Towards writing the encyclopaedia of life: an introduction to DNA barcoding

Vincent Savolainen, Robyn S Cowan, Alfried P Vogler, George K Roderick and Richard Lane

*Phil. Trans. R. Soc. B* 2005 **360**, 1805-1811

**References**

This article cites 46 articles, 7 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1805.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1805.full.html#related-urls

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to *Phil. Trans. R. Soc. B* go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
Towards writing the encyclopaedia of life: an introduction to DNA barcoding

Vincent Savolainen1,2,*, Robyn S. Cowan1, Alfried P. Vogler2,3, George K. Roderick4 and Richard Lane2

1Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK
2The Natural History Museum, Cromwell Road, London SW7 5BD, UK
3Division of Biology, Imperial College London, Silwood Park, Ascot SL5 7PY, UK
4Environmental Science, University of California, Berkeley, CA 94720-3114, USA

An international consortium of major natural history museums, herbaria and other organizations has launched an ambitious project, the ‘Barcode of Life Initiative’, to promote a process enabling the rapid and inexpensive identification of the estimated 10 million species on Earth. DNA barcoding is a diagnostic technique in which short DNA sequence(s) can be used for species identification. The first international scientific conference on Barcoding of Life was held at the Natural History Museum in London in February 2005, and here we review the scientific challenges discussed during this conference and in previous publications. Although still controversial, the scientific benefits of DNA barcoding include: (i) enabling species identification, including any life stage or fragment, (ii) facilitating species discoveries based on cluster analyses of gene sequences (e.g. cox1 = COI, in animals), (iii) promoting development of handheld DNA sequencing technology that can be applied in the field for biodiversity inventories and (iv) providing insight into the diversity of life.

Keywords: DNA barcoding; taxonomy; inventories; cox1; conservation; DNA bank

1. ‘STAR TREK’S TRICORDER’ COMING TO REALITY

In the early 1960s, World War II veteran Gene Roddenberry brought to the air a now famous science fiction drama, Star Trek, in which a handheld ‘tricorder’ device was used to scan and identify alien life forms (www.startrek.com). Four decades later, the first international conference on ‘Barcoding Life’ was held at the Natural History Museum in London (UK), attended by over 200 participants from about 50 countries, and a portable DNA sequencing device to identify all life was claimed to now be within reach (Marshall 2005). Of course, the London conference had nothing to do with Star Trek, but there is a parallel one mitochondial gene as a universal ‘identification’ marker for animal species (Hebert et al. 2003a, b). Building upon the idea of the ‘universal product code’, known as ‘barcodes’ in the retail industry (Brown 1997), a few DNA nucleotides (e.g. the sequences of a short DNA fragment) may well provide an immediate diagnosis for species. As with commercial barcodes, the use of these ‘species barcodes’ first requires the assembly of a comprehensive library that links barcodes and organisms. Recognizing the potential of this approach, the Alfred P. Sloan Foundation funded two meetings at Cold Spring Harbor, the first in March 2003, the second in September the same year. From these meetings came
the idea that major natural history museums should take the lead in connecting diagnostic DNA sequences both to specimen vouchers in collections and to the existing taxonomic system, the so-called Linnean system. In spring 2004, the Sloan Foundation provided another substantial award to establish a secretariat for the ‘Barcode of Life’, based at the Smithsonian’s National Museum of Natural History in Washington, DC, USA. The Consortium for the Barcode of Life (CBOL) was also created and joined by many natural history museums and herbaria, research organizations and private partners (www.barcoding.si.edu).

Barcoding life has quickly attracted much attention and received considerable media coverage. Last autumn the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) sealed a partnership with CBOL whereby barcode standard DNA sequences and relevant supporting data can now be archived in GenBank. This is an important step forward because despite the greatest efforts of the curatorial teams at GenBank (www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html) and elsewhere, much DNA sequencing work has hitherto been done without adhering to standards in taxonomy and data quality, and systematic coverage remains sketchy, precluding a wider use of these molecular tools in taxonomy. Much debate has also been generated; for example at the Partnerships for Enhancing Expertise in Taxonomy’s fifth biennial conference Vincent Smith, Kipling Will and Paul Hebert participated in a vivid debate on ‘Genetic Barcoding’, (available for viewing at www.conferences.uiuc.edu/peet/video.html) which will be published soon in Systematic Biology (Hebert & Gregory in press; Smith in press; Will et al. in press).

The advocates of DNA barcoding say that it will revitalize biological collections and speed up species identification and inventories (Gregory 2005; Schindel & Miller 2005), whereas its opponents argue that it will destroy traditional systematics and turn it into a service industry (Ebach & Holdrege 2005); several papers have provided weighted analyses of the pros and cons (Moritz & Cicero 2004; Marshall 2005).

Last February, in contrast, the spirit of the CBOL conference in London was to provide a scientific and technological forum where an objective examination of the prospects and limitations of DNA barcoding were made possible, and standing pro-actively far away from the often tedious and rather naïve polemics that have surrounded the barcoding initiative. Instead, the main scientific issues debated were (i) is it possible to distinguish a large number of species using short DNA sequence data; (ii) can closely related or fast-evolving species be distinguished with this technique? (iii) what are the appropriate DNA sequences for barcoding various taxa, i.e. will the partial cox1 sequence be useful in groups other than those tested so far, or will we have to use multiple markers from different genomes to identify all life? The technological challenges also included the building of both a simple portable DNA sequencing device and a centralized, and appropriately curated, barcoding-specific database. This themed issue compiles some of the best contributions from the London conference. Here we introduce the broad scope of papers and views that helped in making this meeting a success.

2. DNA-BASED BIODIVERSITY INVENTORIES
The direct benefits of DNA barcoding undoubtedly include:

(i) make the outputs of systematics available to the largest possible community of end-users by providing standardized and high-tech identification tools, e.g. for biomedicine (parasites and vectors), agriculture (pests), environmental assays and customs (trade in endangered species);
(ii) relieve the enormous burden of identifications from taxonomists, so they can focus on more pertinent duties such as delimiting taxa, resolving their relationships and discovering and describing new species;
(iii) pair up various life stages of the same species (e.g. seedlings, larvae);
(iv) provide a bio-literacy tool for the general public.

Perhaps another relatively uncontroversial aspect of DNA barcoding is that it will also facilitate basic biodiversity inventories. Indeed, from the premises of molecular phylogenetics to assembling the tree of life (Blaxter 2003; Cracraft & Donoghue 2004), DNA has proved useful in identifying clades and evolutionary relationships. Whether or not actual species can be identified with DNA (see below), the number of distinct DNA sequences in environmental sampling and reconstruction of phylogenetic trees to place these sequences into an evolutionary context have been used in several inventories of cryptic biodiversity (e.g. soil bacteria or marine/freshwater micro-organisms). Initially referred to as DNA typing or profiling, the DNA barcoding initiative has taken this step forward, and several taxa have now been surveyed in their natural habitats using this technique.

Such an approach has been particularly useful for marine organisms (Shander & Willlassen 2005), including fishes (Mason 2003; Ward et al. 2005), soil meiofauna (Blaxter et al. 1998, 2004), freshwater meioinobenthos (Markmann & Tautz 2005) and even extinct birds (Lambert et al. 2005). In the rainforests, rapid DNA-based entomological inventories have been performed so efficiently (Janzan et al. 2005; Monaghan et al. 2005; Smith et al. 2005) that tropical ecologists have been among the most active advocates of DNA barcoding (Janzan 2004).

More pragmatically, DNA barcodes have proved useful in biosecurity, e.g. for surveillance of disease vectors (Besansky et al. 2003) and invasive insects (Armstrong & Ball 2005), as well as for law enforcement and primatology (Lorenz et al. 2005). Barcoding efforts have also recently received the attention of conservation agencies. For example, the UK Darwin Initiative for the Survival of Species (www.darwin.gov.uk) has funded two projects this year that include DNA barcoding activities to support conservation priorities, capacity building and trade surveillance in meso-American orchids and cacti.

3. BEYOND A UNIVERSAL COX1 BARCODE
The core idea of DNA barcoding is based on the fact that short pieces of DNA can be found that vary only to
a very minor degree within species, such that this
variation is much less than between species (www.
barcodinglife.org). Simplistically, a threshold of vari-
ation could even possibly be characterized for each
taxonomic group (ca 2–12%) above which groups of
individuals do not belong to the same species but
instead form a supra-specific taxon. Clustering ana-
lyses could therefore, be performed based on DNA
sequences, reveal species groups and assign unknown
individuals to species (Hebert et al. 2003a; figure 1).
One such piece of DNA, the mitochondrial cyto-
chrome oxidase subunit I (cox1, usually referred to as
COI in barcoding studies; see White et al. (1998) for a
discussion on gene nomenclature), was proposed to be
a good candidate for barcoding animal species (Hebert
et al. 2003a).

With several early successes of using cox1 (Hebert
et al. 2003b; Remigio & Hebert 2003; Hogg & Hebert
2004), larger sequencing programmes were set up (e.g.
for fishes and birds, see CBOL website), with
concerted massive data production rapidly differenti-
ing the barcoding movement from previous DNA-
concerted massive data production rapidly differentia-
ing the barcoding movement from previous DNA-
concerted massive data production rapidly differentia-
ing the barcoding movement. In its strictest
sense, DNA barcoding addresses only a limited aspect
taxonomic work (Gregory 2005). As DNA barcodes
are applicable to all life stages, it is also especially
useful in cases where larval stages are difficult to
identify with traditional methods, e.g. butterflies
(Janzen et al. 2005) or amphibians (Vences et al.
2005), and social insects in which several casts have
different ‘unrelated’ morphologies (Smith et al. 2005).
In all of these cases, DNA barcoding is applied only in
conjunction with classical approaches. Where species
are simply unknown or no attempts have been made to
delimit them, the barcode approach as originally
intended would be limited in its applicability.

However, it is a widely accepted fact that species,
however defined, are variable for most DNA markers
including the widely used cox1 gene. Hence, the
analogy to commercial barcodes presumes that the
variation within these species is smaller than between
them (www.barcodinglife.org). Therefore, an obvious
contribution that barcoding is making to taxonomy is
helping to discover cryptic species (Hebert et al. 2004).
Using DNA to discover such morphologically similar
but genetically differentiated species is not new or
contentious (Moritz & Cicero 2004); even cryptic
elephant species have been described based largely on
genetic distances and clustering analyses (Roca et al.
2001). However, in these cases the reference to
established species no longer needs to be strict, and
species delimitation is at least partially relying on DNA
data. This is a challenging problem that requires the
characterization of appropriate markers and analytical
tools, i.e. to discriminate clusters of interbreeding
individuals versus those that have experienced an
interruption to gene flow for a long enough period of
time that species recognition is appropriate.

Recent barcoding papers have advocated criteria on
sequence similarity, assuming a cut-off value for
maximum within-species variation (e.g. Lambert et al.
2005). Within a parsimony framework, however,
‘barcoders’ have looked for unique combinations of
autapomorphies in populations (DeSalle et al. 2005).
Others have suggested that because of within species
variation at potentially every nucleotide more sophis-
ticated methods of species assignment are necessary
(Matz & Nielsen 2005). Such approaches are an active
area of research and are being implemented in new
user-friendly software (e.g. Steinke et al. 2005).

Phenetic approaches have also been used where species
are not so easily conceptualized as biological entities, as
in micro-organisms, or where these entities are difficult
to define based on morphology, as in nematodes
(Blaxter et al. 2005) and other meiofauna (Markmann
& Tautz 2005). This has led to the concept of
‘molecular operational taxonomic unit’ that refers to
clusters of individuals that are recognized based on
such analyses of sequence similarity (Blaxter et al.
2005).

4. EXPLORING SPECIES LIMITS

Broadly speaking, taxonomy is concerned with the
identity of organisms and their relationships. The
discipline certainly faces many challenges in this new
century (Godfray 2002; Godfray & Knapp 2004;
Smith in press), and DNA barcodes are likely to play a
major role in the future of taxonomy. In its strictest
sense, DNA barcoding addresses only a limited aspect
of the taxonomic process, by matching DNA
sequences to ‘known’ species, the latter being delimited with traditional (e.g. morphological) meth-
odologies. In this context, the role of barcodes is to
provide a tool to assign unidentified specimens to
already characterized species (Hebert et al. 2003a).
This is of great utility to the end users of taxonomy,
and will help make more rapid progress in traditional
taxonomic work (Gregory 2005). As DNA barcodes
are applicable to all life stages, it is also especially
useful in cases where larval stages are difficult to
identify with traditional methods, e.g. butterflies
(Janzen et al. 2005) or amphibians (Vences et al.
2005), and social insects in which several casts have
different ‘unrelated’ morphologies (Smith et al. 2005).
In all of these cases, DNA barcoding is applied only in
conjunction with classical approaches. Where species
are simply unknown or no attempts have been made to
delimit them, the barcode approach as originally
intended would be limited in its applicability.

However, it is a widely accepted fact that species,
however defined, are variable for most DNA markers
including the widely used cox1 gene. Hence, the
analogy to commercial barcodes presumes that the
variation within these species is smaller than between
them (www.barcodinglife.org). Therefore, an obvious
contribution that barcoding is making to taxonomy is
helping to discover cryptic species (Hebert et al. 2004).
Using DNA to discover such morphologically similar
but genetically differentiated species is not new or
contentious (Moritz & Cicero 2004); even cryptic
elephant species have been described based largely on
genetic distances and clustering analyses (Roca et al.
2001). However, in these cases the reference to
established species no longer needs to be strict, and
species delimitation is at least partially relying on DNA
data. This is a challenging problem that requires the
characterization of appropriate markers and analytical
tools, i.e. to discriminate clusters of interbreeding
individuals versus those that have experienced an
interruption to gene flow for a long enough period of
time that species recognition is appropriate.

Recent barcoding papers have advocated criteria on
sequence similarity, assuming a cut-off value for
maximum within-species variation (e.g. Lambert et al.
2005). Within a parsimony framework, however,
‘barcoders’ have looked for unique combinations of
autapomorphies in populations (DeSalle et al. 2005).
Others have suggested that because of within species
variation at potentially every nucleotide more sophis-
ticated methods of species assignment are necessary
(Matz & Nielsen 2005). Such approaches are an active
area of research and are being implemented in new
user-friendly software (e.g. Steinke et al. 2005).

Phenetic approaches have also been used where species
are not so easily conceptualized as biological entities, as
in micro-organisms, or where these entities are difficult
to define based on morphology, as in nematodes
(Blaxter et al. 2005) and other meiofauna (Markmann
& Tautz 2005). This has led to the concept of
‘molecular operational taxonomic unit’ that refers to
clusters of individuals that are recognized based on
such analyses of sequence similarity (Blaxter et al.
2005).
Some researchers have argued that ultimately the short DNA sequences themselves could potentially provide the basis for a taxonomic system, in which a set of sequences is used to delimit a cohesive group of organisms and the sequence itself represents the species diagnosis—or even the ‘type’. This idea of DNA taxonomy (Tautz et al. 2003) assumes that evolutionary entities in nature are recognizable equally as well from DNA sequences as from any other evidence in traditional taxonomy, see Markmann & Tautz (2005). Preliminary analyses presented in this issue are shown for beetles (Monaghan et al. 2005), meiofauna (Markmann & Tautz 2005) and nematodes (Blaxter et al. 2005). This use of DNA barcoding, however, remains among the most contentious (Will et al. in press).

Despite the firm commitment of the barcoding community to collection-based taxonomic research,
criticisms from systematists have continued (Lipscomb et al. 2003; Wheeler 2004; Ebach & Holderege 2005; Will et al. in press). CBOL has responded to this by noting that barcoding is neither a substitute for alphataxonomy nor about inferring phylogenies (Schindel & Miller 2005). However, we must keep open the possibility that the barcode sequences per se and their ever-increasing taxonomic coverage could become an unprecedented resource for taxonomy and systematics studies in addition to being a diagnostic tool. As little as ten years ago, a standard paper in the top-ranked journals of molecular systematics was likely to be based on no more sequence information per taxon than the barcodes of today and with less dense taxon sampling. Although, phylogenetic support levels were frequently low in these studies, it would be incongruous to ignore the phylogenetic information content of short mitochondrial DNA sequences at appropriate levels of divergence (see Rubinoff & Holland in press, for a critique), especially if in the future these could be supplemented with sequences from a standardized set of nuclear markers. In plants, the need for multiple markers is likely to be a necessity and is already being explored (Chase et al. 2005; Kress et al. 2005), but this approach may be equally useful for most other groups (Monaghan et al. 2005). With sampling of multiple individuals in populations and across geographic ranges, the power of barcodes could well also help resolve several taxonomic problems and assist in establishing the extent of species entities, as several papers in this volume discuss (DeSalle et al. 2005; Janzen et al. 2005; Smith et al. 2005). It also possible that some taxa can be established from the sequence variation alone and re-identified unequivocally in future collections while awaiting morphological analysis and formal species description, i.e. the ‘reverse taxonomy’ of Markmann & Tautz (2005; see also Monaghan et al. 2005).

5. A LIFE BARCODER FOR CONSERVATION

As the technical aspects of large-scale production of molecular barcodes are becoming more refined (Hajibabaei et al. 2005), and the value of the resulting database is increasingly apparent, barcoding of life has now developed into a more complex tool with uses at the interface between population genetics, phylogenetics and taxonomy. This is not new in essence, but perhaps what makes the barcoding of life unique is the large scale of its technological and societal ambitions. Another important factor is the aim of barcoding for standardization of the markers, DNA banking and proper taxonomic vouchering. The urgency of creating DNA and tissue banks has been well recognized (Savolainen & Reeves 2004; Lorenz et al. 2005; Savolainen et al. in press), and solutions for linking DNA samples with taxonomic vouchers are being developed for all sorts of organisms, for example for those ‘barcoded’ nematodes (Blaxter et al. 2005) that at first did not seem to exhibit morphological variation at the species level (De Ley et al. 2005).

Barcoding of life will have to be both integrative and integrated with other worldwide taxonomic initiatives such as the Global Taxonomic Initiative of the Convention on Biological Diversity (www.biodiv.org) or the Global Biodiversity Information Facility (www.gbif.org). Perhaps within three years a handheld DNA sequencer will become available (Rita Colwell, personal communication, Smithsonian Botanical Symposium 2005, see http://persoon.si.edu/sbs/index.cfm), and one can now envision the time when automated DNA barcoding and wireless communications technology will be combined in a portable device. Not only will a ‘Life Barcoder’ be used to identify species, but also be linked via the World Wide Web to other kinds of biodiversity data such as images, uses, conservation status or biology. Surely if every child, politician and scientist has such direct access to life form information, then the importance of preserving biodiversity can only be formidably enhanced. This is not to deny naively the complexity of the problem of biological conservation, but given its urgency, we should welcome and help develop new initiatives that hold promise in that direction, which is the case for the barcoding of life.

We thank Mark Chase, Jim Mallet, Craig Moritz, David Schindel and Quentin Wheeler for useful comments and discussions, as well as the Royal Society, especially James Joseph, Editor, for having agreed to publish this themed issue. We are also grateful to the Consortium for the Barcode of Life, the Alfred P. Sloan and Gordon and Betty Moore Foundations and the Darwin Initiative for funding.

REFERENCES


Godfray, H. C. J. 2002 Challenges for taxonomy—the discipline will have to reinvent itself if it is to survive and flourish. *Nature* **417**, 17–19. (doi:10.1038/417017a)


DNA barcodes for biosecurity: invasive species identification

K.F Armstrong and S.L Ball

Phil. Trans. R. Soc. B 2005 360, 1813-1823

Supplementary data
"Data Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2009/02/12/360.1462.1813.DC1.html

References
This article cites 46 articles, 4 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1813.full.html#ref-list-1

Erratum
An erratum has been published for this article, the contents of which has been appended at the end of this reprint. The erratum is available online at: correction
http://rstb.royalsocietypublishing.org/content/360/1464/2373.2

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
DNA barcodes for biosecurity: invasive species identification

K. F. Armstrong* and S. L. Ball

National Centre for Advanced Bio-Protection Technologies, PO Box 84, Lincoln University, Canterbury, New Zealand

Biosecurity encompasses protecting against any risk through ‘biological harm’, not least being the economic impact from the spread of pest insects. Molecular diagnostic tools provide valuable support for the rapid and accurate identification of morphologically indistinct alien species. However, these tools currently lack standardization. They are not conducive to adaptation by multiple sectors or countries, or to coping with changing pest priorities. The data presented here identifies DNA barcodes as a very promising opportunity to address this. DNA of tussock moth and fruit fly specimens intercepted at the New Zealand border over the last decade were reanalysed using the cox1 sequence barcode approach. Species identifications were compared with the historical dataset obtained by PCR–RFLP of nuclear rDNA. There was 90 and 96% agreement between the methods for these species, respectively. Improvements included previous tussock moth ‘unknowns’ being placed to family, genera or species and further resolution within fruit fly species complexes. The analyses highlight several advantages of DNA barcodes, especially their adaptability and predictive value. This approach is a realistic platform on which to build a much more flexible system, with the potential to be adopted globally for the rapid and accurate identification of invasive alien species.

Keywords: mitochondrial DNA; cytochrome oxidase subunit I; COI; molecular diagnostics; quarantine; insects

1. INTRODUCTION

Biosecurity is emerging as one of the most important issues facing the international community. Traditionally it has been associated with risks from infectious diseases, living modified organisms and biological weapons, but in the very broadest sense it encompasses minimizing risk through ‘biological harm’ (Meyerson et al. 2002). Not least is the economic risk from invasive alien species (IAS) that threaten ecosystem stability, producer livelihoods and consumer confidence (Cock et al. 2003). That risk is facilitated by the movement of exotic species around the world through increasing international tourism and trade, and is influenced by changes in climate and land use. Of those species introduced to novel environments an estimated one percent is anticipated to become invasive and with serious economic impacts (Williamson 1996). An example relevant to the following discussion is provided by Japan, where on average four exotic insect species have become established each year for the last 50 years. Of these 74% were economic pests, but just two, the Oriental fruit fly, Bactrocera dorsalis, and the melon fly, Bactrocera cucurbitae, have cost equivalent to more than EUR200 million to eradicate (Kiritani 1998). Also in the USA, the potential cumulative economic losses from Asian gypsy moth (Lymantria dispar) and nun moth (Lymantria monacha) establishment between 1990 and 2004 were estimated in the range equivalent to EUR28–46 billion (Cock et al. 2003).

New Zealand is very sensitive to the potential impact that such pests could have on the primary industries and natural ecosystems that underpin its economy. This is apparent by internationally having the most comprehensive biosecurity approach based on its Biosecurity Act of 1993 (Meyerson & Reaser 2002). Nevertheless, one of the main weaknesses recognized with this is the difficulty to predict new IAS which limits the implementation of appropriate risk management strategies (Parliamentary Commissioner for the Environment 2000). A critical aspect of prediction, and also monitoring, is the ability to accurately identify any intercepted specimen to the species-level. This is essential for support of early detection systems. It is also a means of collecting complete and accurate data about which species are actually entering for the assessment of risk. However, development of a comprehensive identification capability is hindered by the growing imbalance worldwide between diagnostic needs and the availability of trained taxonomic experts. Long-term research strategies are also required to address the deficiencies in existing taxonomic keys to deal with morphologically indistinct immature life stages, cryptic species and damaged specimens. For a few of the most economically significant and global pests morphothaxonomic keys are now supported by molecular diagnostic technology, e.g. fruit flies (Tephritidae; Armstrong et al. 1997a), tussock moths (Lymantriidae; Armstrong et al. 2003), leafroller moths (Tortricidae; Dugdale et al. 2002) and thrips (Thripidae; Toda & Komazaki 2002). However, such methods are developed on an ad hoc and often reactive basis with immediate local needs in mind and

* Author for correspondence (armstron@lincoln.ac.nz).

