OULU 2016
A 673

ACTA UNIVERSITATIS OULUENSIS

Mikko Pentinsaari

UTILITY OF DNA BARCODES IN IDENTIFICATION AND DELIMITATION OF BEETLE SPECIES, WITH INSIGHTS INTO COI PROTEIN STRUCTURE ACROSS THE ANIMAL KINGDOM

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF SCIENCE

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 673

MIKKO PENTINSAARI

UTILITY OF DNA BARCODES IN IDENTIFICATION AND DELIMITATION OF BEETLE SPECIES, WITH INSIGHTS INTO COI PROTEIN STRUCTURE ACROSS THE ANIMAL KINGDOM

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in Kuusamonsali (YB210), Linnanmaa, on 6 May 2016, at 12 noon

UNIVERSITY OF OULU, OULU 2016

Copyright © 2016
Acta Univ. Oul. A 673, 2016

Supervised by
Docent Marko Mutanen
Docent Lauri Kaila

Reviewed by
Docent Gunilla Ståhls-Mäkelä
Associate Professor Sarah J. Adamowicz

Opponent
Doctor Michael Balke

ISBN 978-952-62-1209-8 (Paperback)
ISBN 978-952-62-1210-4 (PDF)

ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2016

Pentinsaari, Mikko, Utility of DNA barcodes in identification and delimitation of beetle species, with insights into COI protein structure across the animal kingdom.
University of Oulu Graduate School; University of Oulu, Faculty of Science
Acta Univ. Oul. A 673, 2016
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Species are the fundamental units of biological diversity, but their identification and delimitation is often difficult. The difficulties are pronounced in diverse taxa such as insects. DNA barcodes, short standardized segments of the genome, have recently become a popular tool for identifying specimens to species, and are increasingly used as one of the sources of information for species delimitation. In this thesis, I studied the utility of DNA barcodes in species identification and delimitation in beetles (Coleoptera). Beetles are one of the most diverse animal groups, with nearly 400 000 known species. The Nordic beetle fauna is among the most thoroughly studied on the planet, providing excellent conditions for these studies. I also approached barcode sequences from a new angle, exploring amino acid variation and its connections to life history in a sample of the entire animal kingdom. I also studied variation and evolution at the amino acid level in large-scale samples of beetles and moths & butterflies (Lepidoptera). DNA barcodes proved to be a feasible tool for identifying species of Nordic beetles: depending on the criteria for successful identification, 95-98% of specimens could be identified to the species level based on DNA barcodes. Regardless of the delimitation method used, approximately 90% of the currently accepted species were perfectly recovered based on barcode data, and simple rules for forming consensus between delimitations improved the fit between species and barcode clusters even further. Several species that were split into two or more sequence clusters apparently include species new to science that have been previously overlooked. This conclusion is supported by preliminary morphological analysis. The study on amino acid variation revealed both a general pattern of structural conservation throughout the animal kingdom, and some interesting amino acid substitutions with potential to affect enzymatic function. Amino acid variation was more extensive in Coleoptera than in Lepidoptera, potentially due to differences in selection pressure and patterns of molecular evolution in the barcode region between the two orders.

Keywords: Coleoptera, DNA barcoding, protein structure, species delimitation, taxonomy

Pentinsaari, Mikko, DNA-viivakoodien käyttö kovakuoriaisten lajintunnistuksessa ja -rajauksessa, ja eläinten COI-proteiinin rakenteen muuntelu.

Oulun yliopiston tutkijakoulu; Oulun yliopisto, Luonnontieteellinen tiedekunta

Acta Univ. Oul. A 673, 2016
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Laji on luonnon monimuotoisuuden perusyksikkö, mutta lajien tunnistaminen ja rajaaminen on usein vaikeaa. Vaikeudet korostuvat erityisesti hyvin monimuotoisissa eliöryhmissä kuten hyönteisissä. DNA-viivakoodit ovat lyhyitä standardoituja DNA-sekvenssejä, joiden käyttö lajien tunnistamisessa sekä yhtenä tiedon lähteenä lajien rajaamisessa on viime aikoina yleistynyt nopeasti. Tutkin väitöskirjatyössäni DNA-viivakoodien soveltuvuutta lajinmääritykseen ja lajien rajaamiseen kovakuoriaisilla. Kovakuoriaiset ovat yksi maailman lajirikkaimmista eliöryhmistä: lajeja on kuvattu lähes 400000. Pohjois-Euroopan lajisto tunnetaan koko maailman mittakaavassa poikkeuksellisen hyvin, mikä tarjoaa erinomaiset edellytykset tutkia DNA-viivakoodeihin liittyviä kysymyksiä kuoriaisilla. Tutkin DNA-viivakoodeja myös kokonaan uudesta näkökulmasta, selvittäen aminohappotason muuntelua koko eläinkunnan kattavassa otoksessa, sekä laajalla perhos- ja kuoriaisaineistolla. DNA-viivakoodit osoittautuivat erinomaiseksi työkaluksi lajinmääritykseen: riippuen onnistuneen määrityksen kriteereistä 95–98 % kuoriaislajeista voitiin tunnistaa luotettavasti viivakoodien perusteella. Käytetystä menetelmästä riippumatta noin 90 % nykykäsityksen mukaisista lajeista voitiin rajata viivakoodien perusteella oikein, ja soveltamalla yksinkertaisia konsensussääntöjä yhteensopivuus lajien ja viivakoodiklustereiden välillä kasvoi entisestään. Useat kuoriaislajit, jotka jakautuivat kahteen tai useampaan viivakoodiklusteriin, sisältävät alustavien morfologisten tutkimusten perusteella aiemmin huomaamatta jääneitä uusia lajeja. Aminohappo- ja proteiinitason tutkimus osoitti, että viivakoodijakson koodaaman proteiinin rakenne on yleisesti ottaen konservoitunut kautta eläinkunnan. Havaitsin kuitenkin myös useita kiinnostavia aminohappomuutoksia, jotka saattavat vaikuttaa entsyymitoimintaan. Aminohapposekvenssi muuntelee kuoriaisilla paljon enemmän kuin perhosilla, mahdollisesti johtuen taksonien välisistä eroista molekyylievoluutiossa ja viivakoodisekvenssiin kohdistuvassa valintapaineessa.

Asiasanat: DNA-viivakoodit, kovakuoriaiset, lajinrajaus, proteiinirakenne, taksonomia

Acknowledgements

Nature and animals have fascinated me since early childhood. My parents and many other relatives have supported and encouraged this interest from very early on, and still continue to do so, for which I am truly grateful. My interest focused on beetles around age 11 or 12, when I caught some ladybirds (*Adalia bipunctata*, to be specific) and kept them as pets, feeding them aphids and watching them grow from larvae to adults. Without help from more experienced entomologists, getting started with beetle collecting would have been very difficult. Fortunately, my mother got the idea to contact people at the zoological museum at the University of Oulu, specifically Juhani "Jussi" Itämies. I cannot thank Jussi enough for taking my interest in beetles seriously and helping me get started with entomology as a schoolboy. Jussi taught me all the basics about collecting and identifying insects, and his continued support during the past 20 years is one of the most important reasons why I became a professional entomologist and ended up writing this thesis. Another major contributor to my career choice is my former biology teacher Antti Rönkä. The basics of scientific thinking learned during his biology courses in Lyseo and especially the work I did for the Viksu science competition under his supervision confirmed my decision to go for a career in biology.

I do not remember exactly when I first met my supervisor Marko Mutanen, but our first collecting trip together – the first of dozens of trips to all over Finland – was in 2004, a couple of months before I started my studies at the university. Marko's extensive knowledge of the natural world and keen interest in it never ceases to impress me. Much of what I know about plants, birds and of course insects has been learned from Marko during our field trips and numerous discussions besides them. Pretty soon after I started my studies, we already had an initial plan for my master's thesis, which was finished in 2010 and naturally supervised by Marko. Marko's help was crucial in all stages of this project, from funding applications and field work to data analysis and manuscript preparation. Thank you for all your support so far.

My second supervisor Lauri Kaila also co-supervised already my master's thesis. Even though most of the supervision work has fallen to Marko, largely because Lauri's office is hundreds of kilometers further away from mine than Marko's, I am grateful that Lauri agreed to become my supervisor. Lauri's comments on my work have been extremely helpful and all my manuscripts, especially the first ones, became a lot better thanks to him.

Collecting the beetle material for the Finnish Barcode of Life project has been a huge effort and is still ongoing. During this project, I have received invaluable help from many of my coleopterist friends and colleagues in the form of legs or whole beetle specimens for DNA extraction. I wish to thank especially Tom Clayhills, Eero Helve, Sampsa Malmberg, Petri Martikainen, Veli-Matti Mukkala, Juha Salokannel, Mikko Tiusanen and Jussi Vilen. It has been a pleasure to be a member of the Finnish Expert Group on Beetles ─ the meetings and field excursions have been immensely useful and educational, as well as a lot of fun. The same goes for the summer meetings of the Nordic Coleoptera Group during the past few years. Thank you for the memorable moments to all my Finnish and Nordic coleopterist colleagues.

The former Department of Biology has been a very supportive environment for PhD studies. Especially the peer support from the present and former PhD students has been invaluable. A big thank you to Hilde, Vekku, Tuomo, Anni, Jani, Suvi, Nelli, Emma, Netta and all the rest of you. Thanks also to my follow-up group Jouni Aspi, Jukka Forsman & Panu Välimäki, and other members of the more "senior" staff of the former department for their help along the way.

Finally, I want to thank Samuli, Seppo R., Jere, Anni and Risto T. for memorable lab & field moments at the maaelukka field courses (always one of the most enjoyable times of the year), Ida & Osku for providing distraction during the intense final year of this PhD project, and Anne, Harry, Sami, Teppo, Perttu and others in the local entomological community in Oulu for company in the field and the winter meetings, and especially the road trips to the yearly Finnish entomologists' weekend meetings.

This project has been funded by grants from the Jenny and Antti Wihuri Foundation and the Ella and Georg Ehrnrooth Foundation, as well as travel grants for conference trips from the University of Oulu Graduate School. The FinBOL project, in context of which all my material was sequenced, was funded by the Kone Foundation, the Finnish Cultural Foundation and the University of Oulu.

Oulu, April 4, 2016 Mikko Pentinsaari

Abbreviations

ABGD	Automatic Barcode Gap Discovery algorithm
BIN	Barcode Index Number, an interim taxonomic system used in the BOLD database
BOLD	Barcode of Life Data Systems (http://www.boldsystems.org/), portal for sequence data storage and analysis tools
CCDB	The Canadian Centre for DNA Barcoding (http://www.ccdb.ca/)
COI	Cytochrome C oxidase subunit I gene
COX	The cytochrome C oxidase protein complex
FinBOL	The Finnish Barcode of Life project (http://finbol.org/)
GMYC	General Mixed Yule Coalescent model
NJ	Neighbor-Joining
NN	Nearest neighbor species
OTU	Operational Taxonomic Unit
PTP	Poisson Tree Processes model

Original articles and contributions

This thesis is based on the following articles, which are referred to throughout the text by their Roman numerals:

I Pentinsaari M, Hebert PDN & Mutanen M (2014) Barcoding Beetles: A Regional Survey of 1872 Species Reveals High Identification Success and Unusually Deep Interspecific Divergences. PLoS ONE 9(9): e108651.

II Pentinsaari M, Vos R & Mutanen M (2015) Algorithmic single-locus species delimitation: effects of sampling effort, variation and non-monophyly in four methods and 1870 species of beetles. Manuscript.

