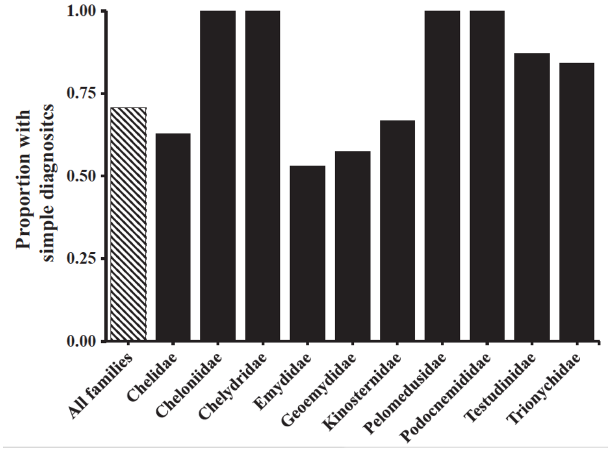


Fig. 2.4.: Proportion of species in the total data set and in each family with identifying characters capable of distinguishing a given species from all others in its family.

2.1.5 Discussion

The barcode sequences assembled here provide a potentially crucial resource for turtle conservation. Barcode records previously existed for only about 50 species; this study more than quadruples that number, allowing approximately two-thirds of extant species to be identified using molecular means and adding entire families to the barcode database that was previously missing. Over the course of the barcoding process, apparent genetic structure was identified in several poorly studied groups, indicating the possible existence of evolutionarily significant units within these putative species that merit further study and possibly extra consideration in conservation efforts. This study also compares distance-based and character-based methods for species identification, and by combining the two highlights a ’third way’ for DNA barcoding that may be useful in improving identification efficiency in taxa for which neither distance nor characters are a perfect fit.

While members of the barcoding community have advanced several different methods of distinguishing species using CO1 sequence information, the distance-based method advanced by Hebert *et al.* (2003) has become and in all probability will remain the standard, workhorse method used in DNA barcoding. Distance-based barcoding uses a 2% divergence (K2P > 0.02) cut-off for vertebrates to determine species identity, implying that individuals should be <2% divergent from members of their own species and more than 2% divergent from members of other species. A maximum of 161 turtle species examined in this study (73%) can be effectively distinguished using this criterion. This is probably an overestimate, as (i) undetected intraspecific divergences >2% may exist in undersampled species and (ii) all closely related species were not sampled for the species examined, leaving open the possibility that some unsampled species could be <2% divergent from the species examined here. In the group of species with more than one individual sampled, the intraspecific divergence criterion was violated about as many times as the interspecific divergence criterion (nine species in Category II vs. 11 species in Category III). As such, raising or lowering the divergence cut-off would probably do little to improve the proportion of species successfully distinguished by a distance-based method.

Species in Category II (high intraspecies divergence) have been targeted as probably examples of cryptic diversity (Hebert *et al.* 2004). Although many of the species identified in this category are rare and/or poorly studied, some evidence points to the existence of cryptic variability within several species. *Elseya novaeguineae*, for example, is regarded as a probably species complex (Georges & Thomson 2009), and the individuals barcoded here fall into three distinct clusters based on CO1 sequence. *Erymnochelys madagascariensis*, another species that is thought to contain multiple population units (Rafeliarisoa *et al.* 2006), also violated the 2% threshold. In the case of the relatively well-studied species *Cuora galbinifrons*, intraspecific divergences of >2% in the publicly available CO1 sequences do indeed map to three distinct clades which Stuart & Parham (2004) argued should be granted full species status based on genetic and morphological divergences. This example from the public data seems to support the possibility that these high intraspecific divergences may represent cryptic diversity. However, the controversy surrounding these designations (Turtle Taxonomic Working Group 2007), and indeed species delimitation based on mitochondrial data alone (Georges & Thomson 2009), reinforces the need for further study including nuclear markers and morphological characteristics to determine the exact nature of this diversity. In some cases, patterns identified in CO1 match biogeographic patterns that have been documented in better-studied species, suggesting that similar evolutionary processes may have been at play in both. For example, *Kinosternon integrum* is broadly sympatric with the Central American iguanid species *Ctenosaura pectinata*, in which high levels of cryptic diversity as well as secondary contact between closely related species have produced patterns of mtDNA structuring (Zarza *et al.* 2008) similar to those noted here.

Observations of low interspecific differentiation (represented here by species in Category III) have been attributed to hybridization and resulting mitochondrial introgression between species, recent speciation or synonymy (Hebert *et al.* 2004). The frequency of low interspecific divergence in turtles can be attributed to several unique aspects of turtle biology. Evidence from marine turtles in the family Cheloniidae (Karl *et al.* 1995; Lara-Ruiz *et al.* 2006) indicates that some turtle species are still able to hybridize after tens of millions years of separation, and instances of intergenus hybridization have been recorded in other turtle families as well (Parham *et al.* 2001; Buskirk *et al.* 2005). Interspecies and even intergenus hybridization may then be possible, if not necessarily frequent, in the wild for many species. Low rates of both molecular evolution and chromosomal rearrangement in turtles (Bickham 1981; Avise *et al.* 1992) may make this hybridization possible by delaying the evolution of genetic barriers to reproduction. Slower rates of molecular evolution may themselves also be an explanation for low levels of differentiation in species that do not hybridize. Because mitochondrial genes tend to accumulate differences at a rate several-fold slower in turtles than in other vertebrates (Avise *et al.* 1992), species considered 'recent radiations' will probably be nearly identical at CO1.

These alternate explanations can be evaluated for some of the well-studied species by using known species ranges to rule out hybridization events. Most of the *Graptemys* species sequenced here are reciprocally allopatric and isolated in separate river drainages (Lamb *et al.* 1994). Only one species sequenced here (*G. gibbonsi*) has a range wide enough to overlap with those of other species (*G. oculifera* and *G. flavimaculata*), and *G. gibbonsi* is relatively well differentiated from these two species within the genus for the barcode fragment. As such, current hybridization is unlikely between the *Graptemys* species examined here. However, hybridization with the more widely distributed *Graptemys* species (*G. ouachitensis* and *G. pseudogeographica*) remains a possibility. Previous molecular work has identified strikingly low differentiation among *Graptemys* in a coding mitochondrial gene and attributed this to recent (<2.5 million years ago) speciation coupled with low rates of molecular evolution (Lamb *et al.* 1994). Similar explanations for low levels of diversification can be invoked for allopatric species in the recently diversified genera *Trachemys* and *Pseudemys*, although hybridization has been noted between *Pseudemys* species in rare cases (Crenshaw 1965). In the family Emydidae, therefore, slow molecular evolution and recent speciation certainly seem to be major causes of low interspecific diversity, although hybridization cannot be ruled out. However, little is known about divergence times or the likelihood of hybridization for other species exhibiting low levels of divergence, and further research will be necessary before these contributing causes can be fully evaluated.

Hebert *et al.* (2004) identified species in Category IV (high intraspecific divergence, low interspecific divergence) as probably examples of sample misidentification. This interpretation, however, assumes that introgression of mitochondrial haplotypes from species more than 2% divergent is either extremely unlikely or impossible. While this assumption may be valid in other taxa, it is demonstrably false for turtles. Several examples from the public data analysed here bear this out. For *Cuora trifasciata*, a species falling into Category IV in our analysis, introgression has produced several highly differentiated mitochondrial clades within the species, even though individuals form only one nuclear clade (Spinks & Shaffer 2007). Feldman & Parham (2004) hypothesize that introgression with *Mauremys annamensis* is a probably cause of high mitochondrial differentiation within another Category IV species in our analysis, *Mauremys mutica*, and hybridization has been recently noted between *Mauremys reevesi* and *Mauremys sinensis* (Fong & Chen 2010). As such, hybridization cannot be ruled out as an explanation for anomalous divergences within species sequenced in this study falling into Category IV (*Trachemys venusta* and *Emydura subglobosa*).

While distance-based barcoding will probably be effective in discriminating the majority of turtle species, this method seems to fail for a fairly large proportion of species. Character-based barcoding provides an attractive complement to distance-based barcoding, especially in turtles where interspecific divergences are probably to fall below the established threshold in closely related species. Relatively, few studies have been performed to date using character-based barcoding methods (Kelly *et al.* 2007; Rach *et al.* 2008; Damm *et al.* 2010; Naro-Maciel *et al.* 2010; Yassin *et al.* 2010). All have used the CAOS algorithm to determine characters that serve as unique species identifiers. This approach was shown to be more successful for differentiating 19 species within a mollusk genus (*Mopalia*) than distance-based barcoding (Kelly *et al.* 2007). A set of pure characters identified by CAOS, combined with several additional characters to form a compound character, was found to be effective for differentiating 54 of 64 species of Odonata (dragonflies and damselflies; Rach *et al.* 2008). The character-based approach had not previously been attempted on a set of species as large as the one examined in this study.

The efficacy of the simple characters identified by CAOS as species identifiers varied between families. The case of the Podocnemididae represents an extremely successful application of character-based barcoding; all species in the family are represented and each possessed simple identifying character states. Even in *Erymnochelys madagascariensis*, a species that displayed >2% intraspecies divergence, the diagnostic characters could unambiguously differentiate each individual in this species from those of other species. In the case of the Trionychidae, 16 of 19 species could be distinguished by simple characters. However, the remaining three species could be identified using the heuristic method of finding a character that unites them with a group containing only species with simple identifiers (all of which can then be distinguished by these characters). In larger families, the number of species for which characters could be found seemed to decline, possibly because of the increased likelihood of homoplasy and back mutations. As such, splitting families into smaller groups and considering compound characters could increase the success of a character-based method. However, a major caveat for all character-based analysis presented here is that, attributed to limited sample size, these character states may not be fixed.

For the species examined here, combining identifying characters with distance-based methods offers an effective means of increasing the proportion of species that can be successfully identified. Twenty-four species violating the distance threshold possessed identifying characters, meaning that incorporating these characters into the identification process would increase the total proportion of species identified by more than 10%. Identifying characters could be incorporated by a stepwise process, as shown in Fig. 2.3, in which species are first identified according to distance-based criteria and then by using identifying characters if ambiguities still remain. The CAOS-based online identification engine described here provides a user-friendly means of carrying out the character-based portion of this approach. However, while characters may aid in species identification, they are not a perfect fix. Species that have extremely similar CO1 haplotypes, such as those in the genus *Graptemys*, often lacked identifying characters simply because of the lack of available variation in CO1. Hybridization and introgression are also serious problems for any mitochondrial identification method. As such, identifying characters provided no resolution for species in Category IV (where introgression was probably an issue). Given the prevalence of introgression among turtle species, the use of a nuclear marker as a supplement to CO1-based barcoding methods may be particularly valuable. Promising candidates for a nuclear barcode marker include the following: recombination activation gene 1 (RAG-1; Krenz *et al.* 2005) and the RNA fingerprint protein 35 intron (R35; Fujita *et al.* 2004). Many of the specimens used to generate the novel CO1 sequences included in this work are currently being sequenced for R35 and RAG-1 as part of separate phylogenetic studies focusing on particular taxa, including the Kinosternidae (Iverson JB, Le M in preparation) and the Australian Chelidae (Georges A, Reid BN, Zhang X, Charlton TR, McCord WP, Le M, in preparation); as such, the utility of both R35 and RAG-1 as complements to the CO1-based barcoding presented here will be assessed in the near future.

While this study shows that accepted barcoding paradigms may be insufficient for species identification in some turtle groups, most species can be effectively discriminated by using a combination of existing methods. The existence of a genetic species identification method for turtles can assist in enforcement of existing laws regulating the traffic of turtles and turtle products and in characterizing the extent of trade in species, especially when these species are traded in otherwise unrecognizable forms. Barcoding could also have a number of possible uses in turtle ecology and conservation beyond its obvious utility in controlling wildlife trade. For example, barcoding of gut contents has been used to elucidate trophic interactions that are hard to observe otherwise (Zeale *et al.* 2011). With the addition of turtle sequences to the barcode database, these studies could detect depredation of turtle eggs, which is extremely high for many turtle species and constitutes one of the most important sources of mortality for a group that is otherwise superbly well armoured (Spencer & Thompson 2003). Turtles are in urgent need of protection, and the barcode sequences provided here will provide a useful tool for conservation and management.

Tab. 2.3.: Example sets of identifying characters for (a) Podocnemididae and (b) Trionychidae. Simple identifying characters are shaded. Characters providing diagnostic information via the heuristic discussed in the text are boxed.

a

	80	89	158	263	308	323	350	368	410	479	527	530	542	545	560
Erymnochelys madagascariensis	A	C	**C**	C	T	A	**A**	**A**	A	**A**	**A**	A	**T**	A	A
Peltocephalus dumerilianus	A	C	**G**	T	**A**	**C**	C	**C**	A	C	**G**	**C**	C	**C**	**T**
Podocnemis erythrocephala	T	C	A	**A**	T	A	C	T	T	C	C	A	C	A	A
Podocnemis expansa	T	**A**	A	C	T	**G**	**T**	T	**G**	**T**	C	A	C	A	**G**
Podocnemis lewyana	**C**	C	T	T	T	A	C	T	**C**	C	C	A	**A**	A	A
Podocnemis sextuberculata	A	**T**	T	C	T	A	C	T	T	C	C	**G**	C	A	A
Podocnemis unifilis	**G**	C	A	C	T	A	C	T	T	C	C	A	C	A	A
Podocnemis vogli	T	C	A	C	**C**	A	C	T	A	C	C	A	C	**T**	A

b

	5	26	121	218	281	290	323	350	512	521	527	536	545	551	614
Amyda cartilaginea	C	A	T	A	A\T	A	C\T	C	A	A	A	A\T	A	A	**T**
Chitra chitra	C	A	T	A	T	A	C	C	A	A	A	A	**T**	A	A
Chitra indica	C	A	T	A	A	A	C	C	**G**	A	A	A	T	G	A
Cyclanorbis elegans	C	A	T	T	A	**G**	T	C	T	C	A	A	**C**	A	**G**
Cyclanorbis senegalensis	C	A	T	T	A	A	T	**A**	T	**T**	**G**	A	A	**T**	C
Cycloderma frenatum	C	A	T	T	A	A	C	C	T	A	**C**	A	A	C	A
Dogania subplana	**T**	A	T	T	A	A	C	C	A	A	A	A	A	A	C
Lissemys punctata	C	A	T	A	A	**T**	T	C	C	C	A	A	A	C	A
Lissemys scuttata	C	**C**	T	A	A	A	T	C	C	C	A	A	A	C	A
Nilssonia formosa	C	A	T	A	**G**	A	C	C	A	A	A	A	A	A	T
Nilssonia gangeticus	C	A	T	A	A	A	T	**T**	A	A	A	A	A	A	T
Nilssonia hurum	C	A	T	**G**	A	A	C	C	A	A	A	A	A	A	T
Palea steindachneri	C	A	T	A	A	C	T	C	A	A	A	**G**	A	A	T
Pelochelys bibroni	C	A	T	A	A	A	**A**	C	C	A	A	A	T	A	A
Pelochelys cantori	C	A	T	A	A	A	**G**	C	C	A	A	A	T	A	A
Pelodiscus sinensis	C	A	**C**	A	T	C	C	C	A	A	A	A	A	G	T
Rafetus euphraticus	C	A	T	A	A	A	T	C	A	A	A	T	A	**G**	A
Rafetus swinhoei	C	A	T	A	A	A	T	C	A	**G**	A	A	A	G	A
Trionyx triunguis	C	A	T	**C**	A	A	T	C	C	A	A	**C**	A	A	C

2.1.6 Acknowledgement

The work described in this report was supported by a generous grant from the Alfred P. Sloan Foundation. TB was funded by H. Wilhelm Schaumann Stiftung (2009-2011) and DAAD (2010). JBI was funded by the Cope Fund through the Joseph Moore Museum of Natural History. The Sackler Institute for Comparative Genomics at the AMNH provided laboratory space, and Sergios-Orestis Kolokotronis provided valuable technical assistance. Blood and tissue samples were generously provided by the Ambrose Monell Cryo Collection at the AMNH, Dave Collins and the Tennessee Zoo, the Berkeley Museum of Vertebrate Zoology, the University of Canberra, the University of Michigan Museum of Zoology, Tag Engstrom and California State University/Chico, and Victoria Milne and the El Paso Zoo.

2.1.7 References

Alacs EA, Georges A, FitzSimmons NN, Robertson J (2010) DNA detective: a review of molecular approaches to wildlife forensics. Forensic Science, Medicine and Pathology, 6, 180-194.

Amer SA, Kumazawa Y (2009) Complete sequence of the mitochondrial genome of the endangered Nile soft-shelled turtle *Trionyx triunguis*. Egyptian Journal of Experimental Biology. Zoology, 5, 43-50.

Avise JC, Bowen BW, Lamb T, Meylan AB, Bermingham E (1992) Mitochondrial DNA evolution at a turtle's pace: evidence for low genetic variability and reduced microevolutionary rate in Testudines. Molecular Biology and Evolution, 9, 457-473.

Bergmann T, Hadrys H, Breves G, Schierwater B (2009) Character-based DNA barcoding: a superior tool for species classification. Berliner und Münchener Tierärztliche Wochenschrift, 122, 446-450.

Bickham JW (1981) Two-hundred-million-year-old chromosomes: deceleration in the rate of karyotypic evolution in turtles. Science, 212, 1291-1293.

Buhlmann KA, Akre TSB, Iverson JB *et al.* (2009) A global analysis of tortoise and freshwater turtle distributions with identification of priority conservation areas. Chelonian Conservation Biology, 8, 116-149.

Buskirk JR, Parham JF, Feldman CR (2005) On the hybridization of two distantly related Asian turtles (Sacalia x Mauremys). Salamandra, 41, 21-26.

Cheung SM, Dudgeon D (2006) Quantifying the Asian turtle crisis: market surveys in southern China, 2000-2003. Aquatic Conservation: Marine and Freshwater Ecosystems, 16, 751-770.

Costa FO, Carvalho GR (2007) The barcode of life initiative: synopsis and prospective societal impacts of DNA barcoding of fish. Genomics, Society and Policy, 3, 29-40.

Crenshaw JW (1965) Serum protein variation in an interspecies hybrid swarm of turtles of the genus Pseudemys. Evolution, 19, 1-15.

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

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

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

van Dijk PP, Stuart BL, Rhodin AGJ (2000) Asian Turtle Trade - Proceedings of a Workshop on Conservation and Trade of Freshwater Turtles and Tortoises in Asia. Chelonian Research Foundation, Lunenburg.

Dove CJ, Rotzel NC, Heacker M, Weigt LA (2008) Using DNA barcodes to identify bird species involved in birdstrikes. Journal of Wildlife Management, 72, 1231-1236.

Eaton MJ, Meyers GL, Kolokotronis S-O, Leslie MS, Martin AP, Amato G (2010) Barcoding bushmeat: molecular identification of Central African and South American harvested vertebrates. Conservation Genetics, 11, 1389-1404.

Feldman CR, Parham JF (2004) Molecular systematics of old world stripe-necked turtles (Testudines: *Mauremys*). Asiatic Herpetological Research, 10, 28-37.

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

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.

Fong JJ, Chen T (2010) DNA evidence for hybridization of wild turtles in Taiwan: possible genetic pollution from trade animals. Conservation Genetics, 11, 2061-2066.

Fujita MK, Engstrom TN, Starkey DE, Shaffer HB (2004) Turtle phylogeny: insights from a novel nuclear intron. Molecular Phylogenetics and Evolution, 31, 1031-1040.

Georges A, Thomson S (2009) Diversity of Australasian freshwater turtles, with an annotated synonymy and keys to species. Zootaxa, 2496, 1-37.

Gibbons JW, Scott DE, Ryan TJ *et al.* (2000) The global decline of reptiles, deja vu amphibians. BioScience, 50, 653-666.

He J, Zhou T, Rao D, Zhang Y (2007) Molecular identification and phylogenetic position of Cuora yunnanensis. Chinese Science Bulletin, 52, 3305-3309.

Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society London B, 270, S96-S99.

Hebert PDN, Stoeckle MY, Zemlak TS, Frances CM (2004) Identification of birds through DNA barcodes. PLOS Biology, 2, 1657-1663.

International Union for the Conservation of Nature (2009) Red List of Threatened Species. http://www.IUCNredlist.org/. Accessed 6 March 2009.

Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primer cocktails for fish DNA barcoding. Molecular Ecology Notes, 6, 998-1002.

Jungt SO, Lee YM, Kartavstev Y, Park IS, Kim DS, Lee JS (2006) The complete mitochondrial genome of the Korean soft-shelled turtle *Pelodiscus sinensis* (Testudines, Trionychidae). DNA Sequencing, 17, 471-483.

Karl SA, Bowen BW, Avise JC (1995) Hybridization among the ancient mariners: characterization of marine turtle hybrids with molecular genetic assays. Journal of Heredity, 86, 262-268.

Kelly RP, Sarkar IN, Eernisse DJ, DeSalle R (2007) DNA barcoding using chitons (genus *Mopalia*). Molecular Ecology Notes, 7, 177-183.

Krenz JG, Naylot GJP, Shaffer HB, Janzen FJ (2005) Molecular phylogenetics and evolution of turtles. Molecular Phylogenetics and Evolution, 37, 178-191.

Kumazawa Y, Nishida M (1999) Complete mitochondrial DNA sequences of the green turtle and blue-tailed mole skink: statistical evidence for archosaurian affinity of turtles. Molecular Biology and Evolution, 16, 782-792.

Lamb T, Lydeard C, Walker RB, Gibbons JW (1994) Molecular systematics of map turtles (*Graptemys*): a comparison of mitochondrial restriction site versus sequence data. Systematic Biology, 43, 543-559.

Lara-Ruiz P, Lopez GG, Santos FR, Soares LS (2006) Extensive hybridization in hawksbill turtles (*Eretmochelys imbricata*) nesting in Brazil revealed by mtDNA analyses. Conservation Genetics, 7, 773-781.

Letunic I, Bork P (2006) Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23, 127-128.

Maddison DR, Maddison WP (2000) MacClade 4: Analysis of phylogeny and character evolution. Version 4.0. Sinauer Associates, Sunderland, Massachusetts. Messing J (1983) New M13 vectors for cloning. Methods in Enzymology, 101, 20-78.

Mindell DP, Sorenson MD, Dimcheff DE, Hasegawa M, Ast JC, Yuri T (1999) Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. Systematic Biology, 48, 138-152.

Naro-Maciel E, Reid B, Fitzsimmons NN, Le M, DeSalle R, Amato G (2010) DNA barcodes for globally threatened marine turtles: a novel registry approach to documenting biodiversity. Molecular Ecology Resources, 10, 252-263.

Nijman V, Shepherd CR (2007) Trade in non-native, CITES-listed, wildlife in Asia, as exemplified by the trade in freshwater turtles and tortoises (Chelonidae) in Thailand. Contributions to Zoology, 76, 207-212.

Ogden R, Dawnay N, McEwing R (2009) Wildlife DNA forensics–bridging the gap between conservation genetics and law enforcement. Endangered Species Research, 9, 179-195.

Parham JF, Simison WB, Kozak KH, Feldman CR, Shi H (2001) New Chinese turtles: endangered or invalid? A reassessment of two species using mitochondrial DNA, allozyme electrophoresis and known-locality specimens. Animal Conservation, 4, 357-367.

Parham JF, Stuart BL, Bour R, Fritz U (2004) Evolutionary distinctiveness of the extinct Yunnan box turtle (*Cuora yunnanensis*) revealed by DNA from an old museum specimen. Proceedings of the Royal Society of London B, 271, S391-S394.

Parham JF, Feldman CR, Boore J (2006a) The complete mitochondrial genome of the enigmatic bigheaded turtle (*Platysternon*): description of unusual genomic features and the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA. BMC Evolutionary Biology, 6, 11.

Parham JF, Macey JR, Papenfuss TJ *et al.* (2006b) The phylogeny of Mediterranean tortoises and their close relatives based on complete mitochondrial genome sequences from museum specimens. Molecular Phylogenetics and Evolution, 38, 50-64.

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

Rafeliarisoa T, Shore G, Engberg S, Louis E, Brenneman R (2006) Characterization of 11 microsatellite marker loci in the Malagasy big-headed turtle (*Erymnochelys madagascariensis*). Molecular Ecology Notes, 6, 1228-1230.

Russell RD, Beckenbach AT (2008) Recoding of translation in turtle mitochondrial genomes: programmed frameshift mutations and evidence of a modified genetic code. Journal of Molecular Evolution, 67, 682-695.

Sarkar IN, Thornton J, Planet PJ, Schierwater B, DeSalle R (2002) A systematic method for classification of novel homeoboxes. Molecular Phylogenetics and Evolution, 24, 388-399.

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

Spencer RJ, Thompson MB (2003) The significance of predation in nest site selection of turtles: an experimental consideration of macro- and microhabitat preferences. Oikos, 102, 592-600.

Spinks PQ, Shaffer HB (2007) Conservation genetics of the Asian box turtles (Geoemydidae, *Cuora*): mitochondrial introgression, numts and inferences from multiple nuclear loci. Conservation Genetics, 8, 641-657.

Spinks PQ, Shaffer HB, Iverson JB, McCord WP (2004) Phylogenetic hypotheses for the turtle family Geoemydidae. Molecular Phylogenetics and Evolution, 32, 164-182.

Stuart BL, Parham JF (2004) Molecular phylogeny of the critically endangered Indochinese box turtle (*Cuora galbinifrons*). Molecular Phylogenetics and Evolution, 31, 164-177.

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.

Turtle Conservation Fund (2002) A global action plan for conservation of tortoises and freshwater turtles. Strategy and funding prospectus 2002-2007. 30 pp. Conservation International and Chelonian Research Foundation, Washington, DC.

Turtle Taxonomic Working Group (2007) An annotated list of modern turtle taxa with comments on areas of taxonomic instability and recent change. Chelonian Research Monographs, 4, 173-199.

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.

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

Zardoya R, Meyer A (1998) Complete mitochondrial genomes suggests diapsid affinities of turtles. Proceedings of the National Academy of Sciences of the United States of America, 95, 14226-14231.

Zarza E, Reynoso VH, Emerson BC (2008) Diversification in the northern neotropics: mitochondrial and nuclear DNA phylogeography of the iguana *Ctenosaura pectinata* and related species. Molecular Ecology, 17, 3259-3275.

Zeale MRK, Butlin RK, Barker GLA, Lees DC, Jones G (2011) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. Molecular Ecology Resources, 11, 236-244.

Zhang L, Nie L, Cao C, Zhan Y (2008) The complete mitochondrial genome of the keeled box turtle *Pyxidea mouhotii* and phylogenetic analysis of major turtle groups. Journal of Genetics and Genomics, 35, 33-40.

2.1.8 Data Accessibility

DNA Sequences: Genbank accessions HQ329587-HQ329787; BOLD accessions BENT102-08-BENT335-09. Alignments and trees: TreeBASE accessions S11480.

2.1.9 Support Information

Additional supporting information may be found in the online version of this article. Table S1 Descriptive data for all taxa and sequences included in this study. 'N' indicates the number of individuals sequences for each species; 'H' indicates the number of haplotypes observed in each species; 'Distance' indicates the species' classification within the distance-based scheme described in the text; 'Diagnostic' indicates the presence ('Y') or absence ('N') of family level simple identifying characters in the species. References and accession numbers are in bold for novel sequences produced in this study. Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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

MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2013) doi: 10.1111/1755-0998.12125

Authors: Tjard Bergmann*, Jessica Rach*, Sandra Damm, Rob DeSalle, Bernd Schierwater and Heike Hadrys

Tjard Bergmann: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; Email address: tjard.bergmann@ecolevol.de

Jessica Rach: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; Email address: jessica.rach@ecolevol.de

Sandra Damm: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; Email address: sandra.damm@ecolevol.de

Rob DeSalle: American Museum of Natural History, Sackler Institute for Comparative Genomics, New York, NY 10024, U.S.A.; Email address: desalle@amnh.org

Bernd Schierwater: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; American Museum of Natural History, Sackler Institute for Comparative Genomics, New York, NY 10024, U.S.A.; Email address: bernd.schierwater@ecolevol.de

Heike Hadrys: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; American Museum of Natural History, Sackler Institute for Comparative Genomics, New York, NY 10024, U.S.A.; Email address: heike.hadrys@ecolevol.de

⋆ = These authors contributed equally to this work.

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

2.2.1 Abstract

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

2.2.2 Introduction

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

The Consortium for the Barcode of Life (CBoL) has agreed on the use of a 648 base-pair fragment at the 5'end of the mitochondrial cytochrome c oxidase subunit 1 gene region (CO1) as default DNA barcode region for vertebrates, insects and as many other animal groups as possible. As it was first promoted as suitable

DNA barcoding marker for many animal groups by Hebert *et al.* (2003b), CO1 has been successfully used for obtaining reliable DNA barcodes and for a broad range of animal groups, such as arthropods (Ekrem *et al.* 2007; Foley *et al.* 2007; Hajibabaei *et al.* 2006; Monaghan *et al.* 2005; Smith *et al.* 2006; Will & Rubinoff 2004; Witt *et al.* 2006), birds (Hebert *et al.* 2004a; Kerr *et al.* 2007; Yoo *et al.* 2006), fishes (Ward *et al.* 2005) and mammals (Clare *et al.* 2007; Dawnay *et al.* 2007). In some animal groups, however, CO1 has failed to deliver reliable DNA barcodes. In cnidarians and sponges, for example, CO1 divergences are extraordinarily low compared to bilaterian animals (Park *et al.* 2007; Shearer *et al.* 2002). On the other hand, in aves, gastropods and amphibians, inter- and also intraspecific variation in CO1 are very high (Hebert *et al.* 2004b; Remigio & Hebert 2003). In 449 dipteran species, the identification success through CO1 ”barcodes” was low due to substantial overlaps in inter- and intraspecific divergences (Meier *et al.* 2006). Moreover, it was shown that the vast majority of nucleotide substitutions within the CO1 fragment occur at the third codon position, which might lead to rapid saturation (Lin & Danforth 2004; Vences *et al.* 2005).

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

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

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

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

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

2.2.3 Material & Methods

a. Sample collection, processing and sequencing

Tissue samples of 271 individuals representing 51 species, 22 genera and 8 families from Europe and Africa (Table S1; electronic supporting material) were collected during 2001 and 2006 by non-invasive sampling (Hadrys *et al.* 2005) and stored in 70% or 98% ethanol prior to DNA extraction. Table 1 lists the analysed species and individuals per species.

Tab. 2.4.: Mean intra- and interspecific divergences of ND1 and CO1 from 51 odonate species; the source of the sequence is shown for ND1 and CO1; mean intra- and interspecific divergences (Kimura 2-parameter distances) are given in %; lowest and highest interspecific distance values for each species are shown

Species	**No. of Individuals**	**ND1 sequence**	**CO1 sequence**	**Mean intraspecific divergence (%) ND1|CO1**	**Mean interspecific divergence (%) ND1|CO1**
A. Anisoptera					
Aeshna cyanea	4	New	New	0 | 0	5.5-30.4 | 7.7-23.4
Aeshna grandis	1	1	New	-	4.8-29.2 | 8.6-24.7
Aeshna mixta	2	New	New	0 | 0.2	4.8-28.9 | 7.5-25.7
Aeshna rileyi	2	1	New	0 | 0	6.6-29.4 | 8.7-25.0
Anaciaeschna triangulifera	1	1	New	-	5.6-31.6 | 7.5-23.7
Anax ephippiger	10	1	New	0.2 | 0.7	7.6-30.3 | 7.4-26.1
Anax imperator	11	1	New	0.3 | 0.2	2.5-30.0 | 5.8-24.1
Anax speratus	6	1	New	0 | 0	2.5-29.6 | 5.8-24.0
Brachytron pratense	2	1	New	0 | 0	8.0-33.0 | 10.1-26.3
Gynacantha usambarica	9	1	-	0 | -	8.9-31.3 | n/c
Gynacantha villosa	1	1	New	-	9.8-31.1 | 10.8-25.6
Paragomphus geneii	5	1	-	0.9 | -	17.9-35.6 | n/c
Crocothemis erythraea	7	1	New	0.3 | 1.0	18.2-37.4 | 14.6-25.1
Crocothemis sanguinolenta	6	New	New	0 | 0.8	16.1-33.0 | 13.2-27.4
Nesciothemis farinosum	5	New	New	0.6 | 0.3	12.2-31.9 | 13.6-25.7

Continue next page

Species	No. of Individuals	ND1 sequence	CO1 sequence	Mean intraspecific divergence (%) ND1|CO1	Mean interspecific divergence (%) ND1|CO1
Orthetrum brachiale	3	1	New	0 | 0.1	8.1-31.2 | 7.1-23.5
Orthetrum chrysostigma	4	1; New	New	0.2 | 0.5	6.0-30.3 | 5.9-24.2
Orthetrum coerulescens	9	1; New	New	0.2 | 0.1	7.4-30.1 | 10.4-23.1
Orthetrum julia falsum	10	1; New	New	0.2 | 0.5	6.0-30.3 | 5.9-23.6
Orthetrum trinacria	5	1	New	0 | 0	11.2-33.1 | 12.0-24.6
Sympetrum sanguineum	2	New	-	0 | -	14.1-32.1 | n/c
Trithemis annulata	3	1	3	0.2 | 0.1	7.9-34.8 | 8.2-23.8
Trithemis arteriosa	1	4	New	-	8.2-32.4 | 9.1-24.1
Trithemis donaldsoni	5	New	New	0.4 | 0.2	12.3-31.9 | 11.8-21.6
Trithemis furva	3	2	3	0.2 | 1.4	9.1-35.2 | 9.7-23.9
Trithemis grouti	2	2	New	1.0 | 0.2	7.9-37.1 | 1.1-25.5
Trithemis hecate	5	1	-	0 | -	13.7-39.2 | n/c
Trithemis kirbyi	4	1; New	New	0.8 | 0.7	15.4-37.3 | 11.3-23.1
Trithemis morrisoni	5	2	3	2.4 | 0.5	4.7-34.8 | 5.0-24.4
Trithemis nuptialis	2	2	3	0 | 0	2.7-34.9 | 1.1-24.8
Trithemis palustris	4	2	3	0.4 | 0.3	4.7-36.7 | 5.0-24.2
Trithemis stictica	7	2; 3	3	0.1 | 0.1	2.7-36.3 | 2.8-24.4
B. Zygoptera					
Calopteryx haemorrhoidales	12	New	-	0.2 | -	15.8-35.5 | n/c
Calopteryx splendens	4	1	-	0 | -	15.8-36.0 | n/c
Platcypha auripes	2	1	New	0.3 | 0	12.0-39.2 | 10.3-25.7
Platcypha caligata	6	1	New	0.3 | 0.2	12.0-36.0 | 10.3-24.8
Ceriagrion tenellum	5	New	New	0 | 0.1	13.4-31.4 | 17.8-23.9
Enallagma cyanthigerum	5	New	New	0 | 0.1	15.1-32.0 | 12.8-23.8
Ischnura graellsii	5	New	New	0 | 0.1	15.1-31.8 | 7.9-25.9
Ischnura senegalensis	5	-	New	-| 0	n/c | 7.9-21.0
Leptagrion elongatum	1	New	New	-	13.4-31.3 | 16.7-25.1
Pseudagrion acaciae	4	1	New	0 | 0.2	0(14.4)-36.0| 0.6-23.4

Continue next page

<table>
<tr><th>Species</th><th>No. of Individuals</th><th>ND1 sequence</th><th>CO1 sequence</th><th>Mean intraspecific divergence (%) ND1|CO1</th><th>Mean interspecific divergence (%) ND1|CO1</th></tr>
<tr><td><i>Pseudagrion bicoerulans</i></td><td>15</td><td>1; New</td><td>New</td><td><b>4.3</b> | <b>4.2</b></td><td>13.1-30.7 | 15.8-23.4</td></tr>
<tr><td><i>Pseudagrion kersteni</i></td><td>11</td><td>1; 6; New</td><td>New</td><td>1.1 | 1.1</td><td>13.1-31.2 |15.8-26.4</td></tr>
<tr><td><i>Pseudagrion massaicum</i></td><td>13</td><td>1; New</td><td>New</td><td>0.6 | 0.7</td><td>14.4-37.6 | 13.6-25.1</td></tr>
<tr><td><i>Pseudagrion niloticum</i></td><td>6</td><td>1; New</td><td>New</td><td>0 | 0.7</td><td>0(14.4)-36.0 | 0.6-23.7</td></tr>
<tr><td><i>Teinobasis alluaudi</i></td><td>6</td><td>New</td><td>New</td><td>0.4 | 0.3</td><td>15.8-29.6 |16.9-27.9</td></tr>
<tr><td><i>Chlorocnemis abbotti</i></td><td>8</td><td>New</td><td>New</td><td>0.3 | 0.2</td><td>16.3-35.5 | 18.2-27.1</td></tr>
<tr><td><i>Coryphagrion grandis</i></td><td>14</td><td>5; New</td><td>New</td><td>2.6 | 2.4</td><td>16.7-30.1 | 18.0-24.2</td></tr>
<tr><td><i>Mecistogaster asticta</i></td><td>1</td><td>New</td><td>New</td><td>-</td><td>12.8-35.3 | 13.9-27.9</td></tr>
<tr><td><i>Mecistogaster martinezi</i></td><td>2</td><td>New</td><td>New</td><td>0 | 0</td><td>12.8-37.1 | 13.9-24.8</td></tr>
</table>

Bold indicates exceptional high values.

1) Rach *et al.* (2008)

2) Damm *et al.* (2010a)

3) Damm *et al.* (2010b)

4) Damm & Hadrys (2012)

5) Groeneveld *et al.* (2007)

6) Dijkstra *et al.* (2007)

DNA was extracted using a standard phenol chloroform method (Hadrys *et al.* 1992). The universal primers LCO1490 (5'-GGTCAACAAATCATAAAGATATTGG-3') and HCO2198 (5'-TAAACTTCAGGGTGACCAAAAAATCA-3') were used to amplify the 'Folmer (CO1) fragment' (Folmer *et al.* 1994) and the primer pair P850 (fw), 5'-TTCAAACCGGTGTAAGCCAGG-3' and P851 (rev) 5'-TAGAATTAGAAGATCAACCAGC-3' was used to amplify a fragment containing a 5' partial fragment of 16S tRNA^{Leu} and a 3' partial fragment of the NADH dehydrogenase 1. PCR amplifications were carried out in 25 μl reactions containing 2.5 μl of 10 X Taq DNA polymerase buffer (Bioline/Invitrogen), 2.5 mM MgCl$_2$, 0.1 mM dNTPs, 7.5 pM each primer and 0.5 U Taq DNA polymerase (either Invitrogen or Bioline). In cases of no immediate amplification success, 0.2 mol/l Trehalose was added to the regular PCR mix (Hajibabaei *et al.* 2005; Spiess *et al.* 2004). Amplification conditions were as follows: initial denaturing at 95 °C 2 min, 30 cycles of 30 s denaturing at 95 °C, 30 s annealing at 48 °C (ND1)/ 50 °C (CO1), 1 min extension at 72 °C, followed by a final extension of 6 min at 72 °C. Amplified products were sequenced either on a MegaBACE 500 sequencer using the DYEnamic ET Dye Terminator Cycle Sequencing Kit (Amersham Bioscience) or on an ABI PRISMTM 310 Genetic Analyzer using ABI BigDyeÆTerminator v1.1 (Applied Biosystems). Sequences were assembled and edited using SEQMANII (v. 5.03; DNASTAR, Inc.). All new sequences were deposited in Genbank (CO1 KC912199-KC912405; ND1 KC912406 - KC912523). In addition sequences from previous publications of our research (Damm *et al.* 2010a; Damm & Hadrys; Damm *et al.* 2010b; Dijkstra *et al.* 2007; Groeneveld *et al.* 2007; Rach *et al.* 2008) were included in our data sets (see Table 1 for details). The complete CO1 and ND1 data sets used in this manuscript are deposited in the CAOS-Library of the CAOS-Workbench website (http://boli-new.uvm.edu/CAOS-workbench/).

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

b. DNA barcode analyses

For distance-based threshold analyses mean distances of CO1 and ND1 sequences within and among species were calculated using the Kimura 2-parameter (K2P) substitution model in MEGA 3.1 (Kumar *et al.* 2004). Mean intraspecific as well as lowest and highest mean interspecific K2P distance values for all species are shown in Table 2.4.

For character-based barcode analyses each dataset of CO1 and ND1 was first aligned with the G-INS-I setting of the Mafft software (Katoh *et al.* 2005) and the alignments were converted into the nexus file format with SeaView version 4 (Gouy *et al.* 2010). A Maximum Likelihood (ML) tree was created for each dataset using RAxML (Stamatakis 2006) with 100 bootstraps. The ML trees served as guide trees for CAOS (Character Attribute Organization System) analyses. Each tree file and the corresponding nexus file were saved as one file using MacClade 4 v. 4.06. (Maddison & Maddison 2000), and processed with the CAOS-Analyzer. The CAOS-Analyzer, which can be run on a web server (http://boli.uvm.edu/CAOS-workbench/) or as a command line program, identifies diagnostic characters, termed "characteristic attributes", for all clades at each branching node within the given guide tree (Bergmann *et al.* 2009; Sarkar *et al.* 2002a; Sarkar *et al.* 2008; Sarkar *et al.* 2002b). In order to produce the character-based barcodes the output files of the CAOS-Analyzer were run through the CAOS-Barcoder. From the reference barcode created by the CAOS-Barcoder we selected by eye twentynine species specific simple "pure" characteristic attributes (shared by all members of a clade and absent from the other clades descending from the same node) for CO1 and ND1 as a representative example for a character-based barcode (Figs 2.5 & 2.6).

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

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

We devised a second test of the robustness of character-based diagnostics for the classification of query sequences that involved creating random substitution datasets based on the real data sets. These simulated data sets were then run through the CAOS-Classifier. For both genes CO1 and ND1 we created 100 random substitution datasets with a 1% nucleotide exchange ratio and 100 random substitution datasets with 5% nucleotide exchange ratio. The substitution included the random selection of either an ”A”, ”T”, ”C”, ”G”, ”-”, ”?” or an ”N” at a random position within the sequences. Each of the 100 random matrices contained all 234 sequences for CO1 and all 266 sequences for ND1.

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

Fig. 2.5.: Character-based DNA barcodes for 45 odonate species based on CO1 sequences; unique combinations of character states at 29 nucleotide positions for each species are shown; grey shaded cells show two different bases at the particular nucleotide position within a species.

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

Fig. 2.6.: Character-based DNA barcodes for 50 odonate species based on ND1 sequences; unique combinations of character states at 29 nucleotide positions for each species are shown; dashed cells indicate the occurrence of three or four character states within a species; grey shaded cells show two different bases at the particular nucleotide position within a species.

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

2.2.4 Results

Distance-based thresholds

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

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

Intraspecific distances. The mean intraspecific K2P distances ranged from 0% to 4.2% in CO1 and 0% to 4.3% in ND1. For six out of 45 species only one sample was analyzed and intraspecific divergences could not be calculated. The highest values were observed for *Pseudagrion bicoerulans* (CO1:4.2%; ND1 4.3%). Here, all four analyzed populations form distinct clusters (see above). High intraspecific distances of at least 1% within one fragment were also detected for *Coryphagrion grandis* (CO1/ND1: 2.6%), *Pseudagrion kersteni* (CO1/ND1: 1.1%), *Trithemis furva* (CO1: 1.4%), *Crocothemis erythreae* (1%) and *Trithemis grouti* (1%). Intraspecific distances of more than 0.5% either within ND1 or CO1 were observed for further eight species (see Table 2.4).

Character-based DNA barcodes

Diagnostic characters for species. A core sequence of 29 nucleotide positions of the CO1 fragment showed the highest number of diagnostic characters for groups at the important nodes and exhibited diagnostic characters for very closely related species (Fig. 2.5). The character states at the chosen nucleotide positions revealed unique base compositions – character-based DNA barcodes – for 43 out of the 45 species. No diagnostic characters were found for differentiating specimens of *Pseudagrion niloticum* from those of *Pseuagrion acaciae*.

Similar to the CO1 sequences, a core region of 29 nucleotide positions of the ND1 fragment was selected (Fig. 2.6). Of the 29 nucleotide positions, 23 were used previously as character-based DNA barcodes in dragonflies. Since the 5' end of the sequences were trimmed by 142 bp, the numbers of nucleotide positions changed and six positions were additionally included. 48 out of 50 species revealed unique combinations of character states at the 29 nucleotide positions. Again, no diagnostic characters were found to distinguish *P. acaciae* and *P. niloticum*.