One contribution of 18 to a Theme Issue ‘DNA barcoding of life’.

© 2005 The Royal Society
little or no coordination between institutes, regions or taxa. The following discussion illustrates how DNA barcodes can provide a very realistic, practical and flexible framework for species identification in the context of biosecurity. Specific examples are given from a New Zealand perspective for two global economically significant agricultural pest insect groups, the fruit flies and tussock moths. However the principles could equally apply to other taxa, sectors and countries. We propose that adoption of the method would enable the international IAS diagnostic community to better cope with changing and localized species priorities, to capitalize on the efforts of others and to address the international standardization of technologies that has been recommended for a more effective and coherent diagnostic effort (e.g. Klijn 2004).

2. LIMITS OF PREVIOUS MOLECULAR DIAGNOSTICS FOR BIOSECURITY

A variety of stand-alone molecular methods exist for identification of regulated pest species. Immunological (e.g. Symondson et al. 1999; Trowell et al. 2000) or protein-based (e.g. Miles 1979; Soares et al. 2000) methods are not widely used being highly taxon-specific, difficult to adapt, vulnerable to environmental factors and reliant on good quality, fresh tissue. The majority of molecular diagnostic methods are instead polymerase chain reaction (PCR) based DNA analyses which are not limited in these ways. However, a decade or more into the application of molecular diagnostics for biosecurity and other identification purposes, two major hurdles exist that preclude the building of a smarter, more co-ordinated and anticipatory IAS identification system.

(a) Finite range of taxa

The number of taxa that can be accommodated by any one method is predetermined and limited to various degrees from one to around fifty species. Due initially to the relative expense of method development, and for other pragmatic reasons, the approach has been to develop tests for those taxa predicted to be the most likely invaders, i.e. for species known to be invasive and spreading elsewhere (Cock et al. 2003). Examples are for species within the fruit flies (Haymer et al. 1994; Armstrong et al. 1997a), tussock moths (Armstrong et al. 2003), leafroller moths (Dugdale et al. 2002) and thrips (Toda & Komazaki 2002). Taxa within these and others are prioritized differently amongst countries according to matching host and climate, access to existing pathways of entry and anticipated economic impact. Modification of protocols to accommodate additional species as the need arises, or to adapt for different sectors or countries with differing taxa priorities may not be practical. This can be especially difficult when diagnosis is reliant on a few single nucleotide polymorphisms that form the basis of primers required for specific PCR or for polymorphic restriction sites or in the design of oligonucleotide arrays. Finding additional informative polymorphisms within such methods can be problematic.

Methods designed out of necessity for a predetermined range of taxa also undermines the potential to cope with the unpredicted arrivals. The entry of species found on inanimate objects, such as vehicles (Armstrong et al. 2003) or solid wood packaging (Wittenburg & Cock 2001), are more difficult to predict compared to those that are closely associated with their host material such as fruit flies (Armstrong et al. 1997a). Others may not be predicted because they are innocuous or minor pests in their native range. However, they can become a significant pest in a new environment with no specific natural enemies or competition. This has been a significant issue in, for example, North America where of the six most devastating forestry pests introduced only the European strain of the gypsy moth was known as a pest in its indigenous range (Cock et al. 2003). Similarly in New Zealand the unanticipated arrival in 1999 of painted apple moth Teia anartoides (Lepidoptera: Lymantriidae) from Australia, where it is a minor localized pest, was predicted to have a significant impact and so an eradication programme was initiated. The cost to New Zealand if it is not eradicated is anticipated to be equivalent to EUR33–205 million over the next 20 years (Case study 3 2002). A significantly more flexible and anticipatory diagnostic system is required to provide timely support for management of these events.

(b) Diverse methodologies

There has been little or no consistency in the PCR-based technologies used. A number of different methods have been designed, such as species-specific PCR (e.g. Kohlmayr et al. 2002; Lu et al. 2002; Liu 2004), PCR restriction fragment length polymorphism (PCR–RFLP); e.g. Armstrong et al. 1997a,b; Brunner et al. 2002), multiplex PCR (Kumar et al. 1999; Kengne et al. 2001), DNA sequencing (e.g. Brown et al. 2002; Dugdale et al. 2002) and oligonucleotide array analyses (Naeole & Haymer 2003). Even the idiosyncrasies of similar methods means that they are rarely directly transferable between laboratories or for use with different taxa and data cannot be shared.

There is also no consistency in the gene or parts of genes used to identify species. For insect identification examples of mitochondrial DNA (mtDNA) used are cytochrome oxidase subunit I (cox1; e.g. Brunner et al. 2002; Kohlmayr et al. 2002), non-transcribed region between cox1 and tRNAphe (Stauffer 1997), 16S rDNA (Brown et al. 2002) and cytochrome B (Khemakhem et al. 2001, 2002). For nuclear gene regions the rDNA internal transcribed spacer regions ITS1 plus ITS2 (Armstrong et al. 1997a,b), ITS1 only (Chiu et al. 2001) and ITS2 only (Pfeifer et al. 1995) have been used, as well as an actin gene intron (He & Haymer 1997) and randomly amplified polymorphic DNA (RAPD; Kengne et al. 2001). To a certain extent choice is dependent on the taxonomic range involved and appropriate evolutionary rate of the gene, but there may also be elements of convenience regarding primers available and in-house experience. The consequences of this disparity have been recognized to be an issue of much broader dimensions across the field of phylogenetics (Caterino et al. 2000).
In essence therefore, molecular diagnostic tests, which are more and more being accepted as an inevitable and essential component of the biosecurity toolbox (Martin et al. 2000), remain very limited. They are not flexible enough to accommodate the growing number of IAS, to identify unanticipated arrivals or to capitalize on the efforts of others that collectively work across a very diverse taxonomic range.

3. IDENTIFICATION USING DNA BARCODES

The emergence of DNA barcoding as a means of species identification (Hebert et al. 2003a) has the potential to address the shortcomings outlined above. In contrast to the molecular diagnostic methods available to date, DNA barcoding proposes to use information within a single gene region common across all taxa and to access that information by DNA sequencing under universal conditions. These features lend it well to standardization across species and laboratories, thus providing a platform for global exchange of homologous data and capitalizing on the efforts of others to build a more flexible system.

There is a growing literature demonstrating that COX1 will reliably discriminate a diverse range of taxa at the species level (e.g. Hebert et al. 2003a,b; Hogg & Hebert 2004; Whiteman et al. 2004; Ball et al. 2005; Shander & Willassen 2005). This gene, along with 16S, 18S, and elongation factor-1α genes, has also been encouraged as a standard target for insect phylogenetics (Caterino et al. 2000). Of enormous benefit to the international diagnostics community is the very large amount of COX1 sequence information that already exists in the literature for a diverse range of insect taxa. However from a biosecurity perspective, where accuracy is critical, the robustness of identifications and genetic limits of this gene need to be established. Potential complications arising from discordance with morphologically established species limits, species sequence overlap or divergence across intra-specific geographic ranges also need to be examined. Even so, if the concept can be verified to operational agencies as sound, it needs to be demonstrated that DNA sequencing is a practical and rapid alternative to the current technologically more accessible methods.

4. TESTING COX1 DNA BARCODES FOR EXOTIC INSECT IDENTIFICATION

To examine the suitability of COX1 sequence as a diagnostic tool for biosecurity, two datasets that exist for molecular identifications of specimens intercepted at the New Zealand border over the past decade have been revisited. The datasets are for the tussock moths (Lepidoptera: Lymantriidae) and fruit flies (Diptera: Tephritidae). Several species within these groups are considered internationally to be significant economic pests. They are not established in New Zealand, but are considered high risk to New Zealand’s forestry and horticultural industries, respectively. DNA from a random selection of specimens that have been intercepted at the border was used. To test accuracy of the barcode method, identifications so determined were compared to those that, to all intents and purposes, had been successfully identified to species using the previously designed PCR–RFLP and specific-PCR methods. To test for improvement on the previous methods, all specimens that were previously unidentifiable by those methods were also included. The latter were either because of failure to PCR amplify the 1.5 kb ITS rDNA region for subsequent RFLP analysis, ambiguous RFLP patterns, RFLP patterns that were not recognized amongst those established for the target list of species or failure to amplify a nested species-specific PCR product.

DNA previously obtained for border specimens and specimens of the morphologically identified species contributing to the ‘profile’ data (see Electronic Appendix, table 3), was PCR amplified and sequenced for the COX1 Folmer region according to established procedure (Hebert et al. 2003a,b). The only variation was use of the Expand High Fidelity (Roche Diagnostics) polymerase system instead of Taq in the PCR. For use with the tussock moth profile data set, sequence data for other Lepidopteran species was also included from the Barcode of Life Database (BOLD) and GenBank (see Electronic Appendix, table 2). Sequences were aligned and truncated to a ca 650 bp homologous region using Sequencher (Gene Codes Corp.). A profile neighbour-joining (NJ) tree of Kimura-2-parameter (K2P) distances was constructed from the sequence data using MEGA v2.1. The K2P model provides a suitable metric model when genetic distances are low (Nei & Kumar 2000) as anticipated with many of the species here. The simple NJ algorithm was considered at this juncture to be an appropriate starting point for the analyses, given that specimen identification is based entirely on sequence similarity, rather than on strictly phylogenetic relationships, and the speed of analysis that is necessary for biosecurity diagnostic purposes.

(a) Case study 1: tussock moths

Background: around 30 species of tussock moths have been determined to be unwanted organisms under the Biosecurity Act (MAF Biosecurity, Unwanted Organisms Register). Based on their pest status, polyphagous nature and invasive potential, seven northern hemisphere species are considered to present the greatest risk to New Zealand forestry. These are Asian and European gypsy moth (L. dispar), nun moth (L. monacha), pink or rose gypsy moth (L. malthusa), vapourer or rusty tussock moth (Orgyia antiqua), white marked tussock moth (Orgyia leucostigma), Douglas fir tussock moth (Orgyia pseudotsugata) and white spotted tussock moth (Orgyia thyellinae; Armstrong et al. 2003). Specific life history strategies, such as long overwintering phases in the egg stage and indiscriminate oviposition on inanimate surfaces, such as containers, ship superstructures, forestry equipment and used vehicles could enable them to arrive in New Zealand.

Of these, the Asian gypsy moth is particularly well equipped to invade as the females are capable of sustained flight (Keena et al. 2001) and are attracted to lights of vehicles and ports (Wallner et al. 1995). The species also has variable occlusion cues enabling hatching to coincide with favourable environmental conditions (Walsh 1993) and has a proven invasive
Table 1. Comparisons of previous molecular species identifications with DNA barcode identifications for New Zealand border specimens.

<table>
<thead>
<tr>
<th>taxon</th>
<th>total n</th>
<th>previously identified by RFLP or specific PCR</th>
<th>previously unidentifiable by RFLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n (% of total) disagreement (% of identified)</td>
<td>n (% of total) identified by barcoding (% of unidentifiable)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>barcode versus RFLP barcode versus ssp&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>

| tussock moth      | 57      | 49 (86.0%) 1 (2.0%)                                      | 8 (7.1%) 5 (62.5%)                           |
| fruit fly         | 81      | 79* (97.5%) 2 (2.5%)                                    | 2 (<0.1%) 2 (100%)                           |

* Includes 10 cases where common RFLP patterns indicated 2–3 possible species; final identifications determined by species and geographic origin of the host fruit.

<sup>b</sup> Species-specific primers.

ability (Savotikov et al. 1995). Consequently the tussock moth egg masses commonly intercepted on imported used vehicles had previously been assumed by quarantine officers to be Asian gypsy moth. Unfortunately, while this species can be readily distinguished from the others based on adult morphology the early life stages cannot. This is compounded by their arrival on inanimate objects with unknown origin, providing limited host or geographic information to indicate their likely identity. Consequently there has been no accurate record of species actually arrive in New Zealand. This has serious implications for the suitability of the post-border quarantine systems that are in place.

To improve interception records, a molecular diagnostic method was developed based on PCR–RFLP of 1.5 kb nuclear ribosomal DNA (rDNA) incorporating partial 18S plus complete ITS1, 5.8S and ITS2 regions. This has since been used to routinely identify the egg masses intercepted on imported used vehicles (Armstrong et al. 2003). Samples also arrive in very poor condition with potentially degraded DNA. Although not a risk in themselves, their accurate identification is necessary for a comprehensive risk analysis. Consequently a species-specific PCR method was designed to supplement the main RFLP diagnoses for specimens failing to amplify the 1.5 kb nuclear rDNA region (unpublished). Specific PCR primers were designed to amplify a 150–300 bp nested region of the ITS1 for the Asian species, L. dispar, L. mathura, L. monacha and O. thyllatina. Used together with a control amplicon of 350 bp of the 18S rDNA, positive amplification indicated a positive identification. Incorporating these original methods into operational procedures the large majority of specimens were confirmed to be gypsy moth, plus two other high risk species, L. monacha and O. thyllatina. Of concern however were the specimens that could not be identified. Some failed to PCR amplify. Others produced novel RFLP patterns for which no species or even genus could be inferred.

Recently, cox1 barcodes have been demonstrated to hold great potential for tussock moth species identification (Ball & Armstrong 2005). In that study 81 ‘test’ specimens were used to interrogate a cox1 sequence profile composed of 18 lymantriid species across four genera. 100% of the cox1 identifications agreed with their prior morphological identification, i.e. in all cases test sequences grouped more closely with their conspecifics than with any other species. This result is consistent with previous DNA barcoding studies of Lepidoptera (Hebert et al. 2003a). Testing this further as a biosecurity tool, new data are presented here for specimens intercepted at the border that had previously been identified to species, or otherwise, by PCR–RFLP or specific PCR.

Results: of the 57 border interception specimens analysed here, 49 had previously been identified to species by RFLP or by L. dispar-specific primers. Eight others had been unidentifiable (table 1). Of the latter, five could be placed with confidence (80–99% bootstrap support) by their cox1 sequence to a genus or species within the profile tree (table 1). This improved the previous RFLP identification rate from 86% to 93%.

Interestingly four of the five additional identifications were not tussock moths. Specimen MAF812 associated closely with two species of Spodoptera and MAF773 with two species of Clostera (Electronic Appendix, figure 1). While these appear to be the most likely congeners, the interspecific divergences suggest that the actual species are not represented in the profile dataset, e.g. 7.3% between Clostera albostriga and Clostera apicalis is of the same order as between MAF773 and each of those species (6.4% and 8.4%, respectively). A third specimen, MAF775, previously unamplifiable for subsequent RFLP or by L. dispar-specific primers, was amplified with the universal species cox1 primers. The sequence identified it as possibly a species of Dasychira, although there was only one species in the profile dataset (Dasychira dorispennata) representing this genus (Electronic Appendix, figure 1). A fourth specimen, MAF913 produced an ambiguous RFLP haplotype, but was identified here as Hyphantria cunea (Electronic Appendix, figure 1) with a mean sequence divergence of 1.4% from three H. cunea profile sequences. A fifth specimen, MAF891, appears to be a divergent form of L. dispar with 2.4% sequence divergence from all other non-Hokkaidoensis L. dispar specimens. The remaining three of the eight previous unidentifiables, MAF816, MAF851 and MAF912, could not be placed with confidence in the profile dataset. They grouped most closely with Hypane humuli, Leuhdorfia japonica

Phil. Trans. R. Soc. B (2005)
and Calophasia lunula, respectively (Electronic Appendix, figure 1), but the bootstrap supports were weak (12, 34 and 33%, respectively). The cox1 sequences suggest that these specimens belong to taxa not represented in this dataset, ranging from other species to other families. It is no surprise therefore that they were outside the diagnostic scope of the original ‘tussock moth specific’ RFLP method and explains why they were difficult to analyse. It also suggests very positively for barcoding that, with greater representation of lepidopteran taxa in the profile dataset, this method could achieve a 100% identification rate which is not possible by other current molecular methods.

Of the 49 previously identified specimens, the barcode identification disagreed with five of them (table 1). Each had been diagnosed as L. dispar. Four of them were identified using the L. dispar specific PCR, as they had been difficult to analyse by the PCR-RFLP. Three of those were identified by barcoding to belong to species not represented in the profile dataset. MAF915 and MAF729 associated with Spilosoma sp. (99% bootstrap support) and Hyphantria sp. (identical sequence) respectively, belonging to another lepidopteran family, the Arctiidae, and MAF839 came out on a long branch within the Lymantriidae, between L. xylina and L. dispar (80% bootstrap support). The fourth, MAF914, associated with high bootstrap support to the Korean haplotype of L. mathura which is genetically divergent from those in Japan (Electronic Appendix, figure 1). This implies the potential use of the barcode data to provide useful geographic origin information that was not possible previously. Data for the four specimens also indicate that the original species-specific primer test (unpublished) was not as broad as it had needed to be and again demonstrates the limitation of previous methods to deal with species outside the anticipated taxonomic range. The fifth specimen amongst those that disagreed, MAF795, was previously identified as L. dispar by RFLP, but 100% cox1 homology to H. cunea suggests otherwise. This specimen has been flagged as one that needs further analysis to determine its true identity.

(b) Case study 2: fruit flies

Background: Of the some 4000 species of fruit fly, around 250 are considered economic pests (White & Elson-Harris 1992). New Zealand remains the only major fruit producing country in the world that is free from them and significant investment has been placed in monitoring and surveillance systems to ensure their early detection and to minimize pathway risk (Cowley & Frampton 1989; Frampton 2000; Stephenson et al. 2003). The species, however, present different degrees of risk to New Zealand based on their host and climatic preferences and differing quarantine actions can result, i.e. to treat, re-ship or destroy the imported produce. Distinction of regulated and non-regulated species groups is the minimum diagnostic requirement, but identification to species is essential for accurate interception data and assessment of pathway risk. As for the tussock moths the majority of these species can be readily distinguished based on their adult morphology (White & Elson-Harris 1992), and late instar larval morphological keys are also becoming available for an increasing number of species (Carroll et al. 2004). Unfortunately it is usually the early instar larvae or eggs that are intercepted at the border in fruit of commercial consignments or accompanying overseas travellers cannot be identified morphologically beyond the family level.

Since 1994 a molecular diagnostic technique based on PCR-RFLP of the ITS1, 5.8S plus ITS2 rDNA regions (Armstrong et al. 1997b), has been used routinely to rapidly identify fruit fly intercepted at the New Zealand border to the species level. Also, in a similar manner to the tussock moth method, a species-specific PCR amplifying a 200 bp nested region has been used for the identification of degraded DNA associated with eggs found in cooked breadfruit (unpublished). Again while these do not present a threat themselves, accurate identification is necessary for comprehensive risk assessment. These methods effectively replaced the need to rear immature stages through to adults for identification, which was often unsuccessful or at best too slow for making timely biosecurity management decisions (Armstrong et al. 1997a). However, this approach has evolved from use with 19 original species to 49 (unpublished) and relies heavily on the host and geographic origin information to limit the list of likely species. Continuing to add more species has increasingly compromised the diagnostic sensitivity of the method as overlapping RFLP patterns become more common. Using DNA barcodes was therefore considered here as a method that might enable accurate identifications amongst a large number of species. The first data towards a fruit fly cox1 species profile, including the only two tephritid barcode region sequences available in Genbank at the time (see Electronic Appendix, table 3), is presented here and used to re-identify specimens intercepted at the border.

Results: one hundred and ninety-three sequences, representing 60 species were used to create a tephritid cox1 reference profile (Electronic Appendix, figure 2). Forty one species were represented by 2–14 specimens taken from across their geographic range where possible. Nineteen species were represented by only one specimen. The profile NJ analysis generally resolved the taxa according to their morphologically derived taxonomy as genera, subgenera, species and species complexes (White & Elson-Harris 1992). Bootstrap support was high at the nodes (greater than 80%) for species that were not part of a species complex (discussed below). There were two exceptions. Bactrocera psidii ‘clustered’ weakly with Bactrocera trinuncola; sequence divergence between them was 2.9%, compared to intra-specific 1.8% and 2.0%, respectively. Also, Bactrocera curvipes was placed within the Bactrocera tryoni complex, but with weak bootstrap support. Interestingly, the latter are also difficult to distinguish by the PCR-RFLP method (Armstrong & Cameron 2000) although the adults are morphologically distinct. Others that came out on long branches, Bactrocera minuta, Bactrocera arecae, Bactrocera distincta and Dacus demmerezii, suggests that there is insufficient taxonomic representation around these species within the current dataset. There were also two discrepancies that warrant further investigation. One is the Bactrocera cognata specimen 1009
which did not cluster with its conspecifics (specimens 975, 976 and 1011) that were correctly located within the *B. dorsalis* species complex (Drew & Hancock 1994; figure 2 in the Electronic Appendix, here considered to include the Asian and Australian species within the branch that includes *B. dorsalis* (specimen 726) through to *B. endiandrae* (specimen 789)). The second was *B. arecae* which is part of the *B. dorsalis* complex but placed distantly from it. These specimens are flagged for morphological and/or molecular re-examination. Importantly, besides some exceptions within the species complexes, there were no sequences that were shared by different species. This is in contrast to the current nuclear rDNA method where some RFLP haplotypes are common to several species.

The **cox1** sequences appear to be limited in their ability to distinguish taxa within the species complexes of *B. dorsalis*, *B. tryoni* (Morrow et al. 2000; figure 2 in the Electronic Appendix, species *B. tryoni*, *B. neohermeralis* and *B. aquilonis*) and *A. fraterculus* (Norrbom et al. 1999; figure 2 in the Electronic Appendix, all species within the branch starting from *A. ludens* through to *A. fraterculus*). This is interesting given that there were cases here, as for the tussock moth data set, where known subspecies could be distinguished. For example, the *B. cucurbitae* strains A and B (3.5% divergence between them here, versus 0.0% and 0.8% respectively within each) are anecdotally separated by host range (unpublished), and *Bactrocera xanthodes* and *Bactrocera paraxanthodes* (7.2% divergence between them here, versus 0.0% within each, respectively) are also separated by host (Drew et al. 1997). Each of these and other aspects of the reference profile data, such as the placement of *B. xanthodes* and *B. paraxanthodes* etc, are the subject of a more in-depth barcoding treatment in a separate publication in preparation.