III Pentinsaari M, Salmela H, Mutanen M & Roslin T (2015) A widely-adopted taxonomic marker sheds light on protein evolution across the animal tree of life. Manuscript.

Contributions

Authors' major contributions to the original articles.

	I	II	III
Original Idea	MP, MM	MP, MM	TR, HS
Data collection	MP, MM	MP	MP, MM
Analyses	MP	MP, RV	HS, MP
Manuscript preparation	MP, PH, MM	MP, MM, RV	MP, TR, HS, MM

MP: Mikko Pentinsaari, MM: Marko Mutanen, PH: Paul D. N. Hebert, RV: Rutger Vos, HS: Heli Salmela, TR: Tomas Roslin

Table of contents

Abstract
Tiivistelmä
Acknowledgements **7**
Abbreviations **9**
Original articles and contributions **11**
Table of contents **13**
1 Introduction **15**
1.1 Species ... 15
1.2 DNA taxonomy ... 16
1.3 DNA barcoding ... 18
1.4 The focal taxon: North European beetles ... 20
1.5 Aims of the study ... 21
2 Material and methods **23**
2.1 Studied material ... 23
2.2 Laboratory procedures ... 23
2.3 Data analyses ... 24
2.3.1 Species identification (I) ... 24
2.3.2 OTU delimitation (II) ... 25
2.3.3 Sampling effects (I, II) ... 29
2.3.4 Effects of non-monophyly and genetic variation on OTU delimitation (II) ... 29
2.3.5 Protein structure modelling (III) ... 30
3 Results and discussion **33**
3.1 Barcode-based identification of North European beetles (I) ... 33
3.2 Accuracy and sensitivity of OTU delimitation methods (II) ... 34
3.3 Sampling effects (I, II) ... 38
3.4 Amino acid variation and structural changes in the COX protein (III) ... 39
4 Conclusions **43**
References **45**
Original articles **55**

1 Introduction

1.1 Species

Diversity is one of the most striking characteristics of life, and the fundamental unit of biological diversity is usually considered to be species (Mayr 1982). The exact nature of species, and if they are even real entities, has for a long time been controversial (Claridge 2009, Mishler 2009). A long list of competing species concepts have been presented: Mayden (1997) lists 22 distinct definitions, which may produce discordant species boundaries when applied to empirical data. In addition, particular species concepts may be inapplicable in some commonly encountered situations, *e.g.* allopatric populations or asexual reproduction in the case of the biological species concept. De Queiroz (2005a, 2005b, 2007) has attempted to find a common element in the competing views and unify them under one inclusive concept (dubbed the General Lineage Concept) which defines species as "separately evolving metapopulation lineages", and the definitive criteria set by the previously presented concepts (reproductive isolation, reciprocal monophyly, niche divergence etc.) are considered merely lines of evidence on lineage separation. The criteria posed by the various concepts are likely to be fulfilled gradually one by one as new species lineages diverge. The diverging lineages will evolve differing ecologies, accumulate fixed diagnostic genetic and morphological characters, their genitalia and gametes will become incompatible and so on, but these changes will most likely not occur at the same time, and not necessarily in the same order in every case (de Queiroz 2007). In recent studies on species delimitation, this view of species has been adopted widely, but not universally (Carstens *et al.* 2013).

Discovery and classification of biodiversity is an important task in and of itself, but a reliable taxonomic frame of reference is also essential for many other branches of biology such as behavioral ecology and community ecology, as well as conservation biology (Gotelli 2004, Mace 2004, Wilson 2004). If reliable taxonomic data are not available, surrogates such as morphospecies are often used for estimating species-level patterns. However, these are prone to errors, potentially compromising the results of the research, and are often not replicable across multiple studies (Krell 2004). Lack of taxonomic information is a very real problem in many study systems as only a minority of the world's species has been described. Estimates of undescribed diversity vary widely depending on the

method applied and assumptions made. However, millions of undescribed species are always inferred, and most of them are likely insects and other arthropods inhabiting the tropical regions (Ødegaard 2000, Chapman 2009, Hamilton *et al.* 2010, Mora *et al.* 2011). Carbayo & Marques (2011) estimate that the description of the remaining uncatalogued animal species with current rates and resources would cost ca. 263 billion USD and take ca. 360 years, assuming a total number of 5.5 million undescribed animal species based on Chapman (2009).

1.2 DNA taxonomy

Traditionally, regardless of the species concept applied (which is usually not explicitly stated in the publications), species delimitation and taxonomic revisions have been based mainly on comparative morphology. This is still the case to a large extent (*e.g.* Assing 2014, Shi & Liang 2015), even though it is already possible to extract and sequence DNA from museum specimens, including old type material, without damaging morphological characters (Gilbert *et al.* 2007, Hernández-Triana *et al.* 2014, Price *et al.* 2015). Due to the relatively high cost and need for a clean lab in sampling DNA sequences from old museum material, genetic analysis of old type specimens is not yet feasible as a standard approach in taxonomic studies. In addition, if only very short sequence fragments are obtained, interpretation of results will likely be difficult especially in closely related species complexes due to limited information content in the short sequences. However, with further development in laboratory techniques, obtaining DNA barcodes and other genetic information from older material may become easier and cheaper in the future. A recent study by Prosser *et al.* (2016) shows great promise for such development.

Using genetic data in taxonomy is becoming increasingly common due to advances in sequencing technology and the resulting rapid decrease in costs of sequencing per base pair (see http://www.genome.gov/sequencingcosts/). For unculturable microbes, DNA sequences have for a long time already been the only widely available source of taxonomic information (Moon-van der Staay *et al.* 2001, Sogin *et al.* 2006). A decade ago, it was suggested that all taxonomy should be primarily or entirely based on sequence data (Tautz *et al.* 2002, 2003, Blaxter 2004). This sparked a heated debate on how taxonomy should be practiced and the role of DNA sequences in it (*e.g.* Seberg *et al.* 2003, Lipscomb *et al.* 2003, Will & Rubinoff 2004, Hebert & Gregory 2005, Will *et al.* 2005; reviewed by Teletchea 2010).

Several elaborate methods for DNA-based species delimitation have been developed recently (*e.g.* Yang & Rannala 2010, Ence & Carstens 2011, Leaché *et al.* 2014, Grummer *et al.* 2014). They generally require multi-locus data, and the approach by Leaché *et al.* is specifically designed for genome-wide single-nucleotide polymorphism data. SpedeSTEM by Ence & Carstens (2011) is designed for validating user-specified division of specimens into pre-defined groups and is thus inapplicable for purely exploratory analysis of species boundaries. Grummer *et al.* (2014) use a Bayes factor approach for quantitative comparison of a limited number of alternative delimitation schemes. Although these methods are very powerful in delimiting species in focused and well-sampled cases, their requirements on the extent of sampling of loci and individuals make them impractical for large-scale surveys of largely unknown taxa (*e.g.* Tänzler *et al.* 2012), or delimitation of very rare species (Lim *et al.* 2012).

For large-scale datasets without prior information on grouping of individuals, the selection of feasible delimitation methods is still very limited. Methods which are able to utilize single-locus data such as DNA barcodes have been developed, but some of these require the computer-intensive work phase of tree construction prior to the actual delimitation analysis (*e.g.* Pons *et al.* 2006, Zhang *et al.* 2013). Purely distance-based approaches such as ABGD (Puillandre *et al.* 2012) are computationally less demanding and thus easier to apply to large datasets. However, genetic divergence within and between species is likely to vary between loci and reflects the age of the species and the history of the loci studied, and no justifiable distance thresholds for species status can be defined (Ferguson 2002, Meier 2008). Accordingly, many recently developed distance-based delimitation methods use more elaborate clustering algorithms instead of simple fixed thresholds (Puillandre *et al.* 2012, Ratnasingham & Hebert 2013). Single-locus data alone is not sufficient for taxonomic study, but when combined with morphology or other independent data sources, it can considerably speed up the process and reveal species that would otherwise go unnoticed (Riedel *et al.* 2013, Mutanen *et al.* 2013). Entities delimited using single-locus data only are better referred to as operational taxonomic units (OTU) than species (Blaxter *et al.* 2005), and this term is used for such entities in this thesis as well.

1.3 DNA barcoding

Like species delimitation, the identification of known species is often complicated. Morphology typically changes during ontogeny (sometimes drastically, as in holometabolic insects and many marine invertebrates) and often varies extensively within species, for example between sexes in animals or according to growing spot conditions in plants. For these reasons, identification keys usually only cover a certain life stage, and sometimes diagnostic characters are only presented for one sex, commonly males in insects.

A major advantage of using genetic data for identification is that DNA sequences do not change during ontogeny, and with a few exceptions, all living cells in an organism contain the same genetic information. DNA-based identification is thus applicable to all life stages without the need to search for different diagnostic characters for adults and immatures. Association of problematic and largely undescribed immature stages to adults becomes possible without having to rear the immatures into adults (Miller *et al.* 2005). DNA can also be used for identifying samples which would otherwise be impossible or at least extremely difficult to assign to species, such as pieces of tissue from the digestive tracts or excrements of predators and parasites (Vesterinen *et al.* 2013, Wirta *et al.* 2014). This has the potential to revolutionize the study of food webs.

Simultaneously with the propositions for entirely DNA-based taxonomy, a DNA-based system for species identification was suggested by Hebert *et al.* (2003). DNA-based identification as such was not a new idea (see *e.g.* Sperling *et al.* 1995, Wells & Sperling 2001). The key point in Hebert *et al.*'s proposal was wide-scale standardization: using the same marker for species identification across the animal kingdom in the same way as product barcodes are used for identifying items at a supermarket checkout counter. Accordingly, Hebert *et al.* (2003) named their proposed identification system 'DNA barcoding'. Even the word "barcode" had been used previously in this context, although for more limited study systems (*Plasmodium* strains: Arnot *et al.* 1993, soil nematodes: Floyd *et al.* 2002). The marker of choice for Hebert *et al.* was a ~650 bp segment of the 5' end of the mitochondrial cytochrome oxidase subunit I gene (COI). It has since been established as the standard DNA barcode region for animals. Another novelty besides standardization is the scale of DNA barcoding: the Barcode of Life Data Systems database (BOLD, http://www.boldsystems.org; Ratnasingham & Hebert 2007) currently includes ca. 4.4 million animal COI barcodes (situation as of April 1, 2016).

A distinct ‘barcode gap’ (Meyer & Paulay 2005) can often be seen between the pairwise genetic distances measured within and among species, with generally very little overlap (*e.g.* Hebert *et al.* 2004b, Hajibabaei *et al.* 2006a). To some extent, this may be a sampling artefact, as increase in geographic scale or denser sampling within clade tends to narrow down the gap and increase overlap (Bergsten *et al.* 2012, Hausmann *et al.* 2013). However, some studies have also reported no significant geographic scale effect on distance distribution or identification success (Lukhtanov *et al.* 2009, Huemer *et al.* 2014b). A distinct barcode gap is not necessary if more elaborate approaches are used instead of simple distance measures. For example, Lou & Golding (2010) developed a Bayesian approach to assigning sequences into species. Their method performed remarkably well in identifying *Drosophila* species despite the existence of numerous problematic sibling species complexes and high frequency of incomplete lineage sorting.