Table 2.5 lists the numbers of pure diagnostic characters for sister species pairs. The lowest number of diagnostic characters within the CO1 fragment was found for *Trithemis nuptialis* and *Trithemis grouti*, which differed by five nucleotide positions. The ND1 fragment revealed 21 pure diagnostic characters for this sister species pair. Very low numbers of diagnostic characters within the ND1 fragment were found for *A. imperator* and *A. speratus* (six diagnostic characters) and *Trithemis stictica* and *T.*

nuptials (eight diagnostic characters). The CO1 fragment exhibited 29 diagnostic characters for the differentiation of *A. imperator* and *A. speratus* and 17 for *T. stictica* and *T. grouti*. For all other pairs of sister species at least 16 diagnostic characters within the CO1 or ND1 fragment have been found.

Tab. 2.5.: Number of pure diagnostic characters identified within the CO1 and ND1 sequences for five sister species pairs. Number of pure diagnostics characters identified for populations or geographical groups of five odonate species. For further explanations, see text.

Sister species pairs		No. Pure diagnostics CO1 / ND1
Anax imperator	*Anax speratus*	29/6
Aeshna cyanea	*Aeshna mixta*	35/19
Trithemis nuptialis	*Trithemis grouti*	5/21
Trithemis stictica	*Trithemis grouti*	17/20
Trithemis stictica	*Trithemis nuptialis*	16/8

Populations		No. Pure diagnostics CO1 / ND1
P. bicoerulans, Mt. Elgon, Kenya	*P. bicoerulans*, Mt. Kenya	2/4
O. julia falsum, Waterberg, Namibia	*O. julia falsum*, Tsauchab, Namibia	5/1
C. grandis, Kenya	*C. grandis*, Tanzania	17/11
O. coerulescens, Germany	*O. coerulescens*, Italy	1/1
T. furva, South Africa	*T. furva*, Ethiopia	9/1

Diagnostic characters identifying geographical clusters or flagging of populations with diagnostics. We also use the DNA barcoding information to group specimens within distinct species according to geographic origin to test for diagnosis of these groups as potential novel species. This process has been called ’flagging’ (Goldstein & DeSalle 2011), where flagging refers to the process of designating populations as potential species worthy of further anatomical, behavioral or other work to determine species existence. Species showing distinct geographical clusters are listed in Table 2.5, and the number of diagnostic characters for each of the geographic clusters are given. For the two German populations of *Orthetrum coerulescens*, one diagnostic character each was found within the CO1 and ND1 sequence to distinguish them from the Italian population. Five diagnostic characters within the CO1 and one within the ND1 fragment could differentiate the two Namibian populations of *Orthetrum julia falsum*. The *Trithemis furva* sample from South Africa shows different character states when comparing it to the two Ethiopian samples (nine nucleotide positions within

CO1 and one within ND1). For *Pseudagrion bicoerulans* distinct clusters for all four populations were observed. Here, the lowest numbers of diagnostic characters were found for the two Kenyan populations from Mount Kenya and Mount Elgon (CO1: 2; ND1: 4). The third Kenyan population and the Tanzanian population differed from the others by at least 14 nucleotide positions within the CO1 and 16 positions within the ND1 fragment. For *Coryphagrion grandis* two distinct clusters were detected, one comprised all three Kenyan and the other all three Tanzanian populations. The clusters revealed pure diagnostic characters at 17 nucleotide positions within CO1 and 11 within ND1.

Leave one out test. In order to test the validity of the CAOS-Classifier for assigning queries to the correct species 234 Odonata reference datasets for CO1 were created all leaving out one of the 234 sequences. For 227 of the 234 left out sequences the best hit was at the same species level with an identity of 98,34-100% (Table S2; electronic supporting material). For the seven remaining query sequences *Mecistogaster asticta*, *Leptagrion elongatum*, *Gynacantha villosa*, *Aeshna grandis*, *Anaciaeschna triangulifera*, *Aeshna mixta* and *Trithemis arteriosa* the best hits were between 82,62% and 90,20%. All of these queries belong to species with specimen sizes of n=1, only for *A. mixta* the number of specimen sequences was n=2. Three queries (*M. asticta*, *L. elongatum*, *T. arteriosa*) were matched with their closest relative in the dataset, while the remaining four queries were assigned to the wrong species.

266 Odonata reference datasets for ND1 were created all leaving out one of the 266 sequences. For 260 of the 266 left out sequences the best hit was at the same species level with an identity of 98,73-100% (Table S3; electronic supporting material). For the six remaining query sequences *Aeshna grandis*, *Anaciaeschna triangulifera*, *Gynacantha villosa*, *Mecistogaster asticta*, *Leptagrion elongatum* and *Trithemis arteriosa* the best hits were between 78.80% and 90.49%. All of these queries belong to species with specimen sizes of n=1. Two queries (*M. asticta*, *T. arteriosa*) were matched with their closest relative in the dataset, while the remaining four queries were assigned to the wrong species.

Random substitution test. In order to test the robustness of diagnostic characters for species identification we created randomly generated sequences and challenged the

CAOS-Classifier with these sequences. The average score of correct species assignments for 100 randomly substituted datasets was evaluated (Table S4; electronic supporting material). For the CO1 datasets with 1% substitution ratio we observed an average score of 233 correct assignments out of 234 (99.5%). Increasing the substitution ratio to 5% led to a reduction of correct assignments to 225 out of 234 (96.1%). For the ND1 datasets with 1% substitution ratio we observed an average score of 249 correct assignments out of 266 (93.8%). Increasing the substitution ratio to 5% led to a reduction of correct assignments to 237 out of 266 (89.1%).

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

2.2.5 Discussion

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

CO1 vs. ND1 vs. CO1/ND1

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

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

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

Comparing character-based barcoding and distance-based thresholds

The majority of DNA barcoding studies have focused on the distance-based approach for analyzing DNA barcodes (Hebert *et al.* 2003a). The accuracy of this method depends on the discrepancy between intra- and interspecific values – the ”barcoding gap” (Meyer & Paulay 2005). In odonates high intra- and low interspecific variability has been observed leading to the conclusion that distance-based methods are ill-suited for DNA barcoding in this insect order (Rach *et al.* 2008). Our data confirm these findings. High mean intraspecific K2P distances of more than 1% are observed for four out of 50 species in the ND1 and for five out of 45 species in the CO1 fragment. The highest intraspecific distance values are seen in *P. bicoerulans* (ND1: 4.3%; CO1: 4.2%), and a rapid speciation in this species has been suspected as in former studies (Dijkstra *et al.* 2007; Hadrys *et al.* 2006). In contrast, in some cases the distance values between sister species are extraordinarily low. For example, the mean K2P divergence of ND1 among *A. speratus* and *A. imperator* (2.5%) is lower than the observed mean intraspecific distance in *C. grandis* (2.6%). The mean interspecific distance between *T. nuptialis* and *T. stictica* is only slightly higher (2.7%). The CO1 distance between *T. nuptialis* and *T. grouti* is only 1.1% and is exceeded by the mean intraspecific CO1 distances in four species. Although we examined only a small part of the worldwide dragonfly diversity we assume that cases of overlapping intra-and interspecific distances are prevalent.

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

Diagnostic characters for geographical clusters; flagging of populations with diagnostics

The ND1 and CO1 sequences can also be examined for diagnostics within distinct geographic clusters of individuals. There are two purposes for searching for such diagnostics. First, the diagnostics can be used to identify populations of origin for unidentified specimens. Such diagnostics can then be used in ecological monitoring studies where samples are hard to identify to population. Second, if diagnostics do exist, then these populations can be flagged for future, integrated taxonomic studies (DeSalle 2006; Rubinoff 2006) that might result in species descriptions for these diagnosable populations (Goldstein *et al.* 2000).

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

Leave one out test

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

Random substitution test

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

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

Comparison of the CAOS-Classifier and BOLD

All test sequences for CO1 were correctly assigned by the CAOS-Classifier to the corresponding species. This result shows that when queries are tested that have at least one representative species sequence within the reference library, an accurate match is identified by the CAOS-Classifier. When we tested the same sequences with BOLD, only 131 of 234 sequences were assigned to a species. In all fairness, we have to mention that at this point, our CO1 sequences were not submitted to BOLD, and the identification was performed using only Odonata data that was included by other researchers. Nevertheless, it also shows that even BOLD has problems with the assignment of sequences when no closely related reference data are available to the program. Of the 131 sequences that were assigned to species, 100 shared the same species as predicted by the CAOS-Classifier. The remaining 31 query sequences were either assigned to a closely related species (22 times) or to a different species than we had assumed (9 times). The first observation can be explained by insufficient data in the BOLD library and a strong similarity of sequences between closely related species. In the second observation, the four specimens we had assigned to *Enallagma cyathigerum* were assigned by BOLD to *Coenagrion hastulatum* with 99.81% identity. Five specimens which we assigned to *Ischnura senegalensis* were assigned by BOLD to *Pseudagrion abyssinica* with 100% identity. Especially in the last case, we can only assume that either the other researchers who added the reference sequences to BOLD made an error in specimen identification or alternatively we have made identification errors. The scenario that both species share the same sequences could be possible but is unlikely.

Conclusions

In this study, we have used 271 odonate samples belonging to 51 species. The analyses of the genetic data reveal that odonates are a challenging test-bed for DNA barcoding. The employment of two combined genetic markers highly enhances the identification of organisms through DNA sequences, even if both markers are of mitochondrial origin. The number of diagnostic characters for the discrimination of taxonomic groups increases substantially with the use of two genetic markers in odonates. The acquisition of an additional marker is not necessarily cost-intensive, but can become a *conditio sine qua non* for many closely related species. A database containing reliable DNA barcodes of as many species as possible highly enhances the discovery of yet unknown species or speciation processes and can be of priceless value for fast biodiversity assessment. It is also clear from this study that diagnostic characters for geographical clusters of specimens are valuable ”flags” for long-time monitoring, speciation studies, conservation management and identification of larval stages.

2.2.6 Acknowledgement

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

2.2.7 Author Contributions

HH and BS initiated and designed the project. TB and JR performed the research. TB and RD wrote and performed the bioinformatics. SD provided additional barcode information. TB, JR, RD and HH wrote the manuscript.

2.2.8 References

Avise JC (2004) Molecular markers, natural history, and evolution, 2nd edition Sinauer Associates, Sunderland, Massachusetts.

Baker RH, DeSalle R (1997) Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst Biol 46, 654-673.

Behura SK (2007) Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. Mol Biol Evol 24, 1492-1505.

Bergmann T, Hadrys H, Breves G, Schierwater B (2009) Character-based DNA barcoding: a superior tool for species classification. Berliner und Münchener Tierärztliche Wochenschrift 122, 446-450.

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

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

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

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

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

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

Dawnay N, Ogden R, McEwing R, Carvalho GR, Thorpe RS (2007) Validation of the barcoding gene COI for use in forensic genetic species identification. Forensic Sci Int.

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.

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

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

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

Ekrem T, Willassen E, Stur E (2007) A comprehensive DNA sequence library is essential for identification with DNA barcodes. Mol Phylogenet Evol 43, 530-542.

Foley DH, Wilkerson RC, Cooper RD, Volovsek ME, Bryan JH (2007) A molecular phylogeny of *Anopheles annulipes* (Diptera: Culicidae) *sensu lato*: The most species-rich anopheline complex. Mol Phylogenet Evol 43, 283-297.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol 3, 294-299.

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

Goldstein PZ, DeSalle R, Amato G, Vogler AP (2000) Conservation genetics at the species boundary. Conserv. Biol. 14, 120-131.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27, 221-224.

Groeneveld LF, Clausnitzer V, Hadrys H (2007) Convergent evolution of gigantism in damselflies of Africa and South America? Evidence from nuclear and mitochondrial sequence data. Mol Phylogenet Evol 42, 339-346.

Hadrys H, Balick M, Schierwater B (1992) Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. Mol Ecol 1, 55-63.

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

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

Hajibabaei M, deWaard JR, Ivanova NV, *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. Philos Trans R Soc Lond B Biol Sci 360, 1959-1967.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PD (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proc Natl Acad Sci U S A 103, 968-971.

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003a) Biological identifications through DNA barcodes. Proc Biol Sci 270, 313-321.

Hebert PD, Ratnasingham S, deWaard JR (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc Biol Sci 270 Suppl 1, S96-99.

Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc Natl Acad Sci U S A 101, 14812-14817.

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

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

Hurst GD, Jiggins FM (2005) Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc Biol Sci 272, 1525-1534.

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

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

Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5, 150-163.

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

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

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

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

Monaghan MT, Balke M, Gregory TR, Vogler AP (2005) DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Philos Trans R Soc Lond B Biol Sci 360, 1925-1933.

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

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

Rach J, DeSalle R, Sarkar IN, Schierwater B, Hadrys H (2008) Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. 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.

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

Remigio EA, Hebert PD (2003) Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. Mol Phylogenet Evol 29, 641-647.

Rubinoff D (2006) DNA barcoding evolves into the familiar. Conserv Biol 20, 1548-1549.

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J Hered 97, 581-594.

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

Sarkar IN, Planet PJ, Bael TE, *et al.* (2002a) Characteristic attributes in cancer microarrays. J Biomed Inform 35, 111-122.

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

Sarkar IN, Thornton JW, Planet PJ, *et al.* (2002b) An automated phylogenetic key for classifying homeoboxes. Mol Phylogenet Evol 24, 388-399.

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

Shearer TL, Van Oppen MJ, Romano SL, Worheide G (2002) Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). Mol Ecol 11, 2475-2487.

Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PD (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proc Natl Acad Sci U S A 103, 3657-3662.

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

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

Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. Philos Trans R Soc Lond B Biol Sci 360, 1859-1868.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005) DNA barcoding Australia's fish species. Philos Trans R Soc Lond B Biol Sci 360, 1847-1857.

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

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

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

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

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

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

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

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

2.3 Some ”ant”swers: Application of a layered barcode approach to problems in ant taxonomy

MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12395

Authors: Omid Paknia*, Tjard Bergmann* and Heike Hadrys

Omid Paknia: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; Email address: omid.paknia@ecolevol.de

Tjard Bergmann: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; Email address: tjard.bergmann@ecolevol.de

Heike Hadrys: Stiftung Tierärztliche Hochschule Hannover, Institut für Tierökologie und Zellbiologie, Hannover, Germany; American Museum of Natural History, Sackler Institute for Comparative Genomics, New York, NY 10024, U.S.A.; Email address: heike.hadrys@ecolevol.de

⋆ = These authors contributed equally to this work.

Keywords: ants, character-based barcoding, cryptic biodiversity, distance-based barcoding, layered barcode approach, taxonomic entities

2.3.1 Abstract

DNA barcoding has emerged as a routine tool in modern taxonomy. Although straightforward, this approach faces new challenges, when applied to difficult situation such as defining cryptic biodiversity. Ants are prime examples for high degrees of cryptic biodiversity due to complex population differentiation, hybridization and speciation processes. Here, we test the DNA barcoding region, cytochrome c oxidase 1 and two supplementary markers, 28S ribosomal DNA and long-wavelength rhodopsin, commonly used in ant taxonomy, for their potential in a layered, character-based barcoding approach across different taxonomic levels. Furthermore, we assess performance of the character-based barcoding approach to determine cryptic species diversity in ants. We found (i) that the barcode potential of a specific genetic marker varied widely among taxonomic levels in ants; (ii) that application of a layered, character-based barcode for identification of specimens can be a solution to taxonomical challenging groups; (iii) that the character-based barcoding approach allows us to differentiate specimens even within locations based on pure characters. In summary, (layered) character-based barcoding offers a reliable alternative for problematic species identification in ants and can be used as a fast and cost-efficient approach to estimate presence, absence or frequency of cryptic species.

2.3.2 Introduction

The original idea amplifying, sequencing and analyzing of one universal gene fragment throughout the animal kingdom although straightforward has brought new challenges to taxonomists. Two challenges stand out in particular: (i) a reliable molecular marker; and (ii) a reliable approach for data analyses. The use of the 658-bp long Folmer region in the CO1 (cytochrome c oxidase 1) gene as a tool for specimen identification in animals, termed DNA barcoding (Hebert *et al.* 2003), has evolved into a routine approach in modern taxonomy. Although many studies have successfully shown that this region is reliable for accurate species barcoding and identification (e.g. Hebert *et al.* 2003; Smith *et al.* 2005), some studies suggest that the application of CO1 does not supply sufficient resolution and could be misleading (e.g. Elias *et al.* 2007; Jansen *et al.* 2009). Consequently, additional gene regions have been suggested as valuable markers to improve species delimitation and identification (e.g. Bergmann *et al.* 2013; Damm *et al.* 2010; DeSalle *et al.* 2005).

In contrast to the use of a specific molecular marker(s), there is no standard method of DNA barcoding analysis (but see Ratnasingham & Hebert 2013). The majority of published studies perform distance-based approaches, for example a neighbour joining (NJ) algorithm, converting DNA sequences into genetic distances (Casiraghi *et al.* 2010). Queries are considered successfully identified when they cluster with conspecific barcode sequences. However, the lack of an appropriate and universal ”threshold of genetic divergence” to assign unknown samples to new or described species remains the main challenge (Collins & Cruickshank 2013; DeSalle *et al.* 2005; Kekkonen & Hebert 2014; Meier *et al.* 2006).

The character-based approach has been first suggested by DeSalle *et al.* (2005) as an alternative to the distance-based approach for DNA barcoding. Character-based DNA barcoding uses the nucleotide variation in each position across DNA regions as diagnostic characters. As a result, formerly founded taxonomic groups are identified through the presence of diagnostic characters or combinations of characters within short strands of DNA (Bergmann *et al.* 2013; Rach *et al.* 2008; Reid *et al.* 2011). Consequently, the character-based DNA barcoding method aims not only to overcome the lack of barcode resolution and universal threshold issues but also be a solution to a greater challenge: while the classical taxonomic studies are character based, employing a similar approach for DNA sequences, makes the combination of classic morphological and DNA-based characters feasible. In other words, DNA characters extracted by this approach can be combined with characters from other disciplines, for example morphology, ecology and geography (e.g. Damm *et al.* 2010), within an integrative taxonomy scheme (DeSalle *et al.* 2005; Schlick-Steiner *et al.* 2009).

In this study – using ants for the first time – we investigate the potential of character-based DNA barcoding as a tool in critical specimen identification at different taxonomic levels. Ants (Hymenoptera: Formicidae) represent a prominent species rich (approx. 13000 described species) insect family. Standing among very few eusocial groups of insects, ants play a ground role in providing ecological services in many terrestrial ecosystems. Despite simplified morphological structure in workers caste, ants pose serious challenges for traditional taxonomy due to high or/and complex intraspecific morphological variations (Blaimer 2012; Ross *et al.* 2010).

Although, there are some successful examples of using distance-based barcodes in ants (e.g. Saux *et al.* 2004; Smith *et al.* 2005), cryptic biodiversity remains a major challenge for their alpha-taxonomy, ecology and conservation (Seifert 2009). Some genera represent hyperdiversity, which makes the identification of their members more challenging (Moreau 2008). For example, of the 77 described *Cardiocondyla* species worldwide the frequency of potential cryptic species has been estimated to be as high as 52% (Seifert 2009). Identifying these potential cryptic species, distance-based DNA barcoding studies using one universal CO1 marker have not shown promising results yet (e.g. Knaden *et al.* 2012; Ueda *et al.* 2012). Integrative taxonomy and the use of more than one genetic marker have been proposed as a viable solution to facilitate reliable identification of ants (Schlick-Steiner *et al.* 2009; Seifert 2009). However, this is a difficult task as the integration of genetic distances into a character-based matrix of morphological, ecological, and geographic characters means to unite two different types of data.

Distance-based and character-based approaches have been directly compared by others (Wong *et al.* 2009; Yassin *et al.* 2010) and us (Bergmann *et al.* 2013; Rach *et al.* 2008; Reid *et al.* 2011) before multiple times. Here, we focus and explore the performance of three DNA markers for character-based barcoding analyses at different levels in ant taxonomy – subfamily, genus and species (cryptic species) level – and introduce the concept of a layered, character-based barcode approach. In theory, by combined analyses of genes with different mutation rates a more refined barcode featuring taxa specific characters at different taxonomic levels could be generated. In a step-by-step layered approach one molecular marker would be used to identify one taxon (e.g. subfamily, genera), while a second or third marker could be consulted to identify deeper taxonomic levels (e.g. species, population).

It has already been shown in earlier studies, that the character-based method is sensitive enough to cluster specimens within a species according to geographical origin. Such clusters could then be tested for diagnostic characters. Absence or presence of diagnostics could be used as markers for potential new species (Bergmann *et al.* 2013). This process of investigating cryptic species by clustering populations according to their geographical origin is called ”flagging”. Flagging refers to the process of designating populations as potential species worthy of future ecological, behavioral and morphological work to determine species existence (Goldstein &

DeSalle 2011). This practice is important especially in some ant genera where potential species – although conservative in their morphological variations – have gone through a rapid radiation (adaptation) in their ecology and behavior (e.g. Andersen 2007). Our target is the diverse Australian *Monomorium rothsteini* complex, which has been suggested to be a group of 'many species' (Greenslade 1979). Andersen (2007) defines it as including up to 50 or more species. A recent integrative study on this species across the Australian continent shows that various *M. rothsteini* lineages can be to some extent identified by using combinations of morphological and molecular characters (Sparks *et al.* 2014). However, a considerable number of lineages could not been diagnosed due to lack of genetic support.

Using a character-based barcoding approach, we address two questions:

1. Do DNA markers perform equally across taxonomic levels? If so, then it is not important which marker is used for each taxonomic level. If the answer is no, then the markers should be used in a correct order across taxonomic levels that could deliver the best resolution for the given taxonomic level.

2. Can ant specimens within a cryptic species complex be designated to their geographical origin to test for diagnosis of potential new species?

We will approach the first question by comparing diagnostic characters of the DNA barcoding marker CO1 (cytochrome c oxidase 1) and two supplementary markers, 28S (28S rDNA) and LWR (long-wavelength rhodopsin), on subfamily, genera and species level. By "flagging" the cryptic *Monomorium rothsteini* complex in the second part of our study, we will address question two.

2.3.3 Material & Methods

Part 1: Data mining ant (Formicid) CO1, 28S rDNA and LWR

Data used in the first part of this study has been mined from GenBank and the Barcode of Life Data System (BOLD). We selected in addition to the DNA barcode marker CO1 (cytochrome c oxidase 1), the two supplementary markers 28S rDNA and long-wavelength rhodopsin (LWR). The latter two gene fragments have been used widely for taxonomic and phylogenetic studies in Hymenoptera. The 28S rDNA marker has been successfully used in recovering phylogenetic relationships among many higher taxonomic groups of Hymenoptera (e.g. Belshaw *et al.* 1998; Dowton & Austin 1998; Saux *et al.* 2004). The LWR gene fragment exhibits relatively high variability at the species level (e.g. Chaubet *et al.* 2013; Derocles *et al.* 2012; Lucky 2011; Lucky & Sarnat 2010).

In total, 1780 Formicid sequences belonging to 259 species, 21 genera, and four subfamilies were retrieved. The 1780 sequences contains 1097 CO1, 397 28s rDNA, and 286 LWR sequences (see supplementary data 1). The number of CO1 sequences, however, decreased to 363, as there were large numbers of identical sequences per species. From these pools all 363 CO1, 397 28S rDNA and 286 LWR sequences were aligned with CLUSTAL W alignment algorithm (Thompson *et al.* 1994), using the default parameters implemented in MEGA 5 software package (Tamura *et al.* 2011) (see supplementary data 2).

To assess and compare the barcoding potential of all three genes in an equal manner, we restricted our analyses to only those taxa that had all three genes available. This step reduced the number from 1046 sequences to 377 sequences (see supplementary data 3). After cropping 5’ and 3’ ends of the alignment to blunt ends and sorting out duplicate sequences the number was further reduced to a final number of 322 sequences (see supplementary data 4). This resulted in the application of 322 sequences (115 CO1; 77 28S rDNA; 130 LWR) belonging to 115 species, three genera (*Camponotus*, *Myrmica*, and *Stenamma*), and two subfamilies (Formicinae and Myrmicinae).

We compared the potential of CO1, 28S rDNA and LWR for assigning the specimens to the subfamily, genera and species levels. On the subfamily level 39 species belonging to Formicinae were compared to 50 species belonging to Myrmicinae. On the genera level 50 species belonging to *Myrmica* were compared to 26 species belonging to *Stenamma*. On the species level 39 *Camponotus* and 50 *Myrmica* species were compared for each gene fragment. Ideally, for the higher taxonomic levels, subfamily and genus, we had to compare closely related taxa. Such data, however, was not available for these three gene markers in BOLD and GenBank.

For all analyses: (i) the subfamilies Formicinae vs. Myrmicinae; (ii) the genera *Myrmica* vs. *Stenamma*; and (iii) the species within the genera *Camponotus* and *Myr-*

mica, we only used sequences of species that were available for all three genes. The sequences were cropped on the 5'prime and the 3'prime end leaving a well-aligned region of 623 bp for CO1, 429 bp for 28S rDNA and 488 bp for Rhodopsin on subfamily level (supplementary data 5). Alignments of 615 bp for CO1, 439 bp for 28S rDNA and 491 bp for Rhodopsin remained on genera level (supplementary data 6). On species level we compared specimen within the genera *Camponotus* and *Myrmica*. Here, the alignments for *Camponotus* were 677/1630/562 bp (CO1/28S/LWR) long. For *Myrmica* the alignments length was 609/420/482 bp (CO1/28S/LWR) long (see Table 2.6 for an overview).

Tab. 2.6.: Shown is the number of relevant barcode positions in each gene region that discriminates at the subfamily, genus and species levels; the length of each alignment used for creating character-based barcodes; and the quality of barcode relevant information within each barcode fragment. On the species level, the CO1 barcode region has significant more characteristic attributes compared to 28S rDNA (28S) and long-wavelength rhodopsin (LWR)

Taxon level	Taxon name	28S	CO1	Rhodopsin
Overview: Number of Barcode Positions				
Subfamily	Camponotus vs Myrmica	31	9	27
Genera	Myrmica vs Stenamma	25	7	30
Species	*Camponotus*	63	307	154
Species	*Myrmica*	88	238	66
Overview: Length of alignments				
Subfamily	Camponotus vs Myrmica	429	623	487
Genera	Myrmica vs Stenamma	439	615	491
Species	*Camponotus*	1630	677	562
Species	*Myrmica*	420	609	482
Overview: Barcode quality (100/alignment * Barcode characters)				
Subfamily	Camponotus vs Myrmica	7%	1%	6%
Genera	Myrmica vs Stenamma	6%	1%	6%
Species	*Camponotus*	4%	45%	27%
Species	*Myrmica*	21%	39%	14%

At the species level, *Camponotus* and *Myrmica* specimens were delineated to species according to the species names given on BOLD in GenBank. When possible this *a priori* naming of species was proofed by publications that these specimens were coming from Saux *et al.* (2004) and Jansen *et al.* (2010). Some of *Camponotus* sequences have been derived from Schluns *et al.* (unpublished data) and left us to not be able to prove the final delineation of those specimens.

Part 1: Character-based barcoding of ant (Formicid) CO1, 28S and LWR

The twelve alignments were barcoded by the application of the CAOS workbench (http://bol.uvm.edu/CAOS-workbench/) resulting in twelve character-based barcode matrices (see supplementary data 7). In short, the aligned sequences of all three genes (CO1, 28S rDNA and LWR) were converted into a nexus file format with SeaView version 4 (Gouy *et al.* 2010). We created a maximum-likelihood tree for each data set using RAxML with 100 bootstraps (Stamatakis 2014). The resulting twelve trees were used as a guide for CAOS (Character Attribute Organization System) analyses by saving each tree and the corresponding sequences as nexus file using MacClade 4 v. 4.06 (Maddison & Maddison 2000) and processed it with the CAOS programs (http://bol.uvm.edu/CAOS-workbench/). To extract the CAs, the twelve matrices were processed with the CAOS-Analyzer. The CAOS-Analyzer extracts CAs unique for each branch at each branching event in the given tree. In a second step, the output data of the Analyzer was converted by the CAOS-Barcoder into character-based barcode matrices (Fig. 2.7A). Only simple pure characters were extracted from the CAOS-Barcoder output files (see supplementary data 7). The efficiency of each character matrix for assigning new queries to the correct group was tested with the CAOS-Classifier (Fig. 2.7B).

Part 2: Data mining *M. rothsteini*

For the second part, we retrieved the CO1 data from the recent published work on the *Monomorium rothsteini* complex (Sparks *et al.* 2014). With a diverse yet cryptic morphology, *M. rothsteini* complex has been represented as an example of the great challenge that exists in systematics of cryptic ants. *Monomorium rothsteini* members show overlaps in both morphological characters and distribution ranges, making their identification difficult. Using a distance-based barcoding approach, Spark *et al.* (2014) were able to identify 38 well-supported clades within the *M. rothsteini* complex. Of all the clades, clade 5a is containing the greatest number of individuals and haplotypes from multiple locations within Australia, however, could not be resolved by morphology or a distance-based barcoding approach, yet. For the purpose of this study, we focused on this clade. The sample set was created

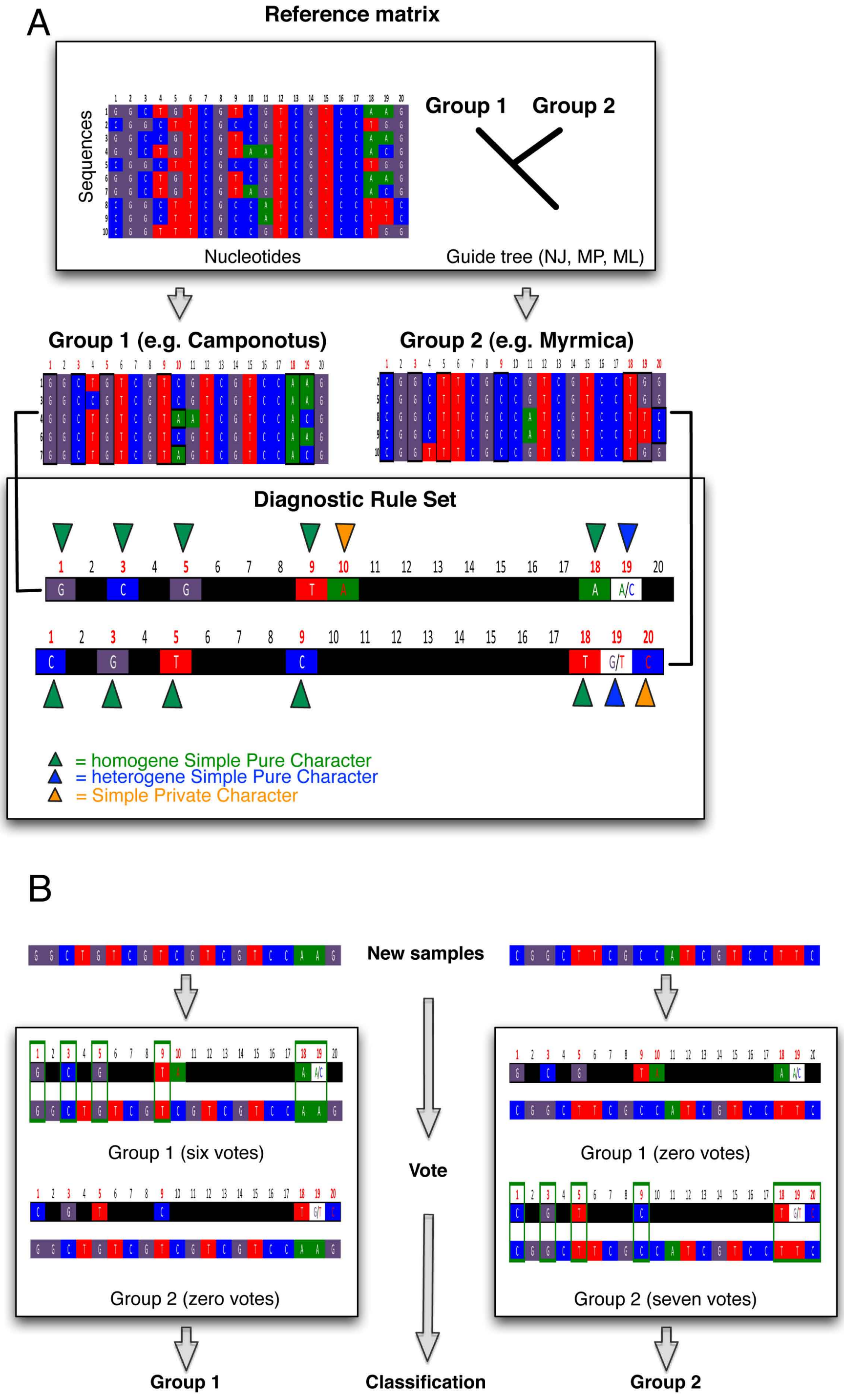


Fig. 2.7.: Character-based barcoding is a two-step process. **Panel A:** Finding characteristic attributes in barcode sequences. The barcode reference sequences are first grouped using a phylogenetic tree. Next, characteristic attributes unique to each group are determined by the CAOS-Analyzer and visualized by the CAOS-Barcoder. These characteristic attributes (CAs) form the basis for a set of diagnostic rules. Simple Pure Character attributes: DNA sequence attributes in these columns are purely diagnostic characters (sensu Davis & Nixon 1992). Simple Private Character attributes: DNA sequence attributes are not purely diagnostic, but rather the character in some individuals of one group are 'private' to that population. **Panel B:** Diagnostic rules can then be used to classify novel samples by a voting process (CAOS-Classifier) in which the new sample is placed in the group for which it has the highest vote total.

and described by Sparks *et al.* (2014) and comprise 42 CO1 sequences of different quality.

Part 2: Character-based barcoding of the *Monomorium rothsteini* complex

Because of the difference in quality of 42 CO1 sequences, we tested two approaches focusing on a) quality of sequence numbers (sequence number > similar length) or b) quantity of sequence length (similar & abundant length > sequence number). In the first approach we used all 42 sequences but reduced the sequence length to a shared length of 545 bp. In the second approach sequence length was prioritized and seven sequences were discarded. Leaving 35 sequences with a length of 934 bp. After trimming both sequence sets, as described before, we identified identical sequences within the data sets. Only one copy of identical sequences was entered in the data set. For the quality approach 20 sequences remained. For the quantity approach 25 sequences remained. As the number of sequences and their length for the quantity approach was superior to the quality approach we continued further analysis focusing on the quantity data set alone. Using the remaining 25 sequences we again tested two approaches: a) sequence similarity, and b) sequence origin. In the first approach we created barcodes focusing on sequence similarity. In the second approach we focused on pooling sequences together based on sequence origin (locality).

All doublet sequences could be pinpointed to a single region. From the ten discarded doublets five came from the same locality, the other five from neighbouring locations.

2.3.4 Results

Part 1: Character-based barcoding of ant (Formicid) CO1, 28S and LWR

As expected, 28S rDNA was the most efficient marker at the subfamily level, while LWR (long-wavelength rhodopsin) performed best at the genus level. For the 90/76 specimens tested (subfamily/genus level), the 28S rDNA region provided in total 31 characters on subfamily - and 25 characters on genus levels. The number of characters for the LWR barcode region was 27 on subfamily - and 30 characters on the genus level (see Fig. 2.8). The 28S rDNA and LWR barcode regions were efficient in clearly discerning the ant groups on subfamily and genus level and used in combination proved to be information rich regions for identifying ant taxa above the species level.

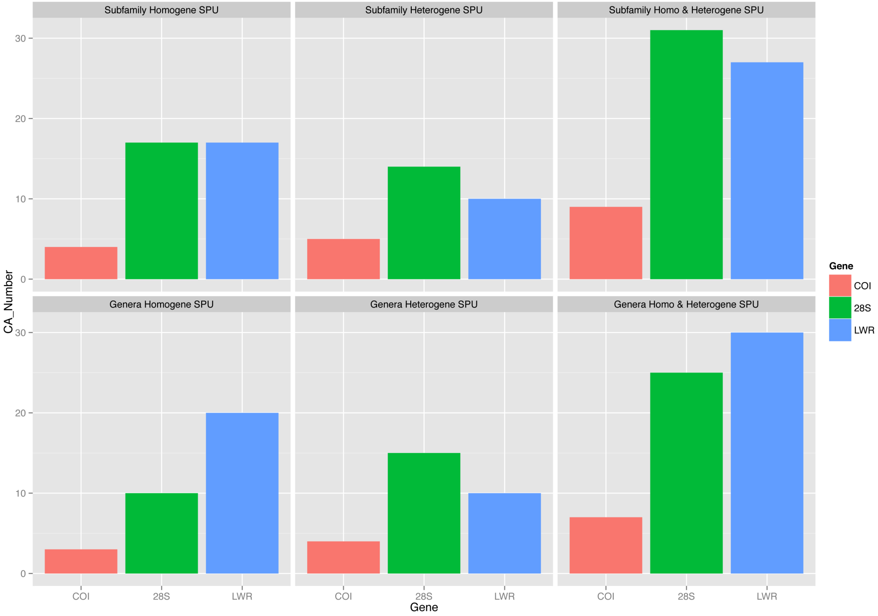
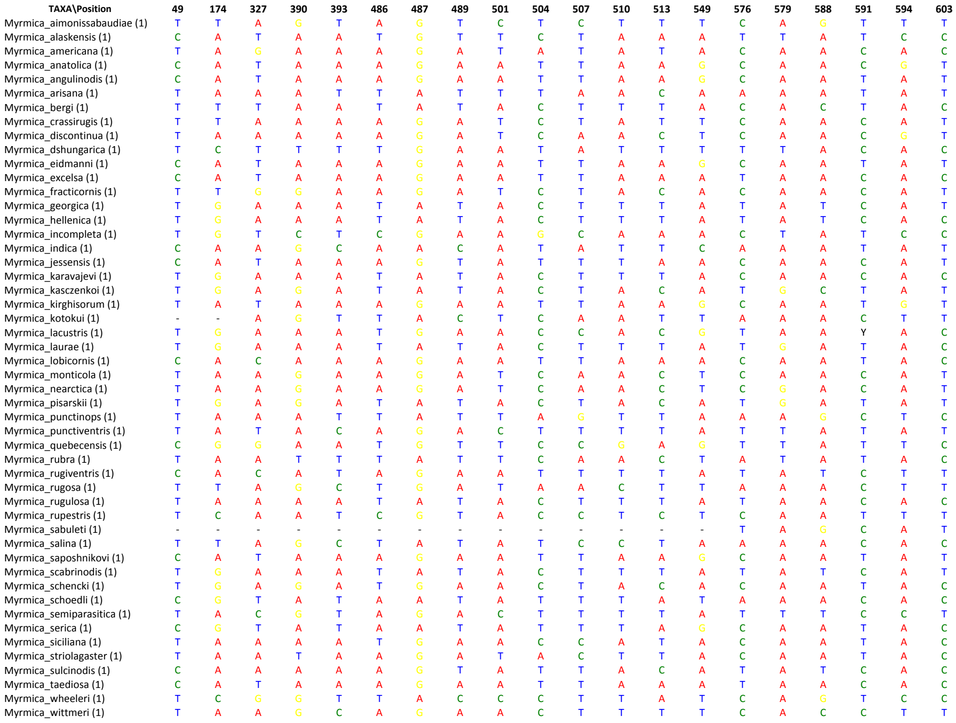


Fig. 2.8.: Character attributes overview. Shown are the numbers of characteristic attributes (CAs) plotted against the tested genes. The 28S rDNA and LWR barcode regions have more CAs on the subfamily and genus level than CO1.

In contrast, the CO1 (cytochrome c oxidase 1) barcode region provided only a small number (9 and 7 respectively) of characters above species levels (see Table 2.6). However, 307 CAs - 45% of the 677 bp long CO1 gene fragment - are informative for species identification within the *Camponotus* genus and 39% of the 609 bp long CO1 barcode region (238 CAs) could be used for differentiating species within the *Myrmica* genus (see Table 2.6 and Fig. 2.9 for details). Here, 28S rDNA and LWR offered considerably less characters for discerning species within the tested genera. For 28S rDNA only 4% (63 CAs in the 1630 bp alignment) and 21% (88 CA in the 420 bp alignment) could be used for discriminating between species within the *Camponotus* and *Myrmica* genera. For LWR only 27% (154 CAs in 562 bp alignment) and 14% (66 CAs in 482 bp alignment) could be used for discerning species within the *Camponotus* and *Myrmica*.



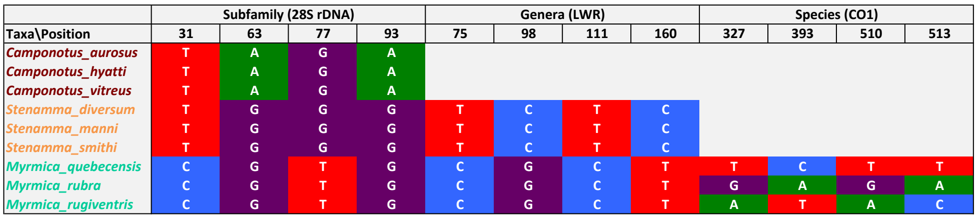
TAXA\Position	49	174	327	390	393	486	487	489	501	504	507	510	513	549	576	579	588	591	594	603
Myrmica_aimonissabaudiae (1)	T	T	A	G	T	A	G	T	C	T	C	T	T	T	C	A	G	T	T	T
Myrmica_alaskensis (1)	C	A	T	A	A	T	G	T	T	C	T	A	A	A	T	T	A	T	C	C
Myrmica_americana (1)	T	A	G	A	A	A	G	A	T	A	T	A	T	A	C	A	A	C	A	C
Myrmica_anatolica (1)	C	A	T	A	A	A	G	A	A	T	T	A	A	G	C	A	A	C	G	T
Myrmica_angulinodis (1)	C	A	T	A	A	A	G	A	A	T	T	A	A	G	C	A	A	T	A	T
Myrmica_arisana (1)	T	A	A	A	T	T	A	T	T	T	A	A	C	A	A	A	A	T	A	T
Myrmica_bergi (1)	T	T	T	A	A	T	A	T	A	C	T	T	T	A	C	A	C	T	A	C
Myrmica_crassirugis (1)	T	T	A	A	A	A	G	A	T	C	T	A	T	T	C	A	A	C	A	T
Myrmica_discontinua (1)	T	A	A	A	A	A	G	A	T	C	A	A	C	T	C	A	A	C	G	T
Myrmica_dshungarica (1)	T	C	T	T	T	T	G	A	A	T	A	T	T	T	T	T	A	C	A	C
Myrmica_eidmanni (1)	C	A	T	A	A	A	G	A	A	T	T	A	A	G	C	A	A	T	A	T
Myrmica_excelsa (1)	C	A	T	A	A	A	G	A	A	T	T	A	A	A	T	A	A	C	A	C
Myrmica_fracticornis (1)	T	T	G	G	A	A	G	A	T	C	T	A	C	A	C	A	A	C	A	T
Myrmica_georgica (1)	T	G	A	A	A	T	A	T	A	C	T	T	T	A	T	A	T	C	A	T
Myrmica_hellenica (1)	T	G	A	A	A	T	A	T	A	C	T	T	T	A	T	A	T	C	A	C
Myrmica_incompleta (1)	T	G	T	C	T	C	G	A	A	G	C	A	A	A	C	T	A	T	C	C
Myrmica_indica (1)	C	A	A	G	C	A	A	C	A	T	A	T	T	C	A	A	A	T	A	T
Myrmica_jessensis (1)	C	A	T	A	A	A	G	T	A	T	T	T	A	A	C	A	A	C	A	T
Myrmica_karavajevi (1)	T	G	A	A	A	T	A	T	A	C	T	T	T	A	C	A	A	C	A	C
Myrmica_kasczenkoi (1)	T	G	A	G	A	T	A	T	A	C	T	A	C	A	T	G	C	T	A	T
Myrmica_kirghisorum (1)	T	A	T	A	A	A	G	A	A	T	T	A	A	G	C	A	A	T	G	T
Myrmica_kotokui (1)	-	-	A	G	T	T	A	C	T	C	A	A	T	T	A	A	A	C	T	T
Myrmica_lacustris (1)	T	G	A	A	A	T	G	A	A	C	C	A	C	G	T	A	A	Y	A	C
Myrmica_laurae (1)	T	G	A	A	A	T	A	T	A	C	T	T	T	A	T	G	A	T	A	C
Myrmica_lobicornis (1)	C	A	C	A	A	A	G	A	A	T	T	A	A	A	C	A	A	T	A	C
Myrmica_monticola (1)	T	A	A	G	A	A	G	A	T	C	A	A	C	T	C	A	A	C	A	T
Myrmica_nearctica (1)	T	A	A	G	A	A	G	A	T	C	A	A	C	T	C	G	A	C	A	T
Myrmica_pisarskii (1)	T	G	A	G	A	T	A	T	A	C	T	A	C	A	T	G	A	T	A	T
Myrmica_punctinops (1)	T	A	A	A	T	T	A	T	T	A	G	T	T	A	A	A	G	C	T	C
Myrmica_punctiventris (1)	T	A	T	A	C	A	G	A	C	T	T	T	T	A	T	T	A	T	A	T
Myrmica_quebecensis (1)	C	G	G	A	A	T	G	T	T	C	C	G	A	G	T	T	A	T	T	C
Myrmica_rubra (1)	T	A	A	T	T	T	A	T	T	C	A	A	C	T	A	T	A	T	A	C
Myrmica_rugiventris (1)	C	A	C	A	T	A	G	A	A	T	T	T	T	A	T	A	T	C	T	T
Myrmica_rugosa (1)	T	T	A	G	C	T	G	A	T	A	A	C	T	T	A	A	A	C	T	T
Myrmica_rugulosa (1)	T	A	A	A	A	T	A	T	A	C	T	T	T	A	T	A	A	C	A	C
Myrmica_rupestris (1)	T	C	A	A	T	C	G	T	A	C	C	T	C	T	C	A	A	T	T	T
Myrmica_sabuleti (1)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	A	G	C	A	T
Myrmica_salina (1)	T	T	A	G	C	T	A	T	A	T	C	C	T	A	A	A	A	C	A	C
Myrmica_saposhnikovi (1)	C	A	T	A	A	A	G	A	A	T	T	A	A	G	C	A	A	T	A	T
Myrmica_scabrinodis (1)	T	G	A	A	A	T	A	T	A	C	T	T	T	A	T	A	T	C	A	T
Myrmica_schencki (1)	T	G	A	G	A	T	G	A	A	C	T	A	C	A	C	A	A	T	A	C
Myrmica_schoedli (1)	C	G	T	A	T	A	A	T	A	T	T	T	A	T	A	A	A	C	A	C
Myrmica_semiparasitica (1)	T	A	C	G	T	A	G	A	C	T	T	T	T	A	T	T	T	C	C	T
Myrmica_serica (1)	C	G	T	A	T	A	A	T	A	T	T	T	A	G	C	A	A	T	A	C
Myrmica_siciliana (1)	T	A	A	A	A	T	G	A	A	C	C	A	T	A	C	A	A	T	A	C
Myrmica_striolagaster (1)	T	A	A	T	A	A	G	A	T	A	C	T	T	A	C	A	A	T	A	C
Myrmica_sulcinodis (1)	C	A	A	A	A	A	G	T	A	T	T	A	C	A	T	A	T	C	A	C
Myrmica_taediosa (1)	C	A	T	A	A	A	G	A	A	T	T	A	A	A	T	A	A	C	A	C
Myrmica_wheeleri (1)	T	C	G	G	T	T	A	C	C	C	T	T	A	T	C	A	G	T	C	C
Myrmica_wittmeri (1)	T	A	A	G	C	A	G	A	A	C	T	T	T	T	C	A	C	C	T	T

Fig. 2.9.: Character-based DNA barcodes of 50 ant species based on CO1 sequences; a subset of unique combinations of character states at 20 nucleotide positions for each species is shown; '-' show missing data.