Eighty one border intercepted specimens were identified by appending their **cox1** sequences to the profile dataset (Electronic Appendix, figure 2). Based on the closest species which they associated with in the NJ tree, 73 (94%) of these identifications were in agreement with the previous RFLP method (table 1) with high bootstrap support (84–100%). The mean sequence divergence between the intercepted specimens and the profile sequences they grouped with was 0.9% (range =0.1%–5.3%). This is consistent with intraspecific **cox1** divergences observed for a variety of insect taxa (Hebert et al. 2003a,b). Four identifications were in disagreement. Three of these were supported by high bootstrap support (Electronic Appendix, figure 2). Specifically, MAF274 was identified within the *B. tyroni* species complex by PCR–RFLP but as *B. facialis* here, MAF665 was *B. passiflorae* but *B. facialis* here and MAF940 was within the *B. tyroni* complex but *Dirioxa pornia* here. The fourth, MAF144, was weakly supported as being within the *B. dorsalis* species complex. The previous PCR–RFLP method had identified it either as *Bactrocera kirki*, *Bactrocera trilineola* or *Bactrocera frauenfeldi* which share common restriction profiles. However, bootstrap support for the entire *B. dorsalis* complex node was generally low in the profile dataset, indicating little confidence in the ability of these sequences to identify ‘unknowns’ within the complex. Further evaluation of these aberrant results as well as the groupings within the species complexes is underway. Finally, of the 81 border intercepted specimens analysed, two were previously classified as ‘unknown’, due to novel RFLP patterns. The **cox1** analysis clearly placed them within the *B. (B.) dorsalis* complex, but interestingly not within the clusters containing the only four species (*B. dorsalis*, *Bactrocera philippinensis*, *Bactrocera papayae* and *Bactrocera carambolae*) for which RFLP profiles had been determined.

**Conclusion**: as the analysis stands the fruit fly **cox1** sequences provide slightly better resolution, and also quantitative support in terms of bootstraps and divergence values, for species-level identification than was previously possible. The dataset however highlights the extent of **cox1** to provide confident identifications within species complexes. For those outside the complexes, the **cox1** data sort the species well in terms of their morphologically-based taxonomy, from genus to species. Ambiguity within the complexes may be a consequence of insufficient variation in **cox1** to accurately reconstruct such recent divergences, but that reasoning does not appear to hold for other sub-species separations, such as that within *B. cucurbitae*. The disparity may be a function of the status of the alpha taxonomy of this genus which is still under scrutiny (Smith et al. 2003) and also the amount of systematic interest that certain taxa have received. In this case, in contrast to *B. cucurbitae*, the *B. dorsalis* complex has been extensively studied taxonomically. Until recently *B. dorsalis* was considered the single most significant fruit fly pest species in Asia. A recent revision by Drew & Hancock (1994) now recognizes it as part of a complex of 52 sibling species of which eight are economically important, and more species continue to be described (see Clarke et al. 2005). In addition, with the highly specialized taxonomic expertise required to distinguish these species with any confidence, their apparent ‘poor resolution’ in the current analysis may in part be a consequence of mistaken specimen identification by the suppliers.

Does this limitation at the level of species complexes invalidate the ability of this method to correctly identify unknowns? As the method stands, it appears to be no less accurate than the existing PCR–RFLP method. For the *B. dorsalis* complex, interspersion of species such as *B. dorsalis sensu stricto*, *B. papayae* and *B. philippinensis* is no less informative than before, and from a biosecurity point this is not an issue as they are all regulated species. In fact there have been no molecular studies that have been able to satisfactorily distinguish these three species to date (Clarke et al. 2005). The method does however promise to be more informative for other species within the complex such as *Bactrocera kandiansis* and *Bactrocera caryae* which form a distinct group; interestingly this identified two border intercepted specimens that previously could only be resolved as far as the species complex. *B. carambolae* also forms a discrete group although it is not well supported. This is consistent with the PCR–RFLP method although it was very confidently distinguished from the others in the complex with that method. Additional notable improvements on the previous method are the separation of the species...
B. kirki, B. trilineola and B. frauenfeldi, which previously shared a common haplotype. Also, specimens that previously gave novel (unknown) PCR–RFLP haplotypes are now identifiable, associating with reasonable bootstrap support to particular profile taxa and providing clues as to the species gaps that need filling.

5. DISCUSSION

(a) Comparative utility of DNA barcodes

In contrast to the other molecular diagnostic methods referred to here, DNA barcoding has some significant advantages: (i) it provides a more accurate and robust approach to diagnosis by using all of the targeted genetic data. Species-specific PCR and PCR–RFLP diagnoses, on the other hand, utilize small windows of the data at priming or restriction sites, ignoring most of the genetic information. (ii) incorporating the range of intra-specific polymorphism by adding as many reference sequences as possible clearly enhances the robustness of any key and assignment of species identification. In contrast, our PCR–RFLP procedures actively avoid using restriction enzymes that detected sub-specific polymorphism because of the ambiguity in interpreting it as species- or population-level variation. (iii) using a tree-based approach enables all the data to be observed at a glance; this is very cumbersome to manage with PCR–RFLP. (iv) the NJ analysis also provides quantitative data with sequence divergences and bootstrap values that give a measure of confidence in the identifications. In general terms, our conclusions about the relative benefits of sequence ‘tree’-based species identification methods, concur with those of other similar molecular identification keys for thrips (Brunner et al. 2002), nematodes (Floyd et al. 2002) and whales (Ross et al. 2003).

From a practical perspective, identifications can be achieved on a par with RFLP analyses, within a 24 h period from extraction to NJ analysis and for around the same cost, if not less. This is given in-house sequencing facilities and an appropriate reference dataset. However, even if the latter is not available, it is much easier to build this up over a relatively short period of time compared to the same for a diagnostic suite of restriction patterns. Assuming that the same care is taken over sequence quality and interpretation as it would be for quality and interpretation of electrophoretic gels, DNA barcodes provide a robust alternative to PCR–RFLP. With the exception of speed, this is also the case with species-specific primer methods. However, that appeared here to be the least predictable method which is not unexpected given the potential ambiguity associated with presence/absence of a PCR product and assumed ‘specificity’ of PCR primers. Brunner et al. (2002) decided not to dispose of the PCR–RFLP profiles for the benefit of other laboratories that do not have convenient DNA sequencing facilities. However, this is unlikely to present a barrier in the future. Even now, DNA sequencing technology is becoming more accessible through a number of dedicated local and offshore commercial sequencing facilities structured for competitive pricing and rapid turnarounds.

(b) Test cases

The two groups of taxa analysed here presented different challenges to using cox1 barcodes for species identification. For the tussock moths there had been difficulty in placing previous RFLP ‘unknowns’ to species. This obstacle was overcome here to a certain extent with subsequent inclusion of a much broader taxonomic range of lepidopteran cox1 sequence data, available through BOLD and to a lesser extent in GenBank. That enabled several unknowns to be assigned to likely genus and species within the Noctuidae and Arctiidae. In retrospect it is not unrealistic to expect that other, non-lymantrid moth species were being intercepted. Females of other lepidoptera, besides gypsy moth (Leonard 1981), oviposit indiscriminately, including on inanimate objects during population outbreaks or when attracted to lights of human settlements. As arrival of these species can not be easily predicted the design of a more comprehensive PCR–RFLP method was not possible. In contrast, placing sequences of the unknowns within a broader taxonomic context was made possible by publically available lepidopteran cox1 data. This now provides a guide as to how best to target further collections and fill the reference species gaps.

Of the previously unknown ‘tussock moth’ species, four were identified as the fall web worm, Hyphantria cunea. This evidence of multiple entry events could elevate this species in terms of risk. In fact, around the same time in 2003, this species was found to have established a localized population in New Zealand. Fortunately, this was eradicated. An active surveillance campaign has operated since that time and interestingly, two more finds have been made recently. Fall webworm had never been found in New Zealand or Australia prior to 2003. It is native to North America and Mexico, but since establishing in Europe and parts of Asia in the 1940s and 1950s, it has become a significant pest of trees in these continents. The identification here of immature life stages on imported used vehicles is evidence that a pathway of entry exists. It also highlights that species other than gypsy moth enter New Zealand via the imported used vehicle pathway more frequently than was originally thought.

In contrast to the tussock moths, the species of immature fruit flies entering and the pathways involved are more predictable due to their close host association. Consequently it has been easier to target species that should be included in an appropriately comprehensive cox1 reference dataset. This was reflected here with considerable confidence in the dataset for making identifications within the Tephritidae. The challenge for barcoding instead was recent evolutionary divergences, at the level of species complexes where blurry species boundaries exist and undermine confidence in identifications. In spite of this, for some cases it was an improvement over the PCR–RFLP method. For example, it was possible to place specimens previously identified as ‘B. dorsalis complex’ with reasonable bootstrap support and minimal sequence divergence (see figure 2 in the Electronic Appendix) to a likely species, B. caryae.

Distinguishing recently diverged taxa is no less an issue for morphologically-based identifications. The reasons for this are varied, but not least is the status of
the alpha taxonomy. Confusing species limits may exist through species being ‘oversplit’ or ‘overlumped’ and the phenotypic boundaries not adequately reflecting the speciation process (Funk & Omland 2003). Nevertheless precise species identification remains very important for biosecurity. Firstly, some species within these complexes are more aggressive in their ability to invade and reach pest status than others. \textit{B. dorsalis} sensu stricto and \textit{B. carambolae} are very successful invaders, but they are difficult to distinguish from, and have overlapping host and geographic ranges with \textit{B. verbascifoliae}, which is not a recognized pest. In reality many fruit fly species, especially tropical ones, are unlikely to be serious pests \emph{per se} in New Zealand, but accurate diagnosis of regulated pests is still necessary to avoid trade restrictions of significant impact. Secondly, some morphologically indistinct regulated species, such as \textit{B. philippinensis} and \textit{B. papayae}, have different host and geographic ranges, which is important information for assessing the specific risk and pathway involved. In that situation, extension of the barcode system to include a second (or more) gene region for resolving what is not possible with the \textit{cox1} Folmer region is justified. This might involve a nested approach to include regions such as \textit{3′ cox1} or the hypervariable 16S rDNA that may evolve at a slightly faster rate. It also warrants considerable effort to include specimens from across the host and geographic range.

\textbf{(c) Barcodes for biosecurity in the future}

Access to ‘historical’ datasets and the same specimens has provided a valuable opportunity to comparatively test the appropriateness and power of the barcoding approach for identifying unknown organisms. The test cases here support the view that \textit{cox1} barcodes offer the best opportunity to date to form the foundation of a flexible and accurate identification system for invasive insect species. Given the potential universality of the application, this approach promises to address the standardization and efficiency needs that are severely lacking at the moment for the international biosecurity community. The approach is even more appealing with the ease of technical transfer between laboratories, enabling consistency of the process from PCR to the alignment of homologous sequence data and sharing of sequence data from diverse and unrelated sources. It lends itself well to automation and as a basis for development of the next generation of diagnostic tools such as micro-array technology (Kochzius \textit{et al.} 2004). DNA barcoding may also contribute to digitized collections more easily than that proposed for morphologically based identifications (Gaston \& O’Neil 2004).

Nevertheless there are a number of issues that will need to be taken into account. There continues to be debate as to the reliability of a \textit{cox1} barcode species identifier given issues of diversity related to mtDNA phylogeography and the rules determining ‘species’ limits (Moritz \& Cicero 2004). Suggesting that just one gene can supply the diagnostic needs for all would be rather naive. Even with our data set there are inconsistencies. For example with the fruit flies \textit{cox1} does not confidently discriminate some of the species within the \textit{B. dorsalis} complex, for which an additional gene region may be appropriate. However it clearly separates races within other species. This implies issues with the alpha taxonomy. It also highlights the fundamental need for continued participation of taxonomists to provide accurately identified reference material and to describe new species potentially discovered during the process.

Employing DNA sequences from known adult specimens to identify their morphologically indistinct immature life stages, as has been considered here, illustrates the power of molecular data to complement (and enhance) the morphological approach to insect diagnoses. Using barcode data in this context, i.e. matching sequences within taxa, should not suffer from the same potential for misidentification as in other diagnostic situations where incongruence with \textit{a priori} predictions based on morphology requires a more comprehensive taxonomic approach (Paquin \& Hedin 2004). However as with other molecular diagnostic methods DNA barcode-based identifications will clearly depend heavily on the availability of appropriate reference taxa to avoid problems associated with inadequate taxon sampling (Moritz \& Cicero 2004). Despite this, given the largely unique nature of \textit{cox1} sequences at the species level demonstrated here (outside of species complexes) and in other studies (e.g. Hebert \textit{et al.} 2003\textit{a,b}; Hogg \& Hebert 2004; Whitman \textit{et al.} 2004; Ball \textit{et al.} 2005; Janzen \textit{et al.} 2005; Shander \& Willassen 2005), missing taxa are likely to lead to non-identification, not misidentification. Non-identification will be recognized by the ‘unknown’ appearing on a relatively long branch with low bootstrap support. Misidentification on the other hand could result from close proximity to a congener and absence of a conspecific reference sequence. This is most likely to occur for recently diverged taxa, and is no less an issue for any other diagnostic method. Robust diagnostic procedures should however be able to pre-empt this in part or completely by making every effort to understand reference taxa from a ‘whole organism’ point of view, i.e. using common sense to incorporate relevant biological aspects into identification of unknowns.

Accuracy of identifications is also dependent on reliability of the simple sequence-similarity approach. In this case using the NJ model the analysis is not constrained by potential rate variation and different base composition amongst taxa as are phylogenetic treatments of the data such as the use of maximum likelihood. In fact the need for rapid identifications in biosecurity would preclude the use of a phylogenetic analysis with the computation of such large datasets taking several days. Conversely, occasional phylogenetic treatment of the data could provide a useful strategy for determining the addition of species when taxon gaps are inferred by anomalous placings and long branches. Indeed, the most important improvement of the barcode method over all other diagnostic methods is the ability to continually add in more species. This introduces a highly desirable, more anticipatory diagnostic approach that will enable the international IAS diagnostic community to better cope with changing and local species priorities, to capitalize on

\textit{Phil. Trans. R. Soc. B} (2005)
the efforts of others and to standardize technologies. There is, nevertheless, unlikely to be a ‘one size fits all’ molecular diagnostic approach to biosecurity which benefits from the use of tools on a case-by-case basis. For example DNA barcoding would not replace routine species-specific portable tests such as the serological Lateral Flow Devices (LFDs) used in the field for confirming plant pathogen identification (Hughes et al. 2005). However inclusion of DNA barcoding in the molecular diagnostics ‘toolbox’ does offer a level of transparency for species identification across countries, and sectors within countries, that is not possible with the current uncoordinated adoption of numerous different diagnostic methods.

Aspects that need to be seriously considered towards adopting this on a global scale will be the ability to obtain validated specimens for building robust reference profiles, an agreed framework for the sharing and quality control of sequence data, confidence in the current user-friendly bioinformatic analyses and rules or standardized criteria for interpreting barcode results. Importantly, the rigorous assessment of these will now be possible through the Barcode of Life initiative (http://www.barcodinglife.org/), assisted by the Consortium for the Barcode of Life (http://www.barcoding.si.edu/). The significant international momentum gathering for this programme will facilitate the emergence of a globally collaborative and less fragmented approach to molecular diagnostics and is entirely appropriate for international biosecurity.

We thank the organizers of the inaugural meeting of the Consortium for the Barcoding of Life (London 2005) for inviting presentation of this work. We thank Robyn Cowan, Ruth Frampton and Barney Stephenson for constructive reviews of the manuscript, Lalitha Karunaratne for technical assistance with sample preparation and DNA sequencing, the many collaborators that have generously made time to provide us with specimens and the University of Guelph for BOLD DNA sequences. This work was supported by the Tertiary Education Commission of New Zealand through the Centre of Research Excellence Fund.

REFERENCES


The supplementary Electronic Appendix is available at [http://dx.doi.org/10.1098/rstb.2005.1713](http://dx.doi.org/10.1098/rstb.2005.1713) or via [http://www.journals.royalsoc.ac.uk](http://www.journals.royalsoc.ac.uk).
Errata


Brain connectivity at different time-scales measured with EEG
T. Koenig, D. Studer, D. Hubl, L. Melie and W. K. Strik

On page 1017, the lower EEG topography and time-course were missing. The corrected figure and its caption appear below:

Figure 1. Left are two EEG topographies, seen from above, nose up. In the middle a (constructed) time-course over 2 s is shown for the two topographies. The two topographies and time-courses overlap in space and time. On the right side, the resulting EEG is displayed. The resulting EEG is a mixture of both activities.


DNA barcodes for biosecurity: invasive species identification
K. F. Armstrong and S. L. Ball

On page 1819, the final sentence of the paragraph ending four lines from the bottom of the right-hand column was incorrect, and should read as follows:

For example, it was possible to place specimens previously identified as ‘B. dorsalis complex’ with reasonable bootstrap support and minimal sequence divergence (see figure 2 in the Electronic Appendix) to a likely species, *B. caryae*. 
DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar

M. Alex Smith, Brian L Fisher and Paul D.N Hebert

Phil. Trans. R. Soc. B 2005 360, 1825-1834

Supplementary data

"Data Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2009/02/12/360.1462.1825.DC1.html

References

This article cites 55 articles, 14 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1825.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1825.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar

M. Alex Smith1,*,†, Brian L. Fisher2,*,† and Paul D. N. Hebert1

1Barcode of Life Initiative, Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, ON, Canada N1G 2W1
2Department of Entomology, California Academy of Sciences, 875 Howard Street, San Francisco, CA 94103, USA

The role of DNA barcoding as a tool to accelerate the inventory and analysis of diversity for hyperdiverse arthropods is tested using ants in Madagascar. We demonstrate how DNA barcoding helps address the failure of current inventory methods to rapidly respond to pressing biodiversity needs, specifically in the assessment of richness and turnover across landscapes with hyperdiverse taxa. In a comparison of inventories at four localities in northern Madagascar, patterns of richness were not significantly different when richness was determined using morphological taxonomy (morphospecies) or sequence divergence thresholds (Molecular Operational Taxonomic Unit(s); MOTU). However, sequence-based methods tended to yield greater richness and significantly lower indices of similarity than morphological taxonomy. MOTU determined using our molecular technique were a remarkably local phenomenon—indicative of highly restricted dispersal and/or long-term isolation. In cases where molecular and morphological methods differed in their assignment of individuals to categories, the morphological estimate was always more conservative than the molecular estimate. In those cases where morphospecies descriptions collapsed distinct molecular groups, sequence divergences of 16% (on average) were contained within the same morphospecies. Such high divergences highlight taxa for further detailed genetic, morphological, life history, and behavioral studies.

Keywords: cox1; CO1; Madagascar; collaborative taxonomy; DNA barcode; biodiversity

1. INTRODUCTION

The increasing loss of biodiversity presents a daunting challenge to taxonomists and requires the discovery and analysis of biodiversity at a greatly accelerated pace. If we are really serious about ‘zero biodiversity loss’ in Madagascar and elsewhere, then conservation planning needs to be based more fundamentally on biodiversity data, and this requires taxonomic knowledge (Brooks et al. 2004a). However, if nothing is done to change the slow pace of current taxonomic efforts and practice, it will take centuries to complete even a preliminary ‘Encyclopedia of life’ on Earth (Wilson 2003). It is clear that if systematics is going to play a practical role in directing the preservation and development of natural systems (Gotelli 2004), changes need to occur throughout the entire taxonomic process, from collecting to description, from publication to dissemination, and from public outreach to advocacy.

In this paper, we show how DNA barcoding (using cytochrome oxidase 1 (cox1 or CO1)—Hebert et al. 2003a,b), enables taxonomic data on hyperdiverse arthropods to be gathered, analysed, and synthesized into useful products in a timeframe that meets the challenge presented by the rate of biodiversity loss. We test a model for accelerating the taxonomic process with the aims of providing the necessary data for effective taxonomy, and—most importantly—the tools for making data accessible and applicable to the conservation agenda. The model is tested on a key taxonomic group, ants, and in an especially threatened area, Madagascar. We describe how cox1 DNA barcoding enables rapid identification of Molecular Operational Taxonomic Units (MOTU—Floyd et al. 2002; Blaxter 2004) for the assessment of richness and turnover across landscapes.

Madagascar has been identified as one of the world’s outstanding biodiversity hotspots, harbouring a unique and threatened biota, whose composition and origins are linked to the breakup of Gondwana (Battistini & Richard-Vindard 1972; Jolly et al. 1984; Storey et al. 1995; Lourenço 1996; Goodman & Patterson 1997; Goodman & Benstead 2003). As in many island environments (Gillespie & Roderick 2002), Madagascar’s indigenous terrestrial arthropods are in severe danger of extinction due to habitat deterioration and invasion of exotic species. Since humans colonized Madagascar approximately 1500–2000 years ago (Burney 1997), it is estimated that as much as 80% of Madagascar’s original habitat has been destroyed (Sussman et al. 1996). Much of the island is now species-poor secondary grassland, which is annually burned and highly eroded.

* Authors for correspondence (salex@uoguelph.ca, bfisher@calacademny.org).
† MAS and BLF have contributed in equal part to this paper and each author reserves the right to be considered first author.

One contribution of 18 to a Theme Issue ‘DNA barcoding of life’.
Never has there been a more supportive political environment in which to address these threats in Madagascar. Over the next five years, the Malagasy government plans to more than triple the number of protected areas and is committed to sustainable conservation planning. To accomplish these goals, areas of conservation importance must be determined. One major obstacle in the identification of areas for protection in Madagascar is incomplete knowledge of the island’s patterns of species richness, turnover, and endemism (Schatz 2002). It is unclear which of the remaining patches of natural vegetation should be of highest priority for conservation. Existing data are often at an inappropriate spatial scale for conservation implementation, not standardized across sites, and focused on vertebrates—which represent only a small proportion of the biota.

There is strong and growing evidence that the investment in collection and compilation of species data yields great value in conservation planning (Brooks et al. 2004a,b). Species data are a precondition of good, efficient conservation (Ferrier et al. 2004). Current knowledge of species distributions, however, is often limited to select vertebrate taxa. Vertebrates, however, are frequently inadequate indicators of biodiversity as a whole; they fail to represent the bulk of diversity, especially the fine-scale patterns of diversity shown by arthropods (Fisher 2000; Ferrier et al. 2004). Arthropods, such as ants, often exhibit high rates of spatial turnover (replacement of species) and therefore provide the essential fine-scale maps for assessing biodiversity at a scale at which conservation decisions are typically made on the ground.

The ant fauna of Madagascar is currently estimated to include approximately 1000 species, of which 96% are endemic (Fisher 1996b; Fisher & Girman 2000). An estimated 75% of the ant fauna in Madagascar, however, remain undescribed (Fisher 1996a). For example, of the 71 species of the genus Strumigenys described in a recent revision, 70 were endemic and newly described (Fisher 2000). Although ants dominate the biomass of most terrestrial communities (Davidson 1997; Davidson et al. 2003), critical pollinators and seed dispersers (Beattie 1985; Ness et al. 2004), and are critical to nutrient cycling and ecosystem function (Thorp 1949; Major 1983; Andersen 1993; Andersen 1997; Diehl et al. 2004), there is a global lack of studies of ant diversity or community structure. This may be largely because of the difficulty of species-level identifications (Bolton 2003).

A DNA-based system of species identification using a single gene (cytochrome oxidase 1) was proposed by Hebert et al. (2003a,b) who coined the term DNA barcoding. Since then, the utility of DNA barcodes for species identification has been successfully demonstrated with several taxonomic groups (Hebert et al. 2004b; Powers 2004; Hajibabaei et al. 2005; Lambert et al. 2005; Ward et al. 2005). The potential for such a system is evident to many who study biodiversity, especially, in smaller (Floyd et al. 2002; Blaxter 2004; Blaxter et al. 2004), understudied, or hyperdiverse (Hebert et al. 2004a) groups or in areas where the estimates of diversity lag well behind what is actually there.