One of the main targets for criticism in DNA barcode studies is the use of distance-based analytical methods, particularly Neighbor-Joining (NJ) trees, which has been seen by some as a resurrection of the phenetic school of systematics (Will & Rubinoff 2004, Will *et al.* 2005, Wheeler 2008). The problems of NJ include sensitivity to rate variation between lineages and to the input order of taxa in the data matrix (Farris *et al.* 1996, Felsenstein 2004). The main reason for its continued use in barcoding contexts is probably the speed of the algorithm compared to *e.g.* maximum likelihood inference, especially on large datasets of several thousand sequences. The goal in barcoding studies is not to construct phylogenies as the short and generally rapidly evolving barcode loci contain limited phylogenetic signal, especially at deeper levels (Hajibabaei *et al.* 2006b). NJ trees are mainly used for fast and easy visualization of sequence clustering. The widespread use of the Kimura two-parameter model of nucleotide substitution, generally without model testing, has also been criticized (Srivathsan & Meier 2012), although the effect of model selection on identification success and the width of the ‘barcode gap’ does not seem to be very dramatic (Srivathsan & Meier 2012).

DNA barcoding can also be based on characters instead of distances. One of the drawbacks of distance-based identification is that a “best match” for any query sequence will always be found in the reference barcode database, and interpreting identification success is far from simple. On the other hand, if identification is based on species-specific combinations of diagnostic characters in the barcode sequence, the presence or absence of those characters makes

interpreting the identification results easier (DeSalle *et al.* 2005, Rach *et al.* 2008).

DNA barcoding is meant to be used primarily as a specimen identification tool and should be distinguished from DNA taxonomy. The confusion between the two initiatives and their goals by both proponents and opponents was one of the reasons for the heated and polarized debate over barcoding (Goldstein & DeSalle 2011). However, barcode sequences can also be used as a line of evidence in species delimitation or as an initial screening tool when searching for species boundaries (Mutanen *et al.* 2013, Huemer *et al.* 2014a, Kekkonen & Hebert 2014).

1.4 The focal taxon: North European beetles

Beetles (Coleoptera) are one of the most species-rich taxa of all animals: approximately 380 000 to 400 000 beetle species have been described to date (Chapman 2009, Slipinski *et al.* 2011). Estimating undescribed diversity requires making a lot of assumptions, and the resulting estimates therefore vary widely (Nielsen & Mound 2000, Ødegaard 2000, Oberprieler *et al.* 2007, Chapman 2009), but it is clear that a vast number of species remain undescribed, especially in the tropics. For example, approximately 62 000 species of weevils (Curculionoidea) had been described by 2007, and the total number of weevil species is estimated to be about 220 000 (Oberprieler *et al.* 2007). Beetles are present in all biogeographical areas except for mainland Antarctica and found in most terrestrial and freshwater habitats, and even in brackish water and the intertidal zone. The diversity of niches and diet reflects the vast number of species: practically every kind of organic resource is consumed by at least some beetle species (Crowson 1981).

The beetle fauna in North Europe has been studied intensively since Linnaeus's days, and the region is among the most thoroughly studied in the world in terms of beetle taxonomy. Approximately 5 400 species of beetles are known from the Nordic and Baltic countries (Silfverberg 2010), and ca. 3 750 from Finland (Rassi *et al.* 2015). New, previously undescribed species are only rarely discovered from this region nowadays (but see Brüstle & Muona 2009 for a recent example). The solid taxonomic framework, established by more than 250 years of study on morphology and ecology, provides excellent conditions for studying the use of DNA barcodes in species identification and as a taxonomist's tool.

1.5 Aims of the study

The first goal of these studies was to build the foundations of a comprehensive North European DNA barcode library for Coleoptera and use it to examine the possibilities of species identification based on DNA barcodes (study I). The vast majority of adult beetles found in the area can be identified based on morphological characters presented in published literature, but larval identification can be extremely difficult even for experts, and the larvae of many species are still undescribed. Barcode-based identification, if feasible, would considerably ease ecological studies where larvae or fragmented remains of adults need to be identified.

The existing OTU-generating methods applicable for DNA barcodes are based on widely different delimitation principles. They also make different assumptions about the raw data (sample size per species etc.) that may not be fulfilled by empirical datasets. Applying multiple approaches with different strengths and weaknesses on the same dataset should increase the reliability of the OTU delimitation, especially if a robust way of dealing with discordance between methods can be found. In study II, I used the North European beetle barcode dataset from study I to test method performance as well as two simple approaches for achieving consensus between delimitations. I also explored the sensitivity of the utilized OTU-generating methods to variation in some features of empirical data: sampling effort, divergence between species, variation within species, and non-monophyly in the gene tree reconstructed from the barcode sequences.

The massive global COI barcode repository in BOLD, with representation from nearly all animal phyla, and most major lineages within each phylum, provides an excellent opportunity to study amino acid and protein-level variation in the COI barcode region at various scales. The goal in study III was to approach DNA barcode sequences from a new angle: to study changes in the amino acid sequence, and the effect of these changes on protein structure. Deletions and amino acid substitutions with potential effects on protein function were identified by 3D modelling of the protein structure, and the appearance of these changes in the evolutionary history and their connections to shifts in ecology were explored.

2 Material and methods

2.1 Studied material

The majority of the beetle material used in studies I and II was collected specifically for DNA barcoding from the Nordic and Baltic countries, mainly Finland, during 2011–2012. As a part of the Finnish Barcode of Life (FinBOL) project, my aim was (and is) to compile a comprehensive DNA barcode library for the Finnish beetle fauna. The specimens were preserved in 70% ethanol as soon as possible after collecting and stored at –20°C until tissue sampling. In addition to this fresh material, I sampled pinned specimens from private collections and from the collection of the Zoological Museum at the University of Oulu. Detailed collecting data on all the beetle material analyzed in I–III are publicly available in BOLD together with photographs of the specimens, barcode sequences and the original sequencing trace files (dataset doi: 10.5883/DS-FBCOL).

In study III, I utilized animal DNA barcode data mined from GenBank or otherwise publicly available in the BOLD database, as well as the FinBOL Lepidoptera library, in addition to the beetle material described above. To get as wide coverage as possible of all major metazoan lineages, at least one full-length high-quality barcode sequence was selected from each class within each phylum in the BOLD hierarchy, provided that such data were publicly available. The Arthropoda, and especially the insects, were sampled more densely than the other phyla due to the great number of species and diversity of life histories.

2.2 Laboratory procedures

Tissue samples of the beetle specimens selected for DNA barcoding were placed in 96-well microplates and sent to the Canadian Centre for DNA Barcoding (CCDB, Guelph, Ontario) for DNA extraction, PCR and sequencing of the COI barcode region. Depending on the size and state (fresh/dry) of the sampled individual, one to three whole leg(s), part of a leg, a piece of the thoracic flight muscles or the whole beetle was used for extraction.

Standard CCDB protocols optimized for largely automated high-throughput barcode sequencing were used in DNA extraction, PCR amplification and sequencing. The extraction protocol was published by Ivanova *et al.* (2006), and

full documentation for the PCR and sequencing protocols is available online at http://ccdb.ca/resources.php. In short, the extraction protocol involves binding the extracted DNA onto glass fiber plates after the usual tissue lysis and washing steps. The method results in high quality extracts even from small tissue samples. Platinum® Taq DNA Polymerase from Invitrogen™ is used in the PCR amplification. A cocktail of the Folmer primer pair LCO1490 / HCO2198 (Folmer *et al.* 1994) and the Lepidoptera primer pair LepF1 / LepR1 (Hebert *et al.* 2004a) was used in the first amplification attempt for most beetle specimens. If resources allowed, amplification of shorter 307 bp and 407 bp sequences was attempted for specimens that failed to produce full-length barcode sequences. The PCR products were sequenced in both directions by BigDye™ cycle sequencing. Details on PCR and sequencing primers for all analyzed specimens are available in BOLD.

I identified all specimens selected for tissue sampling to species (with some few exceptions such as *Mordellistena* spp.) based on morphology. Before more thorough analysis of the data, I constructed a Neighbor-Joining tree of the retrieved barcode sequences and ran basic BOLD analyses (barcode gap analysis and Barcode Index Number discordance report) in order to detect possible cases of misidentification and contamination. Whenever a case of barcode sharing between species, non-monophyly in the COI gene tree or a deep split within species was detected, I re-examined all the specimens involved and corrected any misidentifications noticed.

2.3 Data analyses

In studies I and II, all sequences shorter than 500 bp were excluded from analyses. This was mainly due to the quality requirements of the Barcode Index Number (BIN) system (Ratnasingham & Hebert 2013): shorter sequences are not accepted as founding members of BIN clusters in BOLD. In study III, only full-length, high-quality barcode sequences (658 bp) were included in the analysis as missing data were found to hamper the measurement of variation per amino acid site (see 2.3.5).

2.3.1 Species identification (I)

The Barcode Gap Analysis feature in BOLD was used to assess species identification success. The analysis summarizes genetic distances observed within

and between species. All species found to share haplotypes with another species were interpreted as unidentifiable by DNA barcoding. No additional criteria such as monophyly or distance thresholds were applied in study I (but see section 3.1).

2.3.2 OTU delimitation (II)

Within the last ten years, several methods for delimiting species based on genetic data have been developed. In study II, I used four methods applicable for single-locus data to delimit the North European beetle barcode data into Operational Taxonomic Units (OTU) and tested how the OTU boundaries correspond to known beetle species. Two of the methods (ABGD, BIN) are based on genetic distance measurements and specifically designed for barcode data. The other two (GMYC, PTP) are based on the phylogenetic species concept, and fit models of sequence evolution within and between species onto phylogenetic trees inferred from sequence data. I also tested two simple approaches for forming consensus OTUs from the discordant OTU delimitations resulting from different methods.

Automatic Barcode Gap Discovery

A distinct 'barcode gap' with little or no overlap between the distributions of intra- and interspecific genetic distances can often be seen in DNA barcode datasets (*e.g.* I: Fig. 1), although no universal threshold value for this gap exists. ABGD, or the Automatic Barcode Gap Discovery method, attempts to search for this gap algorithmically based on the distribution of pairwise genetic distances in a dataset with an unknown number of species (Puillandre *et al.* 2012). The method requires two parameters to be set by the user: a prior upper limit for intraspecific variation (P) and a gap width parameter (X). The pairwise sequence divergences in the dataset are ranked from smallest to largest, and the barcode gap is identified as the first 'leap' in the ranked divergences after the limit for intraspecific divergence that is X times larger than any such leaps in the intraspecific distances. After an initial delimitation, the gap search is conducted recursively within each of the initial partitions until no further splitting occurs.

I used the ABGD web service (http://wwwabi.snv.jussieu.fr/public/abgd/) in my analyses. The default setting of 1.5 was used for the barcode gap width parameter. Puillandre *et al.* (Puillandre *et al.* 2012) noted that in four datasets of different animal taxa, the intraspecific divergence parameter P value of 0.01 produced OTU counts and boundaries that were very close to those reported in

the original studies using different methods. For simplicity, I did not extensively explore the effect of the parameter values on the resulting delimitation, and only a small-scale test on the effect of P parameter was performed on the Curculionoidea + Chrysomeloidea subset of the data. The P parameter was set to 0.01 in all other analyses.

Barcode Index Numbers

The Barcode Index Numbers (BIN) serve as an interim taxonomic reference system in the BOLD database, especially for managing the records that lack species-level identification. The Refined Single Linkage (RESL) algorithm used in computing BINs was introduced by Ratnasingham & Hebert (Ratnasingham & Hebert 2013). Initial clusters are formed by employing a fixed sequence divergence threshold of 2.2% uncorrected p-distance. These clusters are then verified and refined into the final BINs by Markov clustering. The BIN assignments of all records in BOLD are updated regularly, and BIN clusters can be split further or merged together as new sequences are added in the BOLD database. In studies I and II, I used BIN assignments downloaded from BOLD on January 24, 2014.