The combination of diagnostic characters from all three genes in hierarchical order (set by taxonomical resolution of the markers) result in a layered barcode (see Figure 2.10). This layered barcode, which is homolog to field guides, should in theory better resolve query sequences to the correct taxon, as all diagnostic characters are combined and used in the most efficient succession.

Part 2: Character-based barcoding of the *Monomorium rothsteini* complex

The character-based barcoding approach identifies diagnostic CAs (Character Attributes) by comparing aligned and unique specimen sequences. The nonsimilar



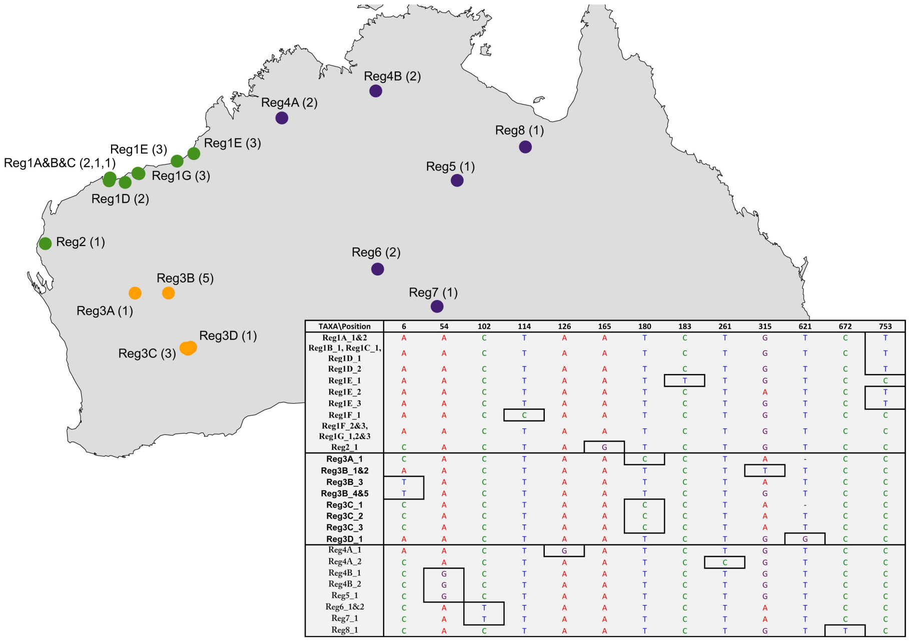
	Subfamily (28S rDNA)				Genera (LWR)				Species (CO1)			
Taxa\Position	31	63	77	93	75	98	111	160	327	393	510	513
Camponotus_aurosus	T	A	G	A								
Camponotus_hyatti	T	A	G	A								
Camponotus_vitreus	T	A	G	A								
Stenamma_diversum	T	G	G	G	T	C	T	C				
Stenamma_manni	T	G	G	G	T	C	T	C				
Stenamma_smithi	T	G	G	G	T	C	T	C				
Myrmica_quebecensis	C	G	T	G	C	G	C	T	T	C	T	T
Myrmica_rubra	C	G	T	G	C	G	C	T	G	A	G	A
Myrmica_rugiventris	C	G	T	G	C	G	C	T	A	T	A	C

Fig. 2.10.: Shown is an example of a layered barcode for three subfamilies (brown, orange and green), two genera (orange and green) and three species (green). On each taxon level, a different gene is used as a barcode marker. For comparing subfamilies, 28S rDNA diagnostics are used, for genera LWR (long-wavelength rhodopsin) and for species CO1 (cytochrome c oxidase) diagnostics. To keep the table transparent, we only choose four exemplary diagnostic characters per barcode region.

sequences are compared by clustering them using a guide tree as described in Figure 1A. Each branching point in the guide tree is a cluster of two groups (left & right branch) where characters diagnostics for each branch are extracted.

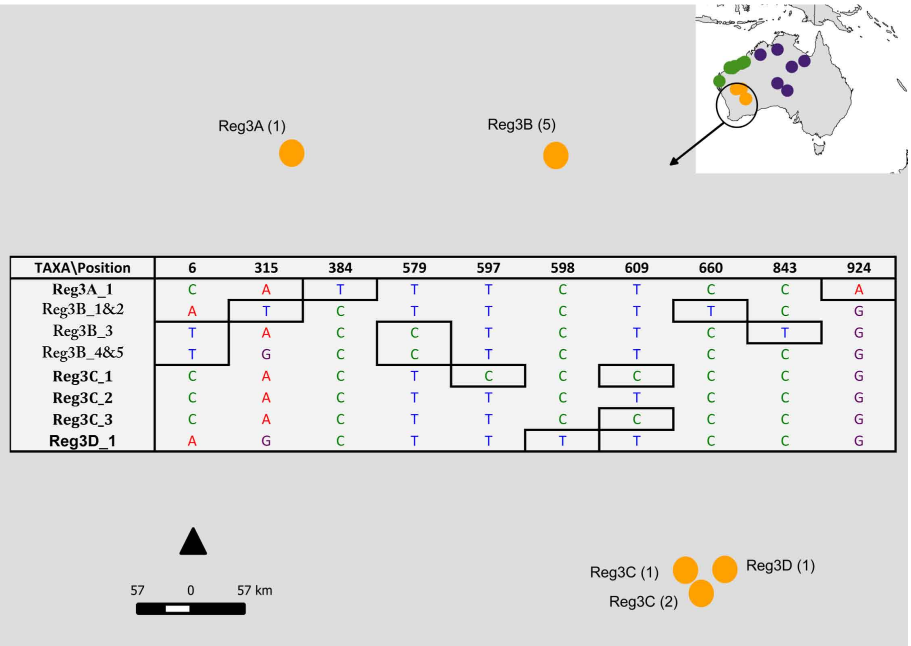
For both ”sequence similarity” and ”sequence origin” approaches, we found a similar number of character-based barcode regions (84 CAs; see supplementary data 8). We summed up all CAs within the 24 branching points in both trees. The resulting number of CAs (simple Pure (sPu) + simple Private (sPr)) was different for both setups (sequence similarity = 363 CAs; sequence origin = 289 CAs; Supplementary data 9). The number of sPu CAs was higher in the second approach (sequence similarity = 70; sequence origin = 87). On the other hand only four branching points showed no sPu in the sequence similarity approach while for the sequence origin approach six branching events exist without sPu.

Both approaches led to character-based barcodes that were able to fully resolve all 25 specimens by their unique characters (see supplementary data 10). Here, we represent only the results of the ”sequence origin” approach, showing that the 25 *Monomorium* specimen sequences tested featured location specific character attributes (CAs; Figs. 2.11 and 2.12). Only private CAs were determined while comparing the three larger regions (e.g. region one and two vs. three separated with 20 private CAs). Pure CAs were responsible for grouping specimen in deeper nodes; usually between individuals within localities. For example, ten specimens within region three were further divided into four groups based on private and pure characters (Fig. 2.12). First, specimens from clusters 3A and 3B were parted from the other two clusters (region 3C and 3D) based on 15 private characters. Then, specimens in clusters 3A and 3B were separated based on eleven pure and two private characters (see supplementary data 11 for a complete breakdown of location specific characters). In two occasions specimens from closely neighbored locations shared the same diagnostic characters. The specimens Reg1B_1, Reg1C_1 and Reg1D_1 for examples were collected from different locations, but shared the same CO1 sequence haplotype. In three occasions specimens sampled from the same location showed identic haplotypes, while on seven occasions multiple haplotypes could be identified. For instance, in region 3B (Fig. 2.12), five specimens were sampled leading to three haplotypes. Specimens one and two (Reg3B_1&2) shared the first haplotype. Specimen three had a unique haplotype (Reg3B_3). Specimens four and five shared another haplotype (Reg3B_4&5).



TAXA\Position	6	54	102	114	126	165	180	183	261	315	621	672	753
Reg1A_1&2	A	A	C	T	A	A	T	C	T	G	T	C	T
Reg1B_1, Reg1C_1, Reg1D_1	A	A	C	T	A	A	T	C	T	G	T	C	T
Reg1D_2	A	A	C	T	A	A	T	C	T	G	T	C	T
Reg1E_1	A	A	C	T	A	A	T	T	T	G	T	C	C
Reg1E_2	A	A	C	T	A	A	T	C	T	A	T	C	T
Reg1E_3	A	A	C	T	A	A	T	C	T	G	T	C	T
Reg1F_1	A	A	C	C	A	A	T	C	T	G	T	C	C
Reg1F_2&3, Reg1G_1,2&3	A	A	C	T	A	A	T	C	T	G	T	C	C
Reg2_1	C	A	C	T	A	G	T	C	T	G	T	C	C
Reg3A_1	C	A	C	T	A	A	C	C	T	A	-	C	C
Reg3B_1&2	A	A	C	T	A	A	T	C	T	T	T	C	C
Reg3B_3	T	A	C	T	A	A	T	C	T	A	T	C	C
Reg3B_4&5	T	A	C	T	A	A	T	C	T	G	T	C	C
Reg3C_1	C	A	C	T	A	A	C	C	T	A	-	C	C
Reg3C_2	C	A	C	T	A	A	C	C	T	A	T	C	C
Reg3C_3	C	A	C	T	A	A	C	C	T	A	T	C	C
Reg3D_1	A	A	C	T	A	A	T	C	T	G	G	C	C
Reg4A_1	A	A	C	T	G	A	T	C	T	G	T	C	C
Reg4A_2	C	A	C	T	A	A	T	C	C	G	T	C	C
Reg4B_1	C	G	C	T	A	A	T	C	T	G	T	C	C
Reg4B_2	C	G	C	T	A	A	T	C	T	G	T	C	C
Reg5_1	C	G	C	T	A	A	T	C	T	G	T	C	C
Reg6_1&2	C	A	T	T	A	A	T	C	T	A	T	C	C
Reg7_1	C	A	T	T	A	A	T	C	T	A	T	C	C
Reg8_1	C	A	C	T	A	A	T	C	T	G	T	T	C

Fig. 2.11.: Distribution of clade 5 of *M. rothsteini’s* complex in Australia. Map showing, 18 sampling locations (Reg1A-G, Reg2, Reg3A-D, Reg4A&B, Reg5, 6, 7, 8). Regions were divided into eight provisional regions (Reg1–8), which were merged into three location-based clusters (green, orange, blue). Although subjective, all these clustering have been based on geographical distance between sampling locations. Number of individuals within each sampling site is shown in parenthesis. Inset shows a small subset of CAs unique for each cluster (for more information see Appendix S12, Supporting information).



TAXA\Position	6	315	384	579	597	598	609	660	843	924
Reg3A_1	C	A	T	T	T	C	T	C	C	A
Reg3B_1&2	A	T	C	T	T	C	T	T	C	G
Reg3B_3	T	A	C	C	T	C	T	C	T	G
Reg3B_4&5	T	G	C	C	T	C	T	C	C	G
Reg3C_1	C	A	C	T	C	C	C	C	C	G
Reg3C_2	C	A	C	T	T	C	T	C	C	G
Reg3C_3	C	A	C	T	T	C	C	C	C	G
Reg3D_1	A	G	C	T	T	T	T	C	C	G

Fig. 2.12.: Distribution of clade 5 of *M. rothsteini's* complex within region 3 (orange points) in Australia. In this region, eight location-specific barcodes (shown in the table) have been divided to four geographical clusters (A, B, C and D). We assigned ten ant specimens to these four geographical clusters using CAOS barcoding. A small fraction of CAs responsible for the assignment has been represented in the table. For cluster 3B, five specimens were sequenced leading to three barcodes.

2.3.5 Discussion

Part 1: Character-based barcoding of ant (Formicid) CO1, 28S and LWR

In this study, we explored two questions with respect to potential problems in ant taxonomy. The first question was "whether DNA markers perform equally across taxonomic levels" using character-based barcoding. We compared performance of the CO1 DNA Barcoding region and two supplementary markers for identification of ant taxonomic entities across three taxonomic levels. Our findings suggest that in ants DNA markers do not perform equally across taxonomic levels. Our analysis of CO1, 28S rDNA and the LWR (long-wavelength rhodopsin) DNA region revealed that, while CO1 proved to be a good barcode marker at the species level, it offered only few CAs (Character Attributes) on higher taxonomic levels. In contrast, 28S rDNA and LWR showed both high numbers of CAs on subfamily and genus level while being short on CAs on species level. This result indicates that a high mutation ratio in a barcode region is useful in disentangling the affiliation of a query to its closest siblings. On the downside, the high frequency of changes introduces high numbers of homoplastic CAs on taxa above species level. In consequence, only few characters remain to determine subfamilies or genera. Using only CO1 as a marker can be problematic in cases where ants cannot be classified to correct higher taxonomic (e.g. group, tribe, or species group) levels by morphology only. For example, in *Cataglyphis* ants, the use of CO1 failed to separate species relationships in detail (Knaden *et al.* 2012). For this and similar cases, we propose a layered barcoding approach where 28S rDNA and LWR markers appear to provide additional resolution in terms of CAs. Although some ants are easy to identify up to the level of subfamily or genera, the layered approach could be used to place the sample to the correct higher taxon when morphological identification is difficult (Lapolla *et al.* 2011; Sosa-Calvo *et al.* 2013). The layered approach could also be used in cases where the specimens are not available or in bad conditions (e.g. stomach contents). Then CAs from CO1 could be used to affirm the lower taxonomic levels such as the species or populations. By using three rRNA markers along CO1 for the revision of Malagasy species of *Anochetus*, Fisher and Smith (2008) found that the CO1 data was the easiest to interpret, but the rRNA markers showed intra-individual variations, which was not present among CO1 sequences. In sum, layered barcodes combine the content of multiple genetic markers into a single key for specimen identification. This layered barcode should also aid in species discovery, as it will help placing a newly recovered barcode within the tree of life by the combined usage of markers. Currently, the character-based barcoding software processes only one marker at a time, creating single marker character-based barcodes that can only by hand be processed into layered barcodes. The next generation of the character-based barcoding software will contain a platform to create and process layered barcodes and ideally should be able to also accept other characters such as morphology, geography and ecology.

Part 2: Character-based barcoding of the *Monomorium rothsteini* complex

Second, we assessed the potential of the character-based barcode approach to determine cryptic species diversity in ants. An earlier study has been successfully using this method for discovery of cryptic species in odonates (Damm *et al.* 2010).

Ants in particular show extreme degrees of cryptic biodiversity due to, possible complex population differentiation, speciation processes and hybridization events (Seifert 1999; Ward 2007). *Monomorium rothsteini* species complex in Australia continent represents an extreme case of a large species complex, in which even integrative morphological and molecular approaches have been unable to fully separate existing lineages. To explore the capability of character-based barcoding in flagging diagnostics for potential new species in cryptic species complexes, we applied the technique to the unresolved clade 5a of the *M. rothsteini's* complex. Here, we have been able to demonstrate the genetic differences among lineages within this clade. Overall, the analyses show the potential existence of eight taxonomic entities defined by geographical distances within this clade. Sparks *et al.* (2014) found the clade 5a with most samples, and the broadest distribution be an assembly of problematic specimens that show overlapping of morphological characters. As suggested by Goldstein and DeSalle (2011), the data generated by DNA barcoding can reveal structures among clusters of sampled individuals and raise questions of whether such clusters represent discrete entities meriting formal description. Consequently, novel sequences can be "flagged" and be further studied to characterize a potential new species (Goldstein & DeSalle 2011). Applying character-based barcoding we were able to cluster the specimens first into three large geographical areas and then into total eight finer regions. For seven out of eight regions, we could even assign specimens to a specific location within a region. In other words, we have been able to find 43 diagnostic positions within the CO1 barcode for distinguishing 25 entities in clade 5a and flagged them for future detailed integrated taxonomic studies, including not just morphology but also ecology and behaviour and other possible information that could results in species descriptions for these diagnosable populations. Considering the high number of flagging populations in the clade 5a of Sparks *et al.* (2014) study, a reanalysis of all clades of this study with character-based barcoding might result in high numbers of flagged populations, suggesting that the potential existing taxonomic entities in this species complex is close to Andersen prediction (2007), up to fifty or more species. Although the case of *M. rothsteini* appears to be an extreme case, fairly similar cases are notable in other taxa and regions. In the Palearctic region, well-studied genera such as *Lasius*, *Cardiocondyla* and *Tetramorium* show high percentage (> 50%) of cryptic diversity

(Schlick-Steiner *et al.* 2006; Seifert 2009). The same is true for the *Solenopsis* genus in the Ecuadorian Andes (Delsinne *et al.* 2012; Sparks *et al.* 2014). Several recent studies have been unsuccessful to fully resolve the existing cryptic diversity by using the distance-based approach (e.g. Schlick-Steiner *et al.* 2006; Ueda *et al.* 2012). Here, particularly, ”distance threshold” and ”barcoding gap” stay as the Achilles' heel of this approach, where the gap between intra- and interspecific variations fades and no reliable distance threshold can be given (Čandek & Kuntner 2014; Meier *et al.* 2006). Notably, using genetic characters to not only identify the species, but also pinpoint the origin of a species is an interesting undertaking as there is no restriction for character-based barcoding to be applied to genetic markers.

Using genetic characters, only one character in theory should be enough as a means for identification given that it is unique to members of one group while the same CA is missing in a second, e.g. closely related group. We suggest that the number of specimens should be adapted to the richness of a taxonomical unit, for example fewer samples are needed if a taxon has a long generation cycle and few offsprings than for a group with short breeding intervals and many offsprings.

The ”quality” problem

The first observation, while creating reference matrices for all three genes by CAs, is that the quality of sequences that we extracted from NCBI and BOLD was ranging from very high to poor. Many sequences included a wide range of gaps (e.g. JN134308, EU525225) or consisted of only a small fragment (e.g. EU042010, EU439638) that we were not able to use within our libraries. Our observations confirm that it is very important to establish quality control routines through various filtering mechanisms in ever growing databases (Pompanon *et al.* 2005; Shen *et al.* 2013; Steinke & Hanner 2011; Vink *et al.* 2012). This step is important for establishing accurate libraries that will help rapid identification of specimens especially for conservational researches.

Conclusion

While cryptic diversity even in well studied arthropods challenges the current knowledge on biodiversity (e.g. Dincă *et al.* 2011), less studied taxa and regions pose greater challenges to taxonomists and ecologists. Current approaches of DNA barcoding, although applicable to some instances of cryptic diversity, often fail to provide reliable performance when it comes to complex situations (Schlick-Steiner *et al.* 2009). The frequent reports of cryptic diversity in ants (e.g. Csősz *et al.* 2014; Ross *et al.* 2010; Seifert 1999, 2009; Sparks *et al.* 2014; Steiner *et al.* 2011) and the failure of molecular markers and analytical approaches to resolve these findings suggest a consideration of other approaches for specimens identification or species discovery. Over the last decade, sophisticated morphological approaches have been used to disclose the unexpected cryptic diversity of many ant taxa (e.g. Bagherian Yazdi *et al.* 2012). Experts in ant morphology/taxonomy, however, can only perform this. The character-based barcode approach can offer a reliable method for precise species identification and flagging of cryptic species, with the crucial advantage of being analogous with traditional taxonomy in a wider context.

2.3.6 Acknowledgement

We thank Kathryn S. Sparks for making the data of *Monomorium rothsteini* available. We thank the editor Dr. Jeremy deWaard and the four anonymous reviewers for their comments and guidance.

2.3.7 Author Contributions

OP, TB and HH conceived the idea. OP and TB mined and analysed the data. OP and TB wrote the paper and HH contributed to the final version.

2.3.8 References

Andersen AN (2007) Ant diversity in arid Australia: a systematic overview. In: Advances in ant systematics (Hymenoptera: Formicidae): homage to E. O.

Wilson - 50 years of contributions. Memoirs of the American Entomological Institute, 80. (eds. Snelling RR, Fisher BL, Ward PS), pp. 91-51.

Bagherian Yazdi A, Muench W, Seifert B (2012) A first demonstration of interspecific hybridization in Myrmica ants by geometric morphometrics (Hymenoptera: Formicidae). Myrmecological News 17, 121-131.

Belshaw R, Fitton M, Herniou E, Gimeno C, Quicke DL (1998) A phylogenetic reconstruction of the Ichneumonoidea (Hymenoptera) based on the D2 variable region of 28S ribosomal RNA. Systematic Entomology 23, 109-123.

Bergmann T, Rach J, Damm S, *et al.* (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.

Blaimer BB (2012) Untangling complex morphological variation: taxonomic revision of the subgenus Crematogaster (Oxygyne) in Madagascar, with insight into the evolution and biogeography of this enigmatic ant clade (Hymenoptera: Formicidae). Systematic Entomology 37, 240-260.

Čandek K, Kuntner M (2014) DNA barcoding gap: Reliable species identification over morphological and geographical scales. Molecular Ecology Resources, DOI: 10.1111/1755-0998.12304.

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

Chaubet B, Derocles SAP, Hullé M, *et al.* (2013) Two new species of aphid parasitoids (Hymenoptera, Braconidae, Aphidiinae) from the high arctic (Spitsbergen, Svalbard). Zoologischer Anzeiger - A Journal of Comparative Zoology 252, 34-40.

Collins R, Cruickshank R (2013) The seven deadly sins of DNA barcoding. Molecular Ecology Resources 13, 969-975.

Csősz S, Seifert B, Müller B, *et al.* (2014) Cryptic diversity in the Mediterranean *Temnothorax lichtensteini* species complex (Hymenoptera:Formicidae). Organisms Diversity & Evolution 14, 75-88.

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

Davis JI, Nixon KC (1992) Populations, Genetic Variation, and the Delimitation of Phylogenetic Species. Systematic Biology 41, 421-435.

Delsinne T, Sonet G, Nagy ZT, *et al.* (2012) High species turnover of the ant genus Solenopsis (Hymenoptera: Formicidae) along an altitudinal gradient in the Ecuadorian Andes, indicated by a combined DNA sequencing and morphological approach. Invertebrate Systematics 26, 457-469.

Derocles SAP, Le Ralec A, Plantegenest M, *et al.* (2012) Identification of molecular markers for DNA barcoding in the Aphidiinae (Hym. Braconidae). Molecular Ecology Resources 12, 197-208.

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 1905-1916.

Dincă V, Lukhtanov VA, Talavera G, Vila R (2011) Unexpected layers of cryptic diversity in wood white Leptidea butterflies. Nat Commun 2, 324.

Dowton M, Austin AD (1998) Phylogenetic relationships among the microgastroid wasps (Hymenoptera: Braconidae): combined analysis of 16S and 28S rDNA genes and morphological data. Molecular Phylogenetics and Evolution 10, 354-366.

Elias M, Hill RI, Willmott KR, *et al.* (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proceedings of the Royal Society B-Biological Sciences 274, 2881-2889.

Fisher BL, Smith MA (2008) A Revision of Malagasy Species of *Anochetus* Mayr and *Odontomachus* Latreille (Hymenoptera: Formicidae). PLoS ONE 3, e1787.

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

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

Greenslade PJM (1979) A Guide to the Ants of South Australia South Australian Museum, Adelaide.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences 270, 313-321.

Jansen G, Savolainen R, Vepsäläinen K (2009) DNA barcoding as a heuristic tool for classifying undescribed Nearctic Myrmica ants (Hymenoptera: Formicidae). Zoologica Scripta 38, 527-536.

Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Molecular Ecology Resources 14, 706-715.

Knaden M, Tinaut A, Stokl J, Cerdà X, Wehner R (2012) Molecular phylogeny of the desert ant genus Cataglyphis (Hymenoptera: Formicidae). Myrmecological News 16, 123-132.

Lapolla JS, Kallal RJ, Brady SG (2011) A new ant genus from the Greater Antilles and Central America, Zatania (Hymenoptera: Formicidae), exemplifies the utility of male and molecular character systems. Systematic Entomology, 37, 200-214.

Lucky A (2011) Molecular phylogeny and biogeography of the spider ants, genus *Leptomyrmex Mayr* (Hymenoptera: Formicidae). Molecular Phylogenetics and Evolution 59, 281-292.

Lucky A, Sarnat EM (2010) Biogeography and diversification of the Pacific ant genus *Lordomyrma Emery*. Journal of Biogeography 37, 624-634.

Maddison DR, Maddison WP (2000) MacClade 4. Sinauer, Sunderland, Massachusetts.

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.

Moreau CS (2008) Unraveling the evolutionary history of the hyperdiverse ant genus *Pheidole* (Hymenoptera: Formicidae). Molecular Phylogenetics and Evolution 48, 224-239.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics 6, 847-859.

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

Ratnasingham S, Hebert PD (2013) A DNA-based registry for all animal species: The

Barcode Index Number (BIN) System. PLoS ONE 8, e66213.

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

Ross KG, Gotzek D, Ascunce MS, Shoemaker DD (2010) Species delimitation: A case study in a problematic ant taxon. Systematic Biology 59, 162-184.

Saux C, Fisher BL, Spicer GS (2004) Dracula ant phylogeny as inferred by nuclear 28S rDNA sequences and implications for ant systematics (Hymenoptera: Formicidae: Amblyoponinae). Molecular Phylogenetics and Evolution 33, 457-468.

Schlick-Steiner BC, Steiner FM, Moder K, *et al.* (2006) A multidisciplinary approach reveals cryptic diversity in Western Palearctic Tetramorium ants (Hymenoptera: Formicidae). Molecular Phylogenetics and Evolution 40, 259-273.

Schlick-Steiner BC, Steiner FM, Seifert B, *et al.* (2009) Integrative taxonomy: A multisource approach to exploring biodiversity. Annual Review of Entomology 55, 421-438.

Seifert B (1999) Interspecific hybridisations in natural populations of ants by example of a regional fauna (Hymenoptera, Formicidae). Insectes Sociaux 46, 45-52.

Seifert B (2009) Cryptic species in ants (Hymenoptera: Formicidae) revisited: we need a change in the alpha-taxonomic approach. Myrmecological News 12, 149-166.

Shen YY, Chen X, Murphy RW (2013) Assessing DNA barcoding as a tool for species identification and data quality control. PLoS ONE 8, e57125.

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: Biological Sciences 360, 1825-1834.

Sosa-Calvo J, Schultz TR, Brandão CRF, *et al.* (2013) *Cyatta abscondita*: taxonomy, evolution, and natural history of a new fungus-farming ant genus from Brazil. PLoS ONE 8, e80498.

Sparks KS, Andersen AN, Donnellan SC, Austin AD (2014) Navigating the mtDNA road map out of the morphological maze: interpreting morphological variation in the diverse *Monomorium rothsteini* (Forel) complex (Hymenoptera: Formicidae). Systematic Entomology 39, 264-278.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Steiner F, Seifert B, Grasso D, *et al.* (2011) Mixed colonies and hybridisation of Messor harvester ant species (Hymenoptera: Formicidae). Organisms Diversity & Evolution 11, 107-134.

Steinke D, Hanner R (2011) The FISH-BOL collaborators' protocol. Mitochondrial DNA 22 Suppl 1, 10-14.

Tamura K, Peterson D, Peterson N, *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28, 2731-2739.

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.

Ueda S, Nozawa T, Matsuzuki T, *et al.* (2012) Phylogeny and phylogeography of *Myrmica rubra* complex (Myrmicinae) in the Japanese Alps. Psyche: A Journal of Entomology 2012.

Vink CJ, Paquin P, Cruickshank RH (2012) Taxonomy and Irreproducible Biological Science. Bioscience 62, 451-452.

Ward PS (2007) Phylogeny, classification, and species-level taxonomy of ants (Hymenoptera: Formicidae). Zootaxa 1668, 549-563.

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.

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

General Discussion 3

3.1 Choosing the best marker

3.1.1 Criteria for a good marker

Two ingredients are necessary in order to reliably identify a specimen by a molecular approach. The first ingredient being the marker constitutes the question: What is a reliable marker? It should be cost and time efficient. Only few, ideally one primer pair should be sufficient to amplify the marker in all specimen of interest. Furthermore, the marker should be easily obtainable, without a risk of amplifying pseudogenes or multiple heterogene alleles. Another important quality is the markers ability to distinguish closely related taxa. Here, a fine balance is of utmost importance. On the one hand, the marker should include highly stable regions or else it is likely that the primers can only bind for a limited range of species. On the other hand, it should provide enough diagnostics to safely differentiate interspecific groups while being conserved when comparing intraspecific specimen.

3.1.2 Comparing markers

Hebert *et al.* (2003) described a good marker, the Folmer region, a 658 bp long fragment of CO1. CO1 is a mitochondrial gene, as such, it has only one haplotype, making sequencing much easier than when sequencing core genome marker. Multiple mitochondria exist in a single cell increasing the number of amplification templates and therefore improving PCR performance. CO1 is part of the respiratory chain and as such present in all animals. CO1's mutation ratio is high enough to distinguish closely related taxa while being conserved in conspecifics. The primer pair described by Hebert *et al.* (2003) can be applied to many animal taxa, although by now many derivations of Hebert's original primers exist being more suitable for problematic groups. Currently, 3.223.815 species level barcode records (194.023 Species/79.002 Interim Species) focussing on the CO1 Folmer Region are available on BOLD (state: 10.28.2018). While the Folmer region has been very successful in discriminating and identifying many species, not all applications of CO1 have been successful. Elias *et al.* (2007) and Jansen *et al.* (2009) stated that CO1 does not supply sufficient resolution and could be misleading. Additional gene regions have been suggested as valuable markers to improve species delimitation and identification (e.g. DeSalle *et al.* 2005; Damm *et al.* 2010). In all three publications presented here CO1 was used. Although the phyla studied in this publications were highly diverse (Testudines, Odonata, Formicidae) the Folmer region provided diagnostic information to most of the sister groups investigated. The Folmer region succeded as a marker especially when investigating taxa on species level and below. While CO1 performed well, it was not perfect and combining the marker with others (ND1 in Bergmann *et al.* 2013; 28S & LWR in Paknia *et al.* 2015) highly improved identification performance. In Bergmann *et al.* (2013), ND1 in addition to CO1 was used to overcome the molecular hurdle of investigating a species rich (~5.000), ancient (~325 million years) phylum. In previous studies of insect groups (Lin & Danforth 2004; Baker & DeSalle 1997; Hadrys *et al.* 2006; Dijkstra *et al.* 2007; Rach *et al.* 2008) it was shown that ND1 is a good marker for low level taxonomic identification. Using 271 Odonata individuals representing 51 species, 22 genera and 8 families, our study confirm that both CO1 and ND1 are suitable markers for taxonomic identification of odonates. The quality of a DNA barcoding marker depends on its availability (easy to sequence) and its discrimination power (intraspecific conserved and interspecific variable). Both attributes are embodied by CO1 and ND1. For both markers a single primer pair was sufficient in obtaining the sequence of most species. While mitochodrial-like sequences frequently occur, only few putative pseudogenes have been observed in our study. Although both markers are of mitochondrial origin, their substitution patterns within and between taxonomic groups differ substantially. When comparing sister groups and geographic clusters, both markers showed complementary density of diagnostic characters. I recommend using both markers when investigating taxonomically challenging insect groups.

For investigating the cryptic ant taxonomy, we choose 28S and LWR in addition to CO1. Hyperdiversity has been reported for some genera (Moreau 2008), one worldwide dispersed species for example (*Cardiocondyla*) has been estimated to include 52% cryptic species (Seifert 2009). In order to resolve such problematic groups we compared the markers' diagnostic potential on three different levels (subfamily, genus and species). The 28S rDNA marker has been successfully used in recovering phylogenetic relationships among many higher taxonomic groups of Hymenoptera (e.g. Belshaw *et al.* 1998; Dowton & Austin 1998; Saux *et al.* 2004). The LWR gene fragment exhibits relatively high variability at the species level (e.g. Lucky & Sarnat 2010; Lucky 2011; Derocles *et al.* 2012; Chaubet *et al.* 2013). 28S was most efficient at subfamily level. LWR performed best at the genus level. In combination, they provide sufficient diagnostics for identification of ant specimen above species level. CO1 provided only few diagnostics above species level, but was highly informative for species identification. High mutation ratio is beneficial when comparing closely related sister species. On the downside, the high frequency of changes introduce equally high numbers of homoplastic CAs making identification on higher taxonomic levels more challenging. This might be a problem when big reference barcode sets are applied and will most likely negatively impact identification success. In this scenario, when specimen ants cannot be classified to their correct higher taxonomic group by morphological traits using only CO1 as a marker can be problematic (Knaden *et al.* 2012). The combination of all three markers in hierarchical order (set by most efficient succession of the markers) results in a layered, character-based barcode and should in theory better resolve query sequences to the correct taxon.

None of the tested DNA markers performed the same. While CO1 and ND1 performed equally well in odonates due to their common mitochondrial nature. 28S and LWR performed better on higher taxa. As the latter two markers are genomic, their mutation ratio is slower and therefore better suited when investigating relationships on family or genus level. The complementary diagnostics discovered in our studies show the importance of selecting the right mix of markers when investigating a phylum. It is advantages to use more than one marker. Especially, when the phylum is known to be challenging.

3.1.3 A unique marker sequence is not equal to a new species

One misconception that has been stimulated by the barcoding community is the idea that BIN's (Barcode Index Numbers) are equal to species identity. The idea that a BIN is treated equally to a defined species is a misconception for several reasons. For once, DNA based markers, as well as they work as identifiers, are in general based on a single gene fragment. For animals, it is the CO1 Folmer region. A single gene will never be able to work as a marker for a complete kingdom. Animals as proven by the implemented studies are highly diverse and have shown to follow different evolutionary rules based upon their generation cycle, number of offspring, ability to hybridize. Not taking into account the diverse ecological pressures different kind of animals under different kind of ecological niches have to endure, genes in themselves follow highly divers conditions of evolution. Genes located in the core genome are usually inherited by both parents, while genes located on the mitochondrial genome are in many cases maternally inherited. A single cell has only one core genome. In contrast, it usually harbors many mitochondria. In conclusion the chance of mutations occurring in mitochondria is much higher than in the cell core. The cell core has a different, more advanced repertoire of repair proteins to encounter mutations than the mitochondria improving its stability further. On top of that, even the strands of in many cases circular mitochondrial genomes underlie different mutation ratios. So if for instance an inversion of a complete gene has occurred and is inherited, this gene will undergo a different evolution and most likely be unqualified as an identifier for distinguishing closely related sister species. The function of the gene and related protein structure are two additional criteria. Highly important genes underlie strong pressure to remain functional. Disrupting mutations can easily lead to a self-destruction of the mitochondria. Depending on the gene length and related protein structure, some genes might be able to better compensate mutations than other (one famous example is sickle cell anemia). CO1 phylogenetic history most likely does not fully reflect the evolutionary history of its host. For all those reasons, it is important not to forget our taxonomical past, but to learn from it and combine it with modern techniques. A good example is the application of a taxonomical circle as it has been proposed by DeSalle *et al.* (2005) and Damm *et al.* (2010). Here, the observation of a new marker sequence should not be treated as equal to finding a new species, but rather used as a clue that a cryptic species might have been identified. Only after the reinvestigation of morphological features, habitat, geographical location, reproductivity conditions and multiple positive arguments, a declaration of a new species should be made. A DNA barcode is an excellent tool for species identification based on using reference specimens that have been identified by traditional means. Newly acquired barcodes that do not match with the BOLD workbench should therefore be treated with care.

3.1.4 Quality control in databases

When we created reference matrices, we realized that the quality of sequences stored in NCBI and BOLD was ranging from very high to poor. Many sequences included a wide range of gaps or consisted of only short fragments. Therefore, it is very important that sufficient quality control routines are established through various filtering mechanisms if usability is supposed to be substained (Pompanon *et al.* 2005; Steinke & Hanner 2011; Vink *et al.* 2012; Shen *et al.* 2013).

3.2 Choosing the right barcoding method (distance- and/or character-based barcoding)

After discussing what a reliable barcoding marker is, the following section will deal with a dependable approach for data analysis. Due to its huge success, the definition of DNA barcoding is linked to the distance-based analysis of specimen. Typically, a NJ algorithm is used to convert DNA sequence data into genetic distances (Casiraghi *et al.* 2010). Queries are considered successfully identified when they cluster with conspecific barcodes. In my studies and research conducted by collaborating scientists (e.g. Damm 2010, Yassin et al. 2010) it was shown that barcoding has limits. These can be complemented and in some cases overcome by application of character-based barcoding. The main advantage of DNA barcoding is its focused approach on a single gene fragment that can easily be obtained by tissue sampling, PCR and sequencing.

The bottleneck of distance-based barcoding lies within its dependency on defined distance-based thresholds (Category 1; Table 3.1). Many studies have proven that a universal "threshold of genetic divergence" to assign unknown specimen to described species does not exist and remain the main challenge (DeSalle *et al.* 2005; Meier *et al.* 2006; Collins & Cruickshank 2013; Kekkonen & Hebert 2014).

Classical taxonomic studies are character-based. Employing a similar approach for DNA sequences is logical and makes the combination of both approaches feasible. Combining diagnostics from different disciplines (morphology, ecology, geography, reproductivity) would therefore agree with the concept of an integrative taxonomy scheme (DeSalle *et al.* 2005; Schlick-Steiner *et al.* 2009; Damm *et al.* 2010).

In instances, where distance-based analysis is not sufficient as an identifier (e.g. taxonomic groups without barcoding gap), it was shown (Reid *et al.* 2011, Bergmann *et al.* 2013, Paknia *et al.* 2015) that character-based barcoding has a better resolution. Distance-based analysis might be faster, however, character-based analysis uses more information encoded within the DNA sequences. Each nucleotide has the potential to be of descriptive value and can in combination with other nucleotides form a unique fingerprint. This fingerprint can be converted into a barcode and as shown can contribute in creating a better and more open identification system than DNA barcoding which is restricted to the Folmer region and the distance-based approach.

3.2.1 Comparing distance- and character-based barcoding in turtles

In Reid et al. (2011) identification success of threatened turtle species by distance-based and character-based barcoding was compared.

Of the 220 species tested in this study 162 species could by application of the barcoding gap be placed into the correct family. Of these 162 species, 130 species showed character-based diagnostics allowing us to successfully distinguish them. From the 58 remaining species, which could not be classified by a distance-based threshold, identifying characters for 23 of these species were found. Sets of simple identifying characters could be established for 153 species of the 220 species tested (70%). The proportion of species in a given family possessing diagnostic characters varied from lower than 60% to 100%. The relatively low number of diagnostic characters within some families could most likely be one of two reasons. One reason described by Hebert *et al.* (2004) attributed observations of low interspecific differentiation as a result of hybridization and mitochondrial introgression between species. Evidence from marine turtles, that support this thesis, has been observed in the family Cheloniidae (Karl *et al.* 1995; Lara-Ruiz *et al.* 2006). Here, it was shown that some turtle species are still able to hybridize even after tens of millions of years of separation. The interspecies and even intergenus hybridization (Parham *et al.* 2001; Buskirk *et al.* 2005) is possible due to low rates of molecular evolution and chromosomal rearrangement in turtles (Bickham 1981; Avise *et al.* 1992). These slow rates of molecular changes might also be the second reason for low levels of differentiation in non-hybridizing species. In contrast to other vertebrates turtle mitochondrial genes undergo evolution several-fold slower (Avise *et al.* 1992) explaining why 'recent radiations' show bad barcoding resolution when focused on CO1 alone.

Reid *et al.* (2011) was the first application of CAOS examining species rich families on this scale. While the efficacy of simple pure characters identified varied between families, Cheloniidae, Chelydridae, Pelomedusidae and Podocnemididae showed an extremely successful application of character-based barcoding. Each species available for these families possessed simple identifying character states. Interestingly, the number of species with diagnostic characters declined in larger families. This observation might be an indication of homoplasy or back mutations and should be monitored/prevented by solid specimen coverage (n>3). Usually, especially in wide ranged approaches, such as this, the sample size is pretty low due to difficulties in sample collection (e.g. rare or protected specimen; difficult to collect specimen such as in Odonata). Therefore, whenever single specimen data show problematic results, bad phylogenetic placement or strong aberration from closely related neighbour species additional information (e.g. morphological, additional molecular markers) should be obtained before the data is placed as reference in BOLD or other databases.

We compared and combined DNA barcoding and character-based barcoding in Reid *et al.* (2011) showing that combining both can improve identification success.

3.2.2 Comparing distance- and character-based barcoding in dragon- and damselflies

In Bergmann *et al.* (2013) an example was given how for taxonomically challenging taxa (here Odonata) the addition of another marker can supplement DNA barcoding. Odonates, are challenging for several reasons: They possess highly skilled flying abilities, making adult animals hard to catch. Odonate larvae, while much easier to monitor, are morphological similar, making them difficult to identify. Fast radiation of odonate species provide significant challenges for application of barcoding gap thresholds. Based on the four barcoding gap categories defined by Hebert *et al.* (2004; Tab. 3.1) most of the species (39 of 44 ND1; 33 of 39 CO1) fulfilled the criteria for category I and can be identified by distance-based barcoding. Using the marker ND1 showed only five species not fullfilling the criteria for category I. Three species (*T. morrisoni*, *P. bicoerulans* & *C. grandis*) displayed intraspecific distance values above 2% (Category II). Two species showed in some instances no interspecific differences (*P. acaciae* & *P. niloticum*; Category III). Six species failed to fulfill the criteria for category I with CO1 as marker. Two species (*P. bicoerulans* & *C. grandis*) as with ND1 were placed into category II, the other four species (*T. grouti*, *T. nuptialis*, *P. acaciae* & *P. niloticum*) showed instances where interspecific distances were below 2%. All in all, using distance-based K2P values, both markers showed great ability in distinguishing odonate species from each other. In two cases (*C. erythreae* & *P. massaicum*), distances between samples of congeneric species were higher than between samples from different taxa (*Anax* & *Ischnura*; for both CO1 & ND1).