There have been great improvements in overcoming other obstacles to including arthropods in biodiversity assessment. Efficient methods exist for their collection (Longino & Colwell 1997; Fisher 1999; Longino et al. 2002) and processing (Fisher 2005). The enormous amount of material collected and processed, however, presents new challenges: how to accelerate their identification and description? In this study, we test if rapid DNA barcoding can be used to provide a surrogate to species diversity patterns. In our analysis we test whether a diversity estimate based on a DNA barcode MOTU is significantly different from estimates based on traditional morphological taxonomy using an understudied taxa (ants) from a part of the world where established taxonomic frameworks are only now emerging (Madagascar). We demonstrate that DNA barcode functional units are an effective surrogate for traditional morphological species and discover the same relative patterns of diversity within and between collection sites. This is important as it allows the rapid and scalable identification of diversity from widespread, numerous and thorough inventories. A sequence-based assessment of diversity can scale to many more sites and specimens in the same amount of time possible for traditional analyses. MOTU will not always be coincident with species (determined using whatever concept one chooses), but rather are categories whose membership is a hypothesis open to re-testing. Thus, this method is not one of strictly DNA-based taxonomy. Our goal is to test whether MOTU perform as an estimate, or surrogate, of species richness patterns derived from morphological based determination of species (morphospecies).

We compare richness and turnover of MOTU from sequence data with species richness and turnover patterns from morphological determinations for ant specimens from four sites in northern Madagascar. Using an operational, tree-based approach to the identification of MOTU, we identified clusters of specimens beneath a threshold of similarity (Hebert et al. 2003a). The identification of MOTU using a sequence divergence threshold represents just one possible means to determine MOTU (for other examples see Floyd et al. 2002; Sites & Marshall 2003; DeSalle & Amato 2004) and is used here only as an example.

2. METHODS
(a) Field sites
We surveyed ants at four localities in November and December 2004 in northeastern Madagascar in the Province of Antsiranana:

(i) Antsahabe: Forêt Antsahabe, 11.4 km 275° Daraina, 550 m, 13°13.7'S, 49°33.4'E 15–18.xi.2004, tropical dry forest.


(iii) Marojejy: Parc National de Marojejy, 28.0 km 38° Andapa, 450 m, 14°26.2'S, 49°46.5'E 23–25.xi.2004, rainforest.

Phil. Trans. R. Soc. B (2005)
At each of the sites, we spent two days searching by hand for ants. This technique differs from standardized inventory protocols for collecting ants (Fisher 1999) but provides a manageable case study for testing the use of DNA barcoding for assessing patterns of richness and turnover. Specimens from every collection from each locality were used for cox1 sequencing and morphological study. All specimen data for material examined in this study are available on AntWeb (www.antweb.org). Material was deposited at California Academy of Sciences, San Francisco, USA (CAS).

(b) Genetic analysis

Specimens were identified to genera immediately upon collection by paratransonomists in Madagascar (Wheeler 1995; Basset et al. 2000) and preserved in 95% ethanol. Upon return to California, specimens were loaded into ScrewTop TrakMates® boxes (Matrix Technologies) and shipped to the University of Guelph, Canada. Here, DNA was extracted from tissues rich in mitochondria (e.g. legs), employing primers with high universality, and amplifying a relatively long PCR product (> 600 base pairs; bp). Total genomic DNA extracts were prepared from small pieces (≤1 mm) of leg using the Nucleospin® 96 Tissue kit (Macherey-Nagel Duren, Germany), following the manufacturer’s protocols. Extracts were resuspended in 30 μl of dH2O, and a 650-base bp region near the 5′ terminus of the cox1 gene was amplified following standard protocol (Hebert et al. 2003a). Briefly, full length sequences were amplified using primers (LF1—ATTCAACCAATCATAAGAAATTTGG and LR1—TGATTTTTGGACATCTCAGGTTTA) (Hebert et al. 2004a)). In cases where a 650 bp product was not successfully generated, internal primer pairs (LF1-ANTMR1D–ATGMCWGGNKYMGG-WACWGWTG) and (MLF1—GCTTTCCACGAAATTTGAAATTTA) (Hajibabaei et al. 2003)—LR) were employed to generate shorter overlapping sequences that allowed the combination of a composite sequence (contig). PCR reactions were carried out in 96-well plates in 12.5 μl reaction volumes containing: 2.5 mM MgCl2, 5 pmol of each primer, 20 μM dNTPs, 10 mM Tris–HCl (pH 8.3), 50 mM KCl, 10–20 ng (1–2 μl) of genomic DNA, and 1 unit of TaqDNA polymerase using a thermocycling profile of one cycle of 2 min at 94°C, five cycles of 40 s at 94°C, 40 s at 45°C, and 1 min at 72°C, followed by 36 cycles of 40 sec at 94°C, 40 sec at 51°C, and 1 min at 72°C, with a final step of 5 min at 72°C. PCR products were easily produced and aligned as no double bands. These observations support the conclusion that the sequences we analysed were mitochondrial DNA and not nuclear pseudogenes (Bensasson et al. 2001).

PCR products were generated from 268 of 280 specimens (95.7%) from 28 genera from four collection sites in northern Madagascar (Electronic Appendix, figure 5). Of these 268, 224 were generated using primers LF_LR (84.6%), 26 were contigs generated using two overlapping primer sites (9.7%), and 18 were shorter sequences generated using primers MLF_LR (6.7%). The average congeneric pairwise divergence was 8.51%. Sequences were heavily AT biased, as expected in insect mtDNA (Crozier & Crozier 1993) (table 1). The observed mutation patterns in the sequence data are consistent with a neutral model of molecular evolution (Tajima’s D: 0.05643, p > 0.10). Each individual was assigned to a morphospecies using morphological characteristics and to a MOTU using sequence thresholds of 2 and 3%.

There was no significant difference between richness (number of morphospecies or MOTU) using the threshold approach or morphological taxonomy (ANOVA, F = 0.216, df = 2, p = 0.811) (figure 1). There was a significant difference in turnover between sites using the different methodologies to assign individuals to morphospecies or MOTU (Shared
Table 1. Sequence statistics for the specimens listed in the Electronic Appendix, table 4.

<table>
<thead>
<tr>
<th>domain</th>
<th>identical pairs</th>
<th>transitional pairs</th>
<th>transversional pairs</th>
<th>average R=s/sv</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg</td>
<td>473</td>
<td>62</td>
<td>71</td>
<td>0.9</td>
<td>39</td>
<td>18.7</td>
<td>30.4</td>
<td>11.9</td>
</tr>
<tr>
<td>1st position</td>
<td>168</td>
<td>18</td>
<td>16</td>
<td>1.2</td>
<td>28.7</td>
<td>17.5</td>
<td>34.6</td>
<td>19.2</td>
</tr>
<tr>
<td>2nd position</td>
<td>193</td>
<td>5</td>
<td>4</td>
<td>1.2</td>
<td>44.8</td>
<td>24</td>
<td>17.1</td>
<td>14.1</td>
</tr>
<tr>
<td>3rd position</td>
<td>113</td>
<td>38</td>
<td>51</td>
<td>0.8</td>
<td>43.5</td>
<td>14.5</td>
<td>39.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

species or MOTUs—ANOVA $F = 4.420$, $df = 2$, $p = 0.033$, Jaccard Index—$F = 7.552$, $df = 2$, $p = 0.006$ (figure 2, table 2). Molecular similarity thresholds tended to emphasize the uniqueness of each site while morphological taxonomy tended to find more overlap between sites.

In general, the comparisons between 2% MOTU with morphology and 3% MOTU with morphology are the same (compare richness patterns in table 3, figure 1). Therefore, for brevity, only the results for the comparison between 3% MOTU and morphology are presented here. There was a strong correlation between 2% ($\sim 13$ substitutions/657 bp) and 3% ($\sim 19$ substitutions/657 bp) MOTU with the species estimates using morphological taxonomy. For example, in Camponotus, individuals were allocated nearly identically using morphology or a 3% threshold (figure 3). Exceptions occurred where deep sequence divergences were apparent between individuals identified as the same morphospecies but were collected at different sites (morphospecies m07, m09), or the same collection sites (m03). There were assignment differences between the 3% MOTU with morphology that are detailed in table 3 and table 4 in the Electronic Appendix. Generally, morphological taxonomy 'lumped' MOTU separated using the threshold approach. The average molecular divergence of morphospecies which contained multiple MOTU was 16.27% ($n = 113$ comparisons, average = 16.27, s.e. = 0.51). On average, there were 2.3 cases per genera where morphology failed to recognize MOTU erected with a 3% threshold (50 times total). Pairwise sequence divergences for only those 50 cases where morphospecies contained multiple 3% MOTU are shown in a frequency histogram in figure 4.

4. DISCUSSION

(a) Biodiversity assessment

Our analysis is unique in that it has compared diversity measures at four sites in Madagascar using both the morphologically defined species units and MOTU based on two different threshold values for DNA barcode sequence divergence. Patterns of richness and turnover of MOTU and morphospecies were not significantly different. The take-home message is not that the values are the same (although for the most case they are remarkably similar), but rather that the patterns of richness within sites and turnover between sites were so similar. Thus, richness and turnover assessments determined using DNA barcode variability, accrued with less than three weeks of preparatory analysis, provided an effective surrogate for species determined through time intensive detailed morphological analyses. It is important to recognize that our analyses do not use the DNA barcode to define an ontological species concept, but rather, to recognize classes of diversity. We test the epistemological hypothesis that the barcode MOTU can be used as a surrogate for the identification of diversity within and between collection localities. Species are spatio-temporally bounded individuals, not a class of items (Baum 1998). However, our recognition of these individuals defined ontologically must frequently come after using epistemological methodology, including category membership. Different epistemological methods of recognizing diversity frequently arrive at different answers (e.g. the Ensatina eschsholtzii complex salamanders of California may include from 1 to 11 different species—Frost & Hillis 1990; Graybeal 1995; Highton 1998; Wake & Schneider 1998) for many inventories of hyperdiverse taxa, the lack of taxonomic expertise, or concentration of expertise to only a few individuals, inhibits morphological assessment of large numbers of specimens. Thus, an ambitious arthropod inventory can quickly overwhelm taxonomists with too many specimens, and thus are often unable to provide fine scale data for conservation for many groups. As an example, the NSF-funded arthropod inventory of Madagascar has shipped over a third of a million specimens to over 100 participating taxonomic collaborators (Fisher 2005). Major taxonomic products from these inventories, that will take decades to produce, represent only a fraction of the diversity collected, and provide no short-term return of biodiversity data to Madagascar. Inventories are essential for documenting global diversity and generating necessary material for taxonomic study. However, for inventories to be relevant in the short term, the...
inventory process must reduce the bottlenecks in returning relevant data for conservation.

Although in this study morphological determination of species only took a matter of months, the morphological protocol would not easily scale to include additional inventories across a wider geographic region. With each new site and each new species added, the documentation and analyses of morphological variation becomes increasingly more difficult. A sequence-based approach to the analysis of diversity, backed by a database of single gene barcodes, allows the exploration of diversity to scale to a rate that is not currently feasible using morphology alone. The DNA barcode provides a surrogate method for identifying units of diversity—a surrogate that will later serve as an additional character set for taxonomic assessment. If taxonomy is to provide a necessary tool to ecology and conservation science in hotspots it must be done at a much faster rate than in the past—especially with small, hyperdiverse or as yet undescribed fauna.

Our analysis provides an example of the speed and complementarity with which DNA barcoding can work in concert with a more conventional morphological taxonomic protocol—neither competing nor replacing (Hebert & Gregory 2005). This result has important ramifications for the argument to include hyperdiverse taxa in conservation planning. Hyperdiverse groups can be included, knowing that in the short term, richness and turnover assessments can be provided and are valuable surrogates of diversity. Most importantly, with each inventory, additional data will be accrued to assist in taxonomic analysis.

Table 2. Jaccard indices of similarity of four collection sites in Madagascar.

<table>
<thead>
<tr>
<th></th>
<th>2%</th>
<th>3%</th>
<th>morphospecies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambato–Marojejy</td>
<td>0</td>
<td>0</td>
<td>0.017</td>
</tr>
<tr>
<td>Ambato–Antsahabe</td>
<td>0.053</td>
<td>0.083</td>
<td>0.158</td>
</tr>
<tr>
<td>Ambato–Binara</td>
<td>0.011</td>
<td>0.036</td>
<td>0.106</td>
</tr>
<tr>
<td>Marojejy–Antsahabe</td>
<td>0</td>
<td>0.021</td>
<td>0.075</td>
</tr>
<tr>
<td>Marojejy–Binara</td>
<td>0.036</td>
<td>0.098</td>
<td>0.163</td>
</tr>
<tr>
<td>Antsahabe–Binara</td>
<td>0.089</td>
<td>0.111</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Table 3. Comparison of the number of morphospecies and MOTU contained within each of the 28 genera examined in this analysis.

<table>
<thead>
<tr>
<th>genera</th>
<th>morphospecies</th>
<th>3% MOTU</th>
<th>2% MOTU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anochetus</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Aphaenogaster</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Camponotus</td>
<td>10</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Cataulcus</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Cerapachys</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Crematogaster</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Eutermorium</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Hypoponera</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Leptogenys</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Leptothorax</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Melissotarsus</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Monomorium</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Myriini</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Odontomachus</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pachycondyla</td>
<td>5</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Paratrechina</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Phidole</td>
<td>9</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Plagiolepis</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Pheidole</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Pionopelta</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pyramica</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Simopone</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Strumigenys</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Tapinoma</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Technomyrmex</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Terataner</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Tetramorium</td>
<td>9</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Tetraponera</td>
<td>7</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>totals</td>
<td>90</td>
<td>117</td>
<td>126</td>
</tr>
</tbody>
</table>
Figure 3. *Camponotus* taxon NJ tree of K2P distance. Coloured bars with names indicate the samples described as a species using morphological taxonomy. Red indicates perfect correlation between 3% MOTU and morphological estimate of species identity. Blue indicates cases where a morphospecies contained multiple 3% MOTU. Accession numbers and collection sites are indicated.
Figure 4. Frequency histogram of the genetic variation (uncorrected p-dist) for only those cases where morphospecies contains multiple 3% MOTU. \(n=113\) comparisons, average = 0.1627, s.e. = 0.51.

(b) What do these results suggest about tropical ant diversity?
Discrepancies between estimates of morphospecies and MOTU identified with \(\text{cox1}\) DNA barcodes occur in restricted geographic areas and usually involve sharp ecological and elevational differentiation (Fisher 1999). For instance, within \(\text{Anochezus madagascarenisis}\), we see subtle phylogeographic structure within the DNA barcode region. Individual \(\text{A. madagascarenisis}\) were collected from the Binara (800 m elevation) on the east coast and Ambato (150 m elevation) on the west coast. These two localities are separated by the highest mountain in Madagascar (2876 m). Individuals from these populations are separated by, on average 1.5% sequence divergence. Are these populations operating as separate species? Are these populations members of the same species but highly divergent? Our data alone cannot answer this question. However, of critical import, our data have identified the surprising level of within-species divergence and lays bare these differences to further study. A standard arthropod molecular clock for \(\text{cox1}\) is 1.2–1.5% per million years (Caccone & Sbordoni 2001; Farrell 2001; Dick et al. 2004). Hymenopterans exhibit rate acceleration for this gene (Hebert et al. 2003b), and therefore average estimates should be interpreted with caution. However, accelerating these rates suggests that populations have been isolated for several hundred thousand years. The opportunity now exists to employ a suite of approaches (behavioural observations, tests of interbreeding, and phylogeographic resolution of more quickly evolving genetic markers) to test species membership. We recognize that a molecular approach to biodiversity estimation may underestimate diversity when collections include quickly evolving species-pairs (Hebert et al. 2003a). Using the DNA barcode based MOTU approach, these pairs would be coded as one MOTU—as we have seen in \(\text{A. madagascarenisis}\). Indeed, ecological demarcation can make significant contributions to species formation and eventual morphological divergence. Distinct and extremely local divergence has been seen in the tropics in Costa Rican salamanders (Garcia-Paris et al. 2000), and the Amazonian poison-dart frog (Loughard et al. 1999). As in these examples, the allopatric \(\text{Anochezus}\) MOTU documented here may represent cryptic species that owe their formation to orogeny: isolative and habitat effects of the development of mountain ranges. Mountains could produce unique lineages and species through the production of alternative refugial habitats through climatic changes (Hewitt 2004). Palynological studies throughout Madagascar have indicated a dynamic environmental history in response to periods of glaciation and interglaciation (Burney 1997; Gasse & Van Campo 1998). These climatic shifts have had a considerable impact on forest vegetation structure and therefore ant habitat.

Our sampling methodology in this preliminary study has resulted in divergent singleton specimens which were summarily coded as 2 or 3% MOTU in our analysis. However, it will only be by sampling multiple individuals from a provisional species, or MOTU, that interspecific variation will be properly assessed, allowing the hypothesis of species-level monophyly to be tested (Funk & Omland 2003—and references within are a thorough review). This is a valid concern in an analysis of MOTU from inventories of hyperdiverse groups such as ants, which often include many taxa known only from single individuals (Fisher 1999; Longino et al. 2002). With additional inventories in the future, many of these singletons will be represented in collections by more specimens. Caution must be expressed regarding any paraphyletic relationship discovered with a MOTU represented by one specimen (Funk & Omland 2003).

In cases where the 2–3% MOTU and the morphological estimation of a species differed, the molecular variants were either isolated geographically or were specimens contained by a species epithet whose actual status is, as yet, undetermined. This discrepancy does not compromise the use of a DNA barcode for species identification; just the opposite. In this case, DNA barcoding has identified regional lineages that can be further tested for species status or phylogeographic structure.

(c) Future
DNA barcoding proved an effective surrogate for morphospecies diversity patterns across localities in northern Madagascar. This study demonstrates how inventories of hyperdiverse taxa such as ants can provide rapid analysis of diversity for conservation assessment. Sequence data generated during the inventory process will also provide an alternative set of characters to assist in inferring species boundaries during future taxonomic studies. Thus, the application of DNA tools during diversity assessment will facilitate and complement taxonomic study. The combination of DNA sequencing data coupled with inventory and traditional taxonomy is a model that can be applied across disciplines and will allow analytical needs to scale to the enormity of the biodiversity crisis (DeSalle & Amato 2004). It will help in the identification and conservation of the evolutionary processes that generate and preserve biodiversity.

Of course, species boundaries are too complex and fuzzy (Green 1996) to be only described by DNA barcode sequence divergence. Rather a suite of informative characters (multi-gene analyses, behavioural studies and taxonomic expertise) is required...
for the delineation of a species. However, we show here that DNA barcodes allow the rapid identification of functional units of diversity that can scale to the magnitude of hyperdiverse arthropods at a timeframe needed by conservation groups responding to habitat destruction and degradation. Ant diversity, measured using MOTU in collaboration with taxonomists should provide the essential fine-scale maps for assessing biodiversity at a scale at which conservation decisions are made.

Little time remains for the documentation of global biodiversity. Taxonomists, equipped with modern tools and collaborations, have a chance to move systematics to the forefront of conservation and the public's attention. With increased taxonomic output and improved public access and visibility, public support for the discovery of life on this planet should follow.

The authors wish to thank the Consortium for Barcoding Life for organizing, the British Natural History Museum for hosting, and the Alfred P. Sloan Foundation for funding the First International Conference on Barcoding Life. Funding for this research was provided by the National Science Foundation awards to BLF and CE Griswold (DEB-0072713) and to BLF and PS Ward (DEB-0344731); the Gordon and Betty Moore Foundation, Natural Sciences and Engineering Research Council (Canada) and Canada Research Chairs program to PDNH, and a Fonds Nature et Technologies B3 postdoctoral fellowship (Quebec) to MAS. We are grateful to April Nobile for image creation, Dylan Burge for specimen preparation and Michele Esposito for specimen data entry. Thanks to Jeremy DeWard and Angela Holliss for sequencing assistance, and Sujeewan Ratnasingham and Rob Dooh for analytical assistance. Fieldwork that provided the basis for this work could not have been completed without the gracious support of the Malagasy people, and the Arthropod Inventory Team (Balsama Rajemison, Helian Ratsirarson, Jean Claude Malagasy people, and the Arthropod Inventory Team have been completed without the gracious support of the Malagasy people, and the Arthropod Inventory Team (Balsama Rajemison, Helian Ratsirarson, Jean Claude

REFERENCES

Andersen, A. N. 1997 Ants as indicators of restoration success at a uranium mine in tropical Australia. Restoration Ecol. 1, 156–167.


Crozier, R. H. & Crozier, Y. C. 1993 The mitochondrial genome of the honeybee Apis mellifera: complete sequence and genome organization. Genetics 133, 97–117.


The supplementary Electronic Appendix is available at http://dx.doi.org/10.1098/rstb.2005.1714 or via http://www.journals.royalsoc.ac.uk.
Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding

Daniel H Janzen, Mehrdad Hajibabaei, John M Burns, Winnie Hallwachs, Ed Remigio and Paul D.N Hebert

*Phil. Trans. R. Soc. B* 2005 **360**, 1835-1845

**Supplementary data**
"Data Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2009/02/12/360.1462.1835.DC1.html

**References**
This article cites 15 articles, 1 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1835.full.html#ref-list-1

**Article cited in:**
http://rstb.royalsocietypublishing.org/content/360/1462/1835.full.html#related-urls

**Email alerting service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](http://rstb.royalsocietypublishing.org/subscriptions)

To subscribe to *Phil. Trans. R. Soc. B* go to: [http://rstb.royalsocietypublishing.org/subscriptions](http://rstb.royalsocietypublishing.org/subscriptions)

This journal is © 2005 The Royal Society
Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding

Daniel H. Janzen¹*, Mehrdad Hajibabaei², John M. Burns³, Winnie Hallwachs¹, Ed Remigio² and Paul D. N. Hebert²

¹Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
²Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada N1G2W1
³Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0127, USA

By facilitating bioliteracy, DNA barcoding has the potential to improve the way the world relates to wild biodiversity. Here we describe the early stages of the use of cox1 barcoding to supplement and strengthen the taxonomic platform underpinning the inventory of thousands of sympatric species of caterpillars in tropical dry forest, cloud forest and rain forest in northwestern Costa Rica. The results show that barcoding a biologically complex biota unambiguously distinguishes among 97% of more than 1000 species of reared Lepidoptera. Those few species whose barcodes overlap are closely related and not confused with other species. Barcoding also has revealed a substantial number of cryptic species among morphologically defined species, associated sexes, and reinforced identification of species that are difficult to distinguish morphologically. For barcoding to achieve its full potential, (i) ability to rapidly and cheaply barcode older museum specimens is urgent, (ii) museums need to address the opportunity and responsibility for housing large numbers of barcode voucher specimens, (iii) substantial resources need be mustered to support the taxonomic side of the partnership with barcoding, and (iv) hand-held field-friendly barcorder must emerge as a mutualism with the taxasphere and the barcoding initiative, in a manner such that its use generates a resource base for the taxonomic process as well as a tool for the user.