General Mixed Yule Coalescent model

The General Mixed Yule Coalescent model (GMYC) was first introduced by Pons *et al.* (2006) and subsequently revised by Monaghan *et al.* (2009) and Fujisawa & Barraclough (2013). It is one of the most frequently used methods of putative species delimitation based on DNA sequence data (*e.g.* Monaghan *et al.* 2009, Ceccarelli *et al.* 2012, Esselstyn *et al.* 2012, Hjalmarsson *et al.* 2013). GMYC combines a Yule model of species birth with a coalescent model of within-species diversification, and attempts to find the point of transition between these two processes in an ultrametric phylogenetic tree where branch lengths represent time. The method allows fitting of both single and multiple transition points to the data. However, the single-threshold version is computationally much less demanding, and it has been found to outperform the multiple-threshold fitting in simulations (Fujisawa & Barraclough 2013).

Test runs of the multiple-threshold version on North European beetles indicated notable oversplitting of species compared to the single-threshold version. Therefore, I only used the single-threshold version in my final analyses.

To obtain the ultrametric trees used as input for GMYC, I used BEAST v. 1.7.5 (Drummond *et al.* 2012) via the CIPRES portal of computing services for phylogenetics (Miller *et al.* 2010). The relaxed lognormal clock model and a coalescent tree prior were used in the BEAST analyses as coalescence is the null model in GMYC, and therefore a conservative choice (Pons *et al.* 2006, Monaghan *et al.* 2009). The maximum likelihood trees generated for PTP analyses (see below) were used as starting trees in inferring the ultrametric trees.

Poisson Tree Processes model

The Poisson Tree Processes model (PTP) introduced by Zhang *et al.* (2013) is similar in principle to GMYC. A model consisting of two Poisson processes, one describing speciation, the other within-species branching, is fitted onto a phylogenetic tree. However, instead of waiting times between branching events, PTP utilizes the number of substitutions directly as raw data, bypassing the potentially error-prone process of dating the phylogeny. The number of substitutions in branches leading to species is expected to be significantly higher than the number of substitutions within species. I used RAxML v. 7.2.8 (Stamatakis 2006) with the Rapid Bootstrap feature (Stamatakis *et al.* 2008) to obtain the input trees for the PTP delimitation analyses. To reduce the computation time required, I divided the data into smaller taxonomically delimited subsets. The best-scoring tree from each RAxML run was used as input for PTP. The PTP web server (http://species.h its.org/ptp/) was used to run the analyses.

Consensus OTUs and delimitation success

Observed incongruence between OTU delimitations by the four methods described above motivated me to search for possible ways to achieve consensus between discordant delimitations. I tried out two simple approaches to deal with discordant OTUs and checked if these approaches improved the compatibility between OTUs and species. Conservative consensus OTUs were formed by simply lumping any discordant cases into one OTU (Fig. 1). I also formed majority consensus OTUs by accepting OTUs delimited identically by three out of four methods, and lumping all other discordant cases (Fig. 1).

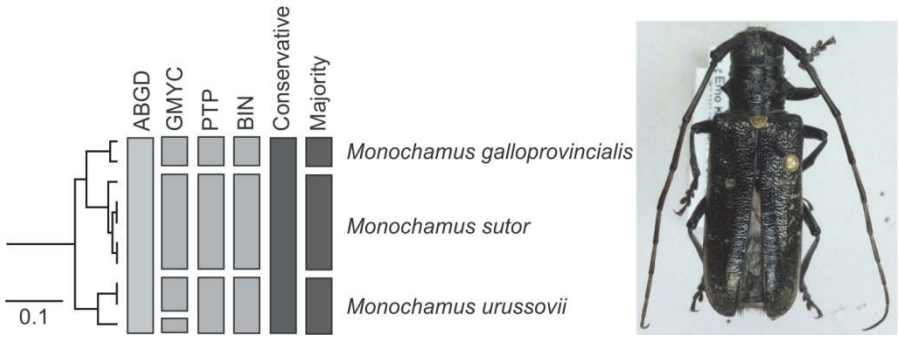

Fig. 1. An example of the consensus approaches used in study II. ABGD lumps all three sampled species of *Monochamus* (Cerambycidae) into one OTU, GMYC splits *M. urussovii* (pictured) in two OTUs, and PTP and BINs both retrieve the same OTU delimitation. The conservative approach lumps all discordant OTUs, in this case combining all three species into one OTU. The majority approach accepts the OTUs including *M. galloprovincialis* and *M. sutor*, delimited identically by three out of four methods, and lumps the remaining discordant specimens into one OTU. PTP, BINs and the majority consensus approach perfectly recover the three traditionally accepted species. The tree shown here is a subtree of the ultrametric Cerambycidae tree used as input for the GMYC analysis (generated with BEAST).

The delimitation outcomes of all four methods, as well as the outcomes of the consensus approaches, were compared to the current knowledge on species boundaries in two ways. For each OTU delimitation, all species were assigned into one of four categories (Match, Split, Merge or Mixture) as described by Ratnasingham & Hebert (2013). A Match is achieved if all specimens of one species (and no representatives of other species) are included in a single OTU. A Split occurs when specimens of one species are divided into two or more OTUs. In a Merge, all specimens of two or more species are included in a single OTU. Mixture refers to a complex case with both a split and a merge involving two or more species.

In addition to this direct comparison, I evaluated the congruence between species and OTU, as well as the similarity of different OTU delimitation outcomes, using the F-measure (Larsen & Aone 1999). The F-measure is a numerical estimate of clustering compatibility and varies on a scale of 0 to 1, where 1 means perfect congruence between cluster boundaries.

2.3.3 Sampling effects (I, II)

Sampling unavoidably affects observations on genetic variation and species identification, as well as the delimitation outcomes of all OTU-generating methods. To estimate the effect of sampling effort on the genetic divergence observed within and between species, I performed resampling and regression analyses (I). A locally weighted polynomial regression (LOESS) curve with 95% confidence intervals was fitted on a scatterplot of the observed maximum genetic divergence within species vs. the number of specimens sampled per species. Spearman's rank order correlation coefficient used to evaluate the association and its significance.

To demonstrate the effect of species-level sampling on observations of divergence between species, I resampled the ground beetles (Carabidae) which are among the best represented families in the FinBOL beetle data (199 species included in the dataset studied in I and II). Randomized sets of 20 to 180 species, with increments of 20 species, were subsampled from the Carabidae and a barcode gap analysis run on the subsample in BOLD. The minimum and mean values of divergence between species in each analysis were recorded. The sampling and analysis was repeated 10 times for each species count, so that altogether 90 barcode gap analyses were conducted. LOESS curves with 95% confidence intervals were fitted onto the data, and Spearman's correlation coefficient used to test the association between species count and divergence.

Most OTU-generating methods require several specimens per species to be sampled in order to produce reliable results (Puillandre *et al.* 2012, Fujisawa & Barraclough 2013). However, the proportion of singletons and doubletons is typically high in empirical datasets (Scharff *et al.* 2003, Lim *et al.* 2012). I plotted the observed species-specific outcomes for each method (match, split, merge, see section 2.3.2) onto the sample size within species to compare the sensitivity of the methods to sampling effort (II).

2.3.4 Effects of non-monophyly and genetic variation on OTU delimitation (II)

Incomplete lineage sorting and/or introgression can cause species to appear non-monophyletic in COI barcode trees. Especially for tree-based methods like GMYC and PTP, non-monophyly will severely hamper OTU delimitation based on barcode data. To assess monophyly of the studied beetle species, I used the

Monophylizer web service (http://monophylizer.naturalis.nl) developed by Rutger Vos (Vos 2015). Monophylizer reads taxon names from a Newick tree, analyzes the tree topology, and classifies each taxon as monophyletic, paraphyletic or polyphyletic.

For all species with observed maximum intraspecific divergence greater than 0.02 (K2P), a haplotype network was constructed with TCS v. 1.21 (Clement *et al.* 2000) in order to check if the high variability was due to a deep split within species or more or less continuous variation with intermediate haplotypes between the extremes. The deepest split observed between haplotypes within species was recorded, and OTU delimitation outcomes plotted on the deepest splits in order to assess the splitting sensitivity of the OTU delimitation methods used.

The beetle data were screened for the highest value of divergence between nearest neighbor species at which merging of species into one OTU was observed. The observed matches and merges were plotted on NN divergence below this limit.

2.3.5 Protein structure modelling (III)

Ambiguous amino acids were found to inflate the entropy measures for each amino acid site in initial analyses. Therefore, only full-length, high-quality DNA barcode sequences (<1% ambiguous bases) were used in the final analyses of amino acid variation and protein structure. Three separate datasets were formed: Metazoa (292 sequences and taxa), Coleoptera (3 208 sequences / 1 764 species) and Lepidoptera (4 628 sequences / 2 547 species). The sequences were collapsed into haplotypes using ALTER (Glez-Peña *et al.* 2010), and aligned and translated in MEGA v. 6.06 (Tamura *et al.* 2013). Due to presence of deletions in many of the Metazoan sequences, the sequences were first aligned algorithmically with ClustalW (Thompson *et al.* 1994) using the default options, and the resulting alignments were manually refined before translation and analysis.

The amino acid variation in each dataset was assessed by calculating entropy (uncertainty, H(x)) values for each amino acid site in BioEdit (Hall 1999). A completely conserved amino acid site has an entropy value of 0, and the value increases with increasing variation in amino acid content. The amino acid sites were divided into (arbitrary) classes based on the entropy value: 0.5–0.7, 0.71–0.9, 0.91–1.1, and >1.1. Amino acid positions with entropy below 0.5 were considered non-variable, and residues that showed no variation at all were defined

as conserved. The amino acids were divided into standard groups based on their biochemical properties (nonpolar aliphatic, polar uncharged, aromatic, positively charged, and negatively charged), and amino acid sites that showed variation within group only were considered non-variable regardless of the entropy value.

After observing considerable difference in amino acid variation between Coleoptera and Lepidoptera, I estimated the pattern of nucleotide substitution in these two datasets using the MCL Substitution Matrix feature in MEGA v. 6.06. Directional biases in mutations can provide information on the main cause of mutations (mispairing of bases vs damage to DNA; Martin 1995).

The position of the variable amino acid sites in the cytochrome oxidase protein was explored by building three-dimensional models of the protein structure. The cattle (*Bos taurus*) COX structure was used as a reference for all structural models as it has been thoroughly studied at a fine resolution (Tsukihara *et al.* 1995, 1996). PyMOL Molecular Graphics System 1.7 (Schrödinger, LLC) was used for visualization of the folded protein and distance measurements between amino acids and enzyme ligands.

The observed amino acid changes with potential to affect enzyme function (amino acid substitutions between biochemical groups, or deletions of multiple amino acids, close to the enzyme ligands) were mapped on phylogenies in order to search for associations between shifts in ecology and the structural changes observed in COI. The consensus tree presented by Dunn *et al.* (2014) was used as a reference for phylum-level relationships. The relationships of flatworms (Platyhelminthes) were based on the work of Park *et al.* (2007). For Coleoptera and Lepidoptera, recent comprehensive molecular phylogenies were used as references (Hunt *et al.* 2007, Mutanen *et al.* 2010, Wahlberg *et al.* 2013, McKenna *et al.* 2015).