Using character-based barcodes, we can distinguish 43 of 45 odonate species (six sequences could not be obtained) by 29 diagnostic nucleotide positions (CO1). 48 of 50 species (one missing species) can be identified by 29 diagnostic positions within the ND1 marker region. Both markers have no diagnostic characters for differentiating specimens of *P. niloticum* from those of *P. acaciae*. As character-based barcoding is independent from distance-based thresholds, but relies on identify-

Tab. 3.1.: Barcoding gap categories

Category	Maximal intraspecific distance	Minimal interspecific distance
I	$<2\%$	$>2\%$
II	$\geq 2\%$	$>2\%$
III	$<2\%$	$\leq 2\%$
IV	$\geq 2\%$	$\leq 2\%$

ing diagnostic characters, comparing both methods shows, that by using the same dataset, we were able to find more diagnostics when using CAOS (K2P: ND1 39/44, CO1 33/39; CAOS: ND1 48/50, CO1 43/45).

3.2.3 Testing character-based barcoding by classifiying "new" queries and by adding random mutations in queries

In order to further investigate the reliability of character-based barcoding through CAOS, I programmed two different programs: One that tests the validity (leave-one-out) and a second (random substitution) that tests the robustness of the CAOS-Classifier (identifies specimen through barcodes). 227/234 (CO1) and 260/266 (ND1) specimen were correctly identified in the "leave-one-out" test. In the "random substitution" test using a substitution ratio of 1% an average score of 233/234 (CO1) and 249/266 (ND1) correct assignments were achieved. Increasing the substitution ratio to 5% still led to 225/234 (CO1) and 237/266 (ND1) correct identifications. Both tests highlight the robustness of classification through the CAOS-Classifier but should also raise awareness that even well developed engines can create false identification results. It is the responsibility of researchers to second-guess their results.

3.2.4 CAOS-Classifier vs BOLD

A good example of false identification is the result of our last test that was conducted in Bergmann *et al.* (2013). Here, the performance of the CAOS-Classifier was compared to the identification success of BOLD using our 234 CO1 sequences as query. While the CAOS-Classifier correctly assigned all 234 queries with BOLD only 131 sequences were assigned with 97-100% accuracy. The remaining 103 queries showed no match and it is very likely that the corresponding species were not part of the BOLD library at this time. Interestingly, some specimen were identified as neighbour species with high support (>99%). In two instances, even a different genus was identified than what we would have expected (*E. cyathigerum* -> *C. hastulatum* & *I. senegalensis* -> *P. abyssinica*). For both species we had collected five specimen (one for *E. cyathigerum* showed the same identification as ours). It is likely that a misidentification has occurred on either the BOLD server or in our study.

3.2.5 Performance conclusion of distance- vs. character-based barcoding

The majority of DNA barcoding studies use the distance-based approach for specimen identification (Hebert *et al.* 2003). The accuracy of this method is highly dependent on the presence of a barcoding gap (Meyer & Paulay 2005). In odonates, high intra- and low interspecific variability has been observed by Rach *et al.* 2008 leading to the conclusion that distance-based methods are ill suited for DNA barcoding in this insect order. This conclusion is exaggerated as many Odonata species can be identified by a distance-based threshold. However, the fact that some species as estimated show high intraspecific and/or low interspecific variation can be agreed on. As we have only investigated a small subset of the complete Odonata community, it can be assumed that more cases of overlapping intra- and interspecific distances exist. The overlapping distances observed for the genera *Crocothemis* and *Pseudagrion* are exemplary for a possible miss-identification of new specimen if critical species are missing from the DNA barcode database. In these occasions character-based barcoding is recommend. Distance thresholds are needless, and diagnostic characters can be easily identified at any needed taxonomic level by means of the CAOS algorithm.

3.2.6 Testing the performance of character-based barcoding in ants

Paknia *et al.* (2015) investigated if character-based barcoding can be used to define cryptic biodiversity in Formicidae (ants). Several studies have been unsuccessful in resolving cryptic diversity by distance-based barcoding (e.g. Schlick-Steiner *et al.* 2006; Ueda *et al.* 2012). The family Formicidae are the ideal phylum to investigate cryptic biodiversity because of their complex population differentiation, hybridization and speciation processes. With more than double the amount of classified species (~13.000), compared to odonates (~5.800), molecular identification of this taxon is even more ambitious than our previous study of taxonomically challenging phyla. Due to high or/and complex intraspecific morphological variations, ants pose a serious challenge for traditional taxonomy (Ross *et al.* 2010; Blaimer 2012). Some distance-based barcode studies on ants have been successful (e.g. Saux *et al.* 2004; Smith *et al.* 2005), but cryptic biodiversity remains a major challenge for defining alpha-taxonomy, ecology and conservation (Seifert 2009). Hyperdiversity has been reported for some genera (Moreau 2008). One worldwide dispersed species for example (Cardiocondyla) has been estimated to include 52% cryptic species (Seifert 2009). Identifying those problematic taxa by distance-based DNA barcoding yielded no promising results (e.g. Knaden *et al.* 2012; Ueda *et al.* 2012). We tested the idea of a layered, character-based barcode approach to solve the identification problem. In theory, combining multiple markers with complementary diagnostic features should increase identification success while each marker working on a different taxonomic level. The idea for this approach came from the observation that ND1 and CO1 complemented each other well in the previous study. While identification of some ants can easily be achieved by morphology up to the level of subfamily or genera, the layered approach could be used as aid in more challenging cases (e.g. Lapolla *et al.* 2011; Sosa-Calvo *et al.* 2013). The layered approach could also be helpful when only tissue is available or animals are in bad condition (e.g. stomach content).

3.3 Character-based flagging as a mean to uncover cryptic species

While we clearly showed the quality of a character-based barcode as a means to identify specimen another advantage of a diagnostic barcode approach is its ability to flag populations. Flagging, in short, is the process of grouping specimen within distinct species according to geographical origin and test if the groups harbor geographic specific traits qualifying them as potential novel species (Goldstein & DeSalle 2011). Using flagging, we were able to distinguish five odonate population based on unique diagnostics specific for the compared geographic clusters. There are two purposes for flagging: Flagging can be used to identify populations of origin for unidentified specimens. Diagnostics specific to geography can then be used in ecological monitoring studies where samples are hard to identify to population. If diagnostics do exist, then these populations can be flagged for future, integrated taxonomic studies (DeSalle 2006; Rubinoff 2006). Later, these observations might result in species descriptions for these diagnosable populations (Goldstein *et al.* 2000).

As flagging was successful in Bergmann *et al.* (2013) we successfully tested the process on the diverse Australian *Monomorium rothsteini* complex. It has been suggested as a "group of many species" by Greenslade *et al.* (1979) and was defined by Anderson *et al.* (2007) as a group of 50 or more species. The *M. rothsteini* complex is one of the great challenges that exist in systematics of cryptic ants. Members of the complex show overlaps in morphological characters and distribution ranges. Using a distance-based barcoding approach, Sparks *et al.* (2014) were able to identify 38 well supported clades within the *M. rothsteini* complex. Clade 5a of the complex contains the greatest number of individuals, and haplotypes from multiple locations within Australia. Clade 5a could neither be resolved by morphology nor distance-based barcoding. As this clade provided the greatest challenge, we chose it for character-based flagging. Using character-based barcoding, we were able to fully resolve all 25 tested specimen by unique geographical diagnostics and could identify eight potential taxonomic entities that might be meriting formal description. Flagging by character-based barcodes allowed us to differentiate ant specimen based on geographical diagnostics. As shown in previous studies character-based barcoding offered a reliable solution for identification of problematic ant species. In Paknia *et al.* 2015 we were able to show that our method is a cost-efficient approach to estimate presence, absence or frequency of potentially cryptic species.

3.4 The future of character-based DNA barcoding

3.4.1 Layered, character-based barcoding

The layered, character-based barcode described in Paknia *et al.* (2015) is a thought model. A real working version can easily be achieved.

My current CAOS-Classifier, when classifying a query, gives grades based upon the number of CAs matching between the query and reference set 1 (left branch in guide tree) versus query and reference set 2 (right branch in guide tree). Whichever branch gains more points will be followed until an ending node is reached. My idea for a layered barcode is keeping this model while adapting it to multiple markers. Currently, only sequences based upon one marker are entered into CAOS. In the layered approach, any desired number of sequences can be entered by just separating the sequences using the symbol "|" between the sequences (e.g. "ACGT|GGGC|CACA" = "28S|LWR|CO1"), other separating symbols could be used to indicate other types of data (e.g. "+" = "amino acids") this way even an integrative barcode could be obtained. Using a concatenated tree as guide, the layered barcode would be working in parallel. For each identification step, the number of hits for each node and barcode element would be compared. Only then, the best match would be followed and this step repeated for the next branch until an ending node is reached. In this way, for each taxonomic level, the best suited barcode would be preferred and used.

3.4.2 Machine learning

While character-based barcoding is currently performed with just sPu and sPr, the addition of compound characters (diagnostic character combinations) could further improve the identification success of CAOS. As screening data for compound characters is processor intensive, a machine learning approach for next generation character-based barcoding could be created. This might be achieved through application of CNN (Convoluted Neuronal Networks) and cloud computing. Cloud computing is the present and future of high performance processing and in combination with a neuronal network, a smartphone application could be created that allows an integrative identification approach in the field. For example, combining high resolution three dimensional scans of reference specimen with a smartphone app (camera; gps sensor; altitude sensor; humidity sensor) and a field PCR/Sequencer would allow an on the fly identification of specimen. I believe *cum granum salis* a prototype tricorder ("Star Trek", mobile identification device) will most likely become reality.

3.4.3 References

Andersen AN (2007) Ant diversity in arid Australia: a systematic overview. In: Advances in ant systematics (Hymenoptera: Formicidae): homage to E. O. Wilson – 50 years of contributions. Memoirs of the American Entomological Institute, 80. (eds. Snelling RR, Fisher BL, Ward PS), pp. 91-51.

Avise JC, Bowen BW, Lamb T, Meylan AB, Bermingham E (1992) Mitochondrial DNA evolution at a turtle's pace: evidence for low genetic variability and reduced microevolutionary rate in the Testudines. Molecular Biology and Evolution 9, 457-473.

Baker RH, DeSalle R (1997) Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst Biol 46, 654-673.

Belshaw R, Fitton M, Herniou E, Gimeno C, Quicke DL (1998) A phylogenetic reconstruction of the Ichneumonoidea (Hymenoptera) based on the D2 variable region of 28S ribosomal RNA. Systematic Entomology 23, 109-123.

Bergmann T, Rach J, Damm S, *et al.* (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.

Bickham JW (1981) Two-Hundred-Million-Year-Old Chromosomes: Deceleration of the Rate of Karyotypic Evolution in Turtles. Science 212, 1291-1293.

Blaimer BB (2012) Untangling complex morphological variation: taxonomic revision of the subgenus Crematogaster (Oxygyne) in Madagascar, with insight into the evolution and biogeography of this enigmatic ant clade (Hymenoptera: Formicidae). Systematic Entomology 37, 240-260.

Buskirk JR, Parham JF, Feldman CR (2005) On the hybridisation between two distantly related Asian turtles (Testudines: *Sacalia* x *Mauremys*). Salamandra 41, 21-26.

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

Chaubet B, Derocles SAP, Hullé M, *et al.* (2013) Two new species of aphid parasitoids (Hymenoptera, Braconidae, Aphidiinae) from the high arctic (Spitsbergen, Svalbard). Zoologischer Anzeiger - A Journal of Comparative Zoology 252, 34-40.

Collins R, Cruickshank R (2013) The seven deadly sins of DNA barcoding. Molecular Ecology Resources 13, 969-975.

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

Derocles SAP, Le Ralec A, Plantegenest M, *et al.* (2012) Identification of molecular markers for DNA barcoding in the Aphidiinae (Hym. Braconidae). Molecular Ecology Resources 12, 197-208.

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

DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 1905-1916.

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

Journal of Odonatology 10, 31-41.

Dowton M, Austin AD (1998) Phylogenetic relationships among the microgastroid wasps (Hymenoptera: Braconidae): combined analysis of 16S and 28S rDNA genes and morphological data. Molecular Phylogenetics and Evolution 10, 354-366.

Elias M, Hill RI, Willmott KR, *et al.* (2007) Limited performance of DNA barcoding in a diverse community of tropical butterflies. Proceedings of the Royal Society B-Biological Sciences 274, 2881-2889.

Goldstein PZ, DeSalle R (2011) Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. BioEssays 33, 135-147.

Goldstein PZ, DeSalle R, Amato G, Vogler AP (2000) Conservation genetics at the species boundary. Conserv. Biol. 14, 120-131.

Greenslade PJM (1979) A Guide to the Ants of South Australia South Australian Museum, Adelaide.

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

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc Biol Sci 270, 313-321.

Jansen G, Savolainen R, Vepsäläinen K (2009) DNA barcoding as a heuristic tool for classifying undescribed Nearctic Myrmica ants (Hymenoptera: Formicidae). Zoologica Scripta 38, 527-536.

Karl SA, Bowen BW, Avise JC (1995) Hybridization among the ancient mariners: characterization of marine turtle hybrids with molecular genetic assays. J Hered 86, s262-268.

Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Molecular Ecology Resources 14, 706-715.

Knaden M, Tinaut A, Stokl J, Cerdà X, Wehner R (2012) Molecular phylogeny of the desert ant genus Cataglyphis (Hymenoptera: Formicidae). Myrmecological News 16, 123-132.

Lapolla JS, Kallal RJ, Brady SG (2011) A new ant genus from the Greater Antilles and Central America, Zatania (Hymenoptera: Formicidae), exemplifies the utility of male and molecular character systems. Systematic Entomology, no-no.

Lara-Ruiz P, Lopez, G.G., Santos, F.R. (2006) Extensive hybridization in hawksbill turtles (*Eretmochelys imbricata*) nesting in Brazil revealed by mtDNA analyses. Conserv Genet 7, 773-781.

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

Lucky A (2011) Molecular phylogeny and biogeography of the spider ants, genus *Leptomyrmex Mayr* (Hymenoptera: Formicidae). Molecular Phylogenetics and Evolution 59, 281-292.

Lucky A, Sarnat EM (2010) Biogeography and diversification of the Pacific ant genus *Lordomyrma Emery*. Journal of Biogeography 37, 624-634.

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.

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

Moreau CS (2008) Unraveling the evolutionary history of the hyperdiverse ant genus *Pheidole* (Hymenoptera: Formicidae). Molecular Phylogenetics and Evolution 48, 224-239.

Paknia O, Bergmann T, Hadrys H (2015) Some ’ant’swers: Application of a layered barcode approach to problems in ant taxonomy. Mol Ecol Resour 15, 1262-1274.

Parham JF, Simison WB, Kozak KH, Feldman CR, Shi H (2001) New Chinese turtles: endangered or invalid? A reassessment of two species using mitochondrial DNA, allozyme electrophoresis and known-locality specimens. Animal Conservation 4, 357-367.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: Causes, consequences and solutions. Nature Reviews Genetics 6, 847-859.

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.

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

Ross KG, Gotzek D, Ascunce MS, Shoemaker DD (2010) Species delimitation: A case study in a problematic ant taxon. Systematic Biology 59, 162-184.

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. J Hered 97, 581-594.

Saux C, Fisher BL, Spicer GS (2004) Dracula ant phylogeny as inferred by nuclear 28S rDNA sequences and implications for ant systematics (Hymenoptera: Formicidae: Amblyoponinae). Molecular Phylogenetics and Evolution 33, 457-468.

Schlick-Steiner BC, Steiner FM, Moder K, *et al.* (2006) A multidisciplinary approach reveals cryptic diversity in Western Palearctic Tetramorium ants (Hymenoptera: Formicidae). Molecular Phylogenetics and Evolution 40, 259-273.

Schlick-Steiner BC, Steiner FM, Seifert B, *et al.* (2009) Integrative taxonomy: A multisource approach to exploring biodiversity. Annual Review of Entomology 55, 421-438.

Seifert B (2009) Cryptic species in ants (Hymenoptera: Formicidae) revisited: we need a change in the alpha-taxonomic approach. Myrmecological News 12, 149-166.

Shen YY, Chen X, Murphy RW (2013) Assessing DNA barcoding as a tool for species identification and data quality control. PLoS One 8, e57125.

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: Biological Sciences 360, 1825-1834.

Sosa-Calvo J, Schultz TR, Brandão CRF, *et al.* (2013) *Cyatta abscondita*: taxonomy, evolution, and natural history of a new fungus-farming ant genus from Brazil. PLoS One 8, e80498.

Sparks KS, Andersen AN, Donnellan SC, Austin AD (2014) Navigating the mtDNA road map out of the morphological maze: interpreting morphological variation in the diverse *Monomorium rothsteini* (Forel) complex (Hymenoptera: Formicidae). Systematic Entomology 39, 264-278.

Steinke D, Hanner R (2011) The FISH-BOL collaborators' protocol. Mitochondrial DNA 22 Suppl 1, 10-14.

Ueda S, Nozawa T, Matsuzuki T, *et al.* (2012) Phylogeny and phylogeography of *Myrmica rubra* complex (Myrmicinae) in the Japanese Alps. Psyche: A Journal of Entomology 2012.

Vink CJ, Paquin P, Cruickshank RH (2012) Taxonomy and Irreproducible Biological Science. Bioscience 62, 451-452.

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

Abbreviations

μl Microliter. 45

AMCC Ambrose Monell Cryo Collection. 16

AMNH American Museum of Natural History. 8, 15, 16, 31

AZA Association of Zoos and Aquariums. 15

BC Before Christ. 2

BIN Barcode Index Number. 94

BOLD Barcode of Life Datasystem. 4, 9, 14–19, 21, 37, 40, 41, 50, 54, 59, 64, 71–73, 84, 92, 95, 97, 99, 100, 122–126

bp Base pair. iv, vi, 3, 5, 13, 46, 51, 73, 76, 77, 91

CA Character Attribute. 5–8, 74, 75, 77–82, 84, 93, 103, 115

CAOS Character Attribute Organization System. 5–7, 9, 18, 19, 23, 27, 28, 35, 42, 46, 47, 56, 57, 65, 74, 81, 97, 99, 100, 103, 115

CBoL Consortium for the Barcode of Life. 39

CNN Convoluted Neuronal Networks. 104

CO1 Cytochrome Oxidase subunit 1. iv–vii, 3, 9, 13, 14, 16, 17, 19, 22, 24–26, 29, 38–48, 50–59, 68, 70–74, 76–83, 86, 91–94, 97–99, 101, 103, 115, 117

DAAD Deutscher Akademischer Austauschdienst. 31

DNA Deoxyribonucleic Acid. vi, vii, 3–5, 9, 13–16, 24, 25, 31–43, 45, 46, 48, 49, 51, 52, 55–57, 60–66, 68–79, 81–83, 85–90, 92–96, 98, 100, 101, 103, 117

e.g. *exempli gratia*. 2, 6, 68–72, 79, 82, 84, 85, 92, 93, 97, 101, 103

i.e. *id est*. 3, 23

ISIS Institute for Science and International Security. 15

IUCN International Union for Conservation of Nature. 13–15, 33

K2P Kimura 2-parameter. 18, 20, 21, 24, 41, 46, 50, 51, 56, 57, 98, 99

LWR Long-Wavelength Rhodopsin. 71–74, 76, 77, 79, 81, 82, 92, 93, 103, 117

m Minutes. 16

MEGA Molecular Evolutionary Genetics Analysis. 46, 72

min Minutes. 46

ML Maximum Likelihood. 46

mM Millimolar. 45

NADH Nicotinamide Adenine Dinucleotide. 39, 40, 45

NCBI National Center for Biotechnology Information. 8, 84, 95

ND1 NADH dehydrogenase subunit 1. iv–vii, 9, 38–40, 42–47, 49–59, 92, 93, 98, 99, 101, 115, 117

NJ Neighbour Joining. 47, 69, 95

numts nuclear mitochondrial DNA segment. 17

PCR Polymerase Chain Reaction. 3, 16, 39, 45, 55, 65, 91, 96

pM Picomolar. 45

RAxML Randomized Axelerated Maximum Likelihood. 65, 74, 90

s Seconds. 16, 46

SD Standard Deviation. 20, 21

sPr simple Private. 79

sPu simple Pure. 79

U Unit. 45

Glossary

CAOS-Analyzer Software that identifies diagnostic characters for a given dataset.. 8, 47, 74, 75

CAOS-Barcoder Software that creates barcodes from the output of the CAOS-Analyzer.. 8, 47, 74, 75

CAOS-Classifier Software that classifies new specimen based on diagnostic characters.. 7, 8, 47, 50, 53, 54, 58, 59, 74, 75, 99, 103

MacClade MacClade is a computer program for phylogenetic analysis written by David Maddison and Wayne Maddison. Its analytical strength is in studies of character evolution. It also provides many tools for entering and editing data and phylogenies, and for producing tree diagrams and charts.. 6, 19, 47, 74

Mafft Software for sequence alignment.. 46

MEGA 4 MEGA is a computer software for conducting statistical analysis of molecular evolution and for constructing phylogenetic trees.. 17, 18, 20, 117

Mesquite Mesquite is modular, extendible software for evolutionary biology, designed to help biologists organize and analyze comparative data about organisms. Its emphasis is on phylogenetic analysis, but some of its modules concern population genetics, while others do non-phylogenetic multivariate analysis.. 6

p-elf P-elf is a computer program for classifying new specimen through diagnostic characters. It classifies a file of query sequences according to the rules generated by p-gnome.. 5

p-gnome P-gnome is a computer program for identifying diagnostic characters. It is a diagnostic rules generator that searches through a given data matrix and establishes diagnostic rule sets for each of the pre-described entities in the data matrix.. 5–7, 9, 19

Phylip PHYLogeny Inference Package (PHYLIP) is a free computational phylogenetics package of programs for inferring evolutionary trees.. 19

R R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.. 18

RAxML Program for creating phylogenetic trees based on maximum likelihood algorythm.. 46

SeaView Software for sequence alignment.. 46, 74

Sequencher Sequencher, is a sequencing software. It is used for DNA sequence assembly and analysis.. 17

List of Figures

List of Tables

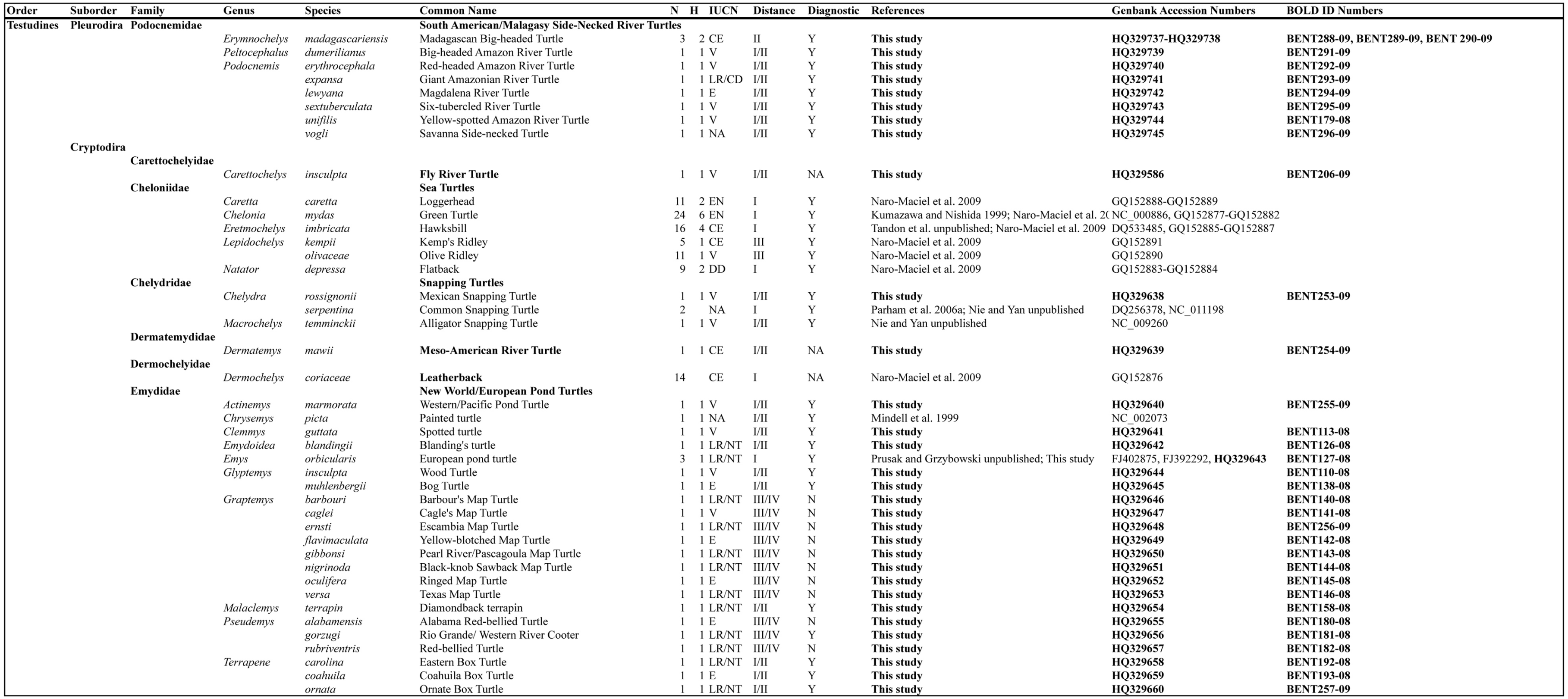
Appendices

A Supplementary Data

A.1 Manuscript 1

Order	Suborder	Family	Genus	Species	Common Name	N	H	IUCN	Distance	Diagnostic	References	Genbank Accession Numbers	BOLD ID Numbers
Testudines													
	Pleurodira												
		Chelidae			**Austro-American Side-necked Turtles**								
			Acanthochelys	*macrocephala*	Big-headed Pantanal Swamp Turtle	3	1	LR/NT	III	Y	**This study**	**HQ329587**	**BENT102-08, BENT 207-09, BENT 208-09**
				pallidipectoris	Chaco Side-necked Turtle	2	1	V	III	Y	**This study**	**HQ329588**	**BENT103-08, BENT 209-09**
				radiolata	Brazilian Radiolated Swamp Turtle	2	2	LR/NT	II	N	**This study**	**HQ329589-HQ329590**	**BENT104-08, BENT 210-09**
			Chelodina	*mccordi*	Roti Island/McCord's Long-necked Turtle	1	1	CE	I/II	Y	**This study**	**HQ329591**	**BENT211-09**
				novaeguineae	New Guinea Long-necked Turtle	1	1	LR/LC	III/IV	Y	**This study**	**HQ329592**	**BENT212-09**
				oblonga	Narrow-breasted Long-necked Turtle	1	1	LR/NT	I/II	N	**This study**	**HQ329593**	**BENT213-09**
				parkeri	Parker's Long-necked turtle	1	1	V	I/II	Y	**This study**	**HQ329594**	**BENT214-09**
				pritchardi	Pritchard's Long-necked Turtle	1	1	E	I/II	Y	**This study**	**HQ329595**	**BENT215-09**
				reimanni	Reimann's Long-necked Turtle	1	1	LR/NT	III/IV	Y	**This study**	**HQ329596**	**BENT216-09**
				rugosa	Northern Long-necked Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329597**	**BENT111-08**
			Chelus	*fimbriata*	Matamata	2	2	NA	I	Y	**This study**	**HQ329598-HQ329599**	**BENT217-09, 218-09**
			Elseya	*albagula*	White-throated Snapping Turtle	1	1	NA	I/II	Y	**This study**	**HQ329600**	**BENT219-09**
				bellii	Western Sawshelled Turtle	1	1	E	I/II	N	**This study**	**HQ329601**	**BENT220-09**
				branderhorsti	Southern New Guinea Snapping Turtle	2	1	V	I	N	**This study**	**HQ329602**	**BENT221-09, BENT 222-09**
				dentata	Northern Australian Snapping Turtle	2	2	NA	III	N	**This study**	**HQ329603-HQ329604**	**BENT223-09, BENT 224-09**
				georgesi	Bellinger River Helmeted Turtle	1	1	DD	I/II	Y	**This study**	**HQ329605**	**BENT225-09**
				irwini	White-headed Snapping Turtle	1	1	NA	III/IV	N	**This study**	**HQ329606**	**BENT226-09**
				lavarackorum	Gulf Snapping Turtle	1	1	NA	III/IV	Y	**This study**	**HQ329607**	**BENT227-09**
												HQ329608-HQ329615	**BENT119-08, BENT120-08, BENT121-08, BENT122-08, BENT123-08, BENT228-09, BENT229-09, BENT230-09, BENT231-09, BENT232-09**
				novaeguineae	New Guinea Spotted Turtle	10	8	LR/LC	II	N	**This study**		
				purvisi	Manning River Helmeted Turtle	1	1	DD	I/II	N	**This study**	**HQ329616**	**BENT233-09**
			Elusor	*macrurus*	Mary River Turtle	2	1	E	I	Y	**This study**	**HQ329617**	**BENT124-08, BENT234-09**
			Emydura	*macquarii*	Southern River Turtle	1	1	NA	I/II	N	**This study**	**HQ329618**	**BENT235-09**
				subglobosa	Red-bellied Short-necked/Painted Turtle	3	3	LR/LC	IV	N	**This study**	**HQ329619-HQ329621**	**BENT125-08, BENT236-09, BENT237-09**
				tanybaraga	Northern Yellow-faced Turtle	2	2	NA	III	N	**This study**	**HQ329623**	**BENT238-09, BENT239-09**
				victoriae	Northern Red-Faced Turtle	2	2	NA	I	Y	**This study**	**HQ329624-HQ329624**	**BENT240-09, BENT241-09**
			Hydromedusa	*maximiliani*	Brazilian Snake-necked Turtle	1	1	V	I/II	Y	**This study**	**HQ329626**	**BENT149-08**
			Mesoclemmys	*heliostemma*	Amazon Toad-headed Turtle	1	1	NA	III/IV	Y	**This study**	**HQ329627**	**BENT107-08**
				raniceps	Amazon Toad-headed Turtle	2	2	NA	III	Y	**This study**	**HQ329628-HQ329629**	**BENT108-08, BENT242-09**
				tuberculata	Tuberculated Toad-headed Turtle	1	1	NA	I/II	Y	**This study**	**HQ329630**	**BENT243-09**
				vanderhaegei	Vanderhaege's Toad-headed Turtle	3	2	LR/NT	I	Y	**This study**	**HQ329631-HQ329632**	**BENT244-09, BENT245-09, BENT246-09**
			Phrynops	*geoffroanus*	Geoffrory's Side-Necked Turtle	1	1	NA	I/II	N	**This study**	**HQ329633**	**BENT247-09**
				williamsi	Williams' Side-necked Turtle	1	1	NA	I/II	N	**This study**	**HQ329634**	**BENT178-08**
			Pseudemydura	*umbrina*	Western Swamp/Short-necked Turtle	1	1	CE	I/II	Y	**This study**	**HQ329635**	**BENT248-09**
			Rheodytes	*leukops*	White-Eyed/Fitzroy River River Turtle	2	1	V	I	Y	**This study**	**HQ329636**	**BENT249-09, BENT250-09**
			Rhinemys	*rufipes*	Red Side-necked Turtle	2	1	LR/NT	I	Y	**This study**	**HQ329637**	**BENT251-09, BENT252-09**
		Pelomedusidae			**African Mud Turtles**								
			Pelomedusa	*subrufa*	African Helmeted Turtle	2	1	NA	I	Y	Zardoya and Meyer 1998	NC_001947, AF039066	
			Pelusios	*broadleyi*	Turkana Mud Turtle	1	1	V	I/II	Y	**This study**	**HQ329725**	**BENT297-09**
				carinatus	African Keeled Mud Turtle	1	1	NA	III/IV	Y	**This study**	**HQ329726**	**BENT298-09**
				castaneus	West African Mud Turtle	1	1	NA	III/IV	Y	**This study**	**HQ329727**	**BENT299-09**
				castanoides	Yellowbelly Mud Turtle	1	1	LR/LC	I/II	Y	**This study**	**HQ329728**	**BENT300-09**
				chapini	Central African Mud Turtle	1	1	NA	III/IV	Y	**This study**	**HQ329729**	**BENT301-09**
				cupulatta	Forest Hinged Turtle	1	1	NA	I/II	Y	**This study**	**HQ329730**	**BENT302-09**
				gabonensis	African Forest Turtle	1	1	NA	I/II	Y	**This study**	**HQ329731**	**BENT303-09**
				marani	Gabon Mud Turtle	1	1	NA	I/II	Y	**This study**	**HQ329732**	**BENT304-09**
				niger	West African Mud/Black Turtle	1	1	NA	I/II	Y	**This study**	**HQ329733**	**BENT305-09**
				rhodesianus	Variable Mud Turtle	1	1	LR/LC	III/IV	Y	**This study**	**HQ329734**	**BENT306-09**
				sinuatus	East African Serrated Mud Turtle	1	1	NA	I/II	Y	**This study**	**HQ329735**	**BENT174-08**
				williamsi	Williams' Mud Turtle	1	1	NA	I/II	Y	**This study**	**HQ329736**	**BENT307-09**

Fig. A.1.: Table S1: Descriptive data for all taxa and sequences included in this study. 'N' indicates the number of individuals sequences for each species; 'H' indicates the number of haplotypes observed in each species; 'Distance' indicates the species' classification within the distance- based scheme described in the text; 'Diagnostic' indicates the presence ('Y') or absence ('N') of familylevel simple identifying characters in the species. References and accession numbers are in BOLD for novel sequences produced in this study.



Order	Suborder	Family	Genus	Species	Common Name	N	H	IUCN	Distance	Diagnostic	References	Genbank Accession Numbers	BOLD ID Numbers
Testudines	**Pleurodira**	**Podocnemidae**			**South American/Malagasy Side-Necked River Turtles**								
			Erymnochelys	*madagascariensis*	Madagascan Big-headed Turtle	3	2	CE	II	Y	**This study**	**HQ329737-HQ329738**	**BENT288-09, BENT289-09, BENT 290-09**
			Peltocephalus	*dumerilianus*	Big-headed Amazon River Turtle	1	1	V	I/II	Y	**This study**	**HQ329739**	**BENT291-09**
			Podocnemis	*erythrocephala*	Red-headed Amazon River Turtle	1	1	V	I/II	Y	**This study**	**HQ329740**	**BENT292-09**
				expansa	Giant Amazonian River Turtle	1	1	LR/CD	I/II	Y	**This study**	**HQ329741**	**BENT293-09**
				lewyana	Magdalena River Turtle	1	1	E	I/II	Y	**This study**	**HQ329742**	**BENT294-09**
				sextuberculata	Six-tubercled River Turtle	1	1	V	I/II	Y	**This study**	**HQ329743**	**BENT295-09**
				unifilis	Yellow-spotted Amazon River Turtle	1	1	V	I/II	Y	**This study**	**HQ329744**	**BENT179-08**
				vogli	Savanna Side-necked Turtle	1	1	NA	I/II	Y	**This study**	**HQ329745**	**BENT296-09**
	Cryptodira												
		Carettochelyidae											
			Carettochelys	*insculpta*	**Fly River Turtle**	1	1	V	I/II	NA	**This study**	**HQ329586**	**BENT206-09**
		Cheloniidae			**Sea Turtles**								
			Caretta	*caretta*	Loggerhead	11	2	EN	I	Y	Naro-Maciel et al. 2009	GQ152888-GQ152889	
			Chelonia	*mydas*	Green Turtle	24	6	EN	I	Y	Kumazawa and Nishida 1999; Naro-Maciel et al. 2(	NC_000886, GQ152877-GQ152882	
			Eretmochelys	*imbricata*	Hawksbill	16	4	CE	I	Y	Tandon et al. unpublished; Naro-Maciel et al. 2009	DQ533485, GQ152885-GQ152887	
			Lepidochelys	*kempii*	Kemp's Ridley	5	1	CE	III	Y	Naro-Maciel et al. 2009	GQ152891	
				olivaceae	Olive Ridley	11	1	V	III	Y	Naro-Maciel et al. 2009	GQ152890	
			Natator	*depressa*	Flatback	9	2	DD	I	Y	Naro-Maciel et al. 2009	GQ152883-GQ152884	
		Chelydridae			**Snapping Turtles**								
			Chelydra	*rossignonii*	Mexican Snapping Turtle	1	1	V	I/II	Y	**This study**	**HQ329638**	**BENT253-09**
				serpentina	Common Snapping Turtle	2		NA	I	Y	Parham et al. 2006a; Nie and Yan unpublished	DQ256378, NC_011198	
			Macrochelys	*temminckii*	Alligator Snapping Turtle	1	1	V	I/II	Y	Nie and Yan unpublished	NC_009260	
		Dermatemydidae											
			Dermatemys	*mawii*	**Meso-American River Turtle**	1	1	CE	I/II	NA	**This study**	**HQ329639**	**BENT254-09**
		Dermochelyidae											
			Dermochelys	*coriaceae*	**Leatherback**	14		CE	I	NA	Naro-Maciel et al. 2009	GQ152876	
		Emydidae			**New World/European Pond Turtles**								
			Actinemys	*marmorata*	Western/Pacific Pond Turtle	1	1	V	I/II	Y	**This study**	**HQ329640**	**BENT255-09**
			Chrysemys	*picta*	Painted turtle	1	1	NA	I/II	Y	Mindell et al. 1999	NC_002073	
			Clemmys	*guttata*	Spotted turtle	1	1	V	I/II	Y	**This study**	**HQ329641**	**BENT113-08**
			Emydoidea	*blandingii*	Blanding's turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329642**	**BENT126-08**
			Emys	*orbicularis*	European pond turtle	3	1	LR/NT	I	Y	Prusak and Grzybowski unpublished; This study	FJ402875, FJ392292, **HQ329643**	**BENT127-08**
			Glyptemys	*insculpta*	Wood Turtle	1	1	V	I/II	Y	**This study**	**HQ329644**	**BENT110-08**
				muhlenbergii	Bog Turtle	1	1	E	I/II	Y	**This study**	**HQ329645**	**BENT138-08**
			Graptemys	*barbouri*	Barbour's Map Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329646**	**BENT140-08**
				caglei	Cagle's Map Turtle	1	1	V	III/IV	N	**This study**	**HQ329647**	**BENT141-08**
				ernsti	Escambia Map Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329648**	**BENT256-09**
				flavimaculata	Yellow-blotched Map Turtle	1	1	E	III/IV	N	**This study**	**HQ329649**	**BENT142-08**
				gibbonsi	Pearl River/Pascagoula Map Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329650**	**BENT143-08**
				nigrinoda	Black-knob Sawback Map Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329651**	**BENT144-08**
				oculifera	Ringed Map Turtle	1	1	E	III/IV	N	**This study**	**HQ329652**	**BENT145-08**
				versa	Texas Map Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329653**	**BENT146-08**
			Malaclemys	*terrapin*	Diamondback terrapin	1	1	LR/NT	I/II	Y	**This study**	**HQ329654**	**BENT158-08**
			Pseudemys	*alabamensis*	Alabama Red-bellied Turtle	1	1	E	III/IV	N	**This study**	**HQ329655**	**BENT180-08**
				gorzugi	Rio Grande/ Western River Cooter	1	1	LR/NT	III/IV	Y	**This study**	**HQ329656**	**BENT181-08**
				rubriventris	Red-bellied Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329657**	**BENT182-08**
			Terrapene	*carolina*	Eastern Box Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329658**	**BENT192-08**
				coahuila	Coahuila Box Turtle	1	1	E	I/II	Y	**This study**	**HQ329659**	**BENT193-08**
				ornata	Ornate Box Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329660**	**BENT257-09**

Fig. A.2.: Table S1: Descriptive data for all taxa and sequences included in this study. 'N' indicates the number of individuals sequences for each species; 'H' indicates the number of haplotypes observed in each species; 'Distance' indicates the species' classification within the distance- based scheme described in the text; 'Diagnostic' indicates the presence ('Y') or absence ('N') of familylevel simple identifying characters in the species. References and accession numbers are in BOLD for novel sequences produced in this study.

Order	Suborder	Family	Genus	Species	Common Name	N	H	IUCN	Distance	Diagnostic	References	Genbank Accession Numbers	BOLD ID Numbers
Testudines	**Cryptodira**	**Emydidae**	*Trachemys*	*decorata*	Hispaniola Slider	1	1	V	III/IV	Y	**This study**	**HQ329661**	**BENT194-08**
				emolli	Nicaraguan Slider	1	1	NA	III/IV	Y	**This study**	**HQ329662**	**BENT195-08**
				gaigeae	Big Bend Slider	1	1	V	III/IV	Y	**This study**	**HQ329663**	**BENT197-08**
				ornata	Ornate Slider	1	1	V	III/IV	N	**This study**	**HQ329665**	**BENT199-08**
				scripta	Pond Slider	2	1	LR/NT	III	Y	Russell and Beckenbach 2008	NC_011573, FJ392294	
				stejnegeri	Puerto Rican Slider	1	1	LR/NT	III/IV	N	**This study**	**HQ329666**	**BENT200-08**
				taylori	Cuatrocienegas Slider	1	1	E	III/IV	Y	**This study**	**HQ329667**	**BENT201-08**
				terrapen	Mexican/ Cat Island Slider	1	1	V	III/IV	N	**This study**	**HQ329668**	**BENT202-08**
				venusta	Meso-American/Belize Slider	2	2	NA	IV	N	**This study**	**HQ329664, HQ329669**	**BENT191-08, BENT 203-08**
				yaquia	Yaqui Slider	1	1	V	III/IV	N	**This study**	**HQ329670**	**BENT205-08**
		Geoemydidae			**Old World Pond Turtles/ American Wood Turtles**								
			Batagur	*baska*	Batagur/ Giant River Turtle	1	1	CE	I/II	N	**This study**	**HQ329671**	**BENT106-08**
				borneoensis	Painted/Three-Striped Batagur	1	1	CE	I/II	N	**This study**	**HQ329672**	**BENT109-08**
				dhongoka	Three-striped Roofed Turtle	1	1	E	I/II	Y	**This study**	**HQ329673**	**BENT258-09**
				kachuga	Red-crowned Roofed Turtle	1	1	CE	I/II	Y	**This study**	**HQ329674**	**BENT151-08**
				trivittata	Burmese Roofed Turtle	1	1	E	I/II	N	**This study**	**HQ329675**	**BENT152-08**
			Cuora	*amboinensis*	Southeast Asian Box Turtle	2	2	V	II	Y	Stuart and Parham 2004; Spinks and Shaffer 2007	AY357738, EF011465	
				aurocapitata	Yellow-headed Box Turtle	4	2	CE	IV	N	Stuart and Parham 2004; Parham et al 2004; Spinks and Shaffer 2007; Nie et al. unpublished	AY357740, AY590463, EF011466, NC_009509	
				flavomarginata	Yellow-margined Box Turtle	3	3	EN	I	Y	Stuart and Parham 2004; Parham et al 2004; Spinks and Shaffer 2007	AY357739, AY590459, EF011467	
				galbinifrons	Indochinese Box Turtle	22	8	CE	II	N	Parham et al. 2001; Parham et al. 2004; Stuart and Parham 2004; Spinks and Shaffer 2007	AF348265-AF348266, AY357742-AY357751, AY357753-AY357754, AY357756-AY357762, EF011477	
				mccordi	McCord's Box Turtle	5	1	CE	I	Y	Stuart and Parham 2004; Parham et al. 2004; Spinks and Shaffer 2007	AY357737, AY590456, EF011470-EF011472	
				mouhotii	Jagged-shelled Turtle	5	4	EN	I	Y	Parham et al. 2001; Spinks and Shaffer 2007; Zhang et al. 2008	AF348272-AF348274, EF011474, NC_010973	
				pani	Pan's Box Turtle	4	2	CE	IV	N	Parham et al. 2001; Stuart and Parham 2004; Spinks and Shaffer 2007	AY35774, AY590457, EF011475-EF011476	
				trifasciata	Chinese Three-striped Box Turtle	20	5	CE	IV	N	Parham et al. 2001; Spinks and Shaffer 2007	AF348270-AF348271, EF011478-EF011485, EF011487-EF011491, EF011494, EF011497-EF011498, EF011500-EF011501	
				yunnanensis	Yunnan Box Turtle	4	3	CE	II	N	Parham et al. 2004; He et al. 2007	AY590460, EF685037-EF685039	
				zhoui	Zhou's Box Turtle	18	1	CE	I	N	Parham et al. 2004; Spinks and Shaffer 2007; He et al. 2007;	AY593968-AY593969, EF011502-EF011515, EF685040, AY590458	
			Cyclemys	*atripons*	Striped Leaf Turtle	1	1	NA	I/II	N	Nie and Zhang unpublished	NC_010970	
				dentata	Asian Leaf Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329676**	**BENT117-08**
			Geoclemys	*hamiltonii*	Spotted Pond Turtle	1	1	V	I/II	Y	**This study**	**HQ329677**	**BENT135-08**
			Geoemyda	*japonica*	Ryukyu/ Okinawan Black-Breasted Leaf Turtle	1	1	E	I/II	Y	**This study**	**HQ329678**	**BENT136-08**
				spengleri	Black-Breasted Hill Turtle	1	1	E	I/II	Y	**This study**	**HQ329679**	**BENT137-08**
			Hardella	*thurjii*	Crowned River Turtle	1	1	V	I/II	Y	**This study**	**HQ329680**	**BENT147-08**
			Heosemys	*annandalii*	Yellow-headed Temple Turtle	1	1	E	I/II	Y	**This study**	**HQ329681**	**BENT259-09**
				depressa	Arakan Forest Turtle	1	1	CE	I/II	N	**This study**	**HQ329682**	**BENT260-09**
				grandis	Giant Asian Pond Turtle	1	1	V	I/II	Y	**This study**	**HQ329683**	**BENT148-08**
				spinosa	Spiny Turtle	1	1	E	I/II	Y	**This study**	**HQ329684**	**BENT261-09**
			Leucocephalon	*yuwonoi*	Sulawesi Forest Turtle	1	1	CE	I/II	Y	**This study**	**HQ329685**	**BENT155-08**
			Malayemys	*subtrijuga*	Malayan Snail-eating Turtle	1	1	V	I/II	Y	**This study**	**HQ329686**	**BENT262-09**
			Mauremys	*annamensis*	Annam Leaf Turtle	1	1	CE	III/IV	N	Feldman and Parham 2004	AY337346	
				caspica	Caspian Pond Turtle	2	1	NA	I	N	Feldman and Parham 2004	AY337347-AY337348	
				japonica	Japanese Pond Turtle	1	1	LR/NT	I/II	Y	Feldman and Parham 2004	AY337349	
				leprosa	Spanish Pond Turtle	2	1	NA	I	N	Feldman and Parham 2004	AY337350-AY337351	
				mutica	Asian Yellow Pond Turtle	5	2	EN	IV	N	Parham et al. 2001; Spinks and Shaffer 2007; Nie and Song unpublished	AF348260-AF348262, EF011464, NC_009330	
				nigricans	Red-necked Pond Turtle	1	1	EN	I/II	N	Parham et al. 2001	AF348264	
				reevesi	Chinese Three-keeled Pond Turtle	2	2	EN	I	N	Parham et al. 2001; Nie et al. unpublished	AF348263, NC_006082	
				rivulata	Balkan Pond Turtle	1	1	NA	I/II	N	Feldman and Parham 2004	AY337352	
				sinensis	Chinese Stripe-necked Turtle	2	2	E	I	N	Feldman and Parham 2004; This study	AY337353, **HQ329687**	**BENT165-08**
			Melanochelys	*tricarinata*	Tricarinate Hill Turtle	1	1	V	I/II	Y	**This study**	**HQ329688**	**BENT263-09**
				trijuga	Indian Black Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329689**	**BENT160-08**
			Morenia	*ocellata*	Bengal Eyed Turtle	1	1	V	I/II	Y	**This study**	**HQ329690**	**BENT264-09**
				petersi	Indian Eyed Turtle	1	1	V	I/II	Y	**This study**	**HQ329691**	**BENT161-08**
			Notochelys	*platynota*	Malayan Flat-shelled Turtle	1	1	V	I/II	N	**This study**	**HQ329692**	**BENT164-08**
			Orlitia	*borneensis*	Bornean River/Malaysian Giant Turtle	1	1	E	I/II	Y	**This study**	**HQ329693**	**BENT166-08**

Fig. A.3.: Table S1: Descriptive data for all taxa and sequences included in this study. 'N' indicates the number of individuals sequences for each species; 'H' indicates the number of haplotypes observed in each species; 'Distance' indicates the species' classification within the distance- based scheme described in the text; 'Diagnostic' indicates the presence ('Y') or absence ('N') of familylevel simple identifying characters in the species. References and accession numbers are in BOLD for novel sequences produced in this study.