**Keywords:** Costa Rica; tropical; Área de Conservación Guanacaste; Hesperiidae; Saturniidae; Sphingidae

1. INTRODUCTION

In 1978, D. H. Janzen and W. Hallwachs began the inventory of the entire caterpillar fauna (exclusive of leaf miners) and their parasitoids of Área de Conservación Guanacaste (ACG) in northwestern Costa Rica (Janzen 2000, 2003, 2004a; Burns & Janzen 2001; Janzen & Hallwachs 2005; Gauld & Janzen 2004; Hebert et al. 2004). Terrestrial ACG is 115 000 ha of dry forest, rain forest, cloud forest, and their intergrades from 0 to 2000 m (http://www.acguanacaste.ac.cr.; Janzen 2000). About 3200 species of caterpillars have now been inventoried (found, reared, photographed, identified, and placed on the project website at http://janzen.sas.upenn.edu.), with approximately 6400 species yet to inventory (as based on a 25-year inventory of adults by Janzen and Hallwachs). This inventory requires a massive ongoing and highly interactive taxonomic platform. It has been provided over five decades by more than 150 members of the taxasphere and their collections, field guides, revisionary papers, and species descriptions, beginning while the senior author was still in high school and visited lowland Mexico to collect butterflies. Interactive revisionary and species-level taxonomy of the inventoried species is the life of the project.

DNA barcoding for the express purpose of identifying species emerged in 2003 (Hebert et al. 2003; www.barcoding.si.edu) as a streamlined, economical, and assembly-line version of the long-established and more general use of DNA sequence information for phylogeny, phylogeography, and population demarcation. We immediately applied it to the taxonomic process underlying the ACG caterpillar inventory. We sought to provide an additional tool for species discovery and identification, as well as to serve as a pilot project for the application of DNA barcoding to complex and species-rich biotas. Byproducts are contributions to the Lepidoptera cytochrome oxidase subunit I (cox1) sequence libraries in BoLD and GenBank, stimulation of the eventual emergence of cheap, field-friendly identification barcoders for the world at large, and promotion of the concept of a low-charge-per-individual identification tollbooth that contributes to the financial maintenance of the taxasphere (Janzen 1993, 2004b).

2. THE CATERPILLAR INVENTORY PROCESS AND DNA BARCODING

Barcoding fits into the logistics of the ACG caterpillar inventory (methodology at http://janzen.sas.upenn.edu and the Janzen powerpoint deposited at the Consortium for the Barcodes of Life (CBOL) website www.barcoding.si.edu/Presentations.htm) as follows. A free-living caterpillar is found in the forest by one
of the project’s 19 resident Costa Rican parataxonomists (Janzen 2004a), brought to one of seven rearing barns scattered across the three primary ACG terrestrial ecosystems, and reared through to adult (or parasitoid) in a plastic bag suspended from a clothesline. Its collateral information is maintained as a single event-based record, with the record and the caterpillar assigned a unique alphanumeric voucher code (e.g. 95-SRNP-5116). On its first encounter(s) by the inventory, the caterpillar is photographed. Care of each individual continues until the newly eclosed adult is killed by freezing in a $-15$ to $-20^\circ$C non-defrosting freezer. Accumulated adults are removed from the freezer at one- to six-month intervals, their field identifications are corroborated and they are: (i) discarded, (ii) pinned, spread, and oven-dried at $50$–$60^\circ$C, or (iii) placed in 100% ethanol and refrozen or refrigerated. At one-to-six-month intervals, the pinned and dried specimens are hand-carried to the University of Pennsylvania (UP) under a formal export permit from the government of Costa Rica, having been collected under a formal research permit issued by the Ministerio de Recursos Naturales y Energía (MINAE). The latter permit explicitly authorizes the collection of specimens for DNA barcoding. At UP they are sorted for later deposition with participating taxonomists in their respective museums. The legs used for sequencing at the University of Guelph CBOL node are taken from these dry specimens. Likewise, the ethanol-preserved specimens are transported at room-temperature to the University of Pennsylvania and stored again in $-20^\circ$C freezers or refrigerators, and then donated to specific taxonomic researchers or the Ambrose Monell Collection for Molecular and Microbial Research in the American Museum of Natural History (http://research.amnh.org/amcc) for public scientific use. At the end of each year, the individual databases are pooled from the seven rearing barns, edited and data-checked, pooled with the master database, and posted on the project website. The project currently generates about 35 000 rearing records per year. At the end of 2004, it had logged about seven million caterpillar rearing days, for 264 370 event-based records.

This assembly-line inventory process provides a strong platform for barcoding because:

(i) many conspecific and individually vouched and databased specimens less than two decades old are museum-available from all ACG ecosystems;
(ii) the inventory voucher specimen is automatically available as the barcode voucher specimen;
(iii) the frozen and then oven-dried specimens have not been field-dried, relaxed at high humidity, and then re-dried when mounted, a treatment that is apparently quite destructive to DNA (occasional specimens are killed with cyanide, but this has had no apparent effect on ease of sequencing) (see Prendini et al. 2002);
(iv) each adult moth or butterfly (or parasitic wasp or fly) has three pairs of dry legs and one member of a pair (and yet another in the case of need) can be removed for sequencing;
(v) the specimens are already identified to some level by standard morphology-based or ecology-based taxonomic protocols before entering into the barcoding process;
(vi) when barcoding generates taxonomic questions, the inventory process is modified (as with morphology-based taxonomic processing) to generate more specimens of the taxon in question, albeit with lag times of six months to a year, owing to the intrinsically slow find–rear–eclose process;
(vii) all species being examined are either sympatric within ACG, or, if restricted to different ecosystems, are parapatric at the interdigitations of the ecosystems over distances of a few hundred metres;
(viii) the specimens being compared and identified morphologically are usually in excellent condition, unlike the worn specimens commonly collected as adults; and
(ix) because they are reared, it is often possible to know if a pair of specimens are sibs, and even to use the barcodes of sibs and parents to explore intra-population variation and confirm the accuracy of sequencing.

3. THE FIRST TRIAL

In March 2003, at the first Sloan Foundation-supported conference at the Banbury Centre, we realized that the ‘barcoding’ initiative (which was to become CBOL at the Smithsonian organizing conference in May 2004) had the potential to be a powerful new tool in the taxonomic toolkit. The ACG inventory sent eight pairs of morphologically similar congeneric skipper butterflies (Hesperiidae) to the Guelph CBOL node. They were found to be easily distinguishable by their $\text{cox1}$ sequences (termed COI sequences at that time). This prompted Janzen, Hallwachs, and Burns to invite the Guelph node to apply barcoding to an estimated seven undescribed, and morphologically very similar, species detected within ACG *Astraptes fulgerator* (Hesperiidae). Barcoding 484 individuals revealed a total of 10 more or less sympatric species in the complex (Hebert et al. 2004).

4. ROUTINE BARCODING IN THE INVENTORY

The clarity of results with *Astraptes fulgerator*, the challenge of applying a new identification tool to the mass of biodiversity information accumulated through nearly three decades of ACG inventory, and the willingness of the Guelph node to barcode tens of thousands of vouchered museum specimens for a few dollars each was irresistible.

(a) Mechanics of barcoding ACG inventory specimens

A dry inventory voucher specimen is selected for analysis, and a single leg broken off at its base with forceps. The forceps are tightly wiped with a portion of unsullied Chemwipe tissue between each use. The dry leg is dropped into a new 2 ml Eppendorf tube or into a tube in a 96-tube MATRIX Box (Matrix Technologies,
Hudson, New Hampshire), with a hand-written (India ink on acid-free bond paper) or laser-printed voucher code placed inside the tube, and couriered to the Guelph node. The museum specimen is flagged with a yellow ‘legs away for DNA’ pin tag, as is the voucher database record. The voucher specimen’s collateral information is uploaded from the inventory database to an Excel form prepared by the Guelph node, and accompanied by two images (upperside and underside), all of which are placed in the specimen’s record in the project databases at Barcode of Life Database (BoLD) at www.barcodinglife.com.

At the Guelph node, DNA is extracted from each leg and cox1 (‘COI’ in previous literature) is PCR amplified and sequenced. The cox1 sequence is placed in BoLD for processing, and later submission to GenBank, along with its collateral information. Residual DNA extracts are preserved in K80 freezers. Specimens that do not sequence well are variously re-sequenced and otherwise processed, depending on the question being asked (see Hajibabaei et al. 2005). The ACG inventory subsequently obtains the placement of this specimen relative to others by constructing a Neighbor Joining (NJ) phenogram (a ‘species identification phenogram’) by using the BoLD website (see examples below and Hebert et al. 2003; Hajibabaei et al. 2005). The NJ phenogram can have bootstrap values placed on it if relevant, and the specimen’s position can be labelled with voucher code, name, geographic location, higher taxon, and/or sequence length as the project wishes. Different subsets of specimens may be differently coloured at the command of the user. The user can also download individual sequence data and collaterals. At the current evolving and developing process at the Guelph node, this entire process costs the ACG inventory $2.50/specimen once it arrives in Guelph. This extremely low price is, however, achieved by subsidy from other grants, most notably from the Gordon and Betty Moore Foundation, the Canadian government, and the University of Guelph.

(b) Typical results
The 2640 specimens of the ACG-reared Hesperiidae of about 350 species barcoded to date produce a manageable NJ phenogram (Electronic Appendix)

Figure 1. The 18 May 2004 portion of the ACG Hesperiidae NJ cox1 phenogram containing a grouping of four species of Dyscophellus (black frame box) and Dyscophellus nicephorus well below that, positioned among Bungalotis and Salatis.
that illustrates many practical aspects of barcoding in this inventory. Figure 1 highlights the portion of this NJ phenogram containing the sequences from 12 specimens of four species of sympatric rain forest Dyscophellus, two of which have similar facies but are readily distinguishable by their genitalia. Two are undescribed and, therefore, bear interim names. Three of the four can be easily distinguished by their caterpillar-and-food-plant combinations. A similar level of separation between congeneric species in the NJ phenogram occurs with about 97% of the 1000-plus morphologically defined ACG species sequenced to date in Hesperiidae, Saturniidae, Sphingidae, Nymphalidae, and Arctiidae. As the sample size for each species increases, the clusters in the NJ phenogram retain their species-level discrete-ness. The placement of a sequence from an unidentified ACG specimen into one of these clusters means that it is very likely to be that species, unless it is a previously unknown species that is among the 3% of confusables (see below).

(c) Phylogenetic signals?
While barcoding does not aim to build phylogenetic trees, it is obvious that morphology-based congeners are often the nearest neighbours in the NJ phenogram. When they are not, it is a signal that the morphological placement may be profitably re-examined. With respect to the example of four species of Dyscophellus given earlier, a fifth sympatric species, Dyscophellus nicephorus, appears well removed in the NJ phenogram, among the array of Bungalotis and Salatis (figure 1). Despite the similarity of adult facies of Dyscophellus nicephorus to three other four Dyscophellus, some members of the inventory staff have long suspected that it was misplaced because its caterpillar has the same colour patterns as do Bungalotis and Salatis, rather than the distinctive colour pattern of the other species.

Figure 2. As in figure 1, but with the colour patterns of last instar caterpillar heads superimposed on the phenogram. Dyscophellus nicephorus is offset on the lower right.

Phil. Trans. R. Soc. B (2005)
four *Dyscophellus* (figure 2). Similarly, the sixth ACG congeneric, *Dyscophellus phraxanor*, has an adult female and a caterpillar that matches well with the four similar *Dyscophellus*, but a very different male; this species also positions far from all of the others in the NJ phenogram (Electronic Appendix). Barcoding unambiguously distinguishes among the six species of ACG *Dyscophellus*, does not confuse them with any other Lepidoptera examined, and suggests that some of their generic placements should be re-examined.

About 3% of the 1000-plus morphological species of ACG Lepidoptera that have been barcoded to date cannot be distinguished from a close relative by their barcodes. An example is three species of *Phocides*. They are distinguishable by wing patterns, genitalia, and caterpillar food plants; but their barcode positions intermingle in the NJ phenogram (Electronic Appendix). However, they neither intermingle with the other three species of ACG *Phocides*, nor with the six other species of look- alike ACG Hesperiidae in two subfamilies and four genera. Other cases of a lack of barcode resolution of ACG hesperiids include *Saliana fusta* and *Saliana triangularis*, and *Cobalus virbius* and *Cobalus fidicula*; two sphingid examples are *Cautethia spuria* and *Cautethia yucatana*, and *Manduca lamuniosa* and *Manduca barnesi* (note added in proof: *Manduca confusion* now appears to be due to sample contamination); there are no saturniid examples (Electronic Appendix).

The morphological species barcoded to date offer a few cases where morphologically similar species possess distinct but very similar clusters (differing by less than 1%) in the NJ phenogram. A dramatic example is offered by *Polyctor cleta* (ACG dry forest) and *Polyctor polyctor* (ACG rain forest). These two medium-sized hesperiids are extremely similar but distinguishable by

---

*Phil. Trans. R. Soc. B* (2005)
facies and genitalia. They differ by just four base pairs in the 648 base pairs region. The single potential sphingid example (Electronic Appendix) is within *Xylophanes crotonis*. Even here, it is unclear as to whether the two clusters within this species in ACG should be viewed as two morphologically identical sympatric species or merely a pair of equally common *cox1* polymorphisms within a single species. Parenthetically, neither of the two clusters represents the newly described other *Xylophanes crotonis* look-alike, *Xylophanes letiranti* (Vaglia & Haxaire 2003), which occurs near but not in ACG. There are no cases of very similar barcodes among morphologically defined ACG-reared Saturniidae.

(f) **Dissolution of one morphological species into several**

Apart from the exceptionally species-rich case of *Astraptes fulgerator* becoming 10 ACG species (Hebert et al. 2004), the barcoding of reared ACG Hesperidae, Saturniidae, and Sphingidae contains significant numbers of examples of an apparent morphological species becoming two or more clusters of adjacent barcodes in the NJ phenogram. On close inspection of their food plants, behaviour, ecosystem or elevation occupied, and/or adult morphology, many—but not all—of these clusters are being found to represent distinct biological entities in ACG.

An example is *Automeris zugana*—a medium-sized, widespread and very well-known saturniid moth (Costa Rica to Ecuador, Lemaire 2002). The first three specimens barcoded, chosen deliberately to span the dry forest and rain forest sides of the ACG, displayed a 2–4% difference in their sequences. While these are substantially smaller differences than those among most morphologically defined species of ACG saturniids (Electronic Appendix), they were large enough to suggest hidden complexity. When 10 *A. zugana* were barcoded, three distinct clusters of sequences emerged. When 42 specimens, chosen to cover the ACG ecosystems, were sequenced, the clusters unambiguously remained (figure 3). The three clusters correlate with subtle differences in adult body weight, facies, genitalia, and ecosystems (the caterpillars are indistinguishable, as are their food plant preferences). The morphological differences had been viewed as intra-specific variation at the time that wild-caught adults were examined by the inventory and by the late Claude Lemaire in the 1980s, though we suspect that Lemaire did not examine the genitalia of more than a few ACG specimens, which happened to be of just one species. One barcode cluster occupies the ACG dry forest, and two occupy the adjoining rain forest—one at 400–600 m elevation and the other at 600–900 m. Ongoing taxonomic efforts will probably link one of these three species to the type specimen of *Automeris zugana* and describe the other two as new. Once described, these three species would fall in the category above of species that differ only slightly in their barcodes but are readily distinguishable by their barcodes.

While examples like that of *A. zugana* are not unusual among the hard-to-catch and often-low-density Hesperidae (more than half of the Hesperidae reared by the caterpillar inventory have never been seen or collected as adults in ACG), they were less expected among Saturniidae and Sphingidae, so loved by collectors and so easily collected with light traps. As mentioned earlier, *Xylophanes crotonis* might turn out to be one of these cases. Other cases still being explored are potential cryptic species within *Xylophanes porcus*, *Xylophanes libya*, *Manduca sexta*, and *Pachyelia fucis*—four seemingly well-known and widespread morphological species. Several other well-known ACG saturniids are experiencing the same fate as described for *Automeris zugana*. There are two unexplored barcode clusters within *Gamelia musta*, *Automeris tridens*, *Automeris postalbida*, and *Hylesia dalina* (Electronic Appendix). All four of these hemileucine saturniids are highly polyphagous as caterpillars (Janzen 2003), and the barcode clusters are parapatric by ecosystem and/or elevation. The most startling of all is the well-known *Eacles imperialis*, which ranges from southeastern Canada to Argentina. The ACG *Eacles imperialis* has two distinct barcode clusters showing an 8% sequence divergence. One cluster occurs in rain forest and the other in the parapatric dry forest. Strikingly, the dry forest cluster only differs by 5% from its morphological conspecific in Great Smoky Mountains National Park, Tennessee, USA, several thousand kilometres to the north (sequences from BoLD). *Eacles imperialis* do not migrate.

If the adult Hesperidae, Sphingidae, and Saturniidae of ACG had not been so thoroughly studied morphologically during the past 100-plus years, there would be many more cases where ‘one’ slightly variable morphological species dissolves into several when it is barcoded.

(g) **Association of sexes**

Associating sexes of wild-caught or reared polyphagous species-rich Lepidoptera can be difficult. Scott Miller and colleagues have already found barcoding to be extremely useful in associating sexes of their reared Tortricidae and Lymantriidae in their extensive caterpillar inventory in Papua New Guinea (www.nmnh.si.edu/new_guinea). In the ACG inventory, the caterpillars of two distinctive ‘species’ of *Saliana* (Hesperiidae) were found at low density, the adults of one being given an interim name and the other tentatively identified as *Saliana severus*. This is an exceptionally dark species of *Saliana*. Barcoding then showed that these two morphological entities had identical *cox1* barcodes. Querying back to the morphological taxonomy, it was noticed that both sexes of *Saliana severus* have dark undersides, and that the interim white-undersided *Saliana* were all females, while the ACG *Saliana severus* were all males. This iterative feedback led to the conclusion that the inventory is not rearing *Saliana severus* but yet some other species of *Saliana* with strong sexual dimorphism.

(h) **Massive interspecific discrimination**

BoLD now contains thousands of vouchered and species-level identified *cox1* sequences from 1000-plus species from the ACG inventory, and has accumulated similar records from another 2000 Lepidopteran species from other parts of the world. This leads to
the obvious experiment of comparing all the ACG specimens in one huge NJ phenogram. We did, and there is no overlap of any species other than those already found with a within-ACG family-level NJ phenogram. Next, we combined all BoLD Sphingidae sequences from Africa \((n=26\) for 11 species), Papua New Guinea \((n=75\) for 28 species), North America \((n=136\) for 32 species), and ACG \((n=614\) for 95 species). Again, there is no overlap of the 166 morphological species clusters in the NJ phenogram other than the less than 3\% already recognized as confusable within a geographic region.

5. CAVEATS AND PROBLEMS
Combining barcoding with the more classical taxonomic process for the inventory in ACG, and serving as a pilot project for barcode library construction, barcoder emergence, and tollbooth development, is a work in progress. Some barriers to progress have emerged.

(a) Sample size per species
It is now commonplace to use mtDNA sequence data to resolve phylogeography of species (e.g. Wuster et al. 2005). However, there has been a strong tendency in barcoding to treat a few sequences as if they were the ‘type’ for a place, potentially missing cryptic species and cases of overlap in the NJ phenogram. This approach was due to an initial desire to maximize species coverage at a time when sequencing costs were still high and analytical protocols were under development. The barcoding done to date with morphologically defined species suggests that if only two to five specimens are barcoded, cases of interspecific overlaps will be recognized; but a significant number of cryptic species that differ by only a few per cent will be missed. While further barcoding is needed to refine this estimate, at least 10 specimens per species should be used from what seems to be one site—assuming that the specimens can be chosen so as to avoid sibling individuals. Samples of this size should expose clues to most cases of sympatric cryptic species that have species-level barcode differences. If such a sample reveals more than one cluster in the NJ phenogram, additional specimens should be barcoded to explore for cryptic species.

(b) Barcoding a morphologically unknown biota
The specimens barcoded in the caterpillar inventory are all sorted to morphospecies (often backed by a species-level name) before the specimens are chosen to be barcoded. This minimizes the number of individuals necessary to barcode in order to know how many clusters there are in the NJ phenogram for any given per cent difference used to define a cluster. It also assists in knowing how to treat singletons that deviate slightly from other members of a cluster but do not form or join a cluster. Are they singletons of a rare species or simply deviant individuals? If barcoding simply examines a cluster. Are they singletons of a rare species or simply deviant individuals? If barcoding simply examines a pool of individuals collected in a Malaise or light trap, a much larger number of individuals would need to be barcoded to reveal all the clusters in the sample. Furthermore, a small fraction of the individuals would remain in taxonomic limbo because it would not be clear if they were the result of intraspecific variation or rare individuals of another species. This is just as it is with morphological sorting of a large sample of unknowns. However, combining barcoding with morphological sorting will give both a more accurate and a more economic result.

(c) Cross-geography barcoding
The ACG inventory and its barcoding is, and will continue to be, a deep sample of a place where any sample point is within flight distance of most other sample points. It does not reveal the extent of intraspecific variation in barcodes that will emerge as widespread species are barcoded across their neotropical ranges (e.g. Dick et al. 2004). However, this work is well underway for moths and butterflies in the eastern half of North America. Early results suggest that between-site intra-specific variation in barcodes will not be a confounding problem in their use for species identification, except in the very small percentage of species whose barcodes overlap.

(d) Developing barcoding versus using barcoding
In the CBOL barcoding initiative (http://www.barcoding.si.edu/), as during the emergence of any new technology, those embedded in the initiative are caught in a tension between full-blast development of barcoding (how to sequence accurately and cheaply, build the sequence libraries, build the barcoder, build and operate the toll booth), and using the new information to solve questions and drive initiatives in other agendas. When do we stop using barcoding to better the caterpillar inventory and be a pilot project, when do we put full time into building the sequence library— with museum and fresh-caught specimens—to barcode the Lepidoptera of the world? The question hinges on availability of funds/technology for each route, on the existence of fellow travellers, and on the personal curiosity yield from each of the two routes. Janzen and Hallwachs are caught up in the mosaic of agendas cocooning the survival of ACG into perpetuity and its pilot project role in biodiversity survival through non-damaging development (Janzen 2000). Burns and the remainder of the taxasphere are caught up in the business of the taxasphere. The CBOL node at Guelph and its occupants (e.g. Hajibabaei et al. 2005) are certainly on the barcoding route, but even they will be distracted from the straight and narrow of developing barcoding as a process and into the application of that process to the real world, if for no other reason than to keep the funding flowing.

This study is a microcosm of this problem. Each time a new array of ACG specimens is barcoded, new taxonomic and biodiversity puzzles are revealed. Each begs for taxonomic, ecological, methodological, and publishing energy for its resolution. Barcoding reveals such puzzles at a far higher rate than they can be treated by the human and financial resources available. This means that, for the sake of barcoding, they are left behind. An example is the publication of the barcoding confirmation and exposure of 10 species in ACG Astraptes fulgerator (Hebert et al. 2004), before the
species have been described and their (known to the
caterpillar inventory) natural histories recorded in print
or website. All signals are that full-scale barcoding will
reveal innumerable questions, as did the microscope
and scanning electron microscope. The ACG cater-
pilar inventory itself has already been heavily exposed
to this conundrum, and has resolved it by using a
website database to record the basic specimen-level
information rather than an interminable series of short
publications. It also refuses to be diverted from the goal
of total inventory. As frequent observers of taxonomists
identifying museum specimens, Janzen and Hallwachs
have particularly noticed the positive feedback when
the barcoding process is applied to previously studied
specimens and identifications. It is a real joy to watch
the outcome of providing a top-flight taxonomist with
a new tool to address long-standing taxonomic tangles
and uncertainties. But that very positive outcome is
also highly seductive away from continued develop-
ment of barcoding as a tool and method.