3 Results and discussion

3.1 Barcode-based identification of North European beetles (I)

Out of the 1 872 species analyzed in study I, 1 842 (*i.e.* 98.3%) shared no haplotypes with other species, and can be liberally interpreted as successfully identifiable. 61 species showed relatively low genetic divergence (<0.02 K2P) from their nearest neighbor species, and 40 of these were even found to be non-monophyletic (II). Even when these potentially problematic species are taken into account, 1 780 / 1 872 studied species (95.1%) can be reliably identified by their COI barcodes by a simple nearest-neighbor blast.

Near-identical haplotypes or barcode sharing between close relatives is not surprising as the COI barcode region is not involved in speciation, and differences accumulate gradually after the lineages diverge (Kwong *et al.* 2012). Even in these ambiguous cases, barcode-based identification is generally accurate to within a species pair or a small group of close relatives — a significant improvement over situations where *e.g.* undescribed larvae or morphologically indistinguishable early larval stages hamper species identification. For example, larvae of only 15 out of the 37 Central European species of Epuraea (Nitidulidae) have been described (Klausnitzer 2001), but 29 of the 31 north and Central European species from this genus sampled for DNA barcoding are unambiguously identifiable by their barcode sequences (combined data from study I and Hendrich *et al.* (2014) reanalyzed here).

The divergence between beetle species was generally very high compared to other large-scale barcode studies, with an average 11.99% K2P distance observed between nearest neighbor (NN) species in study I. The FinBOL Lepidoptera dataset originates from within the same geographic region as the beetle material in study I, and the average NN divergence was 5.73% among 2577 species (unpublished data). In a survey of North American noctuoid moths, the average NN divergence was 3.08% (data from Zahiri *et al.* 2014 reanalyzed here). A comprehensive library of 642 North American bird species had an average NN divergence of 5.9% (Kerr *et al.* 2007). After study I was published, sequence data have been retrieved from an additional 363 North European beetle species, and the average NN divergence has dropped slightly to 10.63% (unpublished data). Even higher divergences have been reported in some studies, but these have generally been based on much less extensive sampling of species (datasets of 20–

150 species studied by Hogg & Hebert 2004, Ball *et al.* 2005, Shaffield *et al.* 2009, Zhou *et al.* 2009, Chang *et al.* 2009). This probably inflates the divergence estimates (see section 3.3).

The observed drastic difference in NN divergences between Coleoptera and Lepidoptera is also reflected in the amino acid variation and may be explained by differences in oxygen metabolism (III). The notably higher variation in amino acid sequences in Coleoptera hints that the barcode sequence may vary more freely in Coleoptera than in Lepidoptera (III, see section 3.4 for details).

3.2 Accuracy and sensitivity of OTU delimitation methods (II)

Approximately 90% of the analyzed species were perfectly recovered regardless of the delimitation method used. The four OTU-generating methods also produced largely congruent results (II: Table 3), with 1752 OTUs including ca. 90% of all studied specimens clustering identically in all four delimitation analyses. However, a significant proportion of OTUs was still discordant between methods and/or with the current taxonomy. Each method makes a different set of simplifying assumptions on the data, and has different requirements on *e.g.* the extent of sampling within species, and may thus fail under some circumstances using empirical data (Carstens *et al.* 2013). Therefore, relying on any single delimitation scheme is unwise, and results from multiple approaches should always be compared and special attention given to cases of discordance between methods (Miralles & Vences 2013, Carstens *et al.* 2013). The simple consensus delimitations devised in II, especially the conservative consensus approach, improved the compatibility between the OTU delimitation and current taxonomy. The main differences between the two consensus delimitations were the notably lower split count and higher merge count in the conservative approach. The consensus approaches are useful in a wide-scale shotgun approach to estimating biodiversity. In more focused studies aiming for resolving taxonomy, a more fruitful approach would be to use the OTU-generating methods to sort out the clear, fully congruent cases, and focus on gathering additional data from the incongruent cases.

All methods used in II occasionally split single divergent specimens within species into separate OTUs. Most of these were captured and corrected by the consensus approaches, which enable the different delimitation approaches to compensate for each other's weaknesses. Based on study II, as well as previous research (*e.g.* Hendrich *et al.* 2010, Miralles & Vences 2013, Hamilton *et al.*

2014, Modica *et al.* 2014), GMYC seems to have an especially strong tendency for oversplitting species. Errors in generating the ultrametric input tree are probably a major reason for the observed oversplitting (Reid & Carstens 2012, Zhang *et al.* 2013). A test run of the ultrametric longhorn beetle tree (Cerambycidae) with PTP revealed that dating error is likely behind at least some of the GMYC oversplits in study II as well.

Apart from sampling or method artefacts, and extensive within-species variation, splits may also be caused by true cases of undetected species. When only DNA barcodes are used for the initial delimitation, support for the observed splits should be found from independent data sources such as morphology or multi-locus genomic data before drawing conclusions on species boundaries. If this is done in a well-organized and efficient way, the process of species description can be speeded up considerably (*e.g.* Riedel *et al.* 2013). Unfortunately, very limited barcode material is still available from most of the splits observed in study II, and only cursory morphological surveys have been performed on those splits with more extensive sampling. However, some obvious or at least very likely cases of overlooked species were detected during these studies. *Dictyoptera aurora* (Herbst, 1784) was given as an example in study II: it is split into two distinct barcode clusters (ca. 6% K2P divergence) with an apparent difference in pronotal shape between them. Another similar case is *Kateretes pusillus* (Thunberg, 1794), which shows a deep barcode split of ca. 10% K2P as well as distinctive differences in coloration and male genitalia between the barcode clusters (Fig. 2). More comprehensive sampling of specimens and geographic variation of these cases is still needed for proper re-evaluation of taxonomy.

Merges are probably the least problematic of the possible incorrect outcomes as excessive lumping of species is easier to correct in subsequent taxonomic studies than excessive splitting (Miralles & Vences 2013). Failure of any one particular delimitation approach or character set to confirm the status of two lineages as separate species does not falsify a hypothesis of two species, as the critical evidence may be found in other characters, or method-specific assumptions on data may be violated (Mayden 2002, Miralles & Vences 2013). Detecting a previously unnoticed independent lineage (falsifying a one-species "null-hypothesis") is more straightforward (Miralles & Vences 2013). Even so, an overly conservative delimitation can also be problematic, for example if two or more species are lumped together *e.g.* in community analysis or in a conservation context. When singleton species were involved, ABGD lumped notably many

species that were correctly delimited by all other methods (II: Fig. 3c). As singletons are very common in empirical data (Scharff *et al.* 2003, Lim *et al.* 2012, see also II: Table 1), this is potentially a significant disadvantage for ABGD and special attention should be paid to singletons when using ABGD.

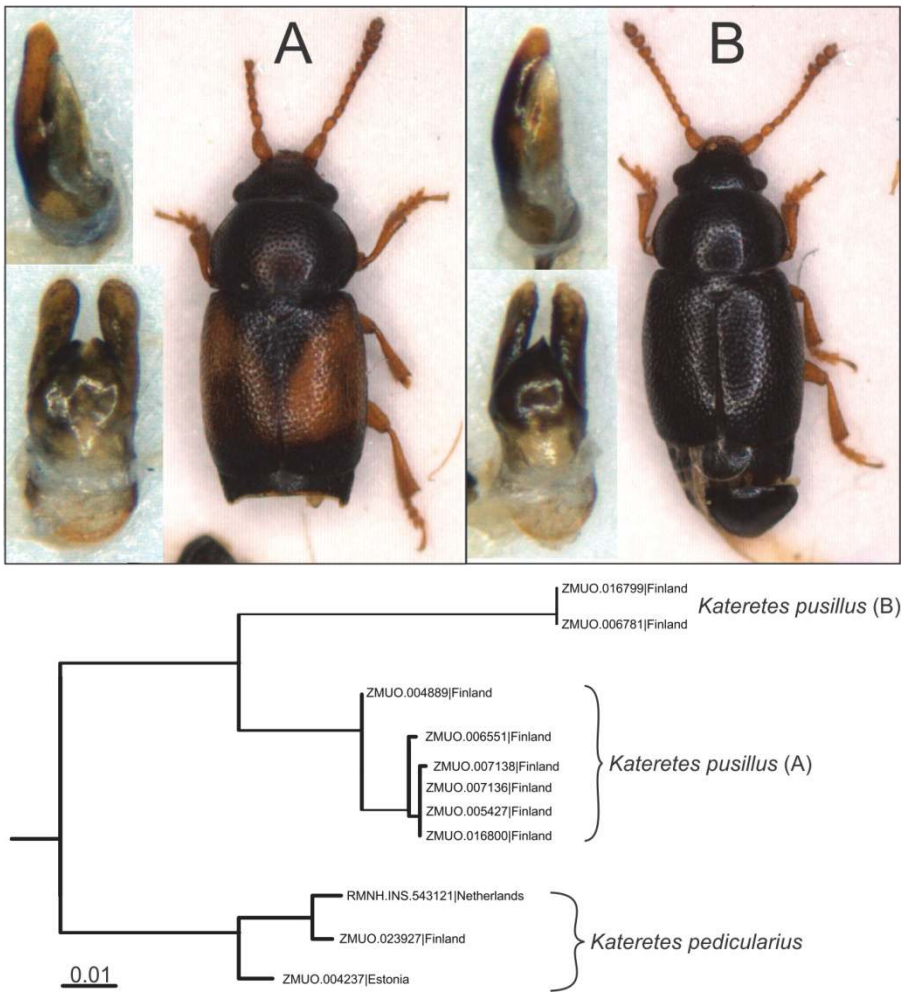

Fig. 2. *Kateretes pusillus* (Thunberg, 1794) is split into two DNA barcode clusters, with distinctive differences in both coloration and male genitalia. The Neigbbor-Joining tree shown here was drawn based on Kimura 2-parameter distances. The closely related *K. pedicularius* (Linnaeus, 1758) is also included in the tree for reference.

Multi-locus data can be used for species delimitation even when gene trees are widely non-monophyletic (Knowles & Carstens 2007), but non-monophyletic species are likely to cause problems for barcode-based delimitation of putative species (Knowles & Carstens 2007, Hendrich *et al.* 2010, Bergsten *et al.* 2012). Not surprisingly, the non-monophyletic species observed in study II were systematically discordant with the delimited OTUs, with a single exception (*Haliplus ruficollis* correctly delimited by ABGD). Most species recovered as non-monophyletic by the Monophylizer tool were tangled with a closely related species with very low interspecific genetic divergences, which is reflected in the high proportion of merge outcomes for these species (II: Table 4).

Non-monophyletic species with deep barcode splits within species are retrieved as Mixtures in the OTU delimitation outcomes. They are the most problematic cases of all the incorrect outcomes, but fortunately seem to be very rare in empirical data (Ratnasingham & Hebert 2013; II: Table 2 & Table 4). Some of these may actually represent overlooked diversity or incorrectly delimited species: deep species-level non-monophyly seems to be mainly caused by either misidentifications in the data or incorrect taxonomy (McKay & Zink 2010, Ross 2014, Mutanen *et al.* unpublished data). The other possible explanations for non-monophyly are incomplete lineage sorting or introgression through hybridization (Funk & Omland 2003, Ross 2014). Distinguishing between these would require more extensive sampling of the genome besides the barcode sequences.

The success rate of any OTU-delineating approach can vary drastically between lineages (Hendrich *et al.* 2010), and it is not easy to predict beforehand which lineages turn out to be problematic. Rapid radiations are likely to result in extensive incongruence between morphological entities and genetic clusters (Monaghan *et al.* 2006). In the Nordic beetle data, the leaf beetle genus Altica provides a good example of a problematic taxon where no barcode haplotypes are shared between species, but all OTU-generating methods failed to delimit many of the species correctly (Fig. 3). Additional independent genetic data would likely help clarify the delimitation in such cases where barcode data is of little use.