Order	Suborder	Family	Genus	Species	Common Name	N	H	IUCN	Distance	Diagnostic	References	Genbank Accession Numbers	BOLD ID Numbers
Testudines	**Cryptodira**	**Geoemydidae**	*Pangshura*	*smithii*	Brown Roofed Turtle	1	1	LR/NT	I/II	N	**This study**	**HQ329694**	**BENT169-08**
				tecta	Indian Roofed Turtle	1	1	LR/LC	I/II	Y	**This study**	**HQ329695**	**BENT170-08**
				tentoria	Indian Tent Turtle	1	1	LR/LC	I/II	Y	**This study**	**HQ329696**	**BENT171-08**
			Rhinoclemmys	*annulata*	Brown Land Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329697**	**BENT184-08**
				areolata	Furrowed Wood Turtle	1	1	NT	I/II	N	**This study**	**HQ329698**	**BENT185-08**
				funerea	Black River Turtle	1	1	LR/NT	I/II	N	**This study**	**HQ329699**	**BENT186-08**
				nasuta	Large-nosed Wood Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329700**	**BENT265-09**
				rubida	Mexican Spotted Terrapin	1	1	NT	I/II	Y	**This study**	**HQ329701**	**BENT187-08**
			Sacalia	*bealei*	Beal's Four-eyed Turtle	1	1	E	I/II	Y	**This study**	**HQ329702**	**BENT188-08**
				quadriocellata	Four-eyed Turtle	1	1	E	I/II	Y	**This study**	**HQ329703**	**BENT189-08**
			Siebenrockiella	*crassicolis*	Malaysian Black Mud Turtle/Smiling Terrapin	1	1	V	I/II	Y	**This study**	**HQ329704**	**BENT190-08**
			Vijayachelys	*silvatica*	Cochin Forest Cane Turtle	1	1	E	I/II	Y	**This study**	**HQ329705**	**BENT266-09**
		Kinosternidae			**American Mud and Musk Turtles**								
			Claudius	*angustatus*	Narrow-bridged Musk Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329706**	**BENT267-09**
			Kinosternon	*acutum*	Tabasco Mud Turtle	1	1	LR/NT	III/IV	N	**This study**	**HQ329707**	**BENT154-08**
				alamosae	Alamos Mud Turtle	1	1	DD	I/II	Y	**This study**	**HQ329708**	**BENT268-09**
				angustipons	Central American Mud Turtle	1	1	V	I/II	N	**This study**	**HQ329709**	**BENT269-09**
				arizonense	Arizona Mud Turtle	2	1	LC	I	Y	**This study**	**HQ329710**	**BENT270-09, BENT282-09**
				chimalhuaca	Jalisco Mud Turtle	1	1	LC	I/II	Y	**This study**	**HQ329711**	**BENT271-09**
				creaseri	Creaser's Mud Turtle	2	1	LC	III	N	**This study**	**HQ329712**	**BENT272-09, BENT273-09**
				dunni	Dunn's Mud Turtle	1	1	V	I/II	N	**This study**	**HQ329713**	**BENT274-09**
				durangoense	Durango Mud Turtle	1	1	DD	I/II	Y	**This study**	**HQ329714**	**BENT275-09**
				flavescens	Yellow Mud Turtle	1	1	NA	I/II	Y	Parham et al. 2006a	DQ256379	
				hirtipes	Mexican Rough-footed Mud Turtle	1	1	LC	I/II	Y	**This study**	**HQ329716**	**BENT277-09**
				integrum	Mexican Mud Turtle	5	5	LC	III	N	**This study**	**HQ329715, HQ329717-HQ329720**	**BENT276-09, BENT278=09, BENT279-09, BENT280-09, BENT281-09**
			Staurotypus	*salvinii*	Pacific Coast Giant Musk Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329722**	**BENT283-09**
				triporcatus	Mexican Giant Musk Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329723**	**BENT285-09**
			Sternotherus	*depressus*	Flattened Musk Turtle	1	1	V	I/II	Y	**This study**	**HQ329724**	**BENT287-09**
		Platysternidae											
			Platysternon	*megacephalum*	**Big-headed Turtle**	2	1	EN	I		Parham et al. 2006a	DQ256377, NC_007970	
		Testudinidae			**Tortoises**								
			Dipsochelys	*dussumieri*	Aldabra Tortoise	1	1	V	I/II	Y	**This study**	**HQ329746**	**BENT131-08**
			Astrochelys	*radiata*	Radiated Tortoise	1	1	CE	I/II	Y	**This study**	**HQ329747**	**BENT308-09**
				yniphora	Angonoka	1	1	CE	I/II	Y	**This study**	**HQ329748**	**BENT309-09**
			Chelonoidis	*chilensis*	Chaco Tortoise	1	1	V	I/II	Y	**This study**	**HQ329749**	**BENT128-08**
				denticulata	South American Yellow-footed Tortoise	1	1	V	I/II	Y	**This study**	**HQ329750**	**BENT129-08**
				nigra	Galapagos Tortoise	1	1	V	I/II	Y	**This study**	**HQ329751**	**BENT132-08**
			Geochelone	*elegans*	Indian Star Tortoise	1	1	LR/LC	I/II	Y	**This study**	**HQ329752**	**BENT130-08**
				platynota	Burmese Star Tortoise	1	1	CE	I/II	N	**This study**	**HQ329753**	**BENT133-08**
				sulcata	African Spurred Tortoise	1	1	V	I/II	Y	**This study**	**HQ329754**	**BENT134-08**
			Gopherus	*agassizii*	Desert Tortoise	2	2	V	I	Y	**This study**	**HQ329755-HQ329756**	**BENT310-09, BENT 311-09**
				berlandieri	Texas Tortoise	5	1	LR/LC	I	Y	**This study**	**HQ329757**	**BENT312-09, BENT313-09, BENT314-09, BENT315-09, BENT316-09**
				flavomarginatus	Bolson's Tortoise	5	1	V	I	Y	**This study**	**HQ329758**	**BENT317-09, BENT318-09, BENT319-09, BENT320-09, BENT321-09**
				polyphemus	Gopher Tortoise	1	1	V	I/II	Y	**This study**	**HQ329759**	**BENT139-08**
			Homopus	*signatus*	Speckled Cape Tortoise	1	1	LR/NT	I/II	Y	**This study**	**HQ329760**	**BENT322-09**
			Indotestudo	*elongata*	Yellow-headed Tortoise	3	1	EN	III	N	Nie et al. unpublished; Parham et al. 2006b	DQ656607, DQ080043, NC_007695	
				forstenii	Forsten's Tortoise	2	1	EN	I	Y	Parham et al. 2006b	DQ080044, NC_007696	
				travancorica	Travancore Tortoise	1	1	V	III/IV	N	**This study**	**HQ329761**	**BENT150-08**
			Kinixys	*homeana*	Home's Hinge-back Tortoise	1	1	V	I/II	Y	**This study**	**HQ329762**	**BENT153-08**
				natalensis	Natal Hinge-back Tortoise	1	1	LR/NT	I/II	Y	**This study**	**HQ329763**	**BENT323-09**
			Malocochersus	*tornieri*	Tornier's Tortoise	2	1	V	I	Y	Parham et al. 2006b	DQ080042, NC_007700	
			Manouria	*emys*	Burmese Mountain Tortoise	2	1	EN	I	Y	Parham et al. 2006b	DQ080040, NC_007693	
				impressa	Impressed Tortoise	2	2	V	I	Y	Nie and Zhang unpublished; this study	NC_011815, **HQ329764**	**BENT159-08**
			Psammobates	*geometricus*	Geometric Tortoise	1	1	E	III/IV	Y	**This study**	**HQ329765**	**BENT324-09**
				pardalis	Leopard Tortoise	2	1	NA	III	N	Parham et al. 2006b	DQ080041, NC_007694	

Fig. A.4.: Table S1: Descriptive data for all taxa and sequences included in this study. 'N' indicates the number of individuals sequences for each species; 'H' indicates the number of haplotypes observed in each species; 'Distance' indicates the species' classification within the distance- based scheme described in the text; 'Diagnostic' indicates the presence ('Y') or absence ('N') of familylevel simple identifying characters in the species. References and accession numbers are in BOLD for novel sequences produced in this study.

Order	Suborder	Family	Genus	Species	Common Name	N	H	IUCN	Distance	Diagnostic	References	Genbank Accession Numbers	BOLD ID Numbers
Testudines	**Cryptodira**	**Testudinidae**	*Pyxis*	*arachnoides*	Common Spider Tortoise	1	1	CE	I/II	Y	**This study**	**HQ329766**	**BENT183-08**
				planicauda	Flat-shelled Spider Tortoise	1	1	CE	I/II	Y	**This study**	**HQ329767**	**BENT325-09**
			Testudo	*graeca*	Spur-thighed Tortoise	3	2	V	II	Y	Parham et al. 2006b	DQ080049-DQ080050, NC_007692	
				hermanni	Hermann's Tortoise	1	1	LR/NT	I/II	Y	Parham et al. 2006b	DQ080046	
				horsfieldii	Central Asian Tortoise	2	1	V	I	Y	Parham et al. 2006b	DQ080045, NC_007697	
				kleinmanni	Kleinmann's Tortoise	2	1	CE	I	Y	Parham et al. 2006b	DQ080048, NC_007699	
				marginata	Marginated Tortoise	2	1	LR/LC	I	Y	Parham et al. 2006b	DQ080047, NC_007698	
		Trionychidae			**Softshell Turtles**								
			Amyda	*cartilaginea*	Asiatic Softshell Turtle	2	2	V	II	N	**This study**	**HQ329768-HQ329769**	**BENT105-08, BENT325-09**
			Chitra	*chitra*	Striped Narrow-headed Softshell Turtle	1	1	CE	I/II	N	**This study**	**HQ329770**	**BENT327-09**
				indica	Narrow-headed Softshell Turtle	1	1	E	I/II	Y	**This study**	**HQ329771**	**BENT328-09**
			Cyclanorbis	*elegans*	Nubian Flapshell Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329772**	**BENT114-08**
				senegalensis	Senegal Flapshell Turtle	2	1	LR/NT	I	Y	**This study**	**HQ329773**	**BENT115-08, BENT116-08**
			Cycloderma	*frenatum*	Zambezi Flapshell Turtle	1	1	LR/NT	I/II	Y	**This study**	**HQ329774**	**BENT118-08**
			Dogania	*subplana*	Malayan Soft-shelled Turtle	2	1	LR/LC	I	Y	Farajallah et al. unpublished	AF366350, NC_002780	
			Lissemys	*punctata*	Indian Flapshell Turtle	2	2	LR/LC	I	Y	Tandon et al. unpublished; This study	EF050073, **HQ329775**	**BENT156-08**
				scutata	Burmese Flapshell Turtle	2	2	DD	I	Y	**This study**	**HQ329776-HQ329777**	**BENT157-08, BENT329-09**
			Nilssonia	*formosa*	Burmese Peacock Softshell Turtle	2	2	E	I	Y	**This study**	**HQ329778-HQ329779**	**BENT162-08, BENT 163-08**
				gangetica	Indian Softshell Turtle	1	1	V	I/II	Y	**This study**	**HQ329780**	**BENT330-09**
				hurum	Indian Peacock Softshell Turtle	2	1	V	I	Y	**This study**	**HQ329781**	**BENT331-09, BENT332-09**
			Palea	*steindachneri*	Wattle-necked Softshell Turtle	3	2	E	I	Y	**This study**	**HQ329782-HQ329783**	**BENT167-08, BENT 168-08, BENT333-09**
			Pelochelys	*bibroni*	New Guinea Giant Softshelled Turtle	1	1	V	I/II	Y	**This study**	**HQ329784**	**BENT172-08**
				cantorii	Cantor's Giant Softshell Turtle	1	1	E	I/II	Y	**This study**	**HQ329785**	**BENT173-08**
			Pelodiscus	*sinensis*	Chinese Softshell Turtle	1	1	V	I/II	Y	Jungt et al. 2006	AY962573	
			Rafetus	*euphraticus*	Euphrates Giant Softshell Turtle	1	1	E	I/II	N	**This study**	**HQ329786**	**BENT334-09**
				swinhoei	Yangtze Giant Softshell Turtle	1	1	CE	I/II	Y	**This study**	**HQ329787**	**BENT335-09**
			Trionyx	*triunguis*	African Softshell turtle	2	1	NA	I	Y	Amer and Kumazawa 2009	NC_012833, AB477345	

Fig. A.5.: Table S1: Descriptive data for all taxa and sequences included in this study. ’N’ indicates the number of individuals sequences for each species; ’H’ indicates the number of haplotypes observed in each species; ’Distance’ indicates the species’ classification within the distance- based scheme described in the text; ’Diagnostic’ indicates the presence (’Y’) or absence (’N’) of familylevel simple identifying characters in the species. References and accession numbers are in BOLD for novel sequences produced in this study.

A.2 Manuscript 2

Species	Family	ID/Sequences	No. Ind.	Country	Locality	Paper ND1	Paper CO1
Aeshna cyanea	Aeshnidae	Acy1- Acy2; Acy4	3	Germany	Hannover	New	New
		Acy04A	1	Germany	Braunschweig	New	New
Aeshna grandis	Aeshnidae	Aegr05A	1	Germany	Hannover	-	New
		Aegr2	1	Germany	Hannover	Rach et al. 2008	-
Aeshna mixta	Aeshnidae	Ami2 - Ami3	2	Germany	Hannover	New	New
Aeshna rileyi	Aeshnidae	Aeri142	2	Tanzania	Kilimanjaro, Machame, Semira Riv.	Rach et al. 2008	New
Anaciaeschna triangulifera	Aeshnidae	Anatri162	1	Tanzania	Pangani River	Rach et al. 2008	New
Anax ephippiger	Aeshnidae	Ae155	2	Tanzania	Pangani River	Rach et al. 2008	New
		Ae3	3	Namibia	Palmwag	Rach et al. 2008	New
		Ae21	5	Namibia	Tsaobis	Rach et al. 2008	New
Anax imperator	Aeshnidae	Ai21	3	Namibia	Tsaobis	Rach et al. 2008	New
		Ai16	4	Namibia	Tsauchab River	Rach et al. 2008	New
		Ai61	1	Namibia	Swakopmund, Swakopm. River, Elke Erb	Rach et al. 2008	New
		Ai98	3	Namibia	Baynes Mts.	Rach et al. 2008	New
Anax speratus	Aeshnidae	As11	4	Namibia	Naukluft	Rach et al. 2008	New
		As16	2	Namibia	Tsauchab River	Rach et al. 2008	New
Brachytron pratense	Aeshnidae	Brpr02	2	France	Saint Martin de Crau	Rach et al. 2008	New
Gynacantha usambarica	Aeshnidae	Gu25	3	Kenya	Buda Forest	Rach et al. 2008	-
		Gu28	2	Kenya	Shimba Hills, Mwele Forest	Rach et al. 2008	-
		Gu49	3	Tanzania	Pemba, Ngezi Forest	Rach et al. 2008	-
		Gu87	1	Tanzania	Zansibar, Jozani Forest	Rach et al. 2008	-
Gynacantha villosa	Aeshnidae	Gyvill60B	1	Kenya	Arabuke Sokoke Forest	Rach et al. 2008	New
Paragomphus geneii	Gomphidae	Pg3	3	Namibia	Palmwag	Rach et al. 2008	-
		Pg98	2	Namibia	Baynes Mts.	Rach et al. 2008	-
Crocothemis erythraea	Libellulidae	Ce3	3	Namibia	Palmwag	Rach et al. 2008	New
		Ce7	1	Namibia	Tsauchab River	Rach et al. 2008	New
		Ce32	3	Namibia	Ongongo	Rach et al. 2008	New
Crocothemis sanguinolenta	Libellulidae	Cs7	3	Namibia	Ongongo	New	New
		Cs98	3	Namibia	Baynes Mts.	New	New
Nesciothemis farinosum	Libellulidae	Nf3	2	Namibia	Palmwag	New	New
		Nf119	3	Namibia	Popa Falls	New	New
Orthetrum brachiale	Libellulidae	Ob1	1	Namibia	Van-Bach-Dam	Rach et al. 2008	New
		Ob32	2	Namibia	Waterberg	Rach et al. 2008	New
Orthetrum chrysostigma	Libellulidae	Oc68	2	Namibia	Fransfontein	New	New
		Oc1	2	Namibia	Van-Bach-Dam	Rach et al. 2008	New
Orthetrum coerulescens	Libellulidae	OcoeRM	3	Italy	Pontecorvo, River Melfa	New	New
		OcoeSZ	3	Germany	Salzgitter Engelnstedt	New	New
		OcoeCE	3	Germany	Marwede, Celle	Rach et al. 2008	New
Orthetrum julia falsum	Libellulidae	Oj16	5	Namibia	Tsauchab River	Rach et al. 2008	New
		Oj32	5	Namibia	Waterberg	New	New
Orthetrum trinacria	Libellulidae	Ot1	2	Namibia	Van-Bach-Dam	Rach et al. 2008	New
		Ot3	3	Namibia	Palmwag	Rach et al. 2008	New
Sympetrum sanguineum	Libellulidae	Ssa1 - Ssa2	2	Spain	Pontevedra	New	-
Trithemis annulata	Libellulidae	Ta119	1	Namibia	Popa Falls	Rach et al. 2008	Damm et al. 2010
		Ta120	2	Namibia	Rehoboth	Rach et al. 2008	Damm et al. 2010
Trithemis arteriosa	Libellulidae	Tart39A	1	Kenya	Tsavo West	Damm et al. 2010c	New
Trithemis donaldsoni	Libellulidae	Tdo1 - Tdo5	5	Namibia	Van-Bach-Dam	New	New
Trithemis furva	Libellulidae	TfurEth	2	Ethiopia	Nekemte	Damm et al. 2010a	Damm et al. 2010
		TfurSAB	1	South Africa	Wakkerstroom	Damm et al. 2010a	Damm et al. 2010
Trithemis grouti	Libellulidae	Tgr44	1	Congo	Lukomete	Damm et al. 2010a	New
		Tgr98	1	Congo	Kimwenza	Damm et al. 2010a	New
Trithemis hecate	Libellulidae	The119	3	Namibia	Popa Falls	Rach et al. 2008	-
		The221	2			Rach et al. 2008	-
Trithemis kirbyi	Libellulidae	Tk3	1	Namibia	Palmwag	Rach et al. 2008	New
		Tk32	2	Namibia	Waterberg	New	New
		Tk74	1	Kenya	Sambu River	Rach et al. 2008	New
Trithemis morrisoni	Libellulidae	Tst119	5	Namibia	Popa Falls	Damm et al. 2010b	Damm et al. 2010
Trithemis nuptialis	Libellulidae	Tnup0017	1	Congo	Lingomo	Damm et al. 2010a	Damm et al. 2010
		Tnup0043	1	Congo	Lukomete	Damm et al. 2010a	Damm et al. 2010
Trithemis palustris	Libellulidae	Tst128	4	Namibia	Kwando	Damm et al. 2010b	Damm et al. 2010
Trithemis stictica	Libellulidae	Tst118	5	Namibia	Zebra River	Damm et al. 2010b	Damm et al. 2010
		Tst11	1	Namibia	Naukluft	Damm et al. 2010b	Damm et al. 2010
		Tst17	1	Kenya	Kiboko River	Damm et al. 2010a	Damm et al. 2010

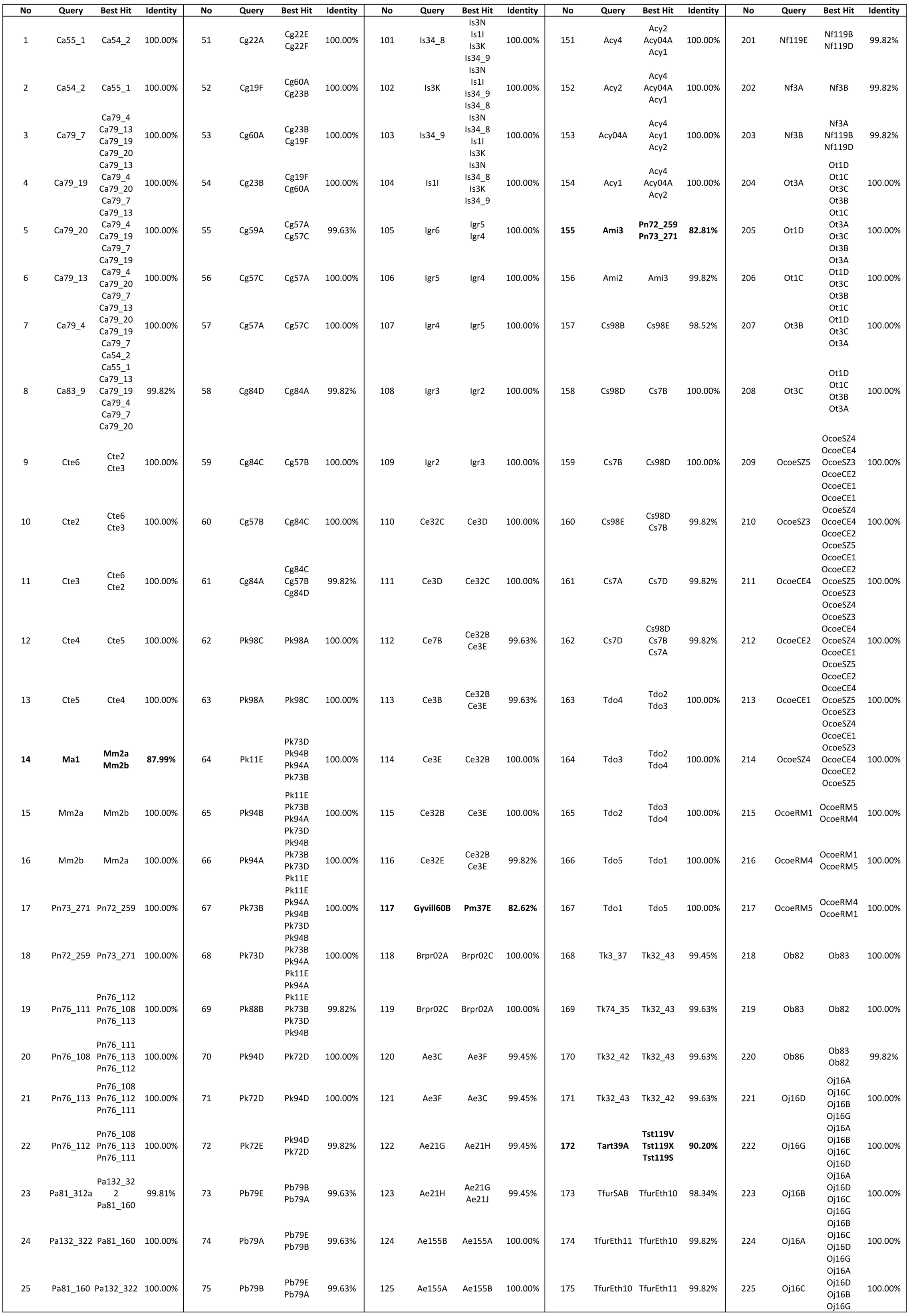
Fig. A.6.: Table S1: Anisoptera

Species	Family	ID/Sequences	No. Ind.	Country	Locality	Paper ND1	Paper CO1
Calopteryx haemorrhoidales	Calopterygidae	ch8	4	France	Saint-Martin-de-Crau	New	-
		ch7	2	Italy	Pontecorvo, Rivor Forum Quesa	New	-
		ch6	2	Spain	Lourizan; Porte v. Luc	New	-
		ch1 - ch3, ch6	4	Italy	Pontecorvo, River Forma Quesa	New	-
Calopteryx splendens	Calopetrygidae	cs2 - cs5	4	Italy	Pontecorvo, River Forma Quesa	Rach et al. 2008	-
Platycypha auripes	Chlorocyphidae	Pau2; Pau4	2	Tanzania	Uzambara Mts,Amani,Sigi Valley	Rach et al. 2008	New
Platycypha caligata	Chlorocyphidae	Pc92	4	Kenya	Lake Chala	Rach et al. 2008	New
		Pc39	1	Kenya	Tsavo West, Mzima	Rach et al. 2008	New
		Pc134	1	Malawi	Lake Malawi	Rach et al. 2008	New
Ceriagrion tenellum	Coenagrionidae	Cte2 - Cte6	5	Spain	Pontevedra	New	New
Enallagma cyathigerum	Coenagrionidae	Ecy1 - Ecy5	5	Spain	Pontevedra	New	New
Ischnura graellsii	Coenagrionidae	IgR1 - Igr6	5	Spain	Pontevedra	New	New
Ischnura senegalensis	Coenagrionidae	Is1	1	Namibia	Van-Bach-Dam	-	New
		Is3	2	Namibia	Palmwag	-	New
		Is34	2	Kenya	Tsavo West NP, L.Jipe	-	New
Leptagrion elongatum	Coenagrionidae	Le4	1	Brasil		New	New
Pseudagrion acaciae	Coenagrionidae	Pa81	3	Tanzania	Pangani River	Rach et al. 2008	New
		Pa132	1	Tanzania	Rufiji, Ruhoi River	Rach et al. 2008	New
Pseudagrion bicoerulans	Coenagrionidae	Pb77	4	Kenya	Mt.Elgon, Rongai River	New	New
		Pb78	4	Kenya	Aberdare Mts, River	New	New
		Pb79	3	Tanzania	Kilimanjaro,Machame,Semira Valley	Rach et al. 2008	New
		Pb113	4	Kenya	Mt.Kenya, Loruku	New	New
Pseudagrion kersteni	Coenagrionidae	Pk72	2	Kenya	Kiboko River, Hunter`s	Dijkstra et al. 2007	New
		Pk73	2	Kenya	Tsavo West, Mzima Springs	New	New
		Pk88	1	Tanzania	Rufiji Delta, Kichi Stream	Dijkstra et al. 2007	New
		Pk94	3	Tanzania	East Usambara Mts,Amani Pond	New	New
		Pk11	1	Namibia	Naukluft	New	New
		Pk98	2	Namibia	Baynes Mts	Rach et al. 2008	New
Pseudagrion massaicum	Coenagrionidae	Pm1	1	Namibia	Van-Bach-Dam	New	New
		Pm15	2	Namibia	Tsauchab River	New	New
		Pm16	2	Namibia	Kuiseb River	New	New
		Pm37	6	Kenya	Shimba Hills, Pemba	New	New
		Pm72	2	Kenya	Kiboko River, Hunter`s	Rach et al. 2008	New
Pseudagrion niloticum	Coenagrionidae	Pn73	1	Kenya	Tsavo West, Mzima	Rach et al. 2008	New
		Pn72	1	Kenya	Kiboko River, Hunter`s	Rach et al. 2008	New
		Pn76	4	Kenya	Ewaso, Nyiro River, Nguruman	Rach et al. 2008	New
Teinobasis alluaudi	Coenagrionidae	Tba25	1	Kenya	Buda Forest	New	New
		Tba49	2	Tanzania	Pemba, Ngezi Forest	New	New
		Tba87	3	Tanzania	Zansibar, Jozani Forest	New	New
Chlorocnemis abbotti	Protoneurinae	Ca54	1	Tanzania	Uluguru Mts,Pandanus For.	New	New
		Ca55	1	Tanzania	Udzungwa Mts, Sonje	New	New
		Ca79	5	Tanzania	Kilimanjaro,Machame,Semira Valley	New	New
		Ca83	1	Tanzania	Uzambara Mts,Amani,Sigi Valley	New	New
Coryphagrion grandis	Pseudostigmatidae	Cg19	2	Kenya	Arabuke Sokoke Forest	Groeneveldet al. 2007	New
		Cg22	3	Kenya	Bandas, Shimba Hills	New	New
		Cg23	1	Kenya	Muhaka Forest, Ukunda	New	New
		Cg57	3	Tanzania	Kichi Hills	Groeneveldet al. 2007	New
		Cg59	1	Tanzania	Kipengoma	New	New
		Cg60	1	Tanzania	Uzambara Mts,Amani,Sigi Valley	New	New
		Cg84	3	Tanzania	Rufiji, Kichi Forest	Groeneveldet al. 2007	New
Mecistogaster asticta	Pseudostigmatidae	Ma1	1	Brasil		New	New
Mecistogaster martinezi	Pseudostigmatidae	Mm2	2	Brasil		New	New

Fig. A.7.: Table S1: Zygoptera

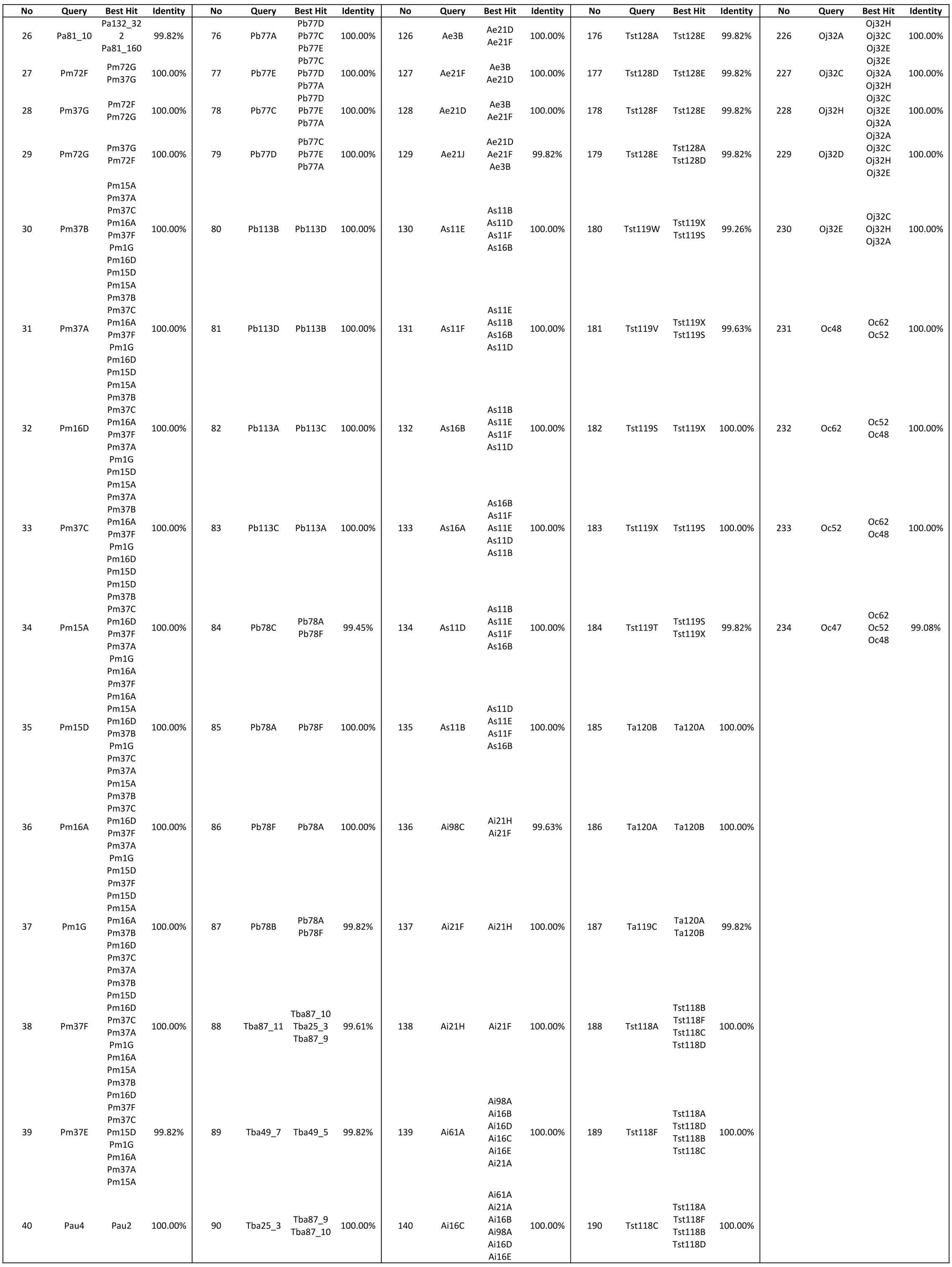
ND1 - Authors	
AUTHORS	Damm, S., Dijkstra, K.D. and Hadrys, H.
TITLE	Red drifters and dark residents: the phylogeny and ecology of a Plio-Pleistocene dragonfly radiation reflects Africa's changing environment (Odonata, Libellulidae, *Trithemis*)
JOURNAL	Mol. Phylogenet. Evol. 54 (3), 870-882 (2010)
AUTHORS	Damm, S., Schierwater, B. and Hadrys, H.
TITLE	An integrative approach to species discovery in odonates: from character-based DNA barcoding to ecology
JOURNAL	Mol. Ecol. 19 (18), 3881-3893 (2010)
AUTHORS	Damm, S. and Hadrys, H.
TITLE	Odonates in the desert: Tracking the dispersal pathways of a desert inhabiting dragonfly, *Trithemis arteriosa*
JOURNAL	Unpublished
AUTHORS	Groeneveld, L.F., Clausnitzer, V. and Hadrys, H.
TITLE	Convergent evolution of gigantism in damselflies of Africa and South America? Evidence from nuclear and mitochondrial sequence data
JOURNAL	Mol. Phylogenet. Evol. 42 (2), 339-346 (2007)
AUTHORS	Dijkstra, K.-D.B., Groeneveld, L.F., Clausnitzer, V. and Hadrys, H.
TITLE	The Pseudagrion split: molecular phylogeny confirms the morphological and ecological dichotomy of Africa's most diverse damselfly genus (Odonata, Coenagrionidae)
JOURNAL	Int. J. Odonatol. (2007) In press
AUTHORS	Rach, J., DeSalle, R., Sarkar, I.N., Schierwater, B. and Hadrys, H.
TITLE	Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata
JOURNAL	Proc. Biol. Sci. 275 (1632), 237-247 (2008)
CO1 -Authors	
AUTHORS	Damm, S., Schierwater, B. and Hadrys, H.
TITLE	An integrative approach to species discovery in odonates: from character-based DNA barcoding to ecology
JOURNAL	Mol. Ecol. 19 (18), 3881-3893 (2010)

Tab. A.1.: Table S1: Authors



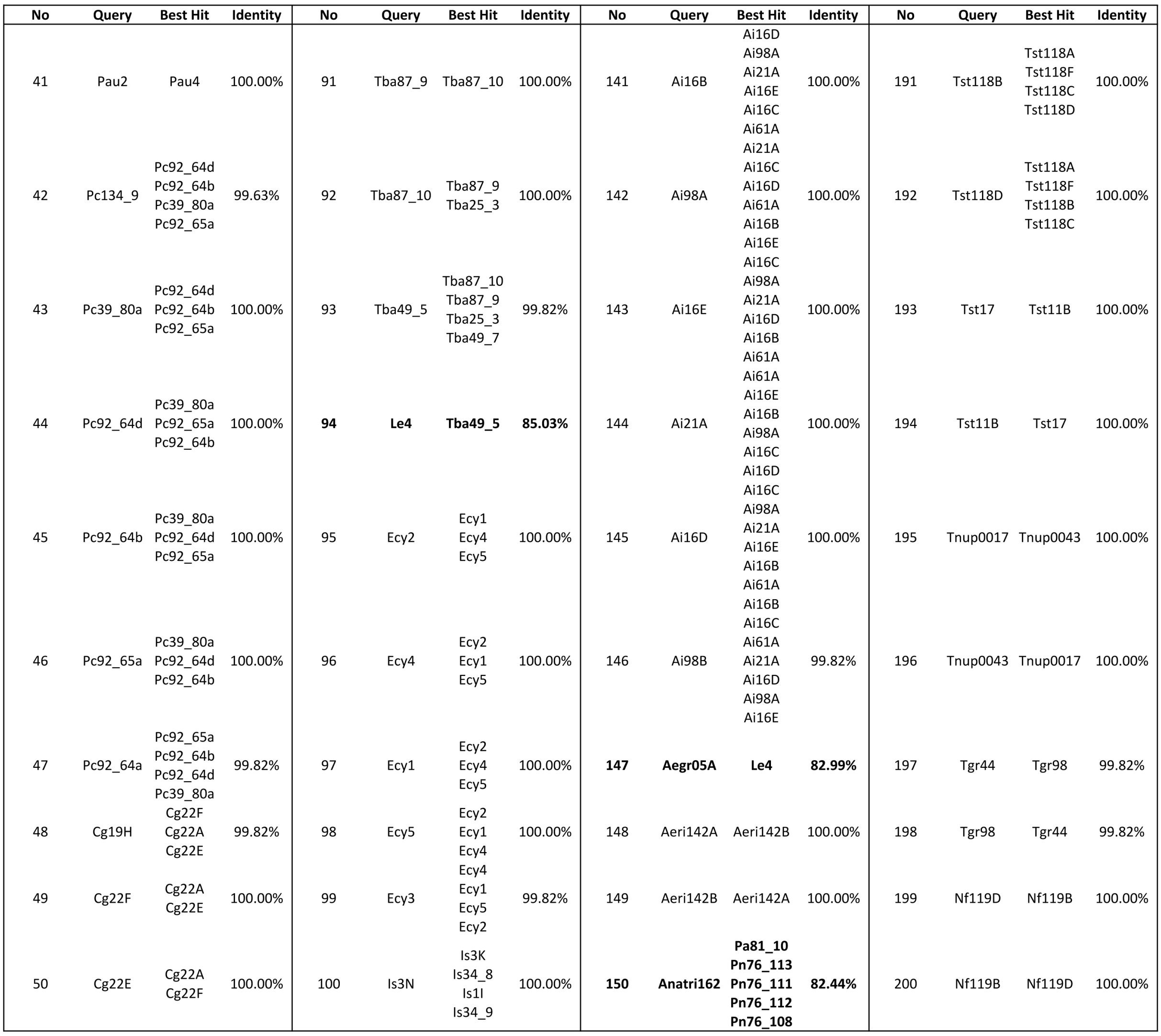
No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity
1	Ca55_1	Ca54_2	100.00%	51	Cg22A	Cg22E Cg22F	100.00%	101	Is34_8	Is3N Is1I Is3K Is34_9	100.00%	151	Acy4	Acy2 Acy04A Acy1	100.00%	201	Nf119E	Nf119B Nf119D	99.82%
2	Ca54_2	Ca55_1	100.00%	52	Cg19F	Cg60A Cg23B	100.00%	102	Is3K	Is3N Is1I Is34_9 Is34_8	100.00%	152	Acy2	Acy4 Acy04A Acy1	100.00%	202	Nf3A	Nf3B	99.82%
3	Ca79_7	Ca79_4 Ca79_13 Ca79_19 Ca79_20	100.00%	53	Cg60A	Cg23B Cg19F	100.00%	103	Is34_9	Is3N Is34_8 Is1I Is3K	100.00%	153	Acy04A	Acy4 Acy1 Acy2	100.00%	203	Nf3B	Nf3A Nf119B Nf119D	99.82%
4	Ca79_19	Ca79_13 Ca79_4 Ca79_20 Ca79_7	100.00%	54	Cg23B	Cg19F Cg60A	100.00%	104	Is1I	Is3N Is34_8 Is3K Is34_9	100.00%	154	Acy1	Acy4 Acy04A Acy2	100.00%	204	Ot3A	Ot1D Ot1C Ot3C Ot3B	100.00%
5	Ca79_20	Ca79_13 Ca79_4 Ca79_19 Ca79_7	100.00%	55	Cg59A	Cg57A Cg57C	99.63%	105	Igr6	Igr5 Igr4	100.00%	**155**	**Ami3**	**Pn72_259** **Pn73_271**	**82.81%**	205	Ot1D	Ot1C Ot3A Ot3C Ot3B	100.00%
6	Ca79_13	Ca79_19 Ca79_4 Ca79_20 Ca79_7	100.00%	56	Cg57C	Cg57A	100.00%	106	Igr5	Igr4	100.00%	156	Ami2	Ami3	99.82%	206	Ot1C	Ot3A Ot1D Ot3C Ot3B	100.00%
7	Ca79_4	Ca79_13 Ca79_20 Ca79_19 Ca79_7	100.00%	57	Cg57A	Cg57C	100.00%	107	Igr4	Igr5	100.00%	157	Cs98B	Cs98E	98.52%	207	Ot3B	Ot1C Ot1D Ot3C Ot3A	100.00%
8	Ca83_9	Ca54_2 Ca55_1 Ca79_13 Ca79_19 Ca79_4 Ca79_7 Ca79_20	99.82%	58	Cg84D	Cg84A	99.82%	108	Igr3	Igr2	100.00%	158	Cs98D	Cs7B	100.00%	208	Ot3C	Ot1D Ot1C Ot3B Ot3A	100.00%
9	Cte6	Cte2 Cte3	100.00%	59	Cg84C	Cg57B	100.00%	109	Igr2	Igr3	100.00%	159	Cs7B	Cs98D	100.00%	209	OcoeSZ5	OcoeSZ4 OcoeCE4 OcoeSZ3 OcoeCE2 OcoeCE1	100.00%
10	Cte2	Cte6 Cte3	100.00%	60	Cg57B	Cg84C	100.00%	110	Ce32C	Ce3D	100.00%	160	Cs98E	Cs98D Cs7B	99.82%	210	OcoeSZ3	OcoeCE1 OcoeSZ4 OcoeCE4 OcoeCE2 OcoeSZ5	100.00%
11	Cte3	Cte6 Cte2	100.00%	61	Cg84A	Cg84C Cg57B Cg84D	99.82%	111	Ce3D	Ce32C	100.00%	161	Cs7A	Cs7D	99.82%	211	OcoeCE4	OcoeCE1 OcoeCE2 OcoeSZ5 OcoeSZ3 OcoeSZ4	100.00%
12	Cte4	Cte5	100.00%	62	Pk98C	Pk98A	100.00%	112	Ce7B	Ce32B Ce3E	99.63%	162	Cs7D	Cs98D Cs7B Cs7A	99.82%	212	OcoeCE2	OcoeSZ3 OcoeCE4 OcoeSZ4 OcoeCE1 OcoeSZ5	100.00%
13	Cte5	Cte4	100.00%	63	Pk98A	Pk98C	100.00%	113	Ce3B	Ce32B Ce3E	99.63%	163	Tdo4	Tdo2 Tdo3	100.00%	213	OcoeCE1	OcoeCE2 OcoeCE4 OcoeSZ5 OcoeSZ3 OcoeSZ4	100.00%
14	**Ma1**	**Mm2a** **Mm2b**	**87.99%**	64	Pk11E	Pk73D Pk94B Pk94A Pk73B	100.00%	114	Ce3E	Ce32B	100.00%	164	Tdo3	Tdo2 Tdo4	100.00%	214	OcoeSZ4	OcoeCE1 OcoeSZ3 OcoeCE4 OcoeCE2 OcoeSZ5	100.00%
15	Mm2a	Mm2b	100.00%	65	Pk94B	Pk11E Pk73B Pk94A Pk73D	100.00%	115	Ce32B	Ce3E	100.00%	165	Tdo2	Tdo3 Tdo4	100.00%	215	OcoeRM1	OcoeRM5 OcoeRM4	100.00%
16	Mm2b	Mm2a	100.00%	66	Pk94A	Pk94B Pk73B Pk73D Pk11E	100.00%	116	Ce32E	Ce32B Ce3E	99.82%	166	Tdo5	Tdo1	100.00%	216	OcoeRM4	OcoeRM1 OcoeRM5	100.00%
17	Pn73_271	Pn72_259	100.00%	67	Pk73B	Pk11E Pk94A Pk94B Pk73D	100.00%	**117**	**Gyvill60B**	**Pm37E**	**82.62%**	167	Tdo1	Tdo5	100.00%	217	OcoeRM5	OcoeRM4 OcoeRM1	100.00%
18	Pn72_259	Pn73_271	100.00%	68	Pk73D	Pk94B Pk73B Pk94A Pk11E	100.00%	118	Brpr02A	Brpr02C	100.00%	168	Tk3_37	Tk32_43	99.45%	218	Ob82	Ob83	100.00%
19	Pn76_111	Pn76_112 Pn76_108 Pn76_113	100.00%	69	Pk88B	Pk94A Pk11E Pk73B Pk73D Pk94B	99.82%	119	Brpr02C	Brpr02A	100.00%	169	Tk74_35	Tk32_43	99.63%	219	Ob83	Ob82	100.00%
20	Pn76_108	Pn76_111 Pn76_113 Pn76_112	100.00%	70	Pk94D	Pk72D	100.00%	120	Ae3C	Ae3F	99.45%	170	Tk32_42	Tk32_43	99.63%	220	Ob86	Ob83 Ob82	99.82%
21	Pn76_113	Pn76_108 Pn76_112 Pn76_111	100.00%	71	Pk72D	Pk94D	100.00%	121	Ae3F	Ae3C	99.45%	171	Tk32_43	Tk32_42	99.63%	221	Oj16D	Oj16A Oj16C Oj16B Oj16G	100.00%
22	Pn76_112	Pn76_108 Pn76_113 Pn76_111	100.00%	72	Pk72E	Pk94D Pk72D	99.82%	122	Ae21G	Ae21H	99.45%	**172**	**Tart39A**	**Tst119V** **Tst119X** **Tst119S**	**90.20%**	222	Oj16G	Oj16A Oj16B Oj16C Oj16D	100.00%
23	Pa81_312a	Pa132_322 Pa81_160	99.81%	73	Pb79E	Pb79B Pb79A	99.63%	123	Ae21H	Ae21G Ae21J	99.45%	173	TfurSAB	TfurEth10	98.34%	223	Oj16B	Oj16A Oj16D Oj16C Oj16G	100.00%
24	Pa132_322	Pa81_160	100.00%	74	Pb79A	Pb79E Pb79B	99.63%	124	Ae155B	Ae155A	100.00%	174	TfurEth11	TfurEth10	99.82%	224	Oj16A	Oj16B Oj16C Oj16D Oj16G	100.00%
25	Pa81_160	Pa132_322	100.00%	75	Pb79B	Pb79E Pb79A	99.63%	125	Ae155A	Ae155B	100.00%	175	TfurEth10	TfurEth11	99.82%	225	Oj16C	Oj16A Oj16D Oj16B Oj16G	100.00%

Fig. A.8.: Table S2A: CO1-L10, Results for CO1 leave one out test.