(e) Variants
Anticipation of problems with barcoding leads immedi-
ately to concern about hybrids. However, a hybrid
should simply cluster with its mother, grandmother,
sisters, etc. in a cox1-based NJ phenogram. This is no
worse than occurs with a morphological search for
hybrids. More puzzling are the moderately frequent
cases in the ACG inventory where a single individual
differs from the remainder of a large sample cluster by
two to eight base pairs, but lacks any morphological or
natural history reason to be suspected as an individual
of a cryptic species, and does not join any other cluster
as the sample size is increased (evident examples in
Electronic Appendix). These cases may simply be
‘deep intraspecific variants’ similar to those encoun-
tered regularly in morphological and behavioural
explorations, but they do beg for a more scientific
explanation.

(f) Laboratory errors
The processing chain from a caterpillar to a sequence in
GenBank (http://www.barcoding.si.edu/CBOLDData-
basesGenBank.htm), with its collateral information
attached, offers a wealth of opportunities for human
and machine errors to creep in. Many of these
opportunities, as well as the specific errors themselves,
are polished out of the system as discovered on a case-
by-case basis. However, there is one general problem
that needs immediate attention. It is essential that the
internet connectivity among the various data and
specimen deposits become so seamless that an error
encountered in a data point or its collateral at one place
in the chain can be corrected, and then that correction
is automatically transmitted through the network to the
other places where the uncorrected data remain. To
emphasize this need is not a great intellectual advance,
but rather a plea for rapid resolution. As we attain
consensus that all DNA sequences, for example, should
be vouchered with specimens, collateral information,
and images, we desperately need to avoid each node in
the chain being a static depository of errors that were
corrected in one place but cannot be corrected
elsewhere without enormous investment of painstaking

human and case-by-case intervention at other nodes.
A specimen of Astraptes TRIGO is simultaneously a
sequence and its collaterals in GenBank and BoLD, on
a pin in the National Museum of Natural History,
Smithsonian Institution, and an event record in the
caterpillar inventory. When it finally gets its scientific
name, it is imperative that with a push of a button
Astraptes TRIGO disappears, other than in audit trails,
and is everywhere replaced by its newly assigned
scientific name. Equally, when it is found that a base
pair read was wrong in its deposited sequence, whether
by a person or an application, the subsequent
 corrections throughout the network need to occur not
by event-specific emails, but by hot linkages, with all
that implies.

(g) Old specimens
The only practical way to rapidly build thorough and
cross-geography global barcode sequence libraries of
millions of species is by barcoding representative
specimens in the world’s museums. The barcode
initiative is not going to recollect the world to build
its sequence library. CBOL has done a magnificent job
of getting the world’s museums politically on board,
but there are two major impediments. First, the funds,
personnel, and energy are not yet available for the
massive taxonomic and physical curatorial process that
is required. Worse, the present process will constantly
be caught in the dilemma described above whereby the
participatory taxonomist is forced to choose between
pursuing the multiple taxonomic puzzles and answers
revealed by barcoding ‘the collection’, and sustaining
the humdrum of minimal curating for barcoding. This
begs for funding for a new kind of curator who largely
carries forward the barcoding process while the
taxonomist energy is applied surgically to select
questions. Even these will quickly exhaust the current
taxonomic human resource. The situation absolutely
demands an absolute increase in the taxonomist guild if
barcoding is to function. Just the questions generated
by barcoding the ACG caterpillar inventory can easily
absorb the full taxonomic capacity of several major
museums for the caterpillar family in question, and
ACG contains no more than 3% of the world’s
Lepidoptera biodiversity.

Second, while sequences can be obtained from a
given old specimen with much work and time (and
money), we are still far from the fast cheap sequencing
that can be done with freshly collected material
(however, see Hajibabaei et al. 2005). This deficit is a
composite of two problems. On the one hand, because
fresh material—such as that reared by the ACG
inventory—is so easy to sequence (and often is fully
databased and vouchered from the beginning), it
seduces the barcoding initiative away from the essential
ability to analyse the old but much more biodiverse
material sitting in museum cabinets and representing a
huge geographic coverage and centuries of effort. On
the other, if a taxonomist does devote extra curatorial
and databasing effort to organize a museum’s holdings
for barcoding, but only a small fraction of samples
successfully sequence, the negative psychological
impact is huge. Equally bad is the damage and cost of
having to sample a very large number of specimens with

Phil. Trans. R. Soc. B (2005)
a hope that a few of them will successfully sequence. Incidentally, one of the huge advantages of being able to sequence for barcoding right at the museum cabinet would be that if it fails, it is known right then and a second sample or specimen can be tried from the same series. This is much better than having to relocate the failed specimen or species months later among its millions of compatriots. Likewise, onsite sequencing will reveal variation and cryptic species at the time they are being curated, allowing sequence sample size to be increased at that moment.

(h) Museums and databases as voucher depositories

It is imperative that barcode sequences be vouchered by specimens, irrespective of whether the specimen has been identified. And as the vouchers become identified, the value of the barcode sequence increases greatly (e.g. De Ley et al. 2005 in this Theme Issue). When the specimen is already in a museum for other purposes, making a barcode voucher of it may mean relatively little change in its cost of permanent maintenance. However, the massive barcoding of new inventory specimens, just as the inventory itself, can easily swamp the holding capacity of our museums. Worse, it can do it with huge series that have large barcoding significance for geographic variation, etc., but are far beyond the traditional reasons and amounts of space allocated in museums to long series of conspecifics. The barcode vouchers from the ACG caterpillar inventory have the potential to consume a substantial amount of the new drawer space in the new expansion and reorganization of the Lepidoptera collection of the National Museum of Natural History at the Smithsonian Institution. Were ACG to take on a total Lepidoptera barcode venture, it would require another depository solution. Equally, the Lepidoptera collection at the Instituto Nacional de Biodiversidad (INBio), Costa Rica’s National Biodiversity Institute, is filled to capacity. To thoroughly barcode the Lepidoptera of Costa Rica would require a doubling of space at INBio just to hold the vouchers. The problem is compounded when barcode vouchers are viewed as stored permanently, which means a huge archival cost with no more scientific return than confirmation capacity for a sequence. It seems clear that true vouchering both for barcode libraries, and for research barcoding once a basic library is established, will require the creation of depositories for that purpose rather than simply squeezing more specimens into currently overcrowded museum facilities. A related question is whether a museum is willing to let a taxonomic specimen be moved into the category of barcode sequence voucher, thereby limiting many of the traditional uses for a specimen. The ACG inventory specimens are gladly given to museum repositories as barcoding and taxonomic vouchers, but trading them, resampling them, displaying them and generally caring for them as individually coded vouchers is a major responsibility not to be entered into lightly for the tens of millions of specimens that true global barcoding implies.

6. CONCLUSIONS

DNA barcoding, as being practised on the ACG caterpillar inventory, is about cheap, mass, and fast sequencing to initially discover and confirm biological species, build reference sequence libraries for the species treated, and eventually use the reference library to aid species-level identifications. The cheaper and quicker it is, the easier it will be to explore the complexity of barcode patterns—and the biology they signal—in time and space.

In a world lacking the taxasphere, the single largest problem with barcoding is the inability to connect the cluster in the NJ phenogram to what is already known about that species by humanity. Barcode reference libraries based on, and connected to, what we already know are essential. But what of the millions of species that can be recognized only through a barcode either because they are very similar morphologically, or because they simply have not been studied enough to know their non-barcode diagnostic traits? These species will simply have to exist in some higher taxonomic rank until they are studied as biological entities, and/or until there truly is a pocket barcorder that is used just as are today the camera, hand lens, dissecting microscope, binoculars, notebook, paper field guide, memory, etc. A barcorder is a DNA microscope with a memory. Given the high potential for the barcorder to store every sequence read, along with the collateral of the moment, there is truly huge potential identification power and ability to connect to what is locally to globally known. Historically, it should be recognized that a barcorder is far from being the first effort for an automated and computer-based species identification tool. Classical keys up through complex web-based interactive keys, though based on a taxasphere-derived terminology, are themselves a kind of NJ phenogram. A recent example is DAISY, an automated identification concept and tool based on image data rather than the DNA barcode sequence (Gauld et al. 2000).

Ongoing integrations of barcoding with field and museum biodiversity studies make clear the need for five ‘libraries’—the ‘literature’, morphology, natural history (food plants, microgeography, phylogeny, etc), taxonomy per se, and DNA barcode sequences—and merge them iteratively to approach reality and bioliteracy. Each of these five libraries is imperfect and variously developed, but when they are merged, they jointly achieve about as good a focus on the biology of a place or taxon as can be obtained.

Apart from the general scientific and public desirability to be able to better, faster, and more cheaply identify organisms for a host of agendas, is there an additional reason to hasten to a realized barcorder and accompanying information? Those of us who would like to see a serious part of today’s surviving biodiversity still with us centuries from now are in a severe race against the multiplex of forces polishing today’s remnants of that biodiversity off the earth. While it is certainly not the solution to end all concerns, a cheap public back-pocket barcorder does have the potential to allow any and all to know what an organism is at the moment that it matters. This essentially allows anyone to ‘read’ biodiversity. As with most literacy, it is
only at certain key times that it matters. But if people can be bioliterate at those key times, humanity’s relationship to wild biodiversity has a high potential of changing for the better. Yes, there will be abusers, just as there are abusers of literacy, but overall, becoming literate has had a highly civilizing impact on humanity. And from the quite selfish viewpoint of the practising biologist, it greatly increases the motivation to collate and organize what we know if the world can get to that information, even if only on the web, at the moment when the actor in the play is biting, stinging, pollinating, munching, or displaying.

However, it is no secret to the world of users and protectors of wild biodiversity that their politico-legislative framework is built on a taxonomic structure that variously defines species (and their subunits), and usually does it morphologically. Barcoding is going to reveal and reinforce a lot of cryptic diversity, and add fuel to the argument of whether we are using or protecting a morphologically defined or a phylogenetically defined biological entity (e.g. Agapow et al. 2004; Debruyne 2005; Simmons et al. 2005). Like any broadly applicable technology, it will be used for bad and good; the barcoding initiative will need to be prepared for that. It was correctly anticipated in 2003 that national permission to barcode thousands of species in the ACG would require years of Costa Rican political debate and permission, legislative interpretation, and explicit enlightenment of social leaders.

In the search for rational support for DNA barcoding—as if any is needed other than its obvious pragmatic usefulness—it has been expressed that a major ‘problem’ with taxonomy is that there are few taxonomists and that one cannot manage more than a few thousand species in his or her head. Both statements are false. There are many taxonomists, but very few jobs for them. Worse, many of these jobs require that they spend substantial time and mental energy on other tasks than taxonomy. We do not need to train more taxonomists so much as we need to hire more of them—the taxasphere combined with individuals who really enjoy doing taxonomy will provide the human resource if there is employment available. Second, we know many taxonomists who handle accurately tens of thousands of names in the combinations of their heads, databases, collections, and literature. Mental capacity is not the problem. The problem is that there is not one of them standing by your left elbow when you need to identify something. And there never will be, no matter how appreciative society becomes of wild biodiversity. A cheap thorough pocket barcoder, and all its supporting information, technology, and linkages, is the only way that the grand bulk of humanity will ever become bioliterate, at least to the degree where living things are generally viewed as more than more biomass to convert or trash.

We close with a reiteration of four speed bumps for the CBOL initiative:

- **Cheap and fast barcode sequencing of old specimens needs to be developed quickly.**
- **Funding is essential for the interactive classical taxonomy and curation to provide and name the specimens that will be used to build the DNA sequence libraries.** Finding people is not nearly as large a problem as is finding the salary and operational support for the people that already have a strong interest in being participants. We need to **HIRE more taxonomists**. They will train and mentor each other and themselves. And every time a leg, feather, or leaf chip goes into a barcoder, a tollbooth has to move a penny into the funding for the taxasphere.
- **Someone has to take up the conversation with the commercial sectors such that while the barcoder is being built, the emerging technology is in a conversation with the sequence libraries and the tollbooth. A marvellous cell phone is of no use if, when you call the number, no one answers, and, when they do, they have no information.**

This study has been supported by grants to DHJ and WH from the Wege Foundation, US National Science Foundation (DEB 0072730), Guanacaste Dry Forest Conservation Fund, and Área de Conservación Guanacaste. JMB was supported by the Smithsonian Institution and the National Museum of Natural History Small Grants Program. PDNH thanks the Gordon and Betty Moore Foundation, the Canada Foundation for Innovation, the Ontario Innovation Trust, the Canada Research Chairs Program, and NSERC for the critical support that they provided to develop and operate a DNA barcoding facility. We also thank the ACG parataxonomists (Janzen 2004a) for assembling specimens, as well as Tanya Dalkey and Stephanie Kirk for acquiring barcode sequences, and Donald J. Harvey for genitalic dissections and Young T. Sohn for drawings. We deeply appreciate the contributions made by Rob Dooh and Sujeewan Ratnasingham to data organization and assembly. We thank M. Stoecklem, S. Miller, V. Savolainen for constructive commentary on the manuscript, and are grateful to CBOL for facilitating the DNA barcoding initiative.

**REFERENCES**


The supplementary Electronic Appendix is available at http://dx.doi.org/10.1098/rstb.2005.1715 or via http://www.journals.royalsoc.ac.uk.
DNA barcoding Australia's fish species
Robert D Ward, Tyler S Zemlak, Bronwyn H Innes, Peter R Last and Paul D.N Hebert

*Phil. Trans. R. Soc. B* 2005 *360*, 1847-1857

**Supplementary data**
http://rstb.royalsocietypublishing.org/content/suppl/2009/02/12/360.1462.1847.DC1.html

**References**
This article cites 30 articles, 5 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1847.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1847.full.html#related-urls

**Email alerting service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](http://rstb.royalsocietypublishing.org/subscriptions)
DNA barcoding Australia’s fish species

Robert D. Ward1.*, Tyler S. Zemlak2, Bronwyn H. Innes1, Peter R. Last1 and Paul D. N. Hebert2

1CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart, Tasmania 7001, Australia
2Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada N1G 2W1

Two hundred and seven species of fish, mostly Australian marine fish, were sequenced (barcoded) for a 655 bp region of the mitochondrial cytochrome oxidase subunit I gene (cox1). Most species were represented by multiple specimens, and 754 sequences were generated. The GC content of the 143 species of teleosts was higher than the 61 species of sharks and rays (47.1% versus 42.2%), largely due to a higher GC content of codon position 3 in the former (41.1% versus 29.9%). Rays had higher GC than sharks (44.7% versus 41.0%), again largely due to higher GC in the 3rd codon position in the former (36.3% versus 26.8%).

1. INTRODUCTION

It has long been recognized that DNA sequence diversity, whether assessed directly or indirectly through protein analysis, can be used to discriminate species. More than 40 years ago, starch gel electrophoresis of proteins was first used to identify species (Manwell & Baker 1963). Nearly 30 years ago, single gene sequence analysis of ribosomal DNA was being used to investigate evolutionary relationships at a high level (Woese & Fox 1977), and mitochondrial DNA approaches dominated molecular systematics in the late 1970s and 1980s (Avise 1994).

Recently, Tautz et al. (2002, 2003) made the case for a DNA-based taxonomic system. DNA sequence analysis has been used for 30 years to assist species identifications, but different sequences have been used for different taxonomic groups and in different laboratories. Hebert et al. (2003) proposed that a single gene sequence would be sufficient to differentiate all, or at least the vast majority of, animal species, and proposed the use of the mitochondrial DNA gene cytochrome oxidase subunit I (cox1) as a global bioidentification system for animals. The sequence was likened to a barcode, with species being delineated by a particular sequence or by a tight cluster of very similar sequences.

Empirical support for the barcoding concept ranges from studies of invertebrates (e.g. springtails and butterflies) to birds (Hebert et al. 2004a,b; Hogg & Hebert 2004). However, the approach is not without controversy (e.g. Lipscomb et al. 2003; Moritz & Cicero 2004). For a barcoding approach to species identification to succeed, within-species DNA sequences need to be more similar to one another than to sequences in different species. Recent studies show that this is generally the case, but there are exceptions. Hybridization among species would create taxonomic uncertainty: mitochondrial DNA is maternally inherited and any hybrid or subsequent generation would have the maternal species DNA only.

Here we examine whether barcoding can be used to discriminate fish species. There are probably close to 30 000 fish species worldwide (the FishBase count of species on March 31 2005 was 28 800—see www.fishbase.org), constituting about 50% of all vertebrate species. They are systematically very diverse, ranging from ancient jawless species (Agnatha: hagfish and lampreys) through to cartilaginous fishes (Chondrichthyes: chimaeras, sharks and rays) and to old and modern bony fish (Osteichthyes: coelacanths, eels, carp, tuna, flatfishes, salmonids, seahorses, etc.). In 2000, fisheries provided more than 15% of total animal protein to the global food supply, employed about

* Author for correspondence (Bob.Ward@csiro.au).

One contribution of 18 to a ‘Theme Issue 'DNA barcoding of life'.

**Keywords:** cytochrome oxidase subunit I; CO1; fish identification; sharks; rays; teleosts

Downloaded from rstb.royalsocietypublishing.org on August 19, 2011
35 million people, and had an estimated first sale value of about US$81 billion (FAO 2002): fish and fish products are important contributors to human food security.

Accurate and unambiguous identification of fish and fish products, from eggs to adults, is important in many areas. It would enable retail substitutions of species to be detected, assist in managing fisheries for long-term sustainability, and improve ecosystem research and conservation. Hitherto, a wide variety of protein- and DNA-based methods have been used for the genetic identification of fish species (see, for example, Ward & Grewe 1994; Pérez-Martín & Sotelo 2003). Here we examine coxl diversity within and among 207 fish species, most of which have been examined from multiple specimens, with the goal of determining whether DNA barcoding can achieve unambiguous species recognition in fish. Most of the species examined were Australian commercial species, but many more such species and bycatch species remain to be barcoded.

2. MATERIAL AND METHODS

Tissue subsamples were isolated from fragments of white muscle of (mostly) Australian fish species that had been stored at −80°C for several years. The majority of these samples had been initially collected for the protein fingerprinting of Australia’s commercial domestic and imported fish species (Yearsley et al. 1999, 2003), others were from species of special interest to scientists at the Commonwealth Scientific and Industrial Research Organization’s division of Marine and Atmospheric Research. Some of the samples were from voucher specimens. Generally we aimed, where possible, to sample five individuals per species (four species were sampled at greater intensity) and achieved this target for 47.5% of the 207 species sequenced. Numbers per species ranged from one to 15 with a mean of 3.66.

DNA extracts were prepared from muscle tissue using Chelex dry release (Hajibabaei et al. 2005). Approximately 655 bp were amplified from the 5′ region of the coxl gene from mitochondrial DNA using different combinations of four newly designed primers: FishF1-5′TCAACCAACCACAAAGACATTTGCCAC3′, FishF2-5′TCGACTATTCAATAAGATCAGGCAC3′, FishR1-5′TAGACTTTCTGGTGGCCAAAGATCA3′, FishR2-5′ACCTCAGGTGTACCGAAGAATCAGAA3′.

In one case, the broadnose shark (Notorhynchus cepedianus), an internal forward primer (5′ATCTTTGTGATGGCAT-GAGCAGGAATGT3′) was used in conjunction with FishR2 to yield a shorter fragment (616 bp). The 25 μl PCR reaction mixes included 18.75 μl of ultrapure water, 2.25 μl of 10× PCR buffer, 1.25 μl of MgCl2 (50 mM), 0.25 μl of each primer (0.01 mM), 0.125 μl of each dNTP (0.05 mM), 0.625 U of Taq polymerase, and 0.5–2.0 μl of DNA template. Amplifications were performed using a Mastercycler® Eppendorf gradient thermal cycler (Brinkmann Instruments, Inc.). The thermal regime consisted of an initial step of 2 min at 95°C followed by 35 cycles of 0.5 min at 94°C, 0.5 min at 54°C, and 1 min at 72°C, followed in turn by 10 min at 72°C and then held at 4°C. PCR products were visualized on 1.2% agarose gels and the most intense products were selected for sequencing. Products were labelled using the BigDye® Terminator v.3.1 Cycle Sequencing Kit (Applied Biosystems, Inc.) and sequenced bidirectionally using an ABI 3730 capillary sequencer following manufacturer’s instructions.

Sequence data, electropherograms, and primer details for specimens are available within the completed project file ‘Fishes of Australia Part 1’ on the Barcode of Life database site at the University of Guelph (see http://www.barcodinglife.org). GenBank numbers are DQ107581 to DQ108334 and are matched against individual specimens in the ‘Fishes of Australia Part 1’ file.

Sequences were aligned using SeqSCAPE v.2.5 software (Applied Biosystems, Inc.). Sequence divergences were calculated using the Kimura two parameter (K2P) distance model (Kimura 1980). Neighbour-joining (NJ) trees of K2P distances were created to provide a graphic representation of the patterning of divergence between species (Saitou & Nei 1987). In the three chosen subgroups of fish, bootstrapping was performed in MEGA3 (Kumar et al. 2004) with 1000 replications.

Nearly all the target species amplified using one or both primer sets. However, there was insufficient product for sequencing from five species: N. cepedianus (n=5), Asymbolus rubiginosus (orange spotted catshark, n=2), Neoplatycephalus conatus (deepwater flathead, n=5), Cephalopholis igarashiensis (goldbar grouper, n=1) and Epinephelus undulatesтратitus (Maori rockcod, n=2). The Maori rockcod samples were in a degraded condition, probably explaining their PCR recalcitrance. For N. cepedianus, a newly designed internal forward primer was used as described above to obtain product that could be sequenced.

3. RESULTS

We present results for all 207 species followed by more detailed examinations of three subgroups of fish. These subgroups were flatheads (14 species from three genera), large tunas (genus Thunnus, eight species), and dogfish or spurdogs of the genus Squallus (which contains several Australian morphs that have not been formally described as species). Flatheads and tunas are teleosts; spurdogs are chondrichthyans.

(a) All species

A total of 207 species were analysed, giving (because of multiple specimens for most species) a total of 754 sequences. The full K2P/NJ tree has been lodged as an Electronic Appendix. It is presented here in summary form as figure 3. All 207 species can be differentiated by coxl barcoding. We did not choose species known to be readily differentiable; indeed, we included several sets of sibling species (for example, in the genera Squallus and Thunnus) and many species were congeneric.

Read lengths were all about 655 bp long, although in some instances some base calls were uncertain. No insertions, deletions or stop codons were observed in any sequence. The lack of stop codons is consistent with all amplified sequences being functional mitochondrial coxl sequences, and that, together with the fact that all amplified sequences were about 655 bp in length, suggests that NUMTs (nuclear DNA sequences originating from mitochondrial DNA sequences) were not sequenced (vertebrate NUMTs are typically smaller than 600 bp; Zhang & Hewitt 1996).

The average K2P distance of individuals within species was 0.39% compared with 9.93% for species within genera (table 1). Overall, therefore, there was a 25× more variation among congeneric species than among conspecific individuals. Mean divergence
among species within families increases to 15.5%, and among species within orders and classes it increases to 22.2% and 23.3%, respectively (table 1, figure 1). The rate of increase declines in the higher taxonomic categories due to substitutional saturation.