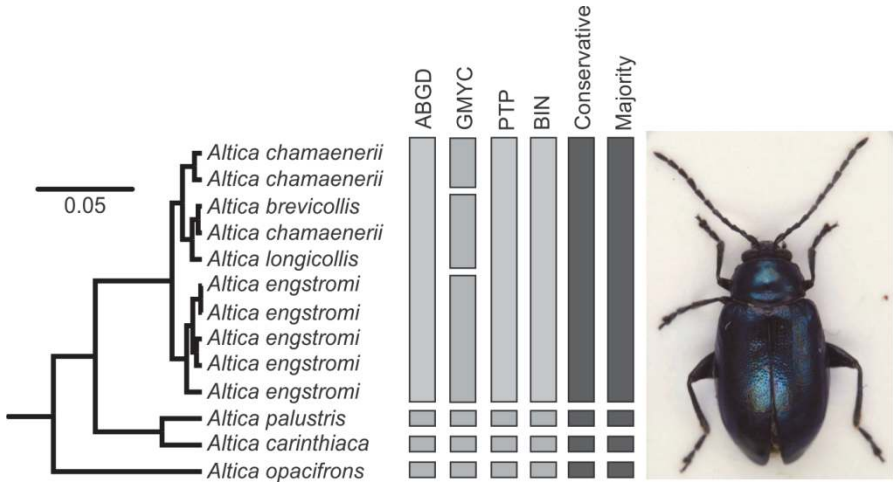

Fig. 3. OTU delimitations in the leaf beetle genus *Altica*. The tree shown here is a subtree of the ultrametric Chrysomelidae tree used as input for the GMYC analysis (generated with BEAST). Note that *A. chamaenerii* was recovered as non-monophyletic in the ultrametric tree.

3.3 Sampling effects (I, II)

The geographic scope of sampling in I and II is rather narrow, with most of the material originating from Finland. This has implications for both species identification (I) and species/OTU delimitation (II). The sample size per species in these studies was generally small (see II: Table 1), and the variation within species was certainly not thoroughly sampled. More species split into multiple OTUs and more non-monophyletic species might be revealed by additional sampling. However, additional within-species sampling in Finland and the rest of the Nordic countries is unlikely to change the results very much: only a small subsample of a species' total genetic variation is likely to occur in a small, restricted part of its range (Bergsten *et al.* 2012), and the majority of the sampled species are widespread in the Palearctic region or at least its Western part. The observed variation within species did increase with sample size in study I, but only slightly (I: Fig. 2).

As the geographic scale of sampling increases, the number of closely related species sampled and the extent of variation within species are both expected to increase (Barraclough & Vogler 2000, Bergsten *et al.* 2012). Bergsten *et al.* (2012) observed exactly this in a study of Agabini diving beetles: The distances

between nearest neighbor species decreased, and the variation within species and proportion of non-monophyletic species increased, as the geographic scale of sampling was extended from local assemblages to a continental scale. A similar pattern was detected by Zahiri *et al.* (2014) in Canadian noctuoids, where the identification success dropped when moving from provincial to Canada-wide scale. However, some studies have reported negligible changes in variation and identification success over thousands of kilometres (Lukhtanov *et al.* 2009, Huemer *et al.* 2014b). A study on geographic variation combining all available European barcode data on beetles is currently in preparation, but unfortunately only scattered data are available from Russia and other parts of Eastern Europe. The increase in within-species variation and numbers of closely related species pairs is likely to affect OTU-generating at least as much as identification success — perhaps more, as closely related species may be reliably identified by distinct barcode haplotypes even though OTU-generating methods might not identify them as separate entities (II).

Completeness of clade-level sampling can affect results even at small geographical scales. In study I, I resampled the family Carabidae to demonstrate the effect of clade-level sampling of the local fauna on the observed divergences between species. The average NN divergence dropped from 10.7% to 7.7% K2P as the number of carabid species included in the analysis increased from 20 to 180 (I: Fig. 3). The remarkably wide "barcoding gaps" and identification successes of 99–100% reported in many studies (*e.g.* Hogg & Hebert 2004, Ball *et al.* 2005, Zhou *et al.* 2009) are likely due to restricted clade-level sampling.

3.4 Amino acid variation and structural changes in the COX protein (III)

The DNA barcode sequence covers a stretch of 219 amino acids right at the center of electron transfer activity in the COX protein. The secondary structure of this stretch consists of six α-helices connected by five loops (III: Fig. 2). Most of the observed variation was concentrated in the loops, which are likely to be functionally redundant. An interesting exception was Loop 3-4 (III: Fig. 2a), which has a stretch of conserved amino acids at the middle. A likely explanation for the conservation can be seen on the 3D model of the protein: this loop is facing the heme ligand (III: Fig. 2b). The other completely conserved amino acids are found in the protein helices, which are relatively rigid structures sitting in a

crowded, lipophilic environment formed by the mitochondrial membrane, and thus likely have limited freedom to vary.

Six of the total 99 variable amino acids in the Metazoa sample showed variation with potential effects on enzymatic reactions, *i.e.* changes from one biochemical group to another within 5 Å of the heme ligands. In notably many cases, unrelated parasitic lineages had experienced convergent amino acid substitutions at these sites, indicating that the transition to parasitism requires changes in cellular respiration. Altogether 11 independent transitions to parasitism were represented in the dataset, and in seven of these cases, transitions between amino acid groups at one or more of these six positions was observed (III: Fig. 5). Many parasites, especially endoparasites, experience hypoxic or anoxic conditions during their life cycles, which may be causing the structural changes observed.

The vast majority of the amino acid deletions observed were found in parasites, and were concentrated in the protein loops. These deletions likely have very little effect on the protein function, with the possible exception of the extensive deletions observed in Dicyemida (small endoparasites of cephalopod molluscs) on both sides of the conserved heme-facing amino acid stretch in Loop 3-4 (five amino acids just before the conserved stretch and four amino acids immediately after it). The length of the mitochondrial genome is known to be associated with the thermal environment inside the host in parasitic nematodes: A shorter genome and thus faster replication rate may be selected for in parasites of endotherms (Lagisz *et al.* 2013). In the Metazoan barcode dataset analysed in III, no association was found between the host type and extent of deletions. However, the DNA barcode region is only a short fragment of the complete genome, and most of the length variation is expected to occur in non-coding regions (Lagisz *et al.* 2013).

There was considerably more amino acid variation in beetle barcodes than in Lepidoptera, even when the potential effect of beetles being an older clade was taken into account. This is in line with the observed difference in between-species divergences between the taxa at the nucleotide level (see 3.1), and may be due to differences in metabolism and evolutionary constraints (see below). As in the wider-scale sample of the animal kingdom, most of the variation was concentrated in the loop structures of the protein. No variation with potential to affect enzyme function was observed in Lepidoptera, which supports the hypothesis of stronger purifying selection in Lepidoptera. In beetles, such variation was found at two amino acid sites (8, 57) and a small handful of

lineages. At both sites, I observed substitutions from smaller amino acids to the bulky phenylalanine, which likely affects the position of the adjacent heme group, potentially changing electron transfer properties of the protein. At position 8, the phenylalanine has appeared at least seven times independently in distantly related taxa, and six of these cases are herbivorous leaf beetles (Chrysomelidae) and weevils (Curculionoidea). No apparent connections were found between the amino acid change and host plant use or other ecological features, however. At position 57, the change to phenylalanine has occurred in two ancestrally fungivorous clades (Phalacridae and Nitidulidae-Kateretidae).

Lepidoptera are generally more eager and active fliers than most beetles, and therefore the COI gene may be under more intensive purifying selection in Lepidoptera. The metabolic rate and the intensity of purifying selection on mitochondrial protein-coding genes show a positive correlation at least in amphibians (Chong & Mueller 2013) and fish (Strohm *et al.* 2015). According to the substitution pattern analysis, both Coleoptera and Lepidoptera have a notable bias towards C to T and G to A substitutions versus T to C and A to G (III: Fig. S3). This is expected if the main cause of mutations is damage to DNA by oxygen radicals generated in cellular respiration (Martin 1995). In actively flying species, the metabolic rate and oxygen consumption (and thereby the rate of oxygen radical generation) are higher than in non-flying species even at rest (Reinhold 1999). The more pronounced mutation bias supports the assumption of higher average rates of metabolism and oxygen consumption in Lepidoptera. However, the variation in metabolic rates of Coleoptera and Lepidoptera and its connections to DNA barcode variation is yet to be properly studied.

4 Conclusions

Extensive tests on the utility of DNA barcodes for species identification have not yet been made on many of the truly diverse animal taxa. Most of the wide-scale studies in insects have focused on Lepidoptera (Hebert *et al.* 2010, Zahiri *et al.* 2014, Huemer *et al.* 2014b). Wide-scale empirical studies are still lacking from two megadiverse insect orders, Diptera and Hymenoptera. Study I, and the even more extensive data release by Hendrich *et al.* (2014) on largely the same fauna, take the first steps in such studies on Coleoptera. The low frequency of ambiguous cases and generally high divergences between species compared to similar-scale studies in Lepidoptera indicate that DNA barcodes could be a feasible tool for species identification in beetles, which form a substantial part of the species-level diversity on Earth. However, the true challenges for beetle barcoding are found in the tropics, where only the surface has been scratched on the topic (*e.g.* Riedel *et al.* 2010).

In a compilation of 79 studies, Krell (2004) found an overall median error in species number estimation of 22% in morphospecies sorting, and the highest recorded error rate was 117% (*i.e.* more than double the number of species found by the morphospecies approach compared to proper taxonomic study). The "morphospecies" sorting error can also be expected to vary between sorters, taxa and samples (Krell 2004). Compared to these figures, the ca. 90% rate of perfect matches between barcode-based OTUs and species achieved in study II gives hope for much more accurate and robust results in beetle community studies on poorly studied taxa if DNA barcoding is adopted. DNA barcodes have the additional advantage over morphospecies that reanalysis of data and utilizing previously generated data is relatively easy. DNA barcodes can already be extracted from bulk samples, such as malaise trap material or kick-net samples of aquatic invertebrates, without any sorting of the material needed before analysis (*e.g.* Hajibabaei *et al.* 2011), and the NGS analysis methods are constantly developing. Despite their obvious utility, barcode-based OTUs still provide a deficient picture of the biological reality, and offer at best a temporary relief for biodiversity studies in the absence of proper taxonomic study. However, they do provide an excellent starting point for more thorough studies on species boundaries.

Even though animal DNA barcodes represent only a small fraction of the mitochondrial genome and not even a complete protein subunit, they offer interesting insights into metabolic protein evolution. The connections between the

observed amino acid substitutions and deletions close to the active site and various life history traits will be an interesting subject for further study – only the surface was scratched in III. Especially the extensive modifications in certain endoparasites seem worth a closer study, including measures of enzymatic activity if possible. The extensive Finnish and Nordic DNA barcode data already available on all major insect orders in the FinBOL project would enable an extension of the Coleoptera-Lepidoptera comparison in III. Paired with a compilation of life history traits and physiological data on the relatively well-known North European insect fauna, this databank might give some insight on why patterns of DNA barcode variation seem to be so different between insect taxa.

References

Arnot DE, Roper C, Bayoumi RAL (1993) Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. Mol Biochem Parasitol 61: 15–24.