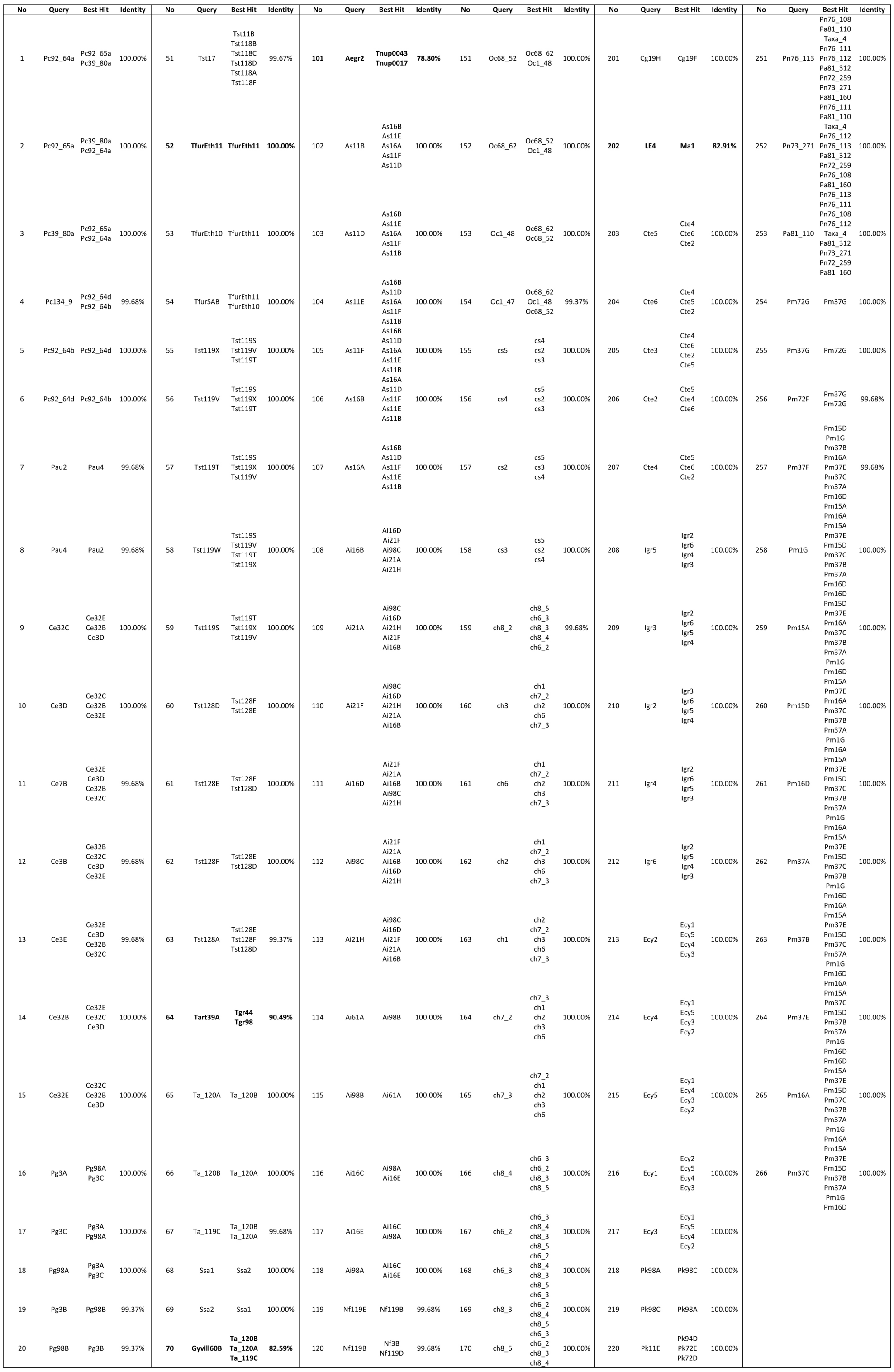
No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity
26	Pa81_10	Pa132_322 Pa81_160	99.82%	76	Pb77A	Pb77D Pb77C Pb77E	100.00%	126	Ae3B	Ae21D Ae21F	100.00%	176	Tst128A	Tst128E	99.82%	226	Oj32A	Oj32H Oj32C Oj32E	100.00%
27	Pm72F	Pm72G Pm37G	100.00%	77	Pb77E	Pb77C Pb77D Pb77A	100.00%	127	Ae21F	Ae3B Ae21D	100.00%	177	Tst128D	Tst128E	99.82%	227	Oj32C	Oj32E Oj32A Oj32H	100.00%
28	Pm37G	Pm72F Pm72G	100.00%	78	Pb77C	Pb77D Pb77E Pb77A	100.00%	128	Ae21D	Ae3B Ae21F	100.00%	178	Tst128F	Tst128E	99.82%	228	Oj32H	Oj32C Oj32E Oj32A	100.00%
29	Pm72G	Pm37G Pm72F	100.00%	79	Pb77D	Pb77C Pb77E Pb77A	100.00%	129	Ae21J	Ae21D Ae21F Ae3B	99.82%	179	Tst128E	Tst128A Tst128D	99.82%	229	Oj32D	Oj32A Oj32C Oj32H Oj32E	100.00%
30	Pm37B	Pm15A Pm37A Pm37C Pm16A Pm37F Pm1G Pm16D Pm15D	100.00%	80	Pb113B	Pb113D	100.00%	130	As11E	As11B As11D As11F As16B	100.00%	180	Tst119W	Tst119X Tst119S	99.26%	230	Oj32E	Oj32C Oj32H Oj32A	100.00%
31	Pm37A	Pm15A Pm37B Pm37C Pm16A Pm37F Pm1G Pm16D Pm15D	100.00%	81	Pb113D	Pb113B	100.00%	131	As11F	As11E As11B As16B As11D	100.00%	181	Tst119V	Tst119X Tst119S	99.63%	231	Oc48	Oc62 Oc52	100.00%
32	Pm16D	Pm15A Pm37B Pm37C Pm16A Pm37F Pm37A Pm1G Pm15D	100.00%	82	Pb113A	Pb113C	100.00%	132	As16B	As11B As11E As11F As11D	100.00%	182	Tst119S	Tst119X	100.00%	232	Oc62	Oc52 Oc48	100.00%
33	Pm37C	Pm15A Pm37A Pm37B Pm16A Pm37F Pm1G Pm16D Pm15D	100.00%	83	Pb113C	Pb113A	100.00%	133	As16A	As16B As11F As11E As11D As11B	100.00%	183	Tst119X	Tst119S	100.00%	233	Oc52	Oc62 Oc48	100.00%
34	Pm15A	Pm15D Pm37B Pm37C Pm16D Pm37F Pm37A Pm1G Pm16A	100.00%	84	Pb78C	Pb78A Pb78F	99.45%	134	As11D	As11B As11E As11F As16B	100.00%	184	Tst119T	Tst119S Tst119X	99.82%	234	Oc47	Oc62 Oc52 Oc48	99.08%
35	Pm15D	Pm37F Pm16A Pm15A Pm16D Pm37B Pm1G Pm37C Pm37A	100.00%	85	Pb78A	Pb78F	100.00%	135	As11B	As11D As11E As11F As16B	100.00%	185	Ta120B	Ta120A	100.00%				
36	Pm16A	Pm15A Pm37B Pm37C Pm16D Pm37F Pm37A Pm1G Pm15D	100.00%	86	Pb78F	Pb78A	100.00%	136	Ai98C	Ai21H Ai21F	99.63%	186	Ta120A	Ta120B	100.00%				
37	Pm1G	Pm37F Pm15D Pm15A Pm16A Pm37B Pm16D Pm37C Pm37A	100.00%	87	Pb78B	Pb78A Pb78F	99.82%	137	Ai21F	Ai21H	100.00%	187	Ta119C	Ta120A Ta120B	99.82%				
38	Pm37F	Pm37B Pm15D Pm16D Pm37C Pm37A Pm1G Pm16A Pm15A	100.00%	88	Tba87_11	Tba87_10 Tba25_3 Tba87_9	99.61%	138	Ai21H	Ai21F	100.00%	188	Tst118A	Tst118B Tst118F Tst118C Tst118D	100.00%				
39	Pm37E	Pm37B Pm16D Pm37F Pm37C Pm15D Pm1G Pm16A Pm37A Pm15A	99.82%	89	Tba49_7	Tba49_5	99.82%	139	Ai61A	Ai98A Ai16B Ai16D Ai16C Ai16E Ai21A	100.00%	189	Tst118F	Tst118A Tst118D Tst118B Tst118C	100.00%				
40	Pau4	Pau2	100.00%	90	Tba25_3	Tba87_9 Tba87_10	100.00%	140	Ai16C	Ai61A Ai21A Ai16B Ai98A Ai16D Ai16E	100.00%	190	Tst118C	Tst118A Tst118F Tst118B Tst118D	100.00%				

Fig. A.9.: Table S2B: CO1-L10, Results for CO1 leave one out test.



No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity
41	Pau2	Pau4	100.00%	91	Tba87_9	Tba87_10	100.00%	141	Ai16B	Ai16D Ai98A Ai21A Ai16E Ai16C Ai61A	100.00%	191	Tst118B	Tst118A Tst118F Tst118C Tst118D	100.00%
42	Pc134_9	Pc92_64d Pc92_64b Pc39_80a Pc92_65a	99.63%	92	Tba87_10	Tba87_9 Tba25_3	100.00%	142	Ai98A	Ai21A Ai16C Ai16D Ai61A Ai16B Ai16E	100.00%	192	Tst118D	Tst118A Tst118F Tst118B Tst118C	100.00%
43	Pc39_80a	Pc92_64d Pc92_64b Pc92_65a	100.00%	93	Tba49_5	Tba87_10 Tba87_9 Tba25_3 Tba49_7	99.82%	143	Ai16E	Ai16C Ai98A Ai21A Ai16D Ai16B Ai61A	100.00%	193	Tst17	Tst11B	100.00%
44	Pc92_64d	Pc39_80a Pc92_65a Pc92_64b	100.00%	**94**	**Le4**	**Tba49_5**	**85.03%**	144	Ai21A	Ai61A Ai16E Ai16B Ai98A Ai16C Ai16D	100.00%	194	Tst11B	Tst17	100.00%
45	Pc92_64b	Pc39_80a Pc92_64d Pc92_65a	100.00%	95	Ecy2	Ecy1 Ecy4 Ecy5	100.00%	145	Ai16D	Ai16C Ai98A Ai21A Ai16E Ai16B Ai61A	100.00%	195	Tnup0017	Tnup0043	100.00%
46	Pc92_65a	Pc39_80a Pc92_64d Pc92_64b	100.00%	96	Ecy4	Ecy2 Ecy1 Ecy5	100.00%	146	Ai98B	Ai16B Ai16C Ai61A Ai21A Ai16D Ai98A Ai16E	99.82%	196	Tnup0043	Tnup0017	100.00%
47	Pc92_64a	Pc92_65a Pc92_64b Pc92_64d Pc39_80a	99.82%	97	Ecy1	Ecy2 Ecy4 Ecy5	100.00%	**147**	**Aegr05A**	**Le4**	**82.99%**	197	Tgr44	Tgr98	99.82%
48	Cg19H	Cg22F Cg22A Cg22E	99.82%	98	Ecy5	Ecy2 Ecy1 Ecy4	100.00%	148	Aeri142A	Aeri142B	100.00%	198	Tgr98	Tgr44	99.82%
49	Cg22F	Cg22A Cg22E	100.00%	99	Ecy3	Ecy4 Ecy1 Ecy5 Ecy2	99.82%	149	Aeri142B	Aeri142A	100.00%	199	Nf119D	Nf119B	100.00%
50	Cg22E	Cg22A Cg22F	100.00%	100	Is3N	Is3K Is34_8 Is1I Is34_9	100.00%	**150**	**Anatri162**	**Pa81_10** **Pn76_113** **Pn76_111** **Pn76_112** **Pn76_108**	**82.44%**	200	Nf119B	Nf119D	100.00%

Fig. A.10.: Table S2C: CO1-L10, Results for CO1 leave one out test.



No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity
1	Pc92_64a	Pc92_65a Pc39_80a	100.00%	51	Tst17	Tst11B Tst118B Tst118C Tst118D Tst118A Tst118F	99.67%	**101**	**Aegr2**	**Tnup0043 Tnup0017**	**78.80%**	151	Oc68_52	Oc68_62 Oc1_48	100.00%	201	Cg19H	Cg19F	100.00%	251	Pn76_113	Pn76_108 Pa81_110 Taxa_4 Pn76_111 Pn76_112 Pa81_312 Pn72_259 Pn73_271 Pa81_160	100.00%
2	Pc92_65a	Pc39_80a Pc92_64a	100.00%	**52**	**TfurEth11**	**TfurEth11**	**100.00%**	102	As11B	As16B As11E As16A As11F As11D	100.00%	152	Oc68_62	Oc68_52 Oc1_48	100.00%	**202**	**LE4**	**Ma1**	**82.91%**	252	Pn73_271	Pn76_111 Pa81_110 Taxa_4 Pn76_112 Pn76_113 Pa81_312 Pn72_259 Pn76_108 Pa81_160	100.00%
3	Pc39_80a	Pc92_65a Pc92_64a	100.00%	53	TfurEth10	TfurEth11	100.00%	103	As11D	As16B As11E As16A As11F As11B	100.00%	153	Oc1_48	Oc68_62 Oc68_52	100.00%	203	Cte5	Cte4 Cte6 Cte2	100.00%	253	Pa81_110	Pn76_113 Pn76_111 Pn76_108 Pn76_112 Taxa_4 Pa81_312 Pn73_271 Pn72_259 Pa81_160	100.00%
4	Pc134_9	Pc92_64d Pc92_64b	99.68%	54	TfurSAB	TfurEth11 TfurEth10	100.00%	104	As11E	As16B As11D As16A As11F As11B	100.00%	154	Oc1_47	Oc68_62 Oc1_48 Oc68_52	99.37%	204	Cte6	Cte4 Cte5 Cte2	100.00%	254	Pm72G	Pm37G	100.00%
5	Pc92_64b	Pc92_64d	100.00%	55	Tst119X	Tst119S Tst119V Tst119T	100.00%	105	As11F	As16B As11D As16A As11E As11B	100.00%	155	cs5	cs4 cs2 cs3	100.00%	205	Cte3	Cte4 Cte6 Cte2 Cte5	100.00%	255	Pm37G	Pm72G	100.00%
6	Pc92_64d	Pc92_64b	100.00%	56	Tst119V	Tst119S Tst119X Tst119T	100.00%	106	As16B	As16A As11D As11F As11E As11B	100.00%	156	cs4	cs5 cs2 cs3	100.00%	206	Cte2	Cte5 Cte4 Cte6	100.00%	256	Pm72F	Pm37G Pm72G	99.68%
7	Pau2	Pau4	99.68%	57	Tst119T	Tst119S Tst119X Tst119V	100.00%	107	As16A	As16B As11D As11F As11E As11B	100.00%	157	cs2	cs5 cs3 cs4	100.00%	207	Cte4	Cte5 Cte6 Cte2	100.00%	257	Pm37F	Pm15D Pm1G Pm37B Pm16A Pm37E Pm37C Pm37A Pm16D Pm15A	99.68%
8	Pau4	Pau2	99.68%	58	Tst119W	Tst119S Tst119V Tst119T Tst119X	100.00%	108	Ai16B	Ai16D Ai21F Ai98C Ai21A Ai21H	100.00%	158	cs3	cs5 cs2 cs4	100.00%	208	Igr5	Igr2 Igr6 Igr4 Igr3	100.00%	258	Pm1G	Pm16A Pm15A Pm37E Pm15D Pm37C Pm37B Pm37A Pm16D	100.00%
9	Ce32C	Ce32E Ce32B Ce3D	100.00%	59	Tst119S	Tst119T Tst119X Tst119V	100.00%	109	Ai21A	Ai98C Ai16D Ai21H Ai21F Ai16B	100.00%	159	ch8_2	ch8_5 ch6_3 ch8_3 ch8_4 ch6_2	99.68%	209	Igr3	Igr2 Igr6 Igr5 Igr4	100.00%	259	Pm15A	Pm16D Pm15D Pm37E Pm16A Pm37C Pm37B Pm37A Pm1G	100.00%
10	Ce3D	Ce32C Ce32B Ce32E	100.00%	60	Tst128D	Tst128F Tst128E	100.00%	110	Ai21F	Ai98C Ai16D Ai21H Ai21A Ai16B	100.00%	160	ch3	ch1 ch7_2 ch2 ch6 ch7_3	100.00%	210	Igr2	Igr3 Igr6 Igr5 Igr4	100.00%	260	Pm15D	Pm16D Pm15A Pm37E Pm16A Pm37C Pm37B Pm37A Pm1G	100.00%
11	Ce7B	Ce32E Ce3D Ce32B Ce32C	99.68%	61	Tst128E	Tst128F Tst128D	100.00%	111	Ai16D	Ai21F Ai21A Ai16B Ai98C Ai21H	100.00%	161	ch6	ch1 ch7_2 ch2 ch3 ch7_3	100.00%	211	Igr4	Igr2 Igr6 Igr5 Igr3	100.00%	261	Pm16D	Pm16A Pm15A Pm37E Pm15D Pm37C Pm37B Pm37A Pm1G	100.00%
12	Ce3B	Ce32B Ce32C Ce3D Ce32E	99.68%	62	Tst128F	Tst128E Tst128D	100.00%	112	Ai98C	Ai21F Ai21A Ai16B Ai16D Ai21H	100.00%	162	ch2	ch1 ch7_2 ch3 ch6 ch7_3	100.00%	212	Igr6	Igr2 Igr5 Igr4 Igr3	100.00%	262	Pm37A	Pm16A Pm15A Pm37E Pm15D Pm37C Pm37B Pm1G Pm16D	100.00%
13	Ce3E	Ce32E Ce3D Ce32B Ce32C	99.68%	63	Tst128A	Tst128E Tst128F Tst128D	99.37%	113	Ai21H	Ai98C Ai16D Ai21F Ai21A Ai16B	100.00%	163	ch1	ch2 ch7_2 ch3 ch6 ch7_3	100.00%	213	Ecy2	Ecy1 Ecy5 Ecy4 Ecy3	100.00%	263	Pm37B	Pm16A Pm15A Pm37E Pm15D Pm37C Pm37A Pm1G Pm16D	100.00%
14	Ce32B	Ce32E Ce32C Ce3D	100.00%	**64**	**Tart39A**	**Tgr44 Tgr98**	**90.49%**	114	Ai61A	Ai98B	100.00%	164	ch7_2	ch7_3 ch1 ch2 ch3 ch6	100.00%	214	Ecy4	Ecy1 Ecy5 Ecy3 Ecy2	100.00%	264	Pm37E	Pm16A Pm15A Pm37C Pm15D Pm37B Pm37A Pm1G Pm16D	100.00%
15	Ce32E	Ce32C Ce32B Ce3D	100.00%	65	Ta_120A	Ta_120B	100.00%	115	Ai98B	Ai61A	100.00%	165	ch7_3	ch7_2 ch1 ch2 ch3 ch6	100.00%	215	Ecy5	Ecy1 Ecy4 Ecy3 Ecy2	100.00%	265	Pm16A	Pm16D Pm15A Pm37E Pm15D Pm37C Pm37B Pm37A Pm1G	100.00%
16	Pg3A	Pg98A Pg3C	100.00%	66	Ta_120B	Ta_120A	100.00%	116	Ai16C	Ai98A Ai16E	100.00%	166	ch8_4	ch6_3 ch6_2 ch8_3 ch8_5	100.00%	216	Ecy1	Ecy2 Ecy5 Ecy4 Ecy3	100.00%	266	Pm37C	Pm16A Pm15A Pm37E Pm15D Pm37B Pm37A Pm1G Pm16D	100.00%
17	Pg3C	Pg3A Pg98A	100.00%	67	Ta_119C	Ta_120B Ta_120A	99.68%	117	Ai16E	Ai16C Ai98A	100.00%	167	ch6_2	ch6_3 ch8_4 ch8_3 ch8_5	100.00%	217	Ecy3	Ecy1 Ecy5 Ecy4 Ecy2	100.00%				
18	Pg98A	Pg3A Pg3C	100.00%	68	Ssa1	Ssa2	100.00%	118	Ai98A	Ai16C Ai16E	100.00%	168	ch6_3	ch6_2 ch8_4 ch8_3 ch8_5	100.00%	218	Pk98A	Pk98C	100.00%				
19	Pg3B	Pg98B	99.37%	69	Ssa2	Ssa1	100.00%	119	Nf119E	Nf119B	99.68%	169	ch8_3	ch6_3 ch6_2 ch8_4 ch8_5	100.00%	219	Pk98C	Pk98A	100.00%				
20	Pg98B	Pg3B	99.37%	**70**	**Gyvill60B**	**Ta_120B Ta_120A Ta_119C**	**82.59%**	120	Nf119B	Nf3B Nf119D	99.68%	170	ch8_5	ch6_3 ch6_2 ch8_3 ch8_4	100.00%	220	Pk11E	Pk94D Pk72E Pk72D	100.00%				

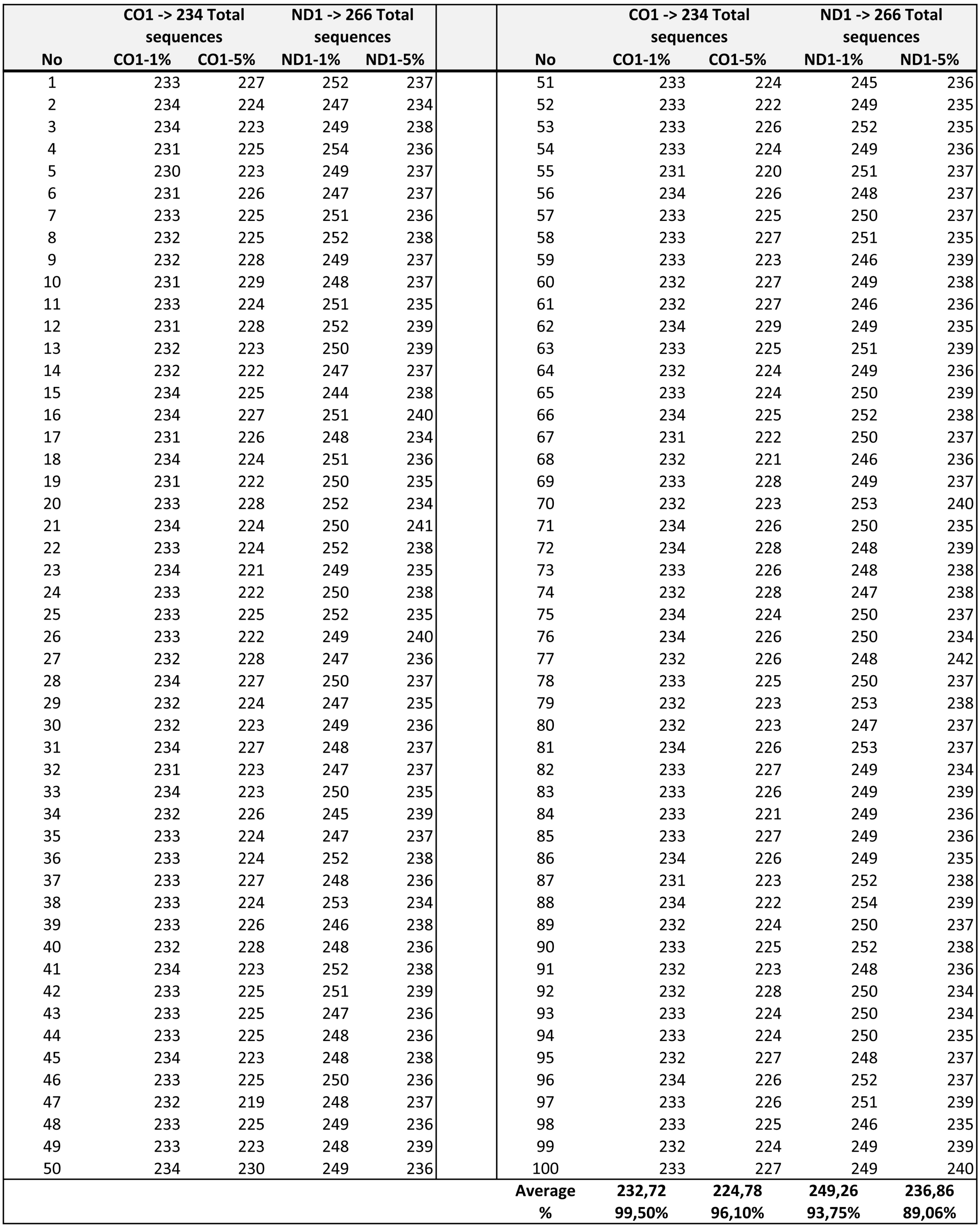
Fig. A.11.: Table S3A: ND1-L10, Results for ND1 leave one out test.

No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity
21	Cs7A	Cs7B Cs7D Cs98E Cs98D Cs98B	100.00%	71	Gu25_40	Gu25_41 Gu25_42 Gu49_31 Gu28_78 Gu28_25 Gu49_47 Gu49_32 Gu87_56	100.00%	121	Nf119D	Nf119B	99.68%	171	Mm2B	Mm2A	100.00%	221	Pk94B	Pk73B	99.68%
22	Cs7B	Cs7A Cs7D Cs98E Cs98D Cs98B	100.00%	72	Gu25_41	Gu25_40 Gu25_42 Gu49_31 Gu28_78 Gu28_25 Gu49_47 Gu49_32 Gu87_56	100.00%	122	Nf3A	Nf3B	99.68%	172	Mm2A	Mm2B	100.00%	222	Pk73B	Pk94B	99.68%
23	Cs7D	Cs7A Cs7B Cs98E Cs98D Cs98B	100.00%	73	Gu25_42	Gu25_40 Gu25_41 Gu49_31 Gu28_78 Gu28_25 Gu49_47 Gu49_32 Gu87_56	100.00%	123	Nf3B	Nf3A	99.68%	**173**	**Ma1**	**Mm2A** **Mm2B**	**88.29%**	223	Pk72E	Pk94D Pk72D	100.00%
24	Cs98B	Cs7A Cs7B Cs98E Cs98D Cs7D	100.00%	74	Gu28_25	Gu25_40 Gu25_41 Gu49_31 Gu28_78 Gu25_42 Gu49_47 Gu49_32 Gu87_56	100.00%	124	Ot1C	Ot1D Ot3B Ot3C Ot3A	100.00%	174	Ca83_9	Ca55_1	100.00%	224	Pk72D	Pk94D Pk72E	100.00%
25	Cs98D	Cs7A Cs7B Cs98E Cs98B Cs7D	100.00%	75	Gu28_78	Gu25_40 Gu25_41 Gu49_31 Gu28_25 Gu25_42 Gu49_47 Gu49_32 Gu87_56	100.00%	125	Ot1D	Ot1C Ot3B Ot3C Ot3A	100.00%	175	Ca55_1	Ca83_9	100.00%	225	Pk94D	Pk72E Pk72D	100.00%
26	Cs98E	Cs7A Cs7B Cs98D Cs98B Cs7D	100.00%	76	Gu49_31	Gu25_40 Gu25_41 Gu28_78 Gu28_25 Gu25_42 Gu49_47 Gu49_32 Gu87_56	100.00%	126	Ot3A	Ot1C Ot3B Ot3C Ot1D	100.00%	176	Ca54_2	Ca55_1 Ca83_9	99.68%	226	Pk94A	Pk73D	100.00%
27	Tk32_43	Tk3_37 Tk74_35	98.73%	77	Gu49_32	Gu25_40 Gu25_41 Gu28_78 Gu28_25 Gu25_42 Gu49_47 Gu49_31 Gu87_56	100.00%	127	Ot3B	Ot1C Ot3A Ot3C Ot1D	100.00%	177	Ca79_13	Ca79_19 Ca79_4 Ca79_20	100.00%	227	Pk73D	Pk94A	100.00%
28	Tk3_37	Tk74_35	100.00%	78	Gu49_47	Gu25_40 Gu25_41 Gu28_78 Gu28_25 Gu25_42 Gu49_32 Gu49_31 Gu87_56	100.00%	128	Ot3C	Ot1C Ot3A Ot3B Ot1D	100.00%	178	Ca79_19	Ca79_13 Ca79_4 Ca79_20	100.00%	228	Pk88B	Pk11E Pk94A Pk73D	99.68%
29	Tk74_35	Tk3_37	100.00%	79	Gu87_56	Gu25_40 Gu25_41 Gu28_78 Gu28_25 Gu25_42 Gu49_32 Gu49_31 Gu49_47	100.00%	129	Ob1_86	Ob32_81 Ob32_83	100.00%	179	Ca79_4	Ca79_13 Ca79_20 Ca79_19	100.00%	229	Pb79A	Pb79B	100.00%
30	Tk32_42	Tk74_35 Tk3_37	99.68%	80	Brpr02A	Brpr02C	100.00%	130	Ob32_81	Ob1_86 Ob32_83	100.00%	180	Ca79_20	Ca79_13 Ca79_4 Ca79_19	100.00%	230	Pb79B	Pb79A	100.00%
31	Tdo2	Tdo3 Tdo4	100.00%	81	Brpr02C	Brpr02A	100.00%	131	Ob32_83	Ob1_86 Ob32_81	100.00%	181	Ca79_7	Ca79_19 Ca79_20 Ca79_13 Ca79_4	100.00%	231	Pb79E	Pb79A Pb79B	99.68%
32	Tdo3	Tdo2 Tdo4	100.00%	82	Aeri142A	Aeri142B	100.00%	132	OcoeRM1	OcoeRM5 OcoeRM4	100.00%	182	Tba49_5	Tba49_7 Tba87_11 Tba25_3	99.37%	232	Pb78F	Pb78A Pb78B	100.00%
33	Tdo4	Tdo2 Tdo3	100.00%	83	Aeri142B	Aeri142A	100.00%	133	OcoeRM4	OcoeRM5 OcoeRM1	100.00%	183	Tba49_7	Tba87_11 Tba25_3	100.00%	233	Pb78A	Pb78F Pb78B	100.00%
34	Tdo1	Tdo5	100.00%	84	Ae21J	Ae155A Ae21F Ae3F Ae21H Ae21D Ae155B Ae21G	99.68%	134	OcoeRM5	OcoeRM4 OcoeRM1	100.00%	184	Tba87_11	Tba49_7 Tba25_3	100.00%	234	Pb78B	Pb78F Pb78A	100.00%
35	Tdo5	Tdo1	100.00%	85	Ae3B	Ae21H Ae21F Ae21D Ae155A Ae155B Ae21G Ae3F	99.68%	135	OcoeSZ5	OcoeSZ4 OcoeSZ3 OcoeCE2 OcoeCE4 OcoeCE1	99.68%	185	Tba25_3	Tba49_7 Tba87_11	100.00%	235	Pb78C	Pb78F Pb78A Pb78B	99.37%
36	The119A	The121C The119B The121A The119C	100.00%	86	Ae155B	Ae21F Ae3F Ae21D Ae155A Ae21H Ae21G	100.00%	136	OcoeCE1	OcoeSZ3 OcoeCE4 OcoeSZ4 OcoeCE2	100.00%	186	Tba87_10	Tba87_9	100.00%	236	Pb113D	Pb113C Pb113A Pb113B	100.00%
37	The119B	The121C The119A The121A The119C	100.00%	87	Ae155A	Ae21F Ae3F Ae21D Ae155B Ae21H Ae21G	100.00%	137	OcoeCE2	OcoeSZ3 OcoeCE4 OcoeSZ4 OcoeCE1	100.00%	187	Tba87_9	Tba87_10	100.00%	237	Pb113A	Pb113D Pb113B Pb113C	100.00%
38	The119C	The121C The119A The121A The119B	100.00%	88	Ae21D	Ae21F Ae3F Ae155B Ae155A Ae21H Ae21G	100.00%	138	OcoeCE4	OcoeSZ3 OcoeCE2 OcoeSZ4 OcoeCE1	100.00%	188	Cg57B	Cg84C	100.00%	238	Pb113B	Pb113D Pb113A Pb113C	100.00%
39	The121A	The121C The119A The119C The119B	100.00%	89	Ae21F	Ae21D Ae3F Ae155B Ae155A Ae21H Ae21G	100.00%	139	OcoeSZ3	OcoeSZ4 OcoeCE4 OcoeCE2 OcoeCE1	100.00%	189	Cg59A	Cg57A	100.00%	239	Pb113C	Pb113D Pb113A Pb113B	100.00%
40	The121C	The121A The119A The119C The119B	100.00%	90	Ae21G	Ae21D Ae3F Ae155B Ae155A Ae21H Ae21F	100.00%	140	OcoeSZ4	OcoeSZ3 OcoeCE4 OcoeCE2 OcoeCE1	100.00%	190	Cg57A	Cg59A	100.00%	240	Pb77D	Pb77A Pb77C Pb77E	100.00%

Fig. A.12.: Table S3B: ND1-L10, Results for ND1 leave one out test.

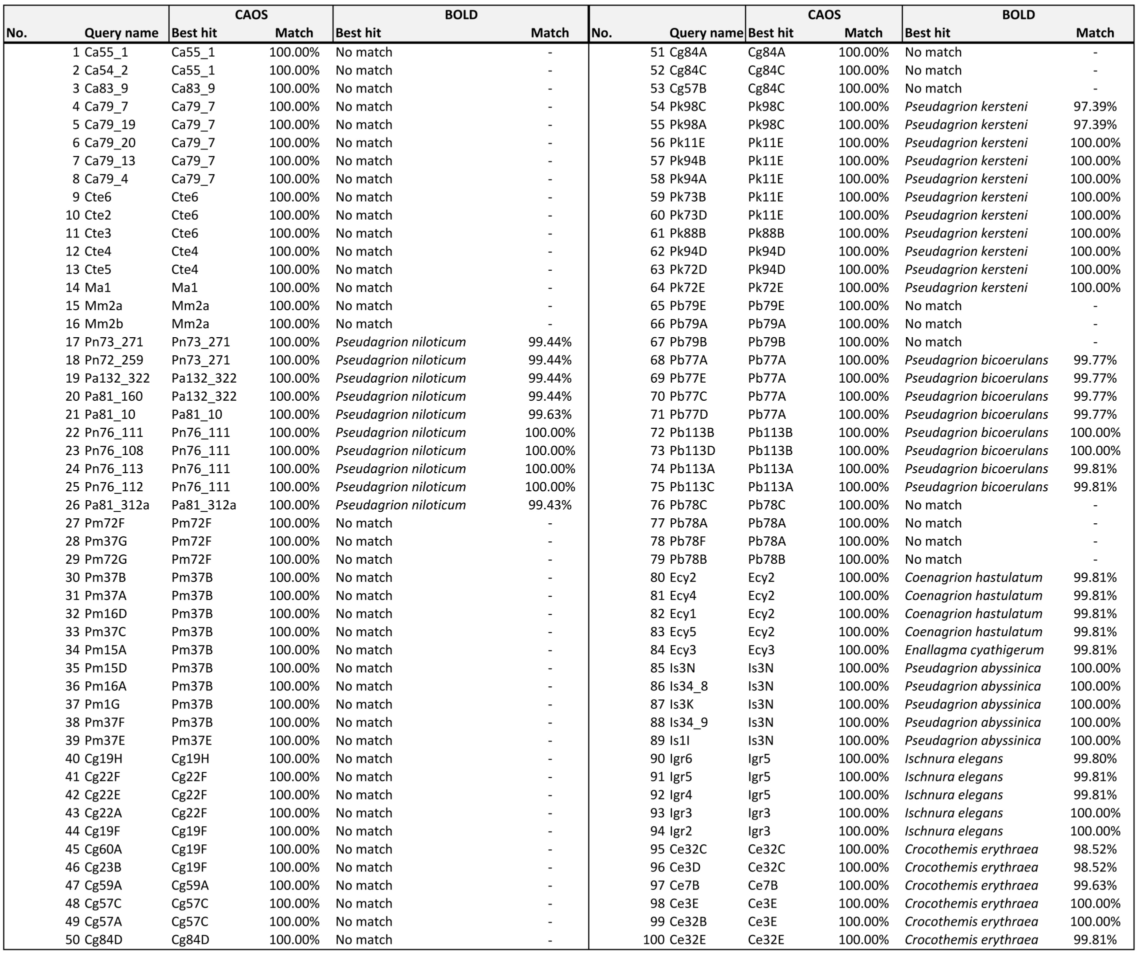
No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity	No	Query	Best Hit	Identity
41	Tgr44	Tgr98	99.02%	91	Ae21H	Ae21D Ae3F Ae155B Ae155A Ae21G Ae21F	100.00%	141	Oj16A	Oj16G Oj16B Oj16D Oj16C	100.00%	191	Cg57C	Cg84A Cg84D	100.00%	241	Pb77C	Pb77A Pb77D Pb77E	100.00%
42	Tgr98	Tgr44	99.02%	92	Ae3F	Ae21H Ae21G Ae155A Ae21F Ae21D Ae155B	100.00%	142	Oj16B	Oj16G Oj16A Oj16D Oj16C	100.00%	192	Cg84A	Cg57C Cg84D	100.00%	242	Pb77A	Pb77C Pb77D Pb77E	100.00%
43	Tnup0017	Tnup0043	100.00%	93	Ae3C	Ae155A Ae21D Ae21H Ae3F Ae21G Ae155B Ae21F	99.68%	143	Oj16C	Oj16G Oj16A Oj16D Oj16B	100.00%	193	Cg84D	Cg57C Cg84A	100.00%	243	Pb77E	Pb77A Pb77C Pb77D	100.00%
44	Tnup0043	Tnup0017	100.00%	94	Acy04A	Acy4 Acy1 Acy2	100.00%	144	Oj16D	Oj16G Oj16A Oj16C Oj16B	100.00%	194	Cg84C	Cg57B	100.00%	244	Pa81_160	Pn76_113 Pn76_111 Pn76_108 Pn76_112 Taxa_4 Pa81_312 Pn73_271 Pn72_259 Pa81_110	100.00%
45	Tst11B	Tst118B Tst118A Tst118D Tst118C Tst118F	100.00%	95	Acy1	Acy4 Acy04A Acy2	100.00%	145	Oj16G	Oj16D Oj16A Oj16C Oj16B	100.00%	195	Cg23B	Cg22E Cg60A	99.05%	245	Pn72_259	Pn76_111 Pa81_110 Taxa_4 Pn76_112 Pn76_113 Pa81_312 Pn73_271 Pn76_108 Pa81_160	100.00%
46	Tst118A	Tst118B Tst118D Tst118C Tst11B Tst118F	100.00%	96	Acy2	Acy4 Acy04A Acy1	100.00%	146	Oj32A	Oj32C Oj32E Oj32H Oj32D	100.00%	196	Cg60A	Cg22A Cg22F Cg22E	99.37%	246	Pa132_322a	Pn76_108 Pa81_110 Pa81_160 Pn76_112 Pn76_111 Pn72_259 Pa81_312 Pn73_271 Pn76_113	100.00%
47	Tst118B	Tst118A Tst118D Tst118C Tst11B Tst118F	100.00%	97	Acy4	Acy2 Acy04A Acy1	100.00%	147	Oj32C	Oj32A Oj32E Oj32H Oj32D	100.00%	197	Cg22A	Cg22F Cg22E	100.00%	247	Pa81_312	Pn76_113 Pn76_111 Pn76_108 Pn76_112 Taxa_4 Pa81_160 Pn73_271 Pn72_259 Pa81_110	100.00%
48	Tst118C	Tst118A Tst118D Tst118B Tst11B Tst118F	100.00%	**98**	**Anatri162**	**Tnup0043** **Tnup0017**	**79.11%**	148	Oj32D	Oj32A Oj32E Oj32H Oj32C	100.00%	198	Cg22E	Cg22F Cg22A	100.00%	248	Pn76_108	Pn76_111 Pa81_110 Taxa_4 Pn76_112 Pn76_113 Pa81_312 Pn72_259 Pn73_271 Pa81_160	100.00%
49	Tst118D	Tst118A Tst118C Tst118B Tst11B Tst118F	100.00%	99	Ami3	Ami2	100.00%	149	Oj32E	Oj32A Oj32D Oj32H Oj32C	100.00%	199	Cg22F	Cg22E Cg22A	100.00%	249	Pn76_111	Pn76_108 Pa81_110 Taxa_4 Pn76_112 Pn76_113 Pa81_312 Pn72_259 Pn73_271 Pa81_160	100.00%
50	Tst118F	Tst118A Tst118C Tst118B Tst11B Tst118D	100.00%	100	Ami2	Ami3	100.00%	150	Oj32H	Oj32A Oj32D Oj32E Oj32C	100.00%	200	Cg19F	Cg19H	100.00%	250	Pn76_112	Pn76_108 Pa81_110 Taxa_4 Pn76_111 Pn76_113 Pa81_312 Pn72_259 Pn73_271 Pa81_160	100.00%

Fig. A.13.: Table S3C: ND1-L10, Results for ND1 leave one out test.