There was a higher overall GC content in the 143 species of bony fish compared with the 61 species of sharks and rays (47.1 ± 0.2% versus 42.2 ± 0.3%, see also table 2). This difference was attributable (table 3, figure 2) to the GC content of the 1st (57.1% versus 53.5%) and, especially, 3rd codon base (41.1% versus 29.9%). The variance of GC content among species of each of the three groups of fishes (chimaerids, sharks and rays, teleosts) was much higher for the 3rd base (GC₃) than the 1st base (GC₁); the 2nd base was nearly invariant. The mean GC content of the 20 barcoded ray species was higher than the 41 shark species (44.7 ± 0.4% versus 41.0 ± 0.3%). This was again mostly attributable to GC₃ (36.3 ± 1.3% versus 26.8 ± 0.9%). Ten of the 13 species with GC₃ > 35% were rays, while all seven with GC₃ < 21% were sharks. This disparity helps to explain the multimodal distribution in GC₃ content in sharks and rays (figure 2).

Originally we had thought that we had two species that were identical with respect to their cox1 barcodes, Centrophorus moluccensis (Endeavour dogfish) and Centrophorus uyato (southern dogfish). However, on receipt of the barcode data we re-checked the identification history of these samples and discovered that the two specimens originally classified as C. uyato had in fact independently been re-identified by a shark taxonomist as C. moluccensis. This experience attests to the absolute necessity of correct species identification in the development of any barcode library. Further, while such a library is being developed, specimens should be retained until fully analysed to facilitate the reappraisal of any potentially spurious results. Voucher collection is a necessity.

Two species of Hydrolagus, Hydrolagus lemures (blackfin ghostshark) and Hydrolagus ogilbyi (Ogilby's ghostshark), were originally regarded as specifically distinct (Last & Stevens 1994) but have been provisionally amalgamated in a world revision of the group (D. Didier, personal communication). In our study, the five samples of H. lemures clustered tightly together (mean genetic distance of 0.21%), as did the four samples of H. ogilbyi (mean genetic distance of 0.19%), with a mean distance between the two taxa of 6.80%. Barcoding (and earlier protein fingerprinting; Yearsley et al. 1999) supports the contention that these are distinct species.

In several instances, we detected deep divergences among individuals that had been assigned to single species. These deep divergences may flag previously
unrecognized species and warrant further study. Two such examples (see figure 3) will be briefly described.

Two individuals of *Hydrolagus novaezelandiae* (short-nose chimaera), both collected off the east coast of the South Island of New Zealand, showed a divergence of 14.08%. One of these individuals was only tentatively ascribed to this species. It is likely, in retrospect, that this identification was incorrect or that an undescribed species exists. These two specimens also had distinct protein fingerprints (Yearsley et al. 2003). Clearly the incorrectly designated specimen is neither *H. lemures* nor *H. ogilbyi*, as both these species are distinct from both specimens of *H. novaezelandiae*.

Deep divergence was also seen between one specimen of the monotypic genus *Pastinachus* (cowtail stingrays) and two further specimens (within- and between-group divergences of 0.61% and 6.43% respectively). Recent studies have shown that *Parupeneus sephen* is a complex of morphologically distinct species in the Indo-Pacific (Last et al. 2005), and the high *cox1* divergence probably reflects the presence of two of these species.

Possible examples of shared haplotypes were seen in the genera *Pristiophorus* and *Plectropomus*. Five samples of each of two species of *Pristiophorus* were sequenced (*Pristiophorus cirratus*, the common sawshark, and *Pristiophorus nudipinnis*, the southern sawshark), but one of the *P. nudipinnis* samples had an identical *cox1* sequence to two of the *P. cirratus* samples. Excluding this aberrant sample, the mean genetic distance within species was 0.35% and the mean genetic distance between species was 10.94%. Similarly, one of two specimens of *Plectropomus leopardus* (common coral trout) clustered very closely with five specimens of *Plectropomus maculatus* (barcheek coral trout; mean sequence divergence of the six samples = 0.19%) and away from the remaining *Pl. leopardus* sequence (divergence = 4.64%). In both genera, the aberrant

---

**Table 2.** Mean percentage base composition (with s.e.), comparing *cox1* sequences of chimaerids, sharks and rays, and teleosts. Where multiple individuals were taken for any one species, a single sequence was selected at random.

<table>
<thead>
<tr>
<th>group</th>
<th>number of species</th>
<th>% of base</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>chimaerids</td>
<td>3</td>
<td>16.67 ± 0.30</td>
<td>28.13 ± 1.13</td>
<td>25.21 ± 0.06</td>
<td>29.98 ± 0.94</td>
<td></td>
</tr>
<tr>
<td>sharks &amp; rays</td>
<td>61</td>
<td>16.75 ± 0.10</td>
<td>25.44 ± 0.25</td>
<td>25.33 ± 0.15</td>
<td>32.48 ± 0.24</td>
<td></td>
</tr>
<tr>
<td>teleosts</td>
<td>143</td>
<td>18.31 ± 0.07</td>
<td>28.75 ± 0.15</td>
<td>23.58 ± 0.09</td>
<td>29.38 ± 0.14</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** GC content of the 1st, 2nd and 3rd codon positions (with s.e.). Where multiple individuals were taken for any one species, a single sequence was selected at random.

<table>
<thead>
<tr>
<th>group</th>
<th>number of species</th>
<th>GC% and codon position</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>chimaerids</td>
<td>3</td>
<td>52.24 ± 0.40</td>
<td>42.33 ± 0.08</td>
<td>34.97 ± 2.40</td>
<td></td>
</tr>
<tr>
<td>sharks &amp; rays</td>
<td>61</td>
<td>53.54 ± 0.19</td>
<td>42.71 ± 0.02</td>
<td>29.89 ± 0.90</td>
<td></td>
</tr>
<tr>
<td>teleosts</td>
<td>143</td>
<td>57.11 ± 0.10</td>
<td>42.63 ± 0.02</td>
<td>41.05 ± 0.50</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2.** Variation in GC content for *cox1* among two groups of fishes. The first, second, third and fourth rows plot the GC content of sharks and rays codon position 1, sharks and rays codon position 3, teleosts codon position 1 and teleosts codon position 3, respectively. Codon position 2 is not shown as this shows very little variation within and among the two fish groups (table 3).
samples might reflect shared haplotypes (perhaps from past introgressive hybridization between species—in both instances the species have overlapping ranges) or, possibly, misidentification of the original specimens.

(b) Flatheads, family Platycephalidae

Flatheads are scorpaeniform fish and members of the family Platycephalidae, of which there are some 60 species worldwide in 18 genera (see http://www.fishbase.org; also see Nelson 1994). We examined 15 mostly commercially significant species from three genera, but one species, *Neoplatycephalus conatus*, failed to amplify. Most of the species were represented by multiple specimens (figure 4). Two genera (*Platyocephalus* and *Neoplatycephalus*) are members of the subfamily Platyocephalinae; the third, *Cymbacephalus*, belongs to the subfamily Onigociinae (Matsubara & Ochiai 1955). Keenan (1988) suggested the subfamily Cymbacephalinae be revived for a small group of species including one examined here, *Cymbacephalus nematophthalmus*.

All assemblages of conspecific individuals had bootstrap values of 95–100%. K2P nucleotide diversity within species was extremely limited, ranging from 0% to 0.93%, with a mean of 0.22%. Ten individuals of *Platyocephalus longispinis* were examined, five each from eastern and western Australia: there was no obvious east–west separation of these individuals in the *P. longispinis* clade.

Divergence between species was high, with average within-genus and within-family K2P distances of 15.55% and 22.07%, respectively. Two distinct clades representing the subfamilies Onigociinae (*Cymbacephalus*, two species) and Platyocephalinae (*Platyocephalus* and *Neoplatycephalus*, ten and two species respectively) were recognized with a bootstrap value of 99%. Four subclades with bootstrap support values of 90–100% can be seen within the Platyocephalinae clade: (i) *P. longispinis/P. bassensis*; (ii) *P. caeruleopunctatus/P. speculator*; (iii) *P. fuscus/P. endrachtensis/P. indicus*; (iv) *N. aurimaculatus/N. richardsoni*.

(c) Tunas of the genus Thunnus

Tunas are large highly migratory fish of the family Scombridae. All are commercially important, some extremely so (FAO 2004). There are eight species in the genus *Thunnus* (we follow Collette *et al.* (2001) in separating *T. thynnus thynnus* and *T. thynnus orientalis* into two species, *T. thynnus* and *T. orientalis*, respectively), six of which are found in Australian waters: *T. obesu* (bigeye tuna), *T. alalunga* (albacore), *T. albacares* (yellowfin tuna), *T. maccoyii* (southern bluefin tuna), *T. orientalis* (northern bluefin tuna) and *T. tonggol* (longtail tuna). We sequenced five individuals of each of these species. We also sequenced four individuals of *T. atlanticus* (Atlantic blackfin tuna) and four individuals of *T. thynnus* (Atlantic bluefin tuna); for the latter, we downloaded data for a further three individuals from GenBank. Finally, we downloaded two additional sequences from GenBank for *T. alalunga* and added three additional sequences that we collected for *T. albacares* from South African waters.

The resulting phenogram (figure 5) of 46 sequences appears very different from the flathead phenogram. Each species clustered as a separate assemblage (no individuals were misplaced), but bootstrap values for species separation were mostly around the 60–70% level. Genetic differences were very small within species, with a mean K2P distance of 0.11%. The three South African individuals of *T. albacares* could not be separated from the five Australian individuals. The mean inter-species distance was very low at 1.04%. Only one clade had a high level of bootstrap support (99%), comprising *T. orientalis* and *T. alalunga*.  

Figure 3. Neighbour-joining tree of 754 *cox1* sequences from 207 fish species, using K2P distances. Multiple specimens of individual species are marked in blue. The three instances of deep intra-specific divergence are identified in orange. The three subgroups examined in more detail are identified.
The genus *Squalus* is represented in Australian waters by three nominal species (*Squalus acanthias*, *Squalus megalops* and *Squalus mitsukurii*) and six species that have not been formally named (*Squalus* spp. A–F). These nine species are described in Last & Stevens (1994). We sequenced five *S. acanthias* (white-spotted dogfish), four *S. megalops* (spiky dogfish), ten *S. mitsukurii* (green-eye dogfish), four *S. sp. B* (eastern highfin spurdog), nine *S. sp. C* (western highfin spurdog), two *S. sp. D* (fatspine spurdog), two *S. sp. E* (western longnose spurdog) and five *S. sp. F* (eastern longnose spurdog), but did not have access to samples of *S. sp. A*.

The resulting phenogram of 41 sequences (figure 6) shows that there are three major clades: *S. acanthias*, *S. megalops*, and a *S. mitsukurii/S. sp. B–F* group. These separate out with bootstrap values of 94–100%. Species B, C, D, E and F form distinct subclades within the *S. mitsukurii/S. sp. B–F* clade, with bootstrap values of 92–100%. Four of the *S. sp. C* specimens were originally identified as *S. mitsukurii*, but were found to be *S. sp. C* following independent re-identification by a shark taxonomist; barcoding supported their identification as *S. sp. C*. All *S. sp. C* specimens have been collected from SW Australia. The single apparent *S. mitsukurii* in clade F was assigned to *S. mitsukurii* based on morphology; but it now appears likely that this specimen was in fact *S. sp. F*. This specimen, along with those of *S. sp. F*, came from the New South Wales coast off eastern Australia. There appear to be two remaining clades of *S. mitsukurii*: the cluster of five specimens with a bootstrap value of 99% were all from the Great Australian Bight, the cluster of four specimens with a bootstrap value of 78% comprised three from New South Wales and one from Western Australia. Whether these clades represent distinct undescribed species or geographic differentiation within a species remains to be assessed.

Omitting the one apparently misclassified specimen, genetic differences were again small within species, with a mean K2P distance of 0.33%. The mean interspecies distance within the genus was quite low at 4.17%.

### 4. Discussion

This study has strongly validated the efficacy of *cox1* barcodes for identifying fish species. We sequenced (usually multiple) specimens of three species of chimaerids, 61 species of sharks and rays and 143 species of teleosts for the barcode region of *cox1*. With no exceptions, all 207 sequenced species could be discriminated. Nearly 98% of all species amplified with the one of two primer sets. Only five of 211 species failed to amplify with these protocols, and one of these amplified with a newly-designed primer set. The four failures came from varied fish groups and included congeners of species that amplified without problem; they may reflect either DNA degradation or primer mismatches. Since our two commonly-used primer sets are extremely similar, we are exploring the possibility of a single, quasi-universal fish primer set that incorporates inosine at the variable positions or has built-in degeneracy.

The GC content of the 655 bp mitochondrial *cox1* region was on average higher in the 143 species of Osteichthyes than in the 61 species of Chondrichthyes: 47.1% versus 42.2%. Saccone et al. (1999) reviewed
Figure 5. K2P distance neighbour-joining tree of 46 cox1 sequences from the eight species of tuna of the genus *Thunnus*. Bootstrap values greater than 50 shown. Specimen numbers for the Barcode of Life Database (BoLD, www.barcodinglife.org) given.

*Phil. Trans. R. Soc. B* (2005)
data from the complete mitochondrial genomes of nine Osteichthyes and three Chondrichthyes species, deriving GC contents of 43.2% and 38.4%, respectively. These values correspond reasonably well to ours, especially with respect to the higher GC content of the teleosts. We observed substantially more nucleotide changes at the 3rd codon position than the 1st, and more at the 1st than the 2nd. For example, the standard

Figure 6. K2P distance neighbour-joining tree of 41 cox1 sequences from eight species of dogfish of the genus Squalus. Bootstrap values greater than 50 shown. Specimen numbers for the Barcode of Life Database (BoLD, www.barcodinglife.org) given.
errors of the GC percentages of the 3rd, 1st and 2nd bases of the Osteichthyes were 0.50, 0.10 and 0.02, respectively (see also figure 2). This reflects the fact that most synonymous mutations occur at the 3rd position, with a few at the 1st position and none at the 2nd. The higher GC content of the Osteichthyes compared with Chondrichthyes was largely due to 3rd base variation, with mean values of 41.1% and 29.9%, respectively, although 1st base differentiation was also observed. Within the Chondrichthyes, GC content of rays was higher than that of sharks (44.7% versus 41.0%), again largely due to GC2 variation. The causes for the GC variation among teleosts, sharks and rays are not known.

No NUMTs (transfers of mtDNA cox1 sequences into the nuclear genome) were observed. A review of the occurrence of NUMTs in plants and animals did not find any evidence of their existence in Actinopterygii (Bensasson et al. 2001), but a comparison of Fugu and nuclear DNA sequences did detect seven or eight NUMTs (Richly & Leister 2004). This confirms the need for vigilance in examining fish amplicons for potential pseudogene status.

Confusion in taxonomic assignments as a result of inter-specific hybridization (Verspoor & Hammar 1991) does not seem to be a major issue—only two of 754 sequences (one in the genus Pristiphorus and one in Plectropomus) appeared in the ‘wrong’ congeneric species. This may represent either introgressive hybridization, or incorrect identification of the original specimen.

Although barcode analysis seeks only to delineate species boundaries, there is clearly some phylogenetic signal in cox1 sequence data. For example, four major clusters were apparent in the NJ phenogram: chimaerids, rays, sharks and dogfish and teleosts. Congeneric species always clustered together and in most cases so did confamilial species. However, methodologies for phylogeny reconstruction from molecular data remain somewhat controversial, with a wide variety of disparate approaches possible (see, for example, Nei & Kumar 2000). We cannot hope to recover the true phylogeny of fishes from a 655 bp fragment of mitochondrial DNA through K2P distance and neighbour joining—rather more gene regions should be used (including nuclear genes) and additional analytic methods deployed including maximum parsimony and maximum likelihood.

Cox1 barcoding for species identification is far more powerful than, for example, protein fingerprinting. Reliable discrimination of Thunnus species using conventional protein electrophoresis is hard if not impossible (e.g. Yearsley et al. 1999), but we found that the same samples were readily identified by cox1 sequencing.

Barcoding discriminated all of the fish species we examined, and would clearly be capable of unambiguously identifying individually isolated fish eggs, larvae, fillets and fins from these species. However, some taxa showed deeper divergence than others. For example, the average within-genus divergence of the flatheads (Platyccephalus, Neoplatyccephalus, Cymbacephalus) was 15.55%, considerably larger than the 4.17% of the genus Squalus, which is itself considerably larger than the within-Thunnus divergence of 1.11%. These differences among genera probably reflect the average age of species divergence, although within genera some species will be older than others. Nevertheless, it seems likely that, for example, the Platyccephalus radiation preceeded the Thunnus radiation. The large tunas have long been suspected of having diverged relatively recently (see, e.g. Elliott & Ward 1995 for allozyme evidence of limited nuclear DNA differentiation).

The Thunnus phenogram (figure 5) only clearly differentiates one clade (T. alalunga and T. orientalis, with 99% bootstrap support). The mitochondrial DNA similarity of T. alalunga and T. orientalis had been earlier described by Chow & Kishino (1995), from cytochrome b and ATPase sequencing. The three species suggested by Collette (1978) to comprise a separate subgenus Neothunnus (Thunnus atlanticus, T. tonggol and T. albacares) form a loosely defined clade in the cox1 phenogram (56% bootstrap support). Our cox1 phenogram is almost identical to one based on sequencing 400 bp of the mtDNA control region of the same eight species (Alvarado Bremer et al. 1997). One minor difference is that the control region tree gives a higher level of bootstrap support, 86%, to the proposed subgenus Neothunnus. The three Neothunnus species differ from the other five species (proposed subgenus Thunnus) in having central heat exchangers rather than lateral heat exchangers, and in being confined to more tropical waters (Collette 1978).

Many of the flathead species that were barcoded (figure 4) were earlier examined allozymically by Keenan (1991). His proposed cladogram of the Platyccephalinae component is very similar to the Platyccephalinae component of figure 3—the four subclades identified from cox1 were also present in the allozyme tree. The mitochondrial and nuclear (allozyme) trees thus compare well. The genera Neoplatyccephalus and Cymbacephalus appear to be monophyletic, Platyccephalus being paraphyletic.

There do not appear to have been any prior surveys of genetic differentiation among species of the genus Squalus, although S. acanthias has been used as a model elasmobranch in some DNA sequencing studies (e.g. Stock & Powers 1995; Hong et al. 1996; Salaneck et al. 2003). Squalus was picked as a genus of particular interest as it included several provisional species (Last & Stevens 1994). The cox1 data clearly supported the biological reality of the species, Squalus spp. B, C, D, E and F, as each of these species had bootstrap values close to 100% for their constituent individuals (figure 6). Indeed, there was greater genetic divergence among these species than among the well-recognized Thunnus species.

The various unresolved questions about specimen identification briefly presented here (for the genera Squalus, Centrophorus, Hydrologus, Pristiphorus) indicate the need to retain whole voucher specimens wherever possible, or at least make an e-voucher from a photograph. While we retained a single voucher specimen for the majority of species discussed here, most other samples are only represented as small tissue samples. Retaining all specimens as vouchers will require significant infrastructure facilities as many fish are large: this may not be practical but it might be...
feasible to retain whole specimens of most species at least until barcoding of those specimens is complete.

In our survey, conspecific samples often (but not always, see the P. longispinis example) came from adjacent areas. Thus we might have somewhat underestimated the extent of within-species diversity. However, any such effect is likely to be minor. Allozyme surveys of marine fish indicate that typically only about 5% percent of genetic variance comes from inter-population differentiation. This percentage is appreciably higher for freshwater fish, around 20% on average (Ward et al. 1994). For freshwater fish, sampling should include individuals from different watersheds whenever possible.

Our results reveal that coax1 barcoding will permit the unambiguous identification of the vast majority of fish species. We now intend to extend our survey to all Australian and all North American fish species. In the longer term, it is hoped that broader collaborations will enable the assembly of a global database of fish coax1 sequences. This will mean collecting sequences from at least 25 000 species. Note that this will inevitably mean that for many species, multiple specimens from widely divergent locations will be sequenced, minimizing the concern expressed above about underestimating genetic diversity. With increasing application of DNA barcoding, many previously unrecognized fish species will be revealed through the discovery of deep divergence of coax1 sequences within currently recognized species. There might also be instances of supposedly distinct species that have identical coax1 sequences, suggesting the possibility of species fusion. Resolution of cases of this nature will require careful morphological analysis from expert taxonomists before any final recommendations can be made. Barcoding and morphological analysis should go hand-in-hand.

Once a global coax1 barcode database has been established for fishes, anyone with direct or indirect access to a DNA sequencer will be able to identify, to a high degree of certainty, any fish egg, larva or carcass fragment. This will be an invaluable tool for fisheries managers, fisheries ecologists and fish retailers, and for those wishing to develop fish identification microarrays. The scientific and practical benefits of fish barcoding are manifold.

We thank those who helped with sequencing and data management at the University of Guelph (Janet Topan, Natalia Ivanova, Sujeewan Ratnasingham, Jeremy de Ward, Rob Dooh) and those that helped with specimen identification and sampling at CSIRO Marine and Atmospheric Research (Ross Daley, Alastair Graham, Tim Fountain, Gordon Yearsley). The Gordon and Betty Moore Foundation provided critical support for both the assembly of barcode sequences and the platform required for their analysis. The CSIRO Wealth from Ocean Flagship program facilitated the assembly and identification of specimens and some of the barcoding. Gordon Yearsley, Jawahar Patil and John Volkman provided useful comments on an earlier version of this manuscript.

REFERENCES


Last, P. R. & Stevens, J. D. 1994 *Sharks and rays of Australia*. Melbourne, Australia: CSIRO Publishing.

Last, P. R., Manjaji, B. M. & Yearsley, G. K. 2005 *Pastinachus solocirostris* sp. nov., a new species of Stingray (Elasmobranchii: Myliobatiformes) from the Indo-Malay Archipelago. *Zootaxa* 1040, 1–16.


The supplementary Electronic Appendix is available at http://dx.doi.org/10.1098/rstb.2005.1716 or via http://www.journals.royalsoc.ac.uk.
Deciphering amphibian diversity through DNA barcoding: chances and challenges

Miguel Vences, Meike Thomas, Ronald M Bonett and David R Vieites

Phil. Trans. R. Soc. B 2005 360, 1859-1868
Deciphering amphibian diversity through DNA barcoding: chances and challenges

Miguel Vences¹,*, Meike Thomas², Ronald M. Bonett³ and David R. Vieites³

¹Institute for Biodiversity and Ecosystem Dynamics, Zoological Museum, University of Amsterdam, Mauritskade 61, 1092 AD Amsterdam, The Netherlands
²Department of Genetics, University of Cologne, Zülpicher Straße 47, 50674 Cologne, Germany
³Museum of Vertebrate Zoology and Department of Integrative Biology, University of California, 3101 Valley Life Sciences Building, Berkeley, CA 94720-3160, USA

Amphibians globally are in decline, yet there is still a tremendous amount of unrecognized diversity, calling for an acceleration of taxonomic exploration. This process will be greatly facilitated by a DNA barcoding system; however, the mitochondrial population structure of many amphibian species presents numerous challenges to such a standardized, single locus, approach. Here we analyse intraspecific patterns of mitochondrial variation in two distantly related groups of amphibians, mantellid frogs and salamanders, to determine the promise of DNA barcoding with cytochrome oxidase subunit I (cox1) sequences in this taxon. High intraspecific cox1 divergences of 7–14% were observed (18% in one case) within the whole set of amphibian sequences analysed. These high values are not caused by particularly high substitution rates of this gene but by generally deep mitochondrial divergences within and among amphibian species. Despite these high divergences, cox1 sequences were able to correctly identify species including disparate geographic variants. The main problems with cox1 barcoding of amphibians are (i) the high variability of priming sites that hinder the application of universal primers to all species and (ii) the observed distinct overlap of intraspecific and interspecific divergence values, which implies difficulties in the definition of threshold values to identify candidate species. Common discords between geographical signatures of mitochondrial and nuclear markers in amphibians indicate that a single-locus approach can be problematic when high accuracy of DNA barcoding is required. We suggest that a number of mitochondrial and nuclear genes may be used as DNA barcoding markers to complement cox1.