Assing V (2014) On the *Bolitochara* species of the West Palaearctic region (Coleoptera: Staphylinidae: Aleocharinae). Stuttgarter Beiträge zur Naturkd A Neue Serie: 33–63.

Ball SL, Hebert PDN, Burian SK, Webb JM (2005) Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. J North Am Benthol Soc 24: 508.

Barraclough TG, Vogler AP (2000) Detecting the Geographical Pattern of Speciation from Species-Level Phylogenies. Am Nat 155: 419–434.

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN, Ribera I, Nilsson AN, Barraclough TG, Vogler AP (2012) The Effect of Geographical Scale of Sampling on DNA Barcoding. Syst Biol 61: 851–869.

Blaxter ML (2004) The promise of a DNA taxonomy. Philos Trans R Soc London B Biol Sci 359: 669–679.

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philos Trans R Soc B Biol Sci 360: 1935–1943.

Brüstle L, Muona J (2009) Life-history studies versus genetic markers - the case of Hylochares cruentatus (Coleoptera, Eucnemidae). J Zool Syst Evol Res 47: 337–343.

Carbayo F, Marques AC (2011) The costs of describing the entire animal kingdom. Trends Ecol Evol 26: 154–155.

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. Mol Ecol 22: 4369–83.

Ceccarelli FS, Sharkey MJ, Zaldívar-Riverón A (2012) Species identification in the taxonomically neglected, highly diverse, neotropical parasitoid wasp genus *Notiospathius* (Braconidae: Doryctinae) based on an integrative molecular and morphological approach. Mol Phylogenet Evol 62: 485–495.

Chang C-H, Rougerie R, Chen J-H (2009) Identifying earthworms through DNA barcodes: Pitfalls and promise. Pedobiologia (Jena) 52: 171–180.

Chapman AD (2009) Numbers of Living Species in Australia and the World. Canberra, Australian Biological Resources Study.

Chong RA, Mueller RL (2013) Low metabolic rates in salamanders are correlated with weak selective constraints on mitochondrial genes. Evolution 67: 894–899.

Claridge MF (2009) Species are real biological entities. In: Ayala FJ, Arp R (eds) Contemporary Debates in Philosophy of Biology. Singapore, Wiley-Blackwell: 91–109.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. Mol Ecol 9: 1657–1659.

Crowson RA (1981) The Biology of the Coleoptera. London, Academic Press.

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.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969–1973.

Dunn CW, Giribet G, Edgecombe GD, Hejnol A (2014) Animal phylogeny and its evolutionary implications. Annu Rev Ecol Evol Syst 45: 371–395.

Ence DD, Carstens BC (2011) SpedeSTEM: A rapid and accurate method for species delimitation. Mol Ecol Resour 11: 473–480.

Esselstyn JA, Evans BJ, Sedlock JL, Anwarali Khan FA, Heaney LR (2012) Single-locus species delimitation: a test of the mixed Yule–coalescent model, with an empirical application to Philippine round-leaf bats. Proc R Soc B Biol Sci 279: 3678–3686.

Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG (1996) Parsimony jackknifing outperforms neighbor-joining. 12: 99–124.

Felsenstein J (2004) Inferring phylogenies. Sunderland, Sinauer Associates, Inc.

Ferguson JWH (2002) On the use of genetic divergence for identifying species. Biol J Linn Soc 75: 509–516.

Floyd R, Abebe E, Papert A, Blaxter M (2002) Molecular barcodes for soil nematode identification. Mol Ecol 11: 839–850.

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.

Fujisawa T, Barraclough TG (2013) Delimiting species using single-locus data and the generalized mixed Yule coalescent approach: a revised method and evaluation on simulated data sets. Syst Biol 62: 707–724.

Funk DJ, Omland KE (2003) Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annu Rev Ecol Evol Syst 34: 397–423.

Gilbert MTP, Moore W, Melchior L, Worobey M (2007) DNA Extraction from Dry Museum Beetles without Conferring External Morphological Damage. PLoS One 2: e272.

Glez-Peña D, Gómez-Blanco D, Reboiro-Jato M, Fdez-Riverola F, Posada D (2010) ALTER: program-oriented conversion of DNA and protein alignments. Nucleic Acids Res 38: W14–W18.

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

Gotelli NJ (2004) A taxonomic wish-list for community ecology. Philos Trans R Soc Lond B Biol Sci 359: 585–97.

Grummer JA, Bryson RW, Reeder TW (2014) Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). Syst Biol 63: 119–33.

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

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One 6: e17497.

Hajibabaei M, Singer GA, Hickey DA (2006b) Benchmarking DNA barcodes: an assessment using available primate sequences. 49: 851–854.

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

Hamilton AJ, Basset Y, Benke KK, Grimbacher PS, Miller SE, Novotný V, Samuelson GA, Stork NE, Weiblen GD, Yen JDL (2010) Quantifying uncertainty in estimation of tropical arthropod species richness. Am Nat 176: 90–95.

Hamilton CA, Hendrixson BE, Brewer MS, Bond JE (2014) An evaluation of sampling effects on multiple DNA barcoding methods leads to an integrative approach for delimiting species: a case study of the North American tarantula genus *Aphonopelma* (Araneae, Mygalomorphae, Theraphosidae). Mol Phylogenet Evol 71: 79–93.

Hausmann A, Godfray HC, Huemer P, Mutanen M, Rougerie R, van Nieukerken EJ, Ratnasingham S, Hebert PDN (2013) Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. PLoS One 8: e84518.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc LondonSeries B Biol Sci 270: 313–321.

Hebert PDN, DeWaard JR, Landry J-F (2010) DNA barcodes for 1/1000 of the animal kingdom. Biol Lett 6: 359–362.

Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Syst Biol 54: 852–859.

Hebert PDN, 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 Biol 2: e312.

Hendrich L, Morinière J, Haszprunar G, Hebert PDN, Hausmann A, Köhler F, Balke M (2014) A comprehensive DNA barcode database for Central European beetles with a focus on Germany: Adding more than 3,500 identified species to BOLD. Mol Ecol Resour 15: 795–818.

Hendrich L, Pons J, Ribera I, Balke M (2010) Mitochondrial Cox1 sequence data reliably uncover patterns of insect diversity but suffer from high lineage-idiosyncratic error rates. PLoS One 5: e14448.

Hernández-Triana LM, Prosser SW, Rodríguez-Perez MA, Chaverri LG, Hebert PDN, Gregory TR (2014) Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. Mol Ecol Resour 14: 508–18.

Hjalmarsson AE, Bukontaite R, Ranarilalatiana T, Randriamihja JH, Bergsten J (2013) Taxonomic revision of Madagascan *Rhantus* (Coleoptera, Dytiscidae, Colymbetinae) with an emphasis on Manjakatompo as a conservation priority. Zookeys 350: 21–45.

Hogg ID, Hebert PDN (2004) Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. Can J Zool 82: 749–754.

Huemer P, Karsholt O, Mutanen M (2014a) DNA barcoding as a screening tool for cryptic diversity: an example from *Caryocolum*, with description of a new species (Lepidoptera, Gelechiidae). Zookeys 404: 91–111.

Huemer P, Mutanen M, Sefc KM, Hebert PDN (2014b) Testing DNA barcode performance in 1000 species of European Lepidoptera: large geographic distances have small genetic impacts. PLoS One 9: e115774.

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

Ivanova N V, deWaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Mol Ecol Notes 6: 998–1002.

Kekkonen M, Hebert PDN (2014) DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Mol Ecol Resour 14: 706–715.

Kerr KCR, Stoeckle MY, Dove CJ, Weigt L a., Francis CM, Hebert PDN (2007) Comprehensive DNA barcode coverage of North American birds. Mol Ecol Notes 7: 535–543.

Klausnitzer B (2001) Die Larven der Käfer Mitteleuropas, 6. Band. Heidelberg, Spektrum Akademischer Verlag.

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst Biol 56: 887–895.

Krell F-T (2004) Parataxonomy vs. taxonomy in biodiversity studies — pitfalls and applicability of "morphospecies" sorting. Biodivers Conserv 13: 795–812.

Kwong S, Srivathsan A, Vaidya G, Meier R (2012) Is the COI barcoding gene involved in speciation through intergenomic conflict? Mol Phylogenet Evol 62: 1009–1012.

Lagisz M, Poulin R, Nakagawa S (2013) You are where you live: parasitic nematode mitochondrial genome size is associated with the thermal environment generated by hosts. J Evol Biol 26: 683–690.

Larsen B, Aone C (1999) Fast and effective text mining using linear-time document clustering. Proc fifth ACM SIGKDD Int Conf Knowl Discov Data Min 16–22.

Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using genome-wide SNP data. Syst Biol 63: 534–542.

Lim GS, Balke M, Meier R (2012) Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Syst Biol 61: 165–169.

Lipscomb D, Platnick N, Wheeler Q (2003) The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol Evol 18: 65–66.

Lou M, Golding GB (2010) Assigning sequences to species in the absence of large interspecific differences. Mol Phylogenet Evol 56: 187–94.

Lukhtanov V, Sourakov A, Zakharov EV, Hebert PDN (2009) DNA barcoding Central Asian butterflies: Increasing geographical dimension does not significantly reduce the success of species identification. Mol Ecol Resour 9: 1302–1310.

Mace GM (2004) The role of taxonomy in species conservation. Philos Trans R Soc Lond B Biol Sci 359: 711–9.

Martin AP (1995) Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol Biol Evol 12: 1124–1131.

Mayden RL (1997) A hierarchy of species concepts: the denouement in the saga of the species problem. In: Claridge MF, Dawah HA, Wilson MR (eds) Species: The units of biodiversity. London, Chapman and Hall: 381–423.

Mayden RL (2002) On biological species, species concepts and individuation in the natural world. Fish Fish 3: 171–196.

Mayr E (1982) The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Cambridge, MA, Harvard University Press.

McKay BD, Zink RM (2010) The causes of mitochondrial DNA gene tree paraphyly in birds. Mol Phylogenet Evol 54: 647–50.

McKenna DD, Wild AL, Kanda K, Bellamy CL, Beutel RG, Caterino MS, Farnum CW, Hawks DC, Ivie MA, Jameson ML, Leschen RAB, Marvaldi AE, McHugh J V, Newton AF, Robertson JA, Thayer MK, Whiting MF, Lawrence JF, Slipinski A *et al.* (2015) The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. Syst Entomol 40: 835–880.

Meier R (2008) DNA sequences in taxonomy - Opportunities and challenges. In: Wheeler QD (ed) The New Taxonomy. Boca Raton FL, CRC Press: 95–126.

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

Miller KB, Alarie Y, Wolfe GW, Whiting MF (2005) Association of insect life stages using DNA sequences: The larvae of *Philodytes umbrinus* (Motschulsky) (Coleoptera: Dytiscidae). Syst Entomol 30: 499–509.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov 2010, New Orleans LA: 1–8.

Miralles A, Vences M (2013) New metrics for comparison of taxonomies reveal striking discrepancies among species delimitation methods in *Madascincus* lizards. PLoS One 8: e68242.

Mishler BD (2009) Species are not uniquely real biological entities. In: Ayala FJ, Arp R (eds) Contemporary Debates in Philosophy of Biology. Singapore, Wiley-Blackwell: 110–122.

Modica MV, Puillandre N, Castelin M, Zhang Y, Holford M (2014) A good compromise: Rapid and robust species proxies for Inventorying biodiversity hotspots using the Terebridae (Gastropoda: Conoidea). PLoS One 9: e102160.