	CO1 -> 234 Total sequences		ND1 -> 266 Total sequences			CO1 -> 234 Total sequences		ND1 -> 266 Total sequences	
No	CO1-1%	CO1-5%	ND1-1%	ND1-5%	No	CO1-1%	CO1-5%	ND1-1%	ND1-5%
1	233	227	252	237	51	233	224	245	236
2	234	224	247	234	52	233	222	249	235
3	234	223	249	238	53	233	226	252	235
4	231	225	254	236	54	233	224	249	236
5	230	223	249	237	55	231	220	251	237
6	231	226	247	237	56	234	226	248	237
7	233	225	251	236	57	233	225	250	237
8	232	225	252	238	58	233	227	251	235
9	232	228	249	237	59	233	223	246	239
10	231	229	248	237	60	232	227	249	238
11	233	224	251	235	61	232	227	246	236
12	231	228	252	239	62	234	229	249	235
13	232	223	250	239	63	233	225	251	239
14	232	222	247	237	64	232	224	249	236
15	234	225	244	238	65	233	224	250	239
16	234	227	251	240	66	234	225	252	238
17	231	226	248	234	67	231	222	250	237
18	234	224	251	236	68	232	221	246	236
19	231	222	250	235	69	233	228	249	237
20	233	228	252	234	70	232	223	253	240
21	234	224	250	241	71	234	226	250	235
22	233	224	252	238	72	234	228	248	239
23	234	221	249	235	73	233	226	248	238
24	233	222	250	238	74	232	228	247	238
25	233	225	252	235	75	234	224	250	237
26	233	222	249	240	76	234	226	250	234
27	232	228	247	236	77	232	226	248	242
28	234	227	250	237	78	233	225	250	237
29	232	224	247	235	79	232	223	253	238
30	232	223	249	236	80	232	223	247	237
31	234	227	248	237	81	234	226	253	237
32	231	223	247	237	82	233	227	249	234
33	234	223	250	235	83	233	226	249	239
34	232	226	245	239	84	233	221	249	236
35	233	224	247	237	85	233	227	249	236
36	233	224	252	238	86	234	226	249	235
37	233	227	248	236	87	231	223	252	238
38	233	224	253	234	88	234	222	254	239
39	233	226	246	238	89	232	224	250	237
40	232	228	248	236	90	233	225	252	238
41	234	223	252	238	91	232	223	248	236
42	233	225	251	239	92	232	228	250	234
43	233	225	247	236	93	233	224	250	234
44	233	225	248	236	94	233	224	250	235
45	234	223	248	238	95	232	227	248	237
46	233	225	250	236	96	234	226	252	237
47	232	219	248	237	97	233	226	251	239
48	233	225	249	236	98	233	225	246	235
49	233	223	248	239	99	232	224	249	239
50	234	230	249	236	100	233	227	249	240
					Average	**232,72**	**224,78**	**249,26**	**236,86**
					%	**99,50%**	**96,10%**	**93,75%**	**89,06%**

Fig. A.14.: Table S4: Random substitution test.

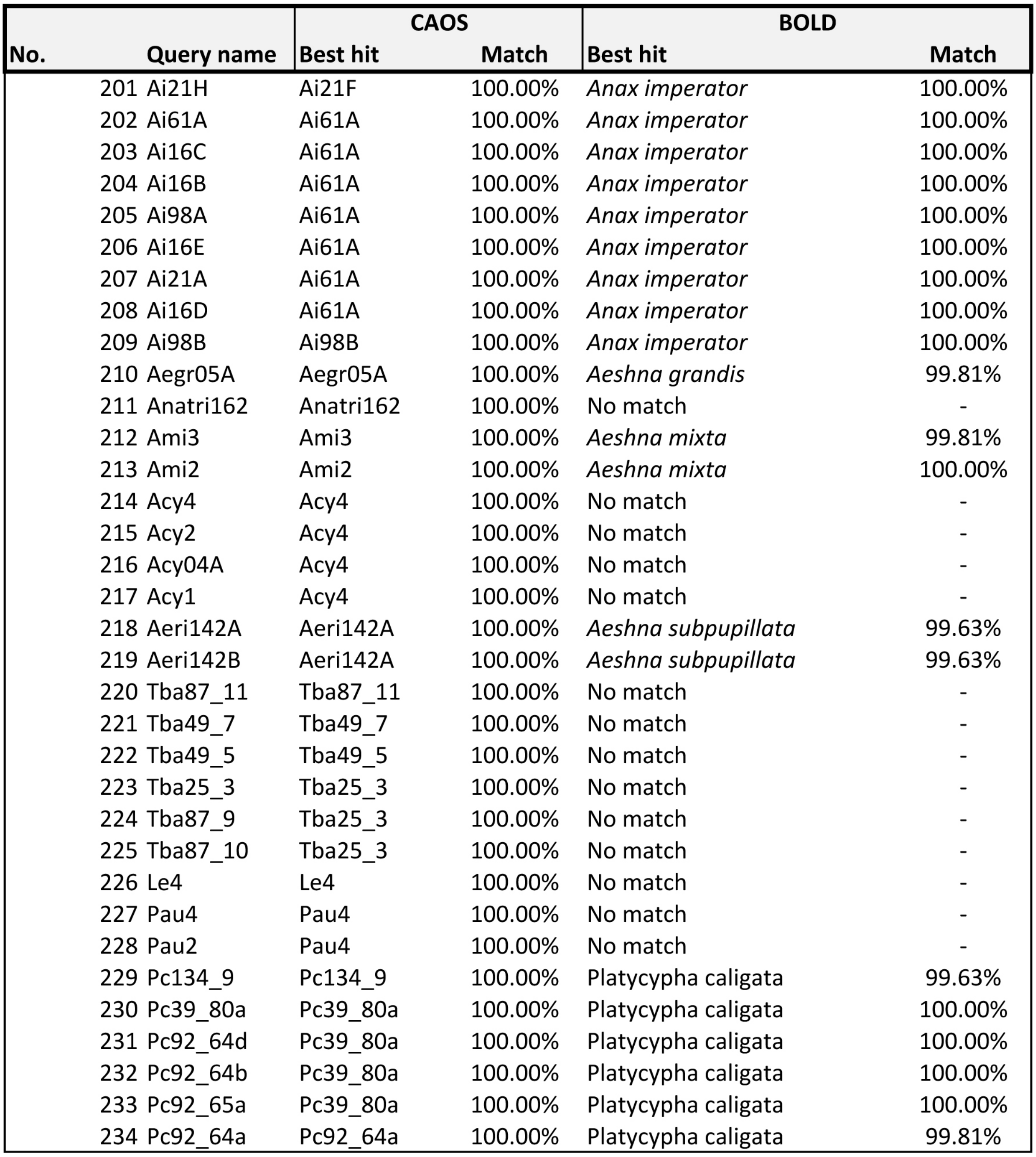


		CAOS		BOLD				CAOS		BOLD	
No.	Query name	Best hit	Match	Best hit	Match	No.	Query name	Best hit	Match	Best hit	Match
1	Ca55_1	Ca55_1	100.00%	No match	-	51	Cg84A	Cg84A	100.00%	No match	-
2	Ca54_2	Ca55_1	100.00%	No match	-	52	Cg84C	Cg84C	100.00%	No match	-
3	Ca83_9	Ca83_9	100.00%	No match	-	53	Cg57B	Cg84C	100.00%	No match	-
4	Ca79_7	Ca79_7	100.00%	No match	-	54	Pk98C	Pk98C	100.00%	*Pseudagrion kersteni*	97.39%
5	Ca79_19	Ca79_7	100.00%	No match	-	55	Pk98A	Pk98C	100.00%	*Pseudagrion kersteni*	97.39%
6	Ca79_20	Ca79_7	100.00%	No match	-	56	Pk11E	Pk11E	100.00%	*Pseudagrion kersteni*	100.00%
7	Ca79_13	Ca79_7	100.00%	No match	-	57	Pk94B	Pk11E	100.00%	*Pseudagrion kersteni*	100.00%
8	Ca79_4	Ca79_7	100.00%	No match	-	58	Pk94A	Pk11E	100.00%	*Pseudagrion kersteni*	100.00%
9	Cte6	Cte6	100.00%	No match	-	59	Pk73B	Pk11E	100.00%	*Pseudagrion kersteni*	100.00%
10	Cte2	Cte6	100.00%	No match	-	60	Pk73D	Pk11E	100.00%	*Pseudagrion kersteni*	100.00%
11	Cte3	Cte6	100.00%	No match	-	61	Pk88B	Pk88B	100.00%	*Pseudagrion kersteni*	100.00%
12	Cte4	Cte4	100.00%	No match	-	62	Pk94D	Pk94D	100.00%	*Pseudagrion kersteni*	100.00%
13	Cte5	Cte4	100.00%	No match	-	63	Pk72D	Pk94D	100.00%	*Pseudagrion kersteni*	100.00%
14	Ma1	Ma1	100.00%	No match	-	64	Pk72E	Pk72E	100.00%	*Pseudagrion kersteni*	100.00%
15	Mm2a	Mm2a	100.00%	No match	-	65	Pb79E	Pb79E	100.00%	No match	-
16	Mm2b	Mm2a	100.00%	No match	-	66	Pb79A	Pb79A	100.00%	No match	-
17	Pn73_271	Pn73_271	100.00%	*Pseudagrion niloticum*	99.44%	67	Pb79B	Pb79B	100.00%	No match	-
18	Pn72_259	Pn73_271	100.00%	*Pseudagrion niloticum*	99.44%	68	Pb77A	Pb77A	100.00%	*Pseudagrion bicoerulans*	99.77%
19	Pa132_322	Pa132_322	100.00%	*Pseudagrion niloticum*	99.44%	69	Pb77E	Pb77A	100.00%	*Pseudagrion bicoerulans*	99.77%
20	Pa81_160	Pa132_322	100.00%	*Pseudagrion niloticum*	99.44%	70	Pb77C	Pb77A	100.00%	*Pseudagrion bicoerulans*	99.77%
21	Pa81_10	Pa81_10	100.00%	*Pseudagrion niloticum*	99.63%	71	Pb77D	Pb77A	100.00%	*Pseudagrion bicoerulans*	99.77%
22	Pn76_111	Pn76_111	100.00%	*Pseudagrion niloticum*	100.00%	72	Pb113B	Pb113B	100.00%	*Pseudagrion bicoerulans*	100.00%
23	Pn76_108	Pn76_111	100.00%	*Pseudagrion niloticum*	100.00%	73	Pb113D	Pb113B	100.00%	*Pseudagrion bicoerulans*	100.00%
24	Pn76_113	Pn76_111	100.00%	*Pseudagrion niloticum*	100.00%	74	Pb113A	Pb113A	100.00%	*Pseudagrion bicoerulans*	99.81%
25	Pn76_112	Pn76_111	100.00%	*Pseudagrion niloticum*	100.00%	75	Pb113C	Pb113A	100.00%	*Pseudagrion bicoerulans*	99.81%
26	Pa81_312a	Pa81_312a	100.00%	*Pseudagrion niloticum*	99.43%	76	Pb78C	Pb78C	100.00%	No match	-
27	Pm72F	Pm72F	100.00%	No match	-	77	Pb78A	Pb78A	100.00%	No match	-
28	Pm37G	Pm72F	100.00%	No match	-	78	Pb78F	Pb78A	100.00%	No match	-
29	Pm72G	Pm72F	100.00%	No match	-	79	Pb78B	Pb78B	100.00%	No match	-
30	Pm37B	Pm37B	100.00%	No match	-	80	Ecy2	Ecy2	100.00%	*Coenagrion hastulatum*	99.81%
31	Pm37A	Pm37B	100.00%	No match	-	81	Ecy4	Ecy2	100.00%	*Coenagrion hastulatum*	99.81%
32	Pm16D	Pm37B	100.00%	No match	-	82	Ecy1	Ecy2	100.00%	*Coenagrion hastulatum*	99.81%
33	Pm37C	Pm37B	100.00%	No match	-	83	Ecy5	Ecy2	100.00%	*Coenagrion hastulatum*	99.81%
34	Pm15A	Pm37B	100.00%	No match	-	84	Ecy3	Ecy3	100.00%	*Enallagma cyathigerum*	99.81%
35	Pm15D	Pm37B	100.00%	No match	-	85	Is3N	Is3N	100.00%	*Pseudagrion abyssinica*	100.00%
36	Pm16A	Pm37B	100.00%	No match	-	86	Is34_8	Is3N	100.00%	*Pseudagrion abyssinica*	100.00%
37	Pm1G	Pm37B	100.00%	No match	-	87	Is3K	Is3N	100.00%	*Pseudagrion abyssinica*	100.00%
38	Pm37F	Pm37B	100.00%	No match	-	88	Is34_9	Is3N	100.00%	*Pseudagrion abyssinica*	100.00%
39	Pm37E	Pm37E	100.00%	No match	-	89	Is1I	Is3N	100.00%	*Pseudagrion abyssinica*	100.00%
40	Cg19H	Cg19H	100.00%	No match	-	90	Igr6	Igr5	100.00%	*Ischnura elegans*	99.80%
41	Cg22F	Cg22F	100.00%	No match	-	91	Igr5	Igr5	100.00%	*Ischnura elegans*	99.81%
42	Cg22E	Cg22F	100.00%	No match	-	92	Igr4	Igr5	100.00%	*Ischnura elegans*	99.81%
43	Cg22A	Cg22F	100.00%	No match	-	93	Igr3	Igr3	100.00%	*Ischnura elegans*	100.00%
44	Cg19F	Cg19F	100.00%	No match	-	94	Igr2	Igr3	100.00%	*Ischnura elegans*	100.00%
45	Cg60A	Cg19F	100.00%	No match	-	95	Ce32C	Ce32C	100.00%	*Crocothemis erythraea*	98.52%
46	Cg23B	Cg19F	100.00%	No match	-	96	Ce3D	Ce32C	100.00%	*Crocothemis erythraea*	98.52%
47	Cg59A	Cg59A	100.00%	No match	-	97	Ce7B	Ce7B	100.00%	*Crocothemis erythraea*	99.63%
48	Cg57C	Cg57C	100.00%	No match	-	98	Ce3E	Ce3E	100.00%	*Crocothemis erythraea*	100.00%
49	Cg57A	Cg57C	100.00%	No match	-	99	Ce32B	Ce3E	100.00%	*Crocothemis erythraea*	100.00%
50	Cg84D	Cg84D	100.00%	No match	-	100	Ce32E	Ce32E	100.00%	*Crocothemis erythraea*	99.81%

Fig. A.15.: Table S5A: CAOS-BOLD: The 234 sequences of the CO1 data set were tested on the CAOS-Classifier and BOLD.

		CAOS		BOLD	
No.	Query name	Best hit	Match	Best hit	Match
101	Ce3B	Ce3B	100.00%	*Crocothemis erythraea*	99.63%
102	Cs98B	Cs98B	100.00%	*Crocothemis sanguinolenta*	99.44%
103	Cs98D	Cs98D	100.00%	*Crocothemis sanguinolenta*	99.07%
104	Cs7B	Cs98D	100.00%	*Crocothemis sanguinolenta*	99.07%
105	Cs7D	Cs7D	100.00%	*Crocothemis sanguinolenta*	98.89%
106	Cs98E	Cs98E	100.00%	*Crocothemis sanguinolenta*	99.07%
107	Cs7A	Cs7A	100.00%	*Crocothemis sanguinolenta*	98.70%
108	Tdo4	Tdo4	100.00%	*Trithemis aconita*	100.00%
109	Tdo3	Tdo4	100.00%	*Trithemis aconita*	100.00%
110	Tdo2	Tdo4	100.00%	*Trithemis aconita*	100.00%
111	Tdo5	Tdo5	100.00%	*Trithemis aconita*	99.63%
112	Tdo1	Tdo5	100.00%	*Trithemis aconita*	99.63%
113	Tk3_37	Tk3_37	100.00%	*Trithemis kirbyi*	99.44%
114	Tk32_43	Tk32_43	100.00%	*Trithemis kirbyi*	100.00%
115	Tk32_42	Tk32_42	100.00%	*Trithemis kirbyi*	99.63%
116	Tk74_35	Tk74_35	100.00%	*Trithemis kirbyi*	99.63%
117	Tart39A	Tart39A	100.00%	*Trithemis kirbyi*	99.81%
118	Tst128A	Tst128A	100.00%	No match	-
119	Tst128E	Tst128E	100.00%	No match	-
120	Tst128D	Tst128D	100.00%	No match	-
121	Tst128F	Tst128F	100.00%	No match	-
122	Tst119W	Tst119W	100.00%	*Trithemis sticta*	99.07%
123	Tst119V	Tst119V	100.00%	*Trithemis sticta*	99.44%
124	Tst119S	Tst119S	100.00%	*Trithemis sticta*	99.81%
125	Tst119X	Tst119S	100.00%	*Trithemis sticta*	99.81%
126	Tst119T	Tst119T	100.00%	*Trithemis sticta*	100.00%
127	Tst118A	Tst118A	100.00%	No match	-
128	Tst118F	Tst118A	100.00%	No match	-
129	Tst118C	Tst118A	100.00%	No match	-
130	Tst118B	Tst118A	100.00%	No match	-
131	Tst118D	Tst118A	100.00%	No match	-
132	Tst17	Tst17	100.00%	No match	-
133	Tst11B	Tst17	100.00%	No match	-
134	Tnup0017	Tnup0017	100.00%	No match	-
135	Tnup0043	Tnup0017	100.00%	No match	-
136	Tgr44	Tgr44	100.00%	No match	-
137	Tgr98	Tgr98	100.00%	No match	-
138	Ta120B	Ta120B	100.00%	*Trithemis annulata*	100.00%
139	Ta120A	Ta120B	100.00%	*Trithemis annulata*	100.00%
140	Ta119C	Ta119C	100.00%	*Trithemis annulata*	99.81%
141	TfurSAB	TfurSAB	100.00%	*Trithemis furva*	99.81%
142	TfurEth11	TfurEth11	100.00%	*Trithemis furva*	99.81%
143	TfurEth10	TfurEth10	100.00%	*Trithemis furva*	100.00%
144	Nf119D	Nf119D	100.00%	No match	-
145	Nf119B	Nf119D	100.00%	No match	-
146	Nf119E	Nf119E	100.00%	No match	-
147	Nf3B	Nf3B	100.00%	No match	-
148	Nf3A	Nf3A	100.00%	No match	-
149	Ot3A	Ot3A	100.00%	*Orthetrum trinacria*	100.00%
150	Ot1D	Ot3A	100.00%	*Orthetrum trinacria*	100.00%
151	Ot1C	Ot3A	100.00%	*Orthetrum trinacria*	100.00%
152	Ot3B	Ot3A	100.00%	*Orthetrum trinacria*	100.00%
153	Ot3C	Ot3A	100.00%	*Orthetrum trinacria*	100.00%
154	OcoeSZ5	OcoeSZ5	100.00%	*Orthetrum coerulescens*	100.00%
155	OcoeSZ3	OcoeSZ5	100.00%	*Orthetrum coerulescens*	100.00%
156	OcoeCE4	OcoeSZ5	100.00%	*Orthetrum coerulescens*	100.00%
157	OcoeCE2	OcoeSZ5	100.00%	*Orthetrum coerulescens*	100.00%
158	OcoeCE1	OcoeSZ5	100.00%	*Orthetrum coerulescens*	100.00%
159	OcoeSZ4	OcoeSZ5	100.00%	*Orthetrum coerulescens*	100.00%
160	OcoeRM1	OcoeRM1	100.00%	*Orthetrum coerulescens*	99.81%
161	OcoeRM4	OcoeRM1	100.00%	*Orthetrum coerulescens*	99.81%
162	OcoeRM5	OcoeRM1	100.00%	*Orthetrum coerulescens*	99.81%
163	Ob82	Ob82	100.00%	*Orthetrum stemmale*	100.00%
164	Ob83	Ob82	100.00%	*Orthetrum stemmale*	100.00%
165	Ob86	Ob86	100.00%	*Orthetrum stemmale*	99.81%
166	Oj16D	Oj16D	100.00%	*Orthetrum julia*	100.00%
167	Oj16G	Oj16D	100.00%	*Orthetrum julia*	100.00%
168	Oj16B	Oj16D	100.00%	*Orthetrum julia*	100.00%
169	Oj16A	Oj16D	100.00%	*Orthetrum julia*	100.00%
170	Oj16C	Oj16D	100.00%	*Orthetrum julia*	100.00%
171	Oj32A	Oj32A	100.00%	*Orthetrum julia*	99.44%
172	Oj32C	Oj32A	100.00%	*Orthetrum julia*	99.44%
173	Oj32H	Oj32A	100.00%	*Orthetrum julia*	99.44%
174	Oj32E	Oj32A	100.00%	*Orthetrum julia*	99.44%
175	Oj32D	Oj32A	100.00%	*Orthetrum julia*	99.43%
176	Oc48	Oc48	100.00%	*Orthetrum julia*	100.00%
177	Oc62	Oc48	100.00%	*Orthetrum julia*	100.00%
178	Oc52	Oc48	100.00%	*Orthetrum julia*	100.00%
179	Oc47	Oc47	100.00%	*Orthetrum chrysostigma*	100.00%
180	Gyvill60B	Gyvill60B	100.00%	No match	-
181	Brpr02A	Brpr02A	100.00%	No match	-
182	Brpr02C	Brpr02A	100.00%	No match	-
183	Ae3C	Ae3C	100.00%	No match	-
184	Ae3F	Ae3F	100.00%	No match	-
185	Ae21G	Ae21G	100.00%	No match	-
186	Ae21H	Ae21H	100.00%	No match	-
187	Ae3B	Ae3B	100.00%	No match	-
188	Ae21F	Ae3B	100.00%	No match	-
189	Ae21D	Ae3B	100.00%	No match	-
190	Ae21J	Ae21J	100.00%	No match	-
191	Ae155B	Ae155B	100.00%	No match	-
192	Ae155A	Ae155B	100.00%	No match	-
193	As11E	As11E	100.00%	No match	-
194	As11F	As11E	100.00%	No match	-
195	As16B	As11E	100.00%	No match	-
196	As11D	As11E	100.00%	No match	-
197	As11B	As11E	100.00%	No match	-
198	As16A	As11E	100.00%	No match	-
199	Ai98C	Ai98C	100.00%	*Anax imperator*	99.63%
200	Ai21F	Ai21F	100.00%	*Anax imperator*	100.00%

Fig. A.16.: Table S5B: CAOS-BOLD: The 234 sequences of the CO1 data set were tested on the CAOS-Classifier and BOLD.



		CAOS		BOLD	
No.	Query name	Best hit	Match	Best hit	Match
201	Ai21H	Ai21F	100.00%	*Anax imperator*	100.00%
202	Ai61A	Ai61A	100.00%	*Anax imperator*	100.00%
203	Ai16C	Ai61A	100.00%	*Anax imperator*	100.00%
204	Ai16B	Ai61A	100.00%	*Anax imperator*	100.00%
205	Ai98A	Ai61A	100.00%	*Anax imperator*	100.00%
206	Ai16E	Ai61A	100.00%	*Anax imperator*	100.00%
207	Ai21A	Ai61A	100.00%	*Anax imperator*	100.00%
208	Ai16D	Ai61A	100.00%	*Anax imperator*	100.00%
209	Ai98B	Ai98B	100.00%	*Anax imperator*	100.00%
210	Aegr05A	Aegr05A	100.00%	*Aeshna grandis*	99.81%
211	Anatri162	Anatri162	100.00%	No match	-
212	Ami3	Ami3	100.00%	*Aeshna mixta*	99.81%
213	Ami2	Ami2	100.00%	*Aeshna mixta*	100.00%
214	Acy4	Acy4	100.00%	No match	-
215	Acy2	Acy4	100.00%	No match	-
216	Acy04A	Acy4	100.00%	No match	-
217	Acy1	Acy4	100.00%	No match	-
218	Aeri142A	Aeri142A	100.00%	*Aeshna subpupillata*	99.63%
219	Aeri142B	Aeri142A	100.00%	*Aeshna subpupillata*	99.63%
220	Tba87_11	Tba87_11	100.00%	No match	-
221	Tba49_7	Tba49_7	100.00%	No match	-
222	Tba49_5	Tba49_5	100.00%	No match	-
223	Tba25_3	Tba25_3	100.00%	No match	-
224	Tba87_9	Tba25_3	100.00%	No match	-
225	Tba87_10	Tba25_3	100.00%	No match	-
226	Le4	Le4	100.00%	No match	-
227	Pau4	Pau4	100.00%	No match	-
228	Pau2	Pau4	100.00%	No match	-
229	Pc134_9	Pc134_9	100.00%	Platycypha caligata	99.63%
230	Pc39_80a	Pc39_80a	100.00%	Platycypha caligata	100.00%
231	Pc92_64d	Pc39_80a	100.00%	Platycypha caligata	100.00%
232	Pc92_64b	Pc39_80a	100.00%	Platycypha caligata	100.00%
233	Pc92_65a	Pc39_80a	100.00%	Platycypha caligata	100.00%
234	Pc92_64a	Pc92_64a	100.00%	Platycypha caligata	99.81%

Fig. A.17.: Table S5C: CAOS-BOLD: The 234 sequences of the CO1 data set were tested on the CAOS-Classifier and BOLD.

A.3 Manuscript 3

Appendix S1 Mined data. Mined sequences from GenBank and BOLD. Unrefined CO1, 28S rDNA and LWR fasta files.

Source: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0001-AppendixS1.zip

Appendix S2 Aligned data. Refined sequence data sets of CO1, 28S rDNA and LWR. Identical sequences were depleted. Raw sequences were aligned with clustal w using MEGA.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0002-AppendixS2.zip

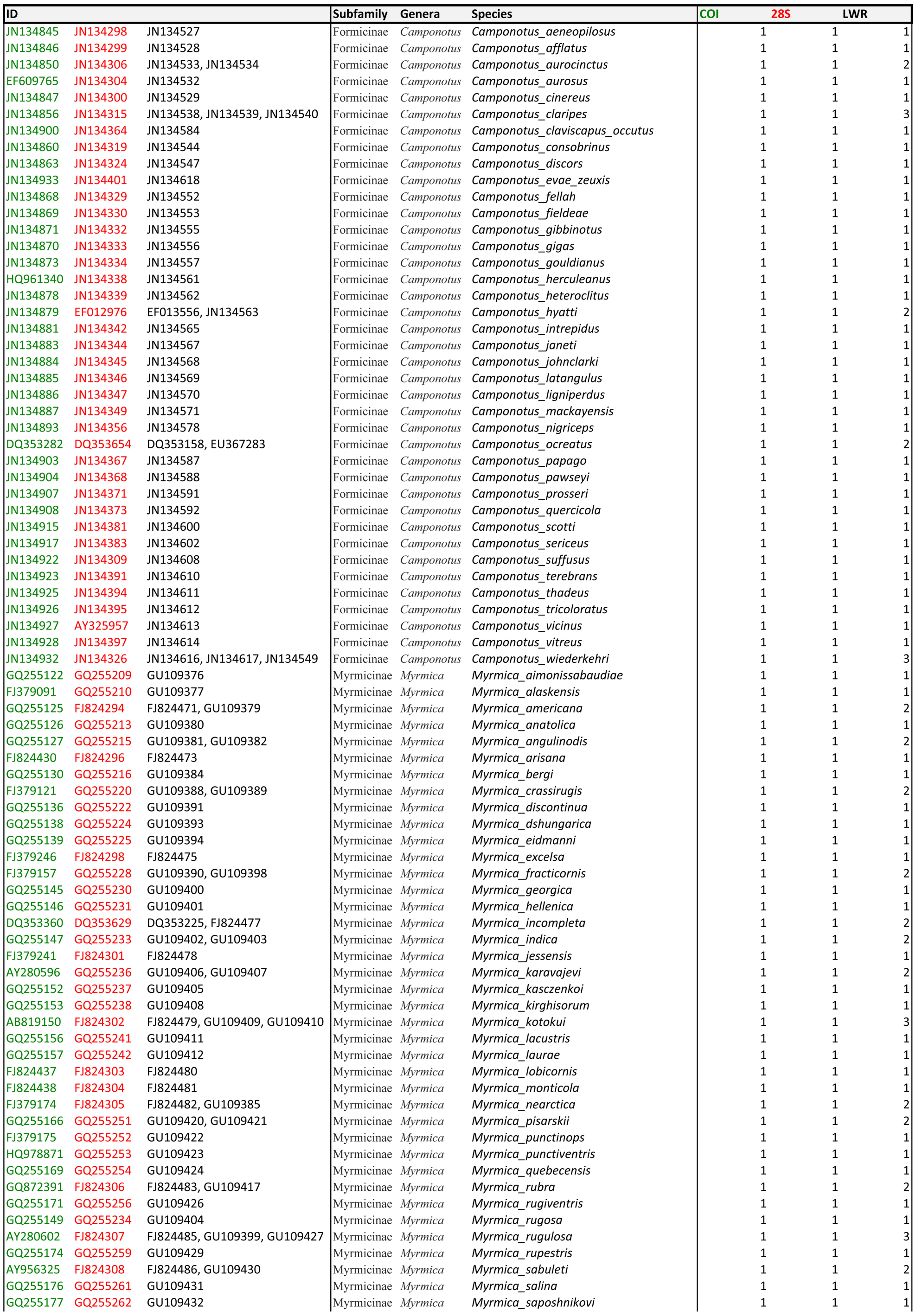
Appendix S5 Subfamily. Fasta files of aligned subfamily sequences.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0005-AppendixS5.zip

Appendix S6 Genera. Fasta files of aligned genera sequences.

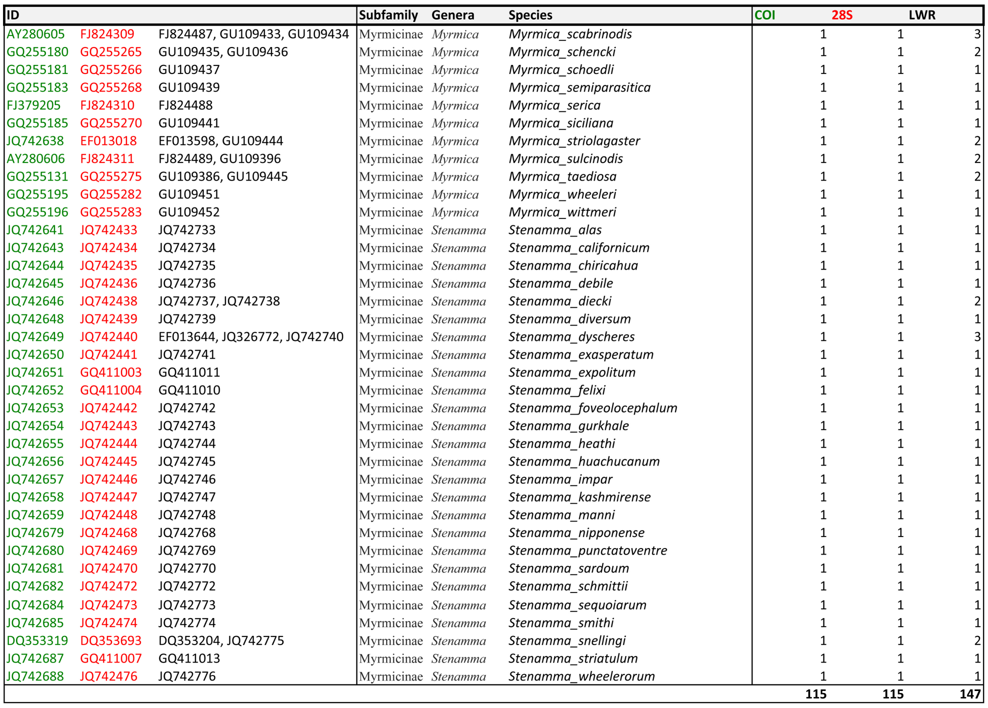
https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0006-AppendixS6.zip

Appendix S7 Twelve CA?Matrices. Contains barcode data of all three genes (CO1, 28S rDNA, LWR) for subfamily, genera and species data sets. In overview1?2 sPu and sPr are listed. In overview3?5 only sPu diagnostics are listed. The 'Total barcode' file contains all diagnostic positions detected within the data set. The "Ref matrix" file can be used with the CAOS?Classifier as a reference barcode file. With this file new specimen can be identified.



ID			Subfamily	Genera	Species	COI	28S	LWR
JN134845	JN134298	JN134527	Formicinae	*Camponotus*	*Camponotus_aeneopilosus*	1	1	1
JN134846	JN134299	JN134528	Formicinae	*Camponotus*	*Camponotus_afflatus*	1	1	1
JN134850	JN134306	JN134533, JN134534	Formicinae	*Camponotus*	*Camponotus_aurocinctus*	1	1	2
EF609765	JN134304	JN134532	Formicinae	*Camponotus*	*Camponotus_aurosus*	1	1	1
JN134847	JN134300	JN134529	Formicinae	*Camponotus*	*Camponotus_cinereus*	1	1	1
JN134856	JN134315	JN134538, JN134539, JN134540	Formicinae	*Camponotus*	*Camponotus_claripes*	1	1	3
JN134900	JN134364	JN134584	Formicinae	*Camponotus*	*Camponotus_claviscapus_occutus*	1	1	1
JN134860	JN134319	JN134544	Formicinae	*Camponotus*	*Camponotus_consobrinus*	1	1	1
JN134863	JN134324	JN134547	Formicinae	*Camponotus*	*Camponotus_discors*	1	1	1
JN134933	JN134401	JN134618	Formicinae	*Camponotus*	*Camponotus_evae_zeuxis*	1	1	1
JN134868	JN134329	JN134552	Formicinae	*Camponotus*	*Camponotus_fellah*	1	1	1
JN134869	JN134330	JN134553	Formicinae	*Camponotus*	*Camponotus_fieldeae*	1	1	1
JN134871	JN134332	JN134555	Formicinae	*Camponotus*	*Camponotus_gibbinotus*	1	1	1
JN134870	JN134333	JN134556	Formicinae	*Camponotus*	*Camponotus_gigas*	1	1	1
JN134873	JN134334	JN134557	Formicinae	*Camponotus*	*Camponotus_gouldianus*	1	1	1
HQ961340	JN134338	JN134561	Formicinae	*Camponotus*	*Camponotus_herculeanus*	1	1	1
JN134878	JN134339	JN134562	Formicinae	*Camponotus*	*Camponotus_heteroclitus*	1	1	1
JN134879	EF012976	EF013556, JN134563	Formicinae	*Camponotus*	*Camponotus_hyatti*	1	1	2
JN134881	JN134342	JN134565	Formicinae	*Camponotus*	*Camponotus_intrepidus*	1	1	1
JN134883	JN134344	JN134567	Formicinae	*Camponotus*	*Camponotus_janeti*	1	1	1
JN134884	JN134345	JN134568	Formicinae	*Camponotus*	*Camponotus_johnclarki*	1	1	1
JN134885	JN134346	JN134569	Formicinae	*Camponotus*	*Camponotus_latangulus*	1	1	1
JN134886	JN134347	JN134570	Formicinae	*Camponotus*	*Camponotus_ligniperdus*	1	1	1
JN134887	JN134349	JN134571	Formicinae	*Camponotus*	*Camponotus_mackayensis*	1	1	1
JN134893	JN134356	JN134578	Formicinae	*Camponotus*	*Camponotus_nigriceps*	1	1	1
DQ353282	DQ353654	DQ353158, EU367283	Formicinae	*Camponotus*	*Camponotus_ocreatus*	1	1	2
JN134903	JN134367	JN134587	Formicinae	*Camponotus*	*Camponotus_papago*	1	1	1
JN134904	JN134368	JN134588	Formicinae	*Camponotus*	*Camponotus_pawseyi*	1	1	1
JN134907	JN134371	JN134591	Formicinae	*Camponotus*	*Camponotus_prosseri*	1	1	1
JN134908	JN134373	JN134592	Formicinae	*Camponotus*	*Camponotus_quercicola*	1	1	1
JN134915	JN134381	JN134600	Formicinae	*Camponotus*	*Camponotus_scotti*	1	1	1
JN134917	JN134383	JN134602	Formicinae	*Camponotus*	*Camponotus_sericeus*	1	1	1
JN134922	JN134309	JN134608	Formicinae	*Camponotus*	*Camponotus_suffusus*	1	1	1
JN134923	JN134391	JN134610	Formicinae	*Camponotus*	*Camponotus_terebrans*	1	1	1
JN134925	JN134394	JN134611	Formicinae	*Camponotus*	*Camponotus_thadeus*	1	1	1
JN134926	JN134395	JN134612	Formicinae	*Camponotus*	*Camponotus_tricoloratus*	1	1	1
JN134927	AY325957	JN134613	Formicinae	*Camponotus*	*Camponotus_vicinus*	1	1	1
JN134928	JN134397	JN134614	Formicinae	*Camponotus*	*Camponotus_vitreus*	1	1	1
JN134932	JN134326	JN134616, JN134617, JN134549	Formicinae	*Camponotus*	*Camponotus_wiederkehri*	1	1	3
GQ255122	GQ255209	GU109376	Myrmicinae	*Myrmica*	*Myrmica_aimonissabaudiae*	1	1	1
FJ379091	GQ255210	GU109377	Myrmicinae	*Myrmica*	*Myrmica_alaskensis*	1	1	1
GQ255125	FJ824294	FJ824471, GU109379	Myrmicinae	*Myrmica*	*Myrmica_americana*	1	1	2
GQ255126	GQ255213	GU109380	Myrmicinae	*Myrmica*	*Myrmica_anatolica*	1	1	1
GQ255127	GQ255215	GU109381, GU109382	Myrmicinae	*Myrmica*	*Myrmica_angulinodis*	1	1	2
FJ824430	FJ824296	FJ824473	Myrmicinae	*Myrmica*	*Myrmica_arisana*	1	1	1
GQ255130	GQ255216	GU109384	Myrmicinae	*Myrmica*	*Myrmica_bergi*	1	1	1
FJ379121	GQ255220	GU109388, GU109389	Myrmicinae	*Myrmica*	*Myrmica_crassirugis*	1	1	2
GQ255136	GQ255222	GU109391	Myrmicinae	*Myrmica*	*Myrmica_discontinua*	1	1	1
GQ255138	GQ255224	GU109393	Myrmicinae	*Myrmica*	*Myrmica_dshungarica*	1	1	1
GQ255139	GQ255225	GU109394	Myrmicinae	*Myrmica*	*Myrmica_eidmanni*	1	1	1
FJ379246	FJ824298	FJ824475	Myrmicinae	*Myrmica*	*Myrmica_excelsa*	1	1	1
FJ379157	GQ255228	GU109390, GU109398	Myrmicinae	*Myrmica*	*Myrmica_fracticornis*	1	1	2
GQ255145	GQ255230	GU109400	Myrmicinae	*Myrmica*	*Myrmica_georgica*	1	1	1
GQ255146	GQ255231	GU109401	Myrmicinae	*Myrmica*	*Myrmica_hellenica*	1	1	1
DQ353360	DQ353629	DQ353225, FJ824477	Myrmicinae	*Myrmica*	*Myrmica_incompleta*	1	1	2
GQ255147	GQ255233	GU109402, GU109403	Myrmicinae	*Myrmica*	*Myrmica_indica*	1	1	2
FJ379241	FJ824301	FJ824478	Myrmicinae	*Myrmica*	*Myrmica_jessensis*	1	1	1
AY280596	GQ255236	GU109406, GU109407	Myrmicinae	*Myrmica*	*Myrmica_karavajevi*	1	1	2
GQ255152	GQ255237	GU109405	Myrmicinae	*Myrmica*	*Myrmica_kasczenkoi*	1	1	1
GQ255153	GQ255238	GU109408	Myrmicinae	*Myrmica*	*Myrmica_kirghisorum*	1	1	1
AB819150	FJ824302	FJ824479, GU109409, GU109410	Myrmicinae	*Myrmica*	*Myrmica_kotokui*	1	1	3
GQ255156	GQ255241	GU109411	Myrmicinae	*Myrmica*	*Myrmica_lacustris*	1	1	1
GQ255157	GQ255242	GU109412	Myrmicinae	*Myrmica*	*Myrmica_laurae*	1	1	1
FJ824437	FJ824303	FJ824480	Myrmicinae	*Myrmica*	*Myrmica_lobicornis*	1	1	1
FJ824438	FJ824304	FJ824481	Myrmicinae	*Myrmica*	*Myrmica_monticola*	1	1	1
FJ379174	FJ824305	FJ824482, GU109385	Myrmicinae	*Myrmica*	*Myrmica_nearctica*	1	1	2
GQ255166	GQ255251	GU109420, GU109421	Myrmicinae	*Myrmica*	*Myrmica_pisarskii*	1	1	2
FJ379175	GQ255252	GU109422	Myrmicinae	*Myrmica*	*Myrmica_punctinops*	1	1	1
HQ978871	GQ255253	GU109423	Myrmicinae	*Myrmica*	*Myrmica_punctiventris*	1	1	1
GQ255169	GQ255254	GU109424	Myrmicinae	*Myrmica*	*Myrmica_quebecensis*	1	1	1
GQ872391	FJ824306	FJ824483, GU109417	Myrmicinae	*Myrmica*	*Myrmica_rubra*	1	1	2
GQ255171	GQ255256	GU109426	Myrmicinae	*Myrmica*	*Myrmica_rugiventris*	1	1	1
GQ255149	GQ255234	GU109404	Myrmicinae	*Myrmica*	*Myrmica_rugosa*	1	1	1
AY280602	FJ824307	FJ824485, GU109399, GU109427	Myrmicinae	*Myrmica*	*Myrmica_rugulosa*	1	1	3
GQ255174	GQ255259	GU109429	Myrmicinae	*Myrmica*	*Myrmica_rupestris*	1	1	1
AY956325	FJ824308	FJ824486, GU109430	Myrmicinae	*Myrmica*	*Myrmica_sabuleti*	1	1	2
GQ255176	GQ255261	GU109431	Myrmicinae	*Myrmica*	*Myrmica_salina*	1	1	1
GQ255177	GQ255262	GU109432	Myrmicinae	*Myrmica*	*Myrmica_saposhnikovi*	1	1	1

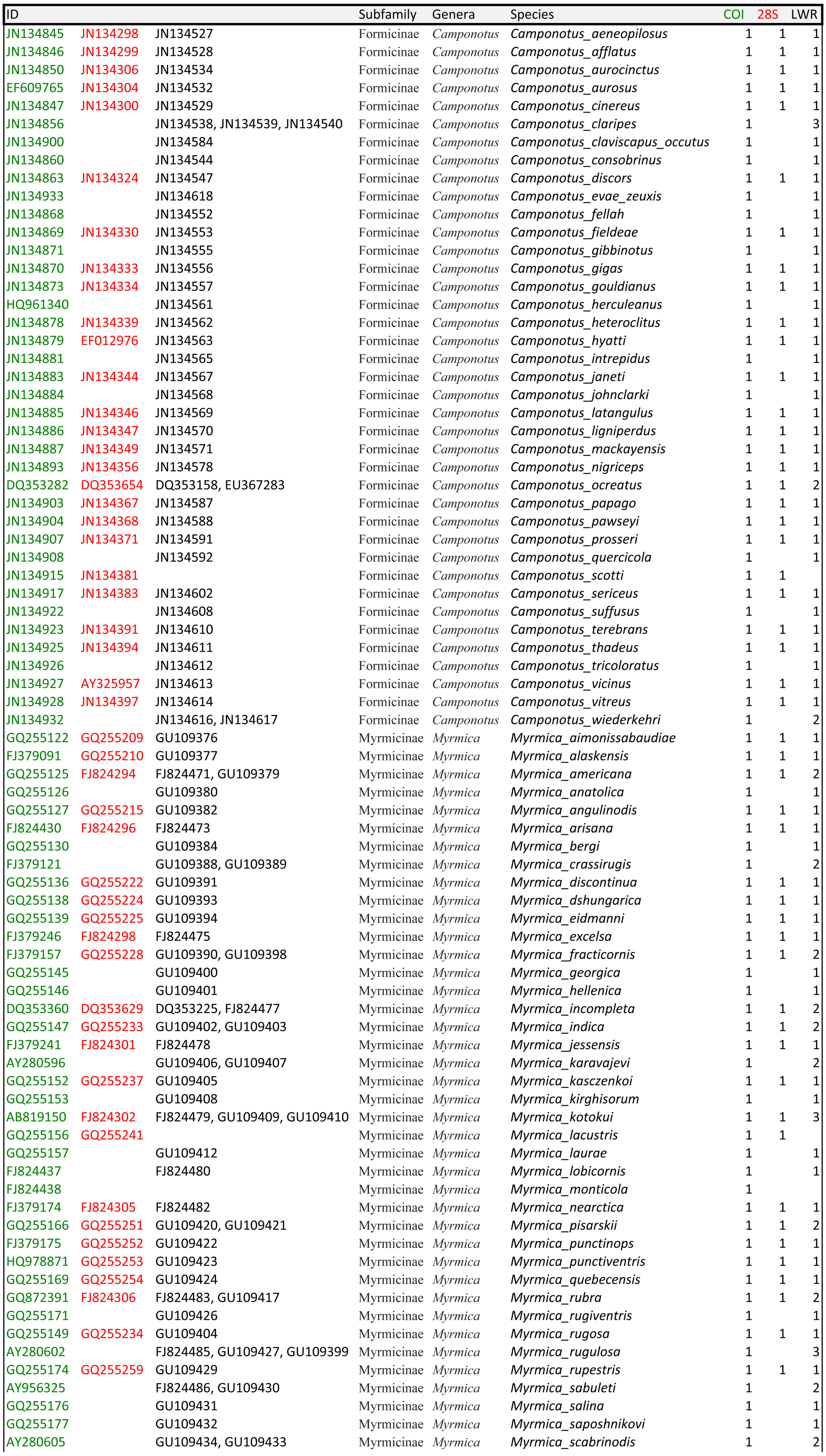
Fig. A.18.: Table S3A: Gene comparison. List of species with all three sequences (CO1, 28S rDNA, LWR) available. The list contains GenBank?IDs, Species names and number of specimen per species.