Monaghan MT, Balke M, Pons J, Vogler AP (2006) Beyond barcodes: complex DNA taxonomy of a South Pacific Island radiation. Proc R Soc London Series B Biol Sci 273: 887–893.

Monaghan MT, Wild R, Elliot M, Fujisawa T, Balke M, Inward DJG, Lees DC, Ranaivosolo R, Eggleton P, Barraclough TG, Vogler AP (2009) Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst Biol 58: 298–311.

Moon-van der Staay SY, De Wachter R, Vaulot D (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature 409: 607–10.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean? PLoS Biol 9: e1001127.

Mutanen M, Kaila L, Tabell J (2013) Wide-ranging barcoding aids discovery of one-third increase of species richness in presumably well-investigated moths. Sci Rep 3: 2901.

Mutanen M, Wahlberg N, Kaila L (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc R Soc B Biol Sci 277: 2839–2848.

Nielsen ES, Mound LA (2000) Global diversity of insects: The problems of estimating numbers. In: Raven PH, Williams T (eds) Nature and Human Society: The Quest for a Sustainable World. Washington DC, National Academy Press: 213–222.

Oberprieler RG, Marvaldi AE, Anderson RS (2007) Weevils, weevils, weevils everywhere. Zootaxa 1668: 491–520.

Park J-K, Kim K-H, Kang S, Kim W, Eom KS, Littlewood DTJ (2007) A common origin of complex life cycles in parasitic flatworms: evidence from the complete mitochondrial genome of *Microcotyle sebastis* (Monogenea: Platyhelminthes). BMC Evol Biol 7: 11.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP (2006) Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55: 595–609.

Price BW, Henry CS, Hall AC, Mochizuki A, Duelli P, Brooks SJ (2015) Singing from the grave: DNA from a 180 year old type specimen confirms the identity of *Chrysoperla carnea* (Stephens). PLoS One 10: e0121127.

Prosser SWJ, DeWaard JR, Miller SE, Hebert PDN (2016) DNA barcodes from century-old type specimens using next-generation sequencing. Mol Ecol Resour 16: 487–497.

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol Ecol 21: 1864–1877.

De Queiroz K (2005a) A unified concept of species and its consequences for the future of taxonomy. Proc Calif Acad Sci 56: 196–215.

De Queiroz K (2005b) Different species problems and their resolution. 27: 1263–1269.

De Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56: 879–886.

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 R Soc Lond B Biol Sci 275: 237–247.

Rassi P, Karjalainen S, Clayhills T, Helve E, Hyvärinen E, Laurinharju E, Malmberg S, Mannerkoski I, Martikainen P, Mattila J, Muona J, Pentinsaari M, Rutanen I, Salokannel J, Siitonen J, Silfverberg H (2015) Kovakuoriaisten maakuntaluettelo 2015 [Provincial List of Finnish Coleoptera 2015]. Sahlbergia 21, Supplement 1: 1–164.

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

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS One 8: e66213.

Reid NM, Carstens BC (2012) Phylogenetic estimation error can decrease the accuracy of species delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. BMC Evol Biol 12: 196.

Reinhold K (1999) Energetically costly behaviour and the evolution of resting metabolic rate in insects. Funct Ecol 13: 217–224.

Riedel A, Daawia D, Balke M (2010) Deep cox1 divergence and hyperdiversity of *Trigonopterus* weevils in a New Guinea mountain range (Coleoptera, Curculionidae). Zool Scr 39: 63–74.

Riedel A, Sagata K, Surbakti S, Tänzler R, Balke M (2013) One hundred and one new species of *Trigonopterus* weevils from New Guinea . Zookeys 280: 1–150.

Ross HA (2014) The incidence of species-level paraphyly in animals: A re-assessment. Mol Phylogenet Evol 76: 10–17.

Scharff N, Coddington JA, Griswold CE, Hormiga G, de Place BP (2003) When to quit? Estimating spider species richness in a northern European deciduous forest. J Arachnol 31: 246–273.

Seberg O, Humphries CJ, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends Ecol Evol 18: 63–65.

Shaffield CS, Hebert PDN, Kevan PG, Packer L (2009) DNA barcoding a regional bee (Hymenoptera: Apoidea) fauna and its potential for ecological studies. Mol Ecol Resour 9: 196–207.

Shi H, Liang H (2015) The genus *Pterostichus* in China II: the subgenus *Circinatus* Sciaky, a species revision and phylogeny (Carabidae, Pterostichini). Zookeys 536: 1–92.

Silfverberg H (2010) Enumeratio renovata Coleopterorum Fennoscandiae, Daniae et Baltiae. Sahlbergia 16: 1–144.

Slipinski SA, Leschen RAB, Lawrence JF (2011) Order Coleoptera Linnaeus, 1758. Zootaxa 3148: 203–208.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc Natl Acad Sci U S A 103: 12115–20.

Sperling FAH, Landry J-F, Hickey DA (1995) DNA-based identification of introduced ermine moth species in North America (Lepidoptera: Yponomeutidae). Ann Entomol Soc Am 88: 155–162.

Srivathsan A, Meier R (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. 28: 190–194.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Syst Biol 57: 758–771.
Strohm JHT, Gwiazdowski RA, Hanner R (2015) Fast fish face fewer mitochondrial mutations: Patterns of dN/dS across fish mitogenomes. Gene 572: 27–34.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol 30: 2725–2729.
Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2002) DNA points the way ahead in taxonomy. Nature 418: 479.
Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. Trends Ecol Evol 18: 70–74.
Teletchea F (2010) After 7 years and 1000 citations: Comparative assessment of the DNA barcoding and the DNA taxonomy proposals for taxonomists and non-taxonomists. Mitochondrial DNA 21: 206–226.
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 Res 22: 4673–4680.
Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S (1995) Structures of metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 Å. Science 269: 1069–1074.
Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S (1996) The whole structure of the 13-Subunit oxidized cytochrome c oxidase at 2.8 Å. Science 272: 1136–1144.
Tänzler R, Sagata K, Surbakti S, Balke M, Riedel A (2012) DNA barcoding for community ecology — how to tackle a hyperdiverse, mostly undescribed Melanesian fauna. PLoS One 7: e28832.
Wahlberg N, Wheat CW, Peña C (2013) Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). PLoS One 8: e80875.
Wells JD, Sperling FAH (2001) DNA-based identification of forensically important Chrysomyinae (Diptera: Calliphoridae). Forensic Sci Int 120: 110–115.
Vesterinen EJ, Lilley T, Laine VN, Wahlberg N (2013) Next generation sequencing of fecal DNA reveals the dietary diversity of the widespread insectivorous predator Daubenton's bat (*Myotis daubentonii*) in southwestern Finland. PLoS One 8: e82168.
Wheeler QD (2008) Undisciplined thinking: Morphology and Hennig's unfinished revolution. Syst Entomol 33: 2–7.
Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Syst Biol 54: 844–851.
Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20: 47–55.
Wilson EO (2004) Taxonomy as a fundamental discipline. Philos Trans R Soc B Biol Sci 359: 739–739.

Wirta HK, Hebert PDN, Kaartinen R, Prosser SW, Várkonyi G, Roslin T (2014) Complementary molecular information changes our perception of food web structure. Proc Natl Acad Sci U S A 111: 1885–1890.

Vos R (2015) Monophylizer: A web and command line tool to assess monophyly. URI: http://monophylizer.naturalis.nl/. Cited 2016/04/04.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci U S A 107: 9264–9269.

Zahiri R, Lafontaine JD, Schmidt BC, DeWaard JR, Zakharov E V, Hebert PDN (2014) A transcontinental challenge — a test of DNA barcode performance for 1,541 species of Canadian Noctuoidea (Lepidoptera). PLoS One 9: e92797.

Zhang J, Kapli P, Pavlidis P, Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29: 2869–2876.

Zhou X, Adamowicz SJ, Jacobus LM, Dewalt RE, Hebert PD (2009) Towards a comprehensive barcode library for arctic life — Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. Front Zool 6: 30.

Ødegaard F (2000) How many species of arthropods? Erwin's estimate revised. Biol J Linn Soc 71: 583–597.

Original articles

I Pentinsaari M, Hebert PDN & Mutanen M (2014) Barcoding Beetles: A Regional Survey of 1872 Species Reveals High Identification Success and Unusually Deep Interspecific Divergences. PLoS ONE 9(9): e108651.

II Pentinsaari M, Vos R & Mutanen M (2015) Algorithmic single-locus species delimitation: effects of sampling effort, variation and non-monophyly in four methods and 1870 species of beetles. Manuscript.

III Pentinsaari M, Salmela H, Mutanen M & Roslin T (2015) A widely-adopted taxonomic marker sheds light on protein evolution across the animal tree of life. Manuscript.

Study I is reprinted under the terms of the Creative Commons Attribution Licence.

Original publications are not included in the electronic version of the dissertation.

658. Kangas, Veli-Matti (2015) Genetic and phenotypic variation of the moose (Alces alces)

659. Prokkola, Hanna (2015) Biodegradation studies of recycled vegetable oils, surface-active agents, and condensing wastewaters

660. Halkola, Eija (2015) Participation in infrastructuring the future school : a nexus analytic inquiry

661. Kujala, Sonja (2015) Dissecting genetic variation in European Scots pine (*Pinus sylvestris* L.) : special emphasis on polygenic adaptation

662. Muilu-Mäkelä, Riina (2015) Polyamine metabolism of Scots pine under abiotic stress

663. Pakanen, Minna (2015) Visual design examples in the evaluation of anticipated user experience at the early phases of research and development

664. Hyry, Jaakko (2015) Designing projected user interfaces as assistive technology for the elderly

665. Varanka, Sanna (2016) Multiscale influence of environmental factors on water quality in boreal rivers : application of spatial-based statistical modelling

666. Luukkonen, Tero (2016) New adsorption and oxidation-based approaches for water and wastewater treatment : studies regarding organic peracids, boiler-water treatment, and geopolymers

667. Tolkkinen, Mari (2016) Multi-stressor effects in boreal streams : disentangling the roles of natural and land use disturbance to stream communities

668. Kaakinen, Juhani (2016) Öljyllä ja raskasmetalleilla pilaantuneita maita koskevan ympäristölainsäädännön ja lupamenettelyn edistäminen kemiallisella tutkimuksella

669. Huttunen, Kaisa-Leena (2016) Biodiversity through time : coherence, stability and species turnover in boreal stream communities

670. Rönkä, Nelli (2016) Phylogeography and conservation genetics of waders

671. Fucci, Davide (2016) The role of process conformance and developers' skills in the context of test-driven development

672. Manninen, Outi (2016) The resilience of understorey vegetation and soil to increasing nitrogen and disturbances in boreal forests and the subarctic ecosystem

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

SERIES EDITORS

A
SCIENTIAE RERUM NATURALIUM
Professor Esa Hohtola

B
HUMANIORA
University Lecturer Santeri Palviainen

C
TECHNICA
Postdoctoral research fellow Sanna Taskila

D
MEDICA
Professor Olli Vuolteenaho

E
SCIENTIAE RERUM SOCIALIUM
University Lecturer Veli-Matti Ulvinen

F
SCRIPTA ACADEMICA
Director Sinikka Eskelinen

G
OECONOMICA
Professor Jari Juga

H
ARCHITECTONICA
University Lecturer Anu Soikkeli

EDITOR IN CHIEF
Professor Olli Vuolteenaho

PUBLICATIONS EDITOR
Publications Editor Kirsti Nurkkala

UNIVERSITY of OULU
OULUN YLIOPISTO

ISBN 978-952-62-1209-8 (Paperback)
ISBN 978-952-62-1210-4 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)