ID			Subfamily	Genera	Species	COI	28S	LWR
AY280605	FJ824309	FJ824487, GU109433, GU109434	Myrmicinae	*Myrmica*	*Myrmica_scabrinodis*	1	1	3
GQ255180	GQ255265	GU109435, GU109436	Myrmicinae	*Myrmica*	*Myrmica_schencki*	1	1	2
GQ255181	GQ255266	GU109437	Myrmicinae	*Myrmica*	*Myrmica_schoedli*	1	1	1
GQ255183	GQ255268	GU109439	Myrmicinae	*Myrmica*	*Myrmica_semiparasitica*	1	1	1
FJ379205	FJ824310	FJ824488	Myrmicinae	*Myrmica*	*Myrmica_serica*	1	1	1
GQ255185	GQ255270	GU109441	Myrmicinae	*Myrmica*	*Myrmica_siciliana*	1	1	1
JQ742638	EF013018	EF013598, GU109444	Myrmicinae	*Myrmica*	*Myrmica_striolagaster*	1	1	2
AY280606	FJ824311	FJ824489, GU109396	Myrmicinae	*Myrmica*	*Myrmica_sulcinodis*	1	1	2
GQ255131	GQ255275	GU109386, GU109445	Myrmicinae	*Myrmica*	*Myrmica_taediosa*	1	1	2
GQ255195	GQ255282	GU109451	Myrmicinae	*Myrmica*	*Myrmica_wheeleri*	1	1	1
GQ255196	GQ255283	GU109452	Myrmicinae	*Myrmica*	*Myrmica_wittmeri*	1	1	1
JQ742641	JQ742433	JQ742733	Myrmicinae	*Stenamma*	*Stenamma_alas*	1	1	1
JQ742643	JQ742434	JQ742734	Myrmicinae	*Stenamma*	*Stenamma_californicum*	1	1	1
JQ742644	JQ742435	JQ742735	Myrmicinae	*Stenamma*	*Stenamma_chiricahua*	1	1	1
JQ742645	JQ742436	JQ742736	Myrmicinae	*Stenamma*	*Stenamma_debile*	1	1	1
JQ742646	JQ742438	JQ742737, JQ742738	Myrmicinae	*Stenamma*	*Stenamma_diecki*	1	1	2
JQ742648	JQ742439	JQ742739	Myrmicinae	*Stenamma*	*Stenamma_diversum*	1	1	1
JQ742649	JQ742440	EF013644, JQ326772, JQ742740	Myrmicinae	*Stenamma*	*Stenamma_dyscheres*	1	1	3
JQ742650	JQ742441	JQ742741	Myrmicinae	*Stenamma*	*Stenamma_exasperatum*	1	1	1
JQ742651	GQ411003	GQ411011	Myrmicinae	*Stenamma*	*Stenamma_expolitum*	1	1	1
JQ742652	GQ411004	GQ411010	Myrmicinae	*Stenamma*	*Stenamma_felixi*	1	1	1
JQ742653	JQ742442	JQ742742	Myrmicinae	*Stenamma*	*Stenamma_foveolocephalum*	1	1	1
JQ742654	JQ742443	JQ742743	Myrmicinae	*Stenamma*	*Stenamma_gurkhale*	1	1	1
JQ742655	JQ742444	JQ742744	Myrmicinae	*Stenamma*	*Stenamma_heathi*	1	1	1
JQ742656	JQ742445	JQ742745	Myrmicinae	*Stenamma*	*Stenamma_huachucanum*	1	1	1
JQ742657	JQ742446	JQ742746	Myrmicinae	*Stenamma*	*Stenamma_impar*	1	1	1
JQ742658	JQ742447	JQ742747	Myrmicinae	*Stenamma*	*Stenamma_kashmirense*	1	1	1
JQ742659	JQ742448	JQ742748	Myrmicinae	*Stenamma*	*Stenamma_manni*	1	1	1
JQ742679	JQ742468	JQ742768	Myrmicinae	*Stenamma*	*Stenamma_nipponense*	1	1	1
JQ742680	JQ742469	JQ742769	Myrmicinae	*Stenamma*	*Stenamma_punctatoventre*	1	1	1
JQ742681	JQ742470	JQ742770	Myrmicinae	*Stenamma*	*Stenamma_sardoum*	1	1	1
JQ742682	JQ742472	JQ742772	Myrmicinae	*Stenamma*	*Stenamma_schmittii*	1	1	1
JQ742684	JQ742473	JQ742773	Myrmicinae	*Stenamma*	*Stenamma_sequoiarum*	1	1	1
JQ742685	JQ742474	JQ742774	Myrmicinae	*Stenamma*	*Stenamma_smithi*	1	1	1
DQ353319	DQ353693	DQ353204, JQ742775	Myrmicinae	*Stenamma*	*Stenamma_snellingi*	1	1	2
JQ742687	GQ411007	GQ411013	Myrmicinae	*Stenamma*	*Stenamma_striatulum*	1	1	1
JQ742688	JQ742476	JQ742776	Myrmicinae	*Stenamma*	*Stenamma_wheelerorum*	1	1	1
						115	**115**	**147**

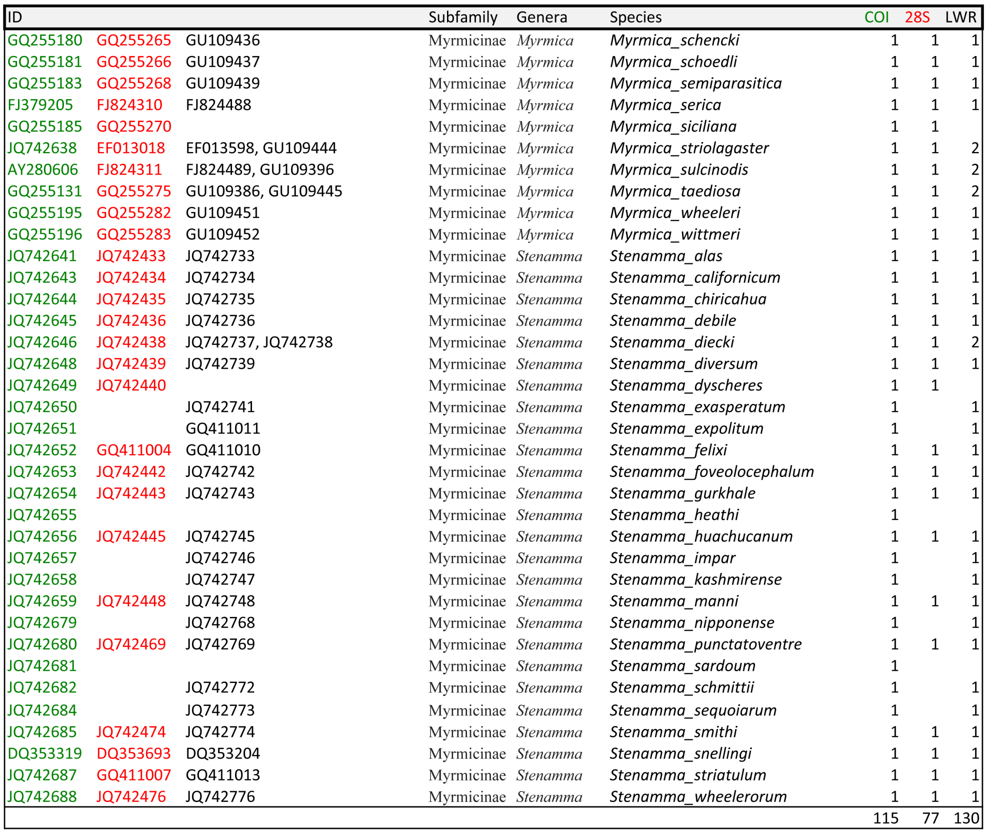
Fig. A.19.: Table S3B: Gene comparison. List of species with all three sequences (CO1, 28S rDNA, LWR) available. The list contains GenBank?IDs, Species names and number of specimen per species.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0007-AppendixS7.zip



ID			Subfamily	Genera	Species	COI	28S	LWR
JN134845	JN134298	JN134527	Formicinae	*Camponotus*	*Camponotus_aeneopilosus*	1	1	1
JN134846	JN134299	JN134528	Formicinae	*Camponotus*	*Camponotus_afflatus*	1	1	1
JN134850	JN134306	JN134534	Formicinae	*Camponotus*	*Camponotus_aurocinctus*	1	1	1
EF609765	JN134304	JN134532	Formicinae	*Camponotus*	*Camponotus_aurosus*	1	1	1
JN134847	JN134300	JN134529	Formicinae	*Camponotus*	*Camponotus_cinereus*	1	1	1
JN134856		JN134538, JN134539, JN134540	Formicinae	*Camponotus*	*Camponotus_claripes*	1		3
JN134900		JN134584	Formicinae	*Camponotus*	*Camponotus_claviscapus_occutus*	1		1
JN134860		JN134544	Formicinae	*Camponotus*	*Camponotus_consobrinus*	1		1
JN134863	JN134324	JN134547	Formicinae	*Camponotus*	*Camponotus_discors*	1	1	1
JN134933		JN134618	Formicinae	*Camponotus*	*Camponotus_evae_zeuxis*	1		1
JN134868		JN134552	Formicinae	*Camponotus*	*Camponotus_fellah*	1		1
JN134869	JN134330	JN134553	Formicinae	*Camponotus*	*Camponotus_fieldeae*	1	1	1
JN134871		JN134555	Formicinae	*Camponotus*	*Camponotus_gibbinotus*	1		1
JN134870	JN134333	JN134556	Formicinae	*Camponotus*	*Camponotus_gigas*	1	1	1
JN134873	JN134334	JN134557	Formicinae	*Camponotus*	*Camponotus_gouldianus*	1	1	1
HQ961340		JN134561	Formicinae	*Camponotus*	*Camponotus_herculeanus*	1		1
JN134878	JN134339	JN134562	Formicinae	*Camponotus*	*Camponotus_heteroclitus*	1	1	1
JN134879	EF012976	JN134563	Formicinae	*Camponotus*	*Camponotus_hyatti*	1	1	1
JN134881		JN134565	Formicinae	*Camponotus*	*Camponotus_intrepidus*	1		1
JN134883	JN134344	JN134567	Formicinae	*Camponotus*	*Camponotus_janeti*	1	1	1
JN134884		JN134568	Formicinae	*Camponotus*	*Camponotus_johnclarki*	1		1
JN134885	JN134346	JN134569	Formicinae	*Camponotus*	*Camponotus_latangulus*	1	1	1
JN134886	JN134347	JN134570	Formicinae	*Camponotus*	*Camponotus_ligniperdus*	1	1	1
JN134887	JN134349	JN134571	Formicinae	*Camponotus*	*Camponotus_mackayensis*	1	1	1
JN134893	JN134356	JN134578	Formicinae	*Camponotus*	*Camponotus_nigriceps*	1	1	1
DQ353282	DQ353654	DQ353158, EU367283	Formicinae	*Camponotus*	*Camponotus_ocreatus*	1	1	2
JN134903	JN134367	JN134587	Formicinae	*Camponotus*	*Camponotus_papago*	1	1	1
JN134904	JN134368	JN134588	Formicinae	*Camponotus*	*Camponotus_pawseyi*	1	1	1
JN134907	JN134371	JN134591	Formicinae	*Camponotus*	*Camponotus_prosseri*	1	1	1
JN134908		JN134592	Formicinae	*Camponotus*	*Camponotus_quercicola*	1		1
JN134915	JN134381		Formicinae	*Camponotus*	*Camponotus_scotti*	1	1	
JN134917	JN134383	JN134602	Formicinae	*Camponotus*	*Camponotus_sericeus*	1	1	1
JN134922		JN134608	Formicinae	*Camponotus*	*Camponotus_suffusus*	1		1
JN134923	JN134391	JN134610	Formicinae	*Camponotus*	*Camponotus_terebrans*	1	1	1
JN134925	JN134394	JN134611	Formicinae	*Camponotus*	*Camponotus_thadeus*	1	1	1
JN134926		JN134612	Formicinae	*Camponotus*	*Camponotus_tricoloratus*	1		1
JN134927	AY325957	JN134613	Formicinae	*Camponotus*	*Camponotus_vicinus*	1	1	1
JN134928	JN134397	JN134614	Formicinae	*Camponotus*	*Camponotus_vitreus*	1	1	1
JN134932		JN134616, JN134617	Formicinae	*Camponotus*	*Camponotus_wiederkehri*	1		2
GQ255122	GQ255209	GU109376	Myrmicinae	*Myrmica*	*Myrmica_aimonissabaudiae*	1	1	1
FJ379091	GQ255210	GU109377	Myrmicinae	*Myrmica*	*Myrmica_alaskensis*	1	1	1
GQ255125	FJ824294	FJ824471, GU109379	Myrmicinae	*Myrmica*	*Myrmica_americana*	1	1	2
GQ255126		GU109380	Myrmicinae	*Myrmica*	*Myrmica_anatolica*	1		1
GQ255127	GQ255215	GU109382	Myrmicinae	*Myrmica*	*Myrmica_angulinodis*	1	1	1
FJ824430	FJ824296	FJ824473	Myrmicinae	*Myrmica*	*Myrmica_arisana*	1	1	1
GQ255130		GU109384	Myrmicinae	*Myrmica*	*Myrmica_bergi*	1		1
FJ379121		GU109388, GU109389	Myrmicinae	*Myrmica*	*Myrmica_crassirugis*	1		2
GQ255136	GQ255222	GU109391	Myrmicinae	*Myrmica*	*Myrmica_discontinua*	1	1	1
GQ255138	GQ255224	GU109393	Myrmicinae	*Myrmica*	*Myrmica_dshungarica*	1	1	1
GQ255139	GQ255225	GU109394	Myrmicinae	*Myrmica*	*Myrmica_eidmanni*	1	1	1
FJ379246	FJ824298	FJ824475	Myrmicinae	*Myrmica*	*Myrmica_excelsa*	1	1	1
FJ379157	GQ255228	GU109390, GU109398	Myrmicinae	*Myrmica*	*Myrmica_fracticornis*	1	1	2
GQ255145		GU109400	Myrmicinae	*Myrmica*	*Myrmica_georgica*	1		1
GQ255146		GU109401	Myrmicinae	*Myrmica*	*Myrmica_hellenica*	1		1
DQ353360	DQ353629	DQ353225, FJ824477	Myrmicinae	*Myrmica*	*Myrmica_incompleta*	1	1	2
GQ255147	GQ255233	GU109402, GU109403	Myrmicinae	*Myrmica*	*Myrmica_indica*	1	1	2
FJ379241	FJ824301	FJ824478	Myrmicinae	*Myrmica*	*Myrmica_jessensis*	1	1	1
AY280596		GU109406, GU109407	Myrmicinae	*Myrmica*	*Myrmica_karavajevi*	1		2
GQ255152	GQ255237	GU109405	Myrmicinae	*Myrmica*	*Myrmica_kasczenkoi*	1	1	1
GQ255153		GU109408	Myrmicinae	*Myrmica*	*Myrmica_kirghisorum*	1		1
AB819150	FJ824302	FJ824479, GU109409, GU109410	Myrmicinae	*Myrmica*	*Myrmica_kotokui*	1	1	3
GQ255156	GQ255241		Myrmicinae	*Myrmica*	*Myrmica_lacustris*	1	1	
GQ255157		GU109412	Myrmicinae	*Myrmica*	*Myrmica_laurae*	1		1
FJ824437		FJ824480	Myrmicinae	*Myrmica*	*Myrmica_lobicornis*	1		1
FJ824438			Myrmicinae	*Myrmica*	*Myrmica_monticola*	1		
FJ379174	FJ824305	FJ824482	Myrmicinae	*Myrmica*	*Myrmica_nearctica*	1	1	1
GQ255166	GQ255251	GU109420, GU109421	Myrmicinae	*Myrmica*	*Myrmica_pisarskii*	1	1	2
FJ379175	GQ255252	GU109422	Myrmicinae	*Myrmica*	*Myrmica_punctinops*	1	1	1
HQ978871	GQ255253	GU109423	Myrmicinae	*Myrmica*	*Myrmica_punctiventris*	1	1	1
GQ255169	GQ255254	GU109424	Myrmicinae	*Myrmica*	*Myrmica_quebecensis*	1	1	1
GQ872391	FJ824306	FJ824483, GU109417	Myrmicinae	*Myrmica*	*Myrmica_rubra*	1	1	2
GQ255171		GU109426	Myrmicinae	*Myrmica*	*Myrmica_rugiventris*	1		1
GQ255149	GQ255234	GU109404	Myrmicinae	*Myrmica*	*Myrmica_rugosa*	1	1	1
AY280602		FJ824485, GU109427, GU109399	Myrmicinae	*Myrmica*	*Myrmica_rugulosa*	1		3
GQ255174	GQ255259	GU109429	Myrmicinae	*Myrmica*	*Myrmica_rupestris*	1	1	1
AY956325		FJ824486, GU109430	Myrmicinae	*Myrmica*	*Myrmica_sabuleti*	1		2
GQ255176		GU109431	Myrmicinae	*Myrmica*	*Myrmica_salina*	1		1
GQ255177		GU109432	Myrmicinae	*Myrmica*	*Myrmica_saposhnikovi*	1		1
AY280605		GU109434, GU109433	Myrmicinae	*Myrmica*	*Myrmica_scabrinodis*	1		2

Fig. A.20.: Table S4A: Gene comparison, reduced data. List of species with all three sequences (CO1, 28S rDNA, LWR) after editing. Editing included cleaving of 5' and 3' ends and reduction of duplicate sequences.



ID			Subfamily	Genera	Species	COI	28S	LWR
GQ255180	GQ255265	GU109436	Myrmicinae	*Myrmica*	*Myrmica_schencki*	1	1	1
GQ255181	GQ255266	GU109437	Myrmicinae	*Myrmica*	*Myrmica_schoedli*	1	1	1
GQ255183	GQ255268	GU109439	Myrmicinae	*Myrmica*	*Myrmica_semiparasitica*	1	1	1
FJ379205	FJ824310	FJ824488	Myrmicinae	*Myrmica*	*Myrmica_serica*	1	1	1
GQ255185	GQ255270		Myrmicinae	*Myrmica*	*Myrmica_siciliana*	1	1	
JQ742638	EF013018	EF013598, GU109444	Myrmicinae	*Myrmica*	*Myrmica_striolagaster*	1	1	2
AY280606	FJ824311	FJ824489, GU109396	Myrmicinae	*Myrmica*	*Myrmica_sulcinodis*	1	1	2
GQ255131	GQ255275	GU109386, GU109445	Myrmicinae	*Myrmica*	*Myrmica_taediosa*	1	1	2
GQ255195	GQ255282	GU109451	Myrmicinae	*Myrmica*	*Myrmica_wheeleri*	1	1	1
GQ255196	GQ255283	GU109452	Myrmicinae	*Myrmica*	*Myrmica_wittmeri*	1	1	1
JQ742641	JQ742433	JQ742733	Myrmicinae	*Stenamma*	*Stenamma_alas*	1	1	1
JQ742643	JQ742434	JQ742734	Myrmicinae	*Stenamma*	*Stenamma_californicum*	1	1	1
JQ742644	JQ742435	JQ742735	Myrmicinae	*Stenamma*	*Stenamma_chiricahua*	1	1	1
JQ742645	JQ742436	JQ742736	Myrmicinae	*Stenamma*	*Stenamma_debile*	1	1	1
JQ742646	JQ742438	JQ742737, JQ742738	Myrmicinae	*Stenamma*	*Stenamma_diecki*	1	1	2
JQ742648	JQ742439	JQ742739	Myrmicinae	*Stenamma*	*Stenamma_diversum*	1	1	1
JQ742649	JQ742440		Myrmicinae	*Stenamma*	*Stenamma_dyscheres*	1	1	
JQ742650		JQ742741	Myrmicinae	*Stenamma*	*Stenamma_exasperatum*	1		1
JQ742651		GQ411011	Myrmicinae	*Stenamma*	*Stenamma_expolitum*	1		1
JQ742652	GQ411004	GQ411010	Myrmicinae	*Stenamma*	*Stenamma_felixi*	1	1	1
JQ742653	JQ742442	JQ742742	Myrmicinae	*Stenamma*	*Stenamma_foveolocephalum*	1	1	1
JQ742654	JQ742443	JQ742743	Myrmicinae	*Stenamma*	*Stenamma_gurkhale*	1	1	1
JQ742655			Myrmicinae	*Stenamma*	*Stenamma_heathi*	1		
JQ742656	JQ742445	JQ742745	Myrmicinae	*Stenamma*	*Stenamma_huachucanum*	1	1	1
JQ742657		JQ742746	Myrmicinae	*Stenamma*	*Stenamma_impar*	1		1
JQ742658		JQ742747	Myrmicinae	*Stenamma*	*Stenamma_kashmirense*	1		1
JQ742659	JQ742448	JQ742748	Myrmicinae	*Stenamma*	*Stenamma_manni*	1	1	1
JQ742679		JQ742768	Myrmicinae	*Stenamma*	*Stenamma_nipponense*	1		1
JQ742680	JQ742469	JQ742769	Myrmicinae	*Stenamma*	*Stenamma_punctatoventre*	1	1	1
JQ742681			Myrmicinae	*Stenamma*	*Stenamma_sardoum*	1		
JQ742682		JQ742772	Myrmicinae	*Stenamma*	*Stenamma_schmittii*	1		1
JQ742684		JQ742773	Myrmicinae	*Stenamma*	*Stenamma_sequoiarum*	1		1
JQ742685	JQ742474	JQ742774	Myrmicinae	*Stenamma*	*Stenamma_smithi*	1	1	1
DQ353319	DQ353693	DQ353204	Myrmicinae	*Stenamma*	*Stenamma_snellingi*	1	1	1
JQ742687	GQ411007	GQ411013	Myrmicinae	*Stenamma*	*Stenamma_striatulum*	1	1	1
JQ742688	JQ742476	JQ742776	Myrmicinae	*Stenamma*	*Stenamma_wheelerorum*	1	1	1
						115	77	130

Fig. A.21.: Table S4B: Gene comparison, reduced data. List of species with all three sequences (CO1, 28S rDNA, LWR) after editing. Editing included cleaving of 5' and 3' ends and reduction of duplicate sequences.

TAXA\Position	6	27	33	37	48	54	84	102	114	126	165	177	180	183	195	207
KC572877_Mor	A	T	C	C	T	A	A	C	T	G	A	T	T	C	C	G
KC572878_Mor	C	C	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572895_Con	C	T	C	C	T	G	A	C	T	A	A	T	T	C	T	A
KC572896_Con	C	T	C	C	T	G	A	C	T	A	A	T	T	C	T	A
KC572907_Ulu	C	T	T	C	C	A	A	T	T	A	A	C	T	C	C	A
KC572922_Car	C	T	C	T	T	A	A	C	T	A	G	T	T	C	C	A
KC572940_Miy	C	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572946_Dav	C	T	C	C	T	G	A	C	T	A	A	T	T	C	C	A
KC572954_Ood	C	T	T	C	T	A	A	T	T	A	A	C	T	C	C	A
KC572964_Bea	A	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572965_Bea	A	T	C	C	T	A	A	C	C	A	A	T	T	C	C	A
KC572967_Hed	A	T	C	C	T	A	A	C	T	A	A	T	T	T	C	A
KC572968_Hed	A	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572969_Hed	A	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572971_Por	A	T	C	C	T	A	T	C	T	A	A	T	T	C	C	A
KC572973_Poi	A	T	T	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572976_Roe	A	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572984_Mee	C	T	C	C	T	A	A	C	T	A	A	T	C	C	C	A
KC572989_Wil	A	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572990_Wil	T	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572991_Wil	T	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572993_Mora	A	T	C	C	T	A	A	C	T	A	A	T	T	C	C	A
KC572994_Leo	C	T	C	C	T	A	A	C	T	A	A	T	C	C	C	A
KC572995_Leo	C	T	C	C	T	A	A	C	T	A	A	T	C	C	C	A
KC572996_Leo	C	T	C	C	T	A	A	C	T	A	A	T	C	C	C	A

TAXA\Position	210	219	223	234	261	282	312	315	384	408	453	507	513	561	567	573
KC572877_Mor	C	C	C	T	T	A	C	G	C	T	C	A	G	C	G	A
KC572878_Mor	C	G	C	T	C	A	T	G	C	C	C	T	G	T	A	A
KC572895_Con	C	G	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572896_Con	C	G	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572907_Ulu	C	A	C	T	T	C	T	A	C	T	T	A	G	C	A	A
KC572922_Car	C	G	T	C	T	A	T	G	C	T	C	A	A	C	A	G
KC572940_Miy	C	G	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572946_Dav	C	G	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572954_Ood	A	A	C	T	T	C	T	A	C	T	C	A	G	C	A	A
KC572964_Bea	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572965_Bea	C	C	C	T	T	C	T	G	C	T	C	A	G	C	A	A
KC572967_Hed	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572968_Hed	C	C	C	T	T	A	T	A	C	T	C	A	G	C	A	A
KC572969_Hed	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572971_Por	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572973_Poi	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572976_Roe	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572984_Mee	C	G	T	T	T	A	T	A	T	T	C	A	G	C	A	A
KC572989_Wil	C	C	C	T	T	A	T	T	C	T	C	A	G	C	A	A
KC572990_Wil	C	G	T	T	T	A	T	A	C	T	C	A	G	C	A	A
KC572991_Wil	C	G	T	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572993_Mora	C	C	C	T	T	A	T	G	C	T	C	A	G	C	A	A
KC572994_Leo	C	G	T	T	T	A	T	A	C	T	C	A	G	C	A	A
KC572995_Leo	C	G	T	T	T	A	T	A	C	T	C	A	G	C	A	A
KC572996_Leo	C	G	T	T	T	A	T	A	C	T	C	A	G	C	A	A

TAXA\Position	579	585	597	598	603	609	615	616	617	618	619	620	621	622	623	624
KC572877_Mor	T	T	T	T	C	T	A	C	A	T	C	C	T	G	A	A
KC572878_Mor	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572895_Con	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572896_Con	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572907_Ulu	T	C	T	C	T	C	G	C	A	T	C	C	T	G	A	A
KC572922_Car	T	T	T	C	C	C	A	C	A	T	C	C	T	G	A	A
KC572940_Miy	T	T	T	C	C	C	A	C	A	T	C	C	T	G	A	A
KC572946_Dav	T	T	T	C	C	C	A	C	A	T	C	C	T	G	A	A
KC572954_Ood	T	T	T	C	C	C	A	C	A	T	C	C	T	G	A	A
KC572964_Bea	T	T	T	T	C	T	A	C	A	T	C	C	T	G	A	A
KC572965_Bea	T	T	T	T	C	T	A	C	A	T	C	C	T	G	A	A
KC572967_Hed	T	T	T	T	C	T	A	C	A	T	C	C	T	G	A	A
KC572968_Hed	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572969_Hed	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572971_Por	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572973_Poi	T	T	T	T	C	T	A	C	A	T	C	C	T	G	A	A
KC572976_Roe	T	T	T	T	C	T	A	C	A	T	C	C	T	G	A	A
KC572984_Mee	T	T	T	C	C	T	-	-	-	-	-	-	-	-	-	-
KC572989_Wil	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572990_Wil	C	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572991_Wil	C	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572993_Mora	T	T	T	T	C	T	-	C	A	T	C	C	G	A	A	G
KC572994_Leo	T	T	C	C	C	C	-	-	-	-	-	-	-	-	-	-
KC572995_Leo	T	T	T	C	C	T	A	C	A	T	C	C	T	G	A	A
KC572996_Leo	T	T	T	C	C	C	A	C	A	T	C	C	T	G	A	A

Fig. A.22.: Table S8A: Barcode Table. Table shows all CAs detected with the CAOS-Analyzer and visualized with the CAOS-Barcoder. CAs were selected from the "sequence?similarity" and "sequence?origin" data sets. The diagnostics found in both data sets are identical. In the first column specimen ID's are listed. In the following columns diagnostics and their position within the barcode are highlighted.

TAXA\Position	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640
KC572877_Mor	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572878_Mor	G	T	C	T	A	T	A	T	T	C	T	A	A	T	C	C
KC572895_Con	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572896_Con	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572907_Ulu	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572922_Car	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572940_Miy	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572946_Dav	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572954_Ood	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572964_Bea	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572965_Bea	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572967_Hed	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572968_Hed	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572969_Hed	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572971_Por	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572973_Poi	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572976_Roe	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572984_Mee	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
KC572989_Wil	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572990_Wil	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572991_Wil	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572993_Mora	-	-	-	T	T	A	A	T	T	C	T	A	A	T	C	C
KC572994_Leo	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
KC572995_Leo	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C
KC572996_Leo	G	T	C	T	A	C	A	T	T	C	T	A	A	T	C	C

TAXA\Position	641	642	643	644	645	654	660	669	672	709	753	754	762	813	840	843
KC572877_Mor	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572878_Mor	T	T	C	C	T	A	C	T	C	T	C	A	A	T	C	T
KC572895_Con	T	T	C	C	T	A	C	C	C	T	C	-	A	T	T	T
KC572896_Con	T	T	C	C	T	A	C	C	C	T	C	A	A	T	T	T
KC572907_Ulu	T	T	C	C	T	A	C	T	C	T	C	A	G	C	T	C
KC572922_Car	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	A
KC572940_Miy	T	T	C	C	T	A	C	T	T	T	C	A	A	T	T	T
KC572946_Dav	T	T	C	C	T	A	C	C	C	C	C	A	A	T	T	T
KC572954_Ood	T	T	C	C	T	A	C	T	C	T	C	A	G	C	T	C
KC572964_Bea	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572965_Bea	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572967_Hed	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572968_Hed	T	T	C	C	T	A	T	T	C	T	T	A	A	T	T	C
KC572969_Hed	T	T	C	C	T	A	T	T	C	T	T	A	A	T	T	C
KC572971_Por	T	T	C	C	T	G	T	T	C	T	T	A	A	T	T	C
KC572973_Poi	T	T	C	C	T	A	C	T	C	T	T	A	A	T	T	C
KC572976_Roe	T	T	C	C	T	A	C	T	C	T	T	A	A	T	T	C
KC572984_Mee	-	-	-	-	-	A	C	T	C	T	C	A	A	T	T	C
KC572989_Wil	T	T	C	C	T	A	T	T	C	T	C	A	A	T	T	C
KC572990_Wil	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	T
KC572991_Wil	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572993_Mora	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572994_Leo	-	-	-	-	-	A	C	T	C	T	C	A	A	T	T	C
KC572995_Leo	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C
KC572996_Leo	T	T	C	C	T	A	C	T	C	T	C	A	A	T	T	C

TAXA\Position	864	882	885	924
KC572877_Mor	A	T	T	G
KC572878_Mor	A	T	T	G
KC572895_Con	A	T	T	G
KC572896_Con	A	T	T	G
KC572907_Ulu	A	T	T	G
KC572922_Car	A	C	T	G
KC572940_Miy	A	T	C	G
KC572946_Dav	A	T	T	G
KC572954_Ood	A	T	T	G
KC572964_Bea	A	T	T	G
KC572965_Bea	G	T	T	G
KC572967_Hed	A	T	T	G
KC572968_Hed	A	T	T	A
KC572969_Hed	A	T	T	A
KC572971_Por	A	T	T	A
KC572973_Poi	A	T	T	G
KC572976_Roe	A	T	T	G
KC572984_Mee	A	T	T	A
KC572989_Wil	A	T	T	G
KC572990_Wil	A	T	T	G
KC572991_Wil	A	T	T	G
KC572993_Mora	A	T	T	G
KC572994_Leo	A	T	T	G
KC572995_Leo	A	T	T	G
KC572996_Leo	A	T	T	G

Fig. A.23.: Table S8B: Barcode Table. Table shows all CAs detected with the CAOS-Analyzer and visualized with the CAOS-Barcoder. CAs were selected from the "sequence?similarity" and "sequence?origin" data sets. The diagnostics found in both data sets are identical. In the first column specimen ID's are listed. In the following columns diagnostics and their position within the barcode are highlighted.

Appendix S9 CA overview 5a. Detailed overview of simple pure (sPu) and simple private (sPr) characters identified at each branching point within the used trees of the ’sequence-similarity’ and ’sequence-origin’ data sets. ’Map-Translation.xlsx’ shows a legend with the region codes translated into locations and geographic positions. ’Overview 2-tables’ show sPu and sPr diagnostics for both data sets. ’Overview 3-table’ highlights sPu only.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0009-AppendixS9.zip

Appendix S10 Classification. 25 specimen with unique diagnostics were character-based barcoded. The CAOS-Classifier was used to test if the barcode matrices based on these 25 specimen could be used to accurately identify the 25 specimen. The ’classifier-output.html’ files show that in both approaches all 25 specimen were identified successfully.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0010-AppendixS10.zip

Appendix S11 Region barcodes. Here, all sPu and sPr characters for each branching event of the origin matrix are listed. Each branching event is shown as a single table. The first column shows the location origin of each specimen. The second column shows the specimen. Specimen within the left branch are colored green. Specimen of the right branch are colored red. The following columns show the position of diagnostics within the barcode and the diagnostic. ’no CA’ means that for this specimen no diagnostic was detected for this position.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0011-AppendixS11.zip

Appendix S12 All location specific CAs. This supplementary tables show all diagnostic characters identified with the CAOS-Analyzer. It also shows how we identified unique diagnostics for specific groups in a step by step approach.

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.12395&file=men12395-sup-0012-AppendixS12.zip

B Curriculum Vitae

Dipl. Biol. Tjard Bergmann

Untere Hauptallee 3, Bad Pyrmont 31812, born 25.01.1980 in Borkum

Academic education

08/2002-12/2007	Diploma degree Biology Leibniz Universität Hannover Major Genetic (1,0) 1. Minor Microbiology (1,0) 2. Minor Immunology (1,0) 3. Minor Biochemistry (1,0) Diploma thesis at the ITZ, Institute for Animal Ecology & Evolution (University of Veterinary Medicine Hannover, Foundation) Topic: "Experimental Studies of Function and Expression of Opsin-Genes in *Trichoplax adhaerens* (Placozoa)"; Supervisor: Prof. Dr. Bernd Schierwater (1,5) Diploma degree in Biology with the final grade "very good" (1,2)
since 12/2008	Working on doctoral thesis at the Institute of Animal Ecology and Evolution Subject: "Character-based barcoding, a symbiosis and potential successor of traditional taxonomy and modern DNA barcoding Supervisor: Prof. Dr. Bernd Schierwater.

Working experience

01/2008-12/2008	Research assistant for research and teaching at the ITZ, Institute of Animal Ecology and Evolution (University of Veterinary Medicine Hannover, Foundation)
08/2010-10/2010	Research stay at the American Museum of Natural History (New York) and the University of Vermont (Burlington) Development of core algorythms for our cooperation project "CAOS" (http://boli.uvm.edu/caos-workbench/) at the AMNH Bioinformatic cooperation in the publication (Reid et al., 2011)
11/2011-12/2011	Participation at the fourth international "Barcode of Life" Conference in Adelaide, Australia Powerpoint-Presentation (Data Analysis Methods: "CAOS-Workbench: A character-based barcoding platform")
07/2012-07/2014	Research assistant for research and teaching at the ITZ, Institute of Animal Ecology and Evolution (University of Veterinary Medicine Hannover, Foundation) Teaching of students in molecular-biologic methods and bioinformatic, data presentation and research
08/2014-12/2018	Technical Assistant for Informatics in the Institute of Zoology (University of Veterinary Medicine Hannover, Foundation) Responsible for bioacoustic analysis, EDV, data management, Supervision of experimental setups

Fellowships

12/2008	H. Wilhelm Schaumann (Duration: 12/2008 - 04/2012)
08/2010	DAAD (Duration: 08/2010 - 10/2010)
12/2011	Graduate Academy of the Leibniz University Hannover
10/2013	University of Veterinary Medicine Hannover, Foundation (Duration:10/2013 - 07/2014)

List of Publications

C

A. Journals

1. Tizard J et al.. (2018) DNA barcoding a unique avifauna: an important tool for evolution, systematics and conservation. BMC Evol Biol. In revision.

2. Rach J, **Bergmann T**, Paknia O, et al. (2017) The marker choice: Unexpected resolving power of an unexplored CO1 region for layered DNA barcoding approaches. PLoS One 12, e0174842. doi: 10.1371/journal.pone.0174842.

3. Paknia O, **Bergmann T**, Hadrys H. (2015) Some ’ant’swers: Application of a layered barcode approach to problems in ant taxonomy. Mol Ecol Resour 15, 1262-1274. doi: 10.1111/1755-0998.12395.

4. **Bergmann T**, Rach J, Damm S, DeSalle R, Schierwater B and Hadrys H. (2013) The potential of distance-based thresholds and character-based DNA barcoding for defining problematic taxonomic entities by CO1 and ND1. MER. doi: 10.1111/1755-0998.12125.

5. Reid BN, Le M, McCord WP, Iverson JB, Georges A, **Bergmann T**, Amato G, DeSalle R, Naro-Maciel E. (2011) Comparing and combining distance-based and character-based approaches for barcoding Turtles. Mol Ecol Resour. doi: 10.1111/j.1755-0998.2011.03032.x.

6. **Bergmann T**, Hadrys H, Breves G, Schierwater B. (2009) Character-based DNA barcoding: a superior tool for species classification. Berl Munch Tierarztl Wochenschr. Nov-Dec; 122(11-12):446-50. doi: 10.2376/0005-9366-122-446.

B. Book chapter

1. Eitel M, Jakob W, Osigus HJ, Paknia O, von der Chevallerie K, **Bergmann T**, and Schierwater B. (2014) Phylogenetics and phylogenomics at the root of Metazoa. Pages 23-49 in J. W. Wägele and T. Bartolomaeus, editors. Deep Metazoan Phylogeny: The Backbone of the Tree of Life, New insights from analyses of molecules, morphology, and theory of data analysis. DE Gruyter Berlin.

2. DeSalle R, Sun T-T, **Bergmann T**, Garcia-Espana A. (2013) The Evolution of Tetraspanins Through a Phylogenetic Lens. Pages 31-45 in F. Berditchevski and E. Rubinstein, editors. Tetraspanins. Springer Netherlands. doi: 10.1007/978-94-007-6070-7_2.

3. Schierwater B, Eitel M, Osigus HJ, von der Chevallerie K, **Bergmann T**, Hadrys H, Cramm M, Heck L, LMR, and DeSalle R. (2010) *Trichoplax* and Placozoa: one of the crucial keys to understanding metazoan evolution. Pp. 289-326 in Key transitions in animal evolution, R. DeSalle and B. Schierwater, eds. CRC Press.

D Acknowledgement

This thesis would not have been possible without the help of various people that contributed to the completion of my thesis in different ways. In this section, I will express my appreciation for their support.

First and foremost I'd like to express my deep gratitude to my supervisor Prof. Dr. Bernd Schierwater for taking me in. Thanks to Prof. Dr. Bernd Schierwater guidance, I changed from a shy boy into a confident scientist. He taught me to think outside of the box, that good sciences is not beautiful, but efficient and how to write manuscripts. I learned to perform routine as well as next generation molecular methods.

I am furthermore thankful to Prof. Dr. Rob DeSalle, and Prof. Dr. Felix Felmy who willingly agreed to review this thesis and to Prof. Dr. Hansjörg Küster and Prof. Dr. Jutta Papenbrock for being part of the commission and hosting my disputation.

I thank PD Dr. Heike Hadrys for entrusting me with multiple projects that all bear fruit in great publications.

I thank all my (former and present) colleagues at the ITZ who have supported me over the years and where always there for me when I needed advise or a friendly word. With them, Annkathrin Acktun, Nicole Bartkowiak, Jutta Bunnenberg, Dr. Sandra Damm, Dr. Michael Eitel, Angie Faust, Wiebke Feindt, PD Dr. Heike Hadrys, Rebecca Herzog, Tina Herzog, Dr. Eckhard Holtorf, Dr. Wolfgang Jakob, Ulrike Oberjatzas, Haju Osigus, Dr. Omid Paknia, Sarah Rolfes, Dr. Sven Sagasser, Dr. Dasa Schleicherova, Björn Seegebarth and Dr. Sabrina Simon the time in the institute always felt pleasant. I also thank the members of the Institute of Zoology, especially Prof. Dr. Zimmermann and Prof. Dr. Felmy, for their support.

One dream that was realized thanks to Prof. Dr. Schierwater was my research stay in New York. It was my first flight and journey to a different continent. In New York I was open heartedly welcomed by Prof. Dr. Rob DeSalle. Prof. Dr. Rob DeSalle, one of the most hard-working, knowledgeable scientists I know, took the time to show me around his museum. Through his guidance and the help of Prof. Dr. Indra Neil Sarkar, I was able to realize our web-based CAOS-Workbench. Thanks to him, even while I was only staying for three month in New York, we were even able to publish a manuscript together, which is now part of this thesis (Reid et al. 2011).

This thesis would not have been possible without the financial support from the Friedrich Schaumann fellowship, a fellowship by the University of Veterinary Medicine Hannover and the help of Prof. Dr. Breves. Thanks to the DAAD, I was able to work for three month in New York. Thanks to the Graduate Academy of the LUH, I could visit the "Barcode of Life" Conference in Adelaide and present character-based barcoding to a broad audience of fellow barcoders.

This thesis is dedicated to my family and lovely wife. They always stood behind me, sometimes with fire and forks, when needed. My last words go out to my lovely daughter Emma Johanna Bergmann. With her kind and equally adventurous nature she never fails to put a smile on my face.