Keywords: DNA barcoding; amphibia; cox1; 16S rRNA; Mantellidae; Aneides

1. INTRODUCTION

Amphibians are a vertebrate class that recently has been in the centre of research and public attention. This partly results from the phenomenon of multi-causal global amphibian declines, which in the most alarming cases occur through the spread of novel pathogens into undisturbed and even protected areas (Berger et al. 1998; Daszak et al. 2003), but in many cases the causes of decline are unknown (Stuart et al. 2004). While on one hand amphibian species are disappearing, on the other hand a large number of new species are being described every year (Hanken 1999). The rise of new species is remarkable (Glaw & Köhler 1998) and does not reflect taxonomic inflation in which known subspecies or variants are elevated to species status but is due to true first-hand discoveries (Köhler et al. 2005). Due to their dependence, in most cases, on both aquatic and terrestrial habitats, amphibians may be especially sensitive to environmental change (Stuart et al. 2004), and have been used as indicator species for habitat degradation (e.g. Welsh & Ollivier 1998; Alford & Richards 1999; Welsh & Droege 2001; Davic & Welsh 2004).

The morphology of amphibians is plagued with homoplasy (e.g. Emerson 1986; Wake 1991), and molecular phylogenetic analyses have uncovered several remarkable radiations, demonstrating that similar ecological and morphological adaptations have occurred in parallel, often in different regions of the world (e.g. Bossuyt & Milinkovitch 2000; Parra-Olea & Wake 2001; Mueller et al. 2004; Van der Meijden et al. in press).

A particular challenge to amphibian taxonomists is when patterns of convergence and parallelism occur among relatively closely related taxa (Wiens et al. 2003), which can completely mask species diversity. Many frog species are morphologically similar to other, partly sympatric taxa, but are strongly differentiated by advertisement calls and genetic divergences. Similarly, many species of salamanders are only reliably distinguishable by molecular methods (Larson & Chippindale 1993). Also, most amphibians have complex life cycles (Wilbur 1980), with a larval phase radically different in morphology (i.e. tadpoles of anurans). This complicates the identification and description of larvae, as they cannot be easily assigned to an adult phase based on their morphology (Thomas et al. 2005).

The plethodontid salamanders of North America have been in the centre of debate of species concepts
and the value of genetic distances for species delimitation (e.g. Highton 2000 versus Wake & Jokusch 2000). As summarized below, analysis of diverse data sets in amphibians indicates common discordances between nuclear and mitochondrial signatures, and distinct phylogeographic structuring of species disrupted by at least occasional hybridization and mitochondrial introgression. Mitochondrial divergences among unanimously recognized species are usually large, but in some cases intraspecific divergence values approach or overlap with interspecific values (Vences et al. 2005), potentially confounding species identifications based on a single sequence.

Altogether, these patterns define amphibians as a challenging group for DNA barcoding purposes. On one hand, the spread of amphibian declines and the indicator value of amphibians claim for a rapid assessment of their species diversity, and for fast and reliable species identification tools. On the other hand, the genetic structure of amphibian species may not in all cases allow reliable species identification through DNA barcodes based on cytochrome oxidase subunit I (cox1) sequences. The usefulness of molecular identification of amphibians has already been explored in the allozyme age, with electrophoretic methods developed to identify embryonic stages in ecological studies (Arntzen 1989) and frog leg meat in the international trade (Veith et al. 2000). Here we (i) briefly review patterns of mitochondrial variation in amphibians, (ii) define two major goals of DNA barcoding in amphibians, the identification of larval stages and of candidate species and (iii) provide novel data on cox1 variation in two intensively studied amphibian groups, mantellid frogs and plethodontid salamanders.

2. METHODS

A fragment of cox1 was sequenced in an array of frogs of the family Mantellidae using a pair of primers proposed for arthropods (Hebert et al. 2003): LCO1490, 5'-GGTCAAATCTATAAGATTG-3', and HCO2198, 5'-TAACTCAGGGTAGCAAAATAC-3' that amplify a region near the 5'-terminus of this gene. As expected (Vences et al. 2005) this pair of primers worked in a large proportion of species but not in all. In salamanders, amplification success was poor, and new primers were required; two primers that amplify a region of 1259 bp starting from the 5'-terminus of this gene were designed by DRV, based on available complete mitochondrial sequences: MVZ_201, 5'-TCAACAAYATAAATGAGGCACC-3' and MVZ_202, 5'-GGGTCTCTGGGTARCTCTGAAATCGTGC-3'. For comparative purposes, the fragment of the large mitochondrial ribosomal subunit 16S proposed by Vences et al. (2005) for amphibian DNA barcoding was sequenced from the same individuals, using the primers 16Sa-L and 16Sb-H (Palumbi 1991). Sequences obtained in this study were deposited in Genbank (accession numbers DQ105329-DQ105345, DQ116461-DQ116497). Our dataset was complemented by 16S and cox1 sequences obtained in a previous study (Vences et al. 2005): AY83978-AY83995, AY847959-AY848683, and by complete mitochondrial genome sequences available from Genbank.

To assess patterns of cox1 evolution, and intra- and interspecific variation in amphibians, we used three separate data sets. Selection of taxa was as follows.

(a) Mantellid frogs (family Mantellidae)

These are an endemic and species-rich clade from Madagascar that has been intensively studied using different data sets (morphology, bioacoustics and genetics). The occurrence of sibling species (species that are sister to each other, or very close relatives, but morphologically very similar) is very common in this clade and adds to the value of molecular approaches to understand their diversity. We, here, compared mantellids of three genera (Aglyptodactylus, Boophis, Mantidactylus) at three levels, extending the data set of Vences et al. (2005): (i) intrapopulation variation was assessed for six species (Aglyptodactylus madagascariensis, Boophis goudoti, Mantidactylus blommersae, M. brevipalma, M. enki, M. tichenki) by sequencing two to five specimens per population. (ii) Interpopulation variation was assessed by comparing individuals from mainly the two eastern rainforest regions Andasibe and Ranomafana; only species were selected where morphological and bioacoustic uniformity, and probably non-fragmented distribution areas, strongly suggest that populations from the study sites are conspecific (B. goudoti, B. luteus, B. septentrionalis, Mantidactylus asper, M. blommersae, M. liber, M. melanopleura, M. redimitus, M. tornieri); (iii) interspecific divergence was assessed by comparing several pairs of sibling species (B. erythrodactylus versus B. taymena; B. goudoti versus B. cf. perigeetes; B. luteus versus B. septentrionalis; B. sibilans versus an undescribed species B. sp.; M. blommersae versus M. domerguei; M. depressiceps versus M. tornieri).

(b) Climbing salamanders (genus Aneides)

We selected four species of Aneides, which present several characteristics that make them suitable for comparative analysis of cox1 variation. All belong to the same clade on the basis of their morphology and genetic relationships (Wake 1963, 1966; Larson et al. 1981; Jackman 1993; Mahoney 2001). They are morphologically distinguishable and their distributional ranges are well known, occurring in the same biogeographic region (western North America), with several recognized contact zones between the species. We selected one specimen per species from the type locality or the closest locality available in the Museum of Vertebrate Zoology (University of California at Berkeley, USA) tissue collection. One or several specimens were selected from the contact zones between species (A. vagrans versus A. flavipunctatus, A. vagrans versus A. lugubris in Mendocino County; A. flavipunctatus versus A. lugubris in Mendocino and Santa Clara counties; A. ferreus versus A. vagrans in Del Norte County (California, USA)). When possible, one specimen per species from a population distant to the distribution range of the other species was also sequenced. Aneides shardi was also included for comparison as it is currently recognized as the sister species to all other Aneides (Wake 1966; Mahoney 2001).

(c) Salamanders of which complete mitochondrial genomes have been sequenced

Four mitochondrial genes, 16S, cox1, NADH dehydrogenase subunit 4 (nad4) and cytochrome b (cob) were aligned from 33 mitochondrial genomes available from Genbank. These include mostly plethodontid salamanders, but also five closely related ambystomatid species, and altogether representatives of six salamander families: Cryptobranchidae,
Amphibians are known to be able to disperse across oceanic barriers (e.g. Vences et al. 2002). Discordance between the geographical signatures of mitochondrial and nuclear markers has also been extensively observed in amphibians, in the phylogeographical structure of one species (Garcia-Paris et al. 2003; Monsen & Blouin 2003; Kuchta & Tan 2005), contact zones of phylogeographical subgroups of one species (Sequeira et al. 2005) and hybrid zones between two species (Babik et al. 2003).

A special situation is that observed in hybridogentic and gynogenetic amphibians. Such phenomena are known in Palearctic water frogs (genus *Rana*, subgenus *Pelophylax*: hybridogenesis) and North American salamanders (genus *Ambystoma*: gynogenesis). These highly complex genetic systems may also lead to situations where mitochondrial phylogeny does not correspond to species phylogeny, and DNA barcoding would thus lead to wrong identifications, but these patterns are still insufficiently studied.

However, the available data indicate that problems of real non-monophyly of mitochondrial haplotypes in a species, through introgression, incomplete lineage sorting or other phenomena, are exceptions rather than the rule in amphibians, and—as in other animal groups—may be unusual enough to only rarely confound species identification. For example, in a large mitochondrial screening of mantellid frogs (Vences et al. 2005), introgression was observed in three out of 200–300 species (1–1.5%), which might be considered as an acceptable error margin, and the affected taxa were always closely related and largely allopatric species pairs. The lesson to be learnt from discordances of geographical signatures of nuclear and mitochondrial markers is that DNA barcoding in closely related, allopatric, and hybridizing taxa (be they considered as species or subspecies) should never rely on mitochondrial markers alone but should always include nuclear markers. Exploring the usefulness of nuclear ribosomal genes for this purpose appears to be promising (Tautz et al. 2003; Markmann & Tautz 2005; Monaghan et al. 2005).

Hence, the one major query to DNA barcoding in amphibians refers to their large mitochondrial variability within and especially among populations. Three potential problems could affect mitochondrial markers under these conditions: (i) the priming sites may be too variable to allow the use of universal primers in all species, or even in all populations of one species, (ii) the gene fragment may be too saturated with mutations to allow reliable assignment of genetically divergent populations to the correct species, (iii) the gene fragment may be too saturated to allow a distinction of conspecific sequences from sequences of a potentially new species. In a previous paper we have shown that the 16S rRNA gene appears to be a suitable barcoding marker for amphibians, and predicted some difficulties in the universal use of *cox1* in this animal group (Vences et al. 2005). In the following we will present patterns of *cox1* variation in amphibians and discuss major applications of DNA barcoding in amphibians.

### 4. THE COX1 GENE IN AMPHIBIANS

When compared with other mitochondrial genes, *cox1* has not been used in a particularly large number of
studies in amphibian phylogeny or phylogeography (e.g. in James & Moritz 2000; Rissler & Taylor 2001; Symula et al. 2001, 2003; Goldberg et al. 2004), and rarely as the only marker. Without doubt, the most commonly used genes include 16S, the small mitochondrial ribosomal subunit (12S) and cob. In April 2005 there were 537 hits in Genbank when searching for \( \text{cox1} \) in amphibians, whereas there were 3641 hits for cob, 3301 for 16S and 2316 for 12S.

An important question addressed by the present study is how variable is \( \text{cox1} \) in amphibians. Several lines of evidence indicate that this gene is highly variable. Vences et al. (2005) analysed standard priming sites for the 3′-terminal segment based on 10 complete mitochondrial sequences of frogs, salamanders and caecilians, and found that the variability (restricted to 3rd codon positions) among these amphibians was higher compared to 59 sequences from an array of taxa spanning across all vertebrates. James & Moritz (2000) regularly observed, in the 3′-terminal \( \text{cox1} \) fragment of the Australian sedge frog \( \text{Litoria fallax} \), pairwise divergences higher than 5% among haplotypes from neighbouring populations, and 11–12% between two major haplotype clades within the species. The high variability of this gene was also obvious from the fact that, in the study of James & Moritz (2000), 87 individuals showed 84 unique haplotypes. Data presented here (figure 1) indicate that in mantellid frogs, 10–14% divergence is regularly found within species and up to 18% might be possible. Our data also indicate up to 7.8% uncorrected divergence within species of climbing salamanders (figure 1). Compared to the 16S rRNA gene, and below 20% uncorrected divergence, the \( \text{cox1} \) substitution rates in both salamanders and frogs appear to be about two times higher than the 16S divergences (figure 2).

A striking example of the species definition issue in amphibians is \( \text{Ensatina eschscholtzii} \), a salamander that in many respects can be understood as a ring species (e.g. Moritz et al. 1992; Wake 1997). Genetic admixture is observed throughout the ring, except for some areas of secondary contact where strongly differentiated subspecies behave as distinct species in sympathy. We compared the 3′ terminal portion of the available \( \text{cox1} \) sequence from the complete mitochondrial of \( \text{Ensatina e. eschscholtzii} \) with that of newly obtained sequences for \( E. \) eschscholtzii platensis, and found an uncorrected pairwise divergence of 12%. This provides a further example of very high \( \text{cox1} \) divergences within amphibian units that are considered as species, although this is not unanimous in this case (e.g. Highton 1998).

Compared to the situation at least in birds (Hebert et al. 2004b), and probably also in some insects (e.g. Hebert et al. 2004a), these high intraspecific mitochondrial divergences are certainly a striking character of amphibians. However, several other animal groups may exhibit equivalent rates of variation as amphibians. Very high intraspecific haplotype divergences have been observed in pulmonate snails (Thomaz et al. 1996), and among lizard populations assumed to be conspecific, e.g. in \( \text{Tarentola} \) (Harris et al. 2004), or in insular populations of the gecko \( \text{Cyrto-}
\text{dactylus kotschyi} \) (up to 20% \( \text{cox1} \) divergence; Kasapidis et al. 2005). Although such lineages may turn out to be separate species in the future, there is little doubt that they are closely related to each other and morphologically largely conserved. This latter study also found intrapopulational divergences of up to 7% in these geckos, which is higher than any value observed in amphibians so far.

To test whether the large \( \text{cox1} \) divergences encountered in amphibians were due to an especially fast molecular evolution of this gene in this taxon, we compared substitution rates in this gene with those in two other mitochondrial genes commonly used in amphibian phylogenetics. Analysis of the data set of complete mitochondrial genomes of salamanders (figures 2 and 3) indicates that mutations at first and second positions are rare as compared to other mitochondrial protein-coding genes. Indeed, \( \text{cox1} \) reaches a plateau of saturation faster and, therefore, is less variable than the two other mitochondrial protein-
Putative comparatively faster evolution, in this taxon, of divergence values found in amphibians are not due to a rates of amphibians there are distinct differences in substitution coding genes.

Also in divergences of other mitochondrial protein-
coding genes. In divergences of other animals (Saccone et al. 1999), hence, the high mitochondrial variability of amphibians is certainly the species identification of larvae. Especially the tadpoles of frogs are morphologically highly divergent from their adult stages (Altig & McDiarmid 1999), a situation comparable with that in holometabolous insects. These larvae can be significant components of aquatic ecosystems (e.g. Ranvestel et al. 2004), and at certain periods of the year are the only available evidence for the occurrence of certain secretive species at a site. Amphibian larvae may also contain relevant taxonomic and, especially, phylogenetic information (e.g. Haas 2003). For such studies, it is of course necessary to collect some basic information: which tadpole belongs to which species, how can it be recognized, and which morphological adaptations does it have? Such fundamental knowledge is scarce in species-rich tropical amphibian communities, because identifying tadpoles to species, or even genus, is an extremely time-consuming task. It involves either (i) laborious rearing of eggs laid by a well-identified pair of adults, or (ii) even more laborious rearing of tadpoles collected in the wild, to obtain metamorphosed juveniles, which then are tentatively assigned to one of the species known to occur at the site. Parmelee et al. (2002) discuss these problems of morphological tadpole identification, and anticipate that rapid molecular DNA techniques suitable for field identification will be available within the next few decades. DNA barcoding clearly offers these tools and has been applied in several studies to successfully identify tadpoles (Malkmus & Kosuch 2000; Ziegler 2002; Ziegler & Vences 2002; Thomas et al. 2005).

The high mitochondrial variability of amphibians may be the source of wrong identifications of tadpoles, if reference and test specimens have different geographical origins. If the within-species differentiation approaches saturation, phenetic and phylogenetic comparisons may not be able to accurately assign

Figure 3. Boxplots of uncorrected pairwise distances for cox1 and 16S in mantellid frogs, and for cox1, 16S and two other genes commonly used in amphibian studies (cob and nd4), based on analysis of 33 complete salamander mitochondrial genomes. Boxes represent mean (black or white line) plus/minus standard deviation; error bars represent 5 and 95% percentiles; stars represent outliers.

5. IDENTIFYING LARVAL AMPHIBIANS

One potential major application of DNA barcoding in amphibians is certainly the species identification of larvae. Especially the tadpoles of frogs are morphologically highly divergent from their adult stages (Altig & McDiarmid 1999), a situation comparable with that in holometabolous insects. These larvae can be significant components of aquatic ecosystems (e.g. Ranvestel et al. 2004), and at certain periods of the year are the only available evidence for the occurrence of certain secretive species at a site. Amphibian larvae may also contain relevant taxonomic and, especially, phylogenetic information (e.g. Haas 2003). For such studies, it is of course necessary to collect some basic information: which tadpole belongs to which species, how can it be recognized, and which morphological adaptations does it have? Such fundamental knowledge is scarce in species-rich tropical amphibian communities, because identifying tadpoles to species, or even genus, is an extremely time-consuming task. It involves either (i) laborious rearing of eggs laid by a well-identified pair of adults, or (ii) even more laborious rearing of tadpoles collected in the wild, to obtain metamorphosed juveniles, which then are tentatively assigned to one of the species known to occur at the site. Parmelee et al. (2002) discuss these problems of morphological tadpole identification, and anticipate that rapid molecular DNA techniques suitable for field identification will be available within the next few decades. DNA barcoding clearly offers these tools and has been applied in several studies to successfully identify tadpoles (Malkmus & Kosuch 2000; Ziegler 2002; Ziegler & Vences 2002; Thomas et al. 2005).

The high mitochondrial variability of amphibians may be the source of wrong identifications of tadpoles, if reference and test specimens have different geographical origins. If the within-species differentiation approaches saturation, phenetic and phylogenetic comparisons may not be able to accurately assign
When comparing various vertebrate taxa to each other, the **cox1** gene was not able to recover most major clades (Vences et al. 2005), and would not have been able, for example, to reliably identify an unknown amphibian sequence as belonging to this class. To test the performance of **cox1** in phenetic identification of amphibians, we carried out two analyses using the BLAST algorithm (Altschul et al. 1990). We used a database that contained all vertebrate **cox1** sequences available from Genbank as of April 2005 (N = 1563). This database contained a number of amphibian sequences, in particular those obtained by Vences et al. (2005). We removed all 13 sequences of mantellid frogs, and then 'BLASTed' our complete set of mantellid frog sequences against the database. In 23 out of 40 searches, the first hit did not refer to an amphibian sequence, results spanning from fishes over birds to mammals. In a second search we added our mantellid sequences from one locality in Madagascar to the vertebrate database, and 'BLASTed' the conspecific sequences from other localities against this expanded database. In 21 out of 22 searches, the searches were successful in identifying a conspecific sequence from another locality as the most similar to the query sequence. These results confirm that **cox1** does not perform well in assigning specimens to major vertebrate lineages when taxon sampling is poor, but corroborates that identification becomes reliable with dense taxon sampling in the reference database. As a conclusion, 16S (Vences et al. 2005) and **cox1** are suitable markers to identify unknown life-history stages of amphibians (eggs, larvae, juveniles, the opposite sex) to species. A crucial aspect for such applications is the development of new software tools to allow fast and reliable comparisons across large sequence sets. Besides pairwise alignments of indel-rich rRNA sequences (e.g. Steinke et al. 2005) this could encompass a better visualization of BLAST output, e.g. in the form of guiding trees or scatterplots, and simultaneous consideration of multiple markers.

### 6. SCREENING FOR CANDIDATE SPECIES

Considering the high rate of discovery of new species of amphibians in times of global amphibian declines, DNA barcoding can be a useful tool to speed up the initial recognition of new units that may represent undescribed species, here termed candidate species (see figure 5). For this purpose, it is necessary to define threshold values that ideally provide a sharp distinction between intraspecific and interspecific divergence values. If an unknown sequence differs from the closest reference sequence by a divergence above the threshold, the individual from whom the sequences were obtained belongs to a candidate species, which means that its taxonomic status merits further investigation. Bradley & Baker (2001), for mammals, set this threshold at 11% for the cytochrome b gene, whereas Hebert et al. (2004b) propose a **cox1** threshold of only 2.7% for birds. These authors propose to calculate a standard sequence threshold as 10 times the mean intraspecific variation observed. In our data sets, mean intraspecific **cox1** divergence was 5.4% in mantellids and 4.3% in *Aneides*, whereas the mean interspecific divergences were 20.7% and 13.3%, respectively.
A calculation of threshold values as proposed by Hebert et al. (2004b) would yield unrealistic values of 43–54% pairwise divergence, which is much above the saturation plateau of cox1 and exceeds the highest divergence values observed among any pair of amphibian species (e.g. figure 3). Threshold values for amphibians can therefore only be tentative, and may be placed around a value of 5% 16S divergence (Vences et al. 2005) or 10% cox1 divergence (figure 1).

The major problem with defining this threshold value is the wide range of overlap between intra- and interspecific divergence values, which seems to be a generalized problem in amphibians (figure 1). In climbing salamanders, the two species with the smallest interspecific divergences are Anelides ferreus vs. A. vagrans (6.5–7.9%), whereas the largest interspecific divergences (above 7%) were found in comparisons of the Santa Clara sequence with those from other populations of A. flavipunctatus. In mantellid frogs, many species had divergence values among populations that were largely overlapping with those of closely related allopatric species pairs.

The selection of taxa and specimens to calculate averages is a further issue. Figure 1b is based on all pairwise comparisons in the set of available mantellid sequences. The amount of overlap between intraspecific and interspecific divergences appears relatively limited and far from the respective average values. In contrast, figure 1c shows only the mean values for intraspecific comparisons (one value per species), and only the interspecific comparisons of closely related sibling species where identification problems are most acute. As easily visible from the graph, the overlap of values is much more pronounced in figure 1c, which suggests that thorough statistical tests are needed to estimate the probabilities of correct identification of species and candidate species, using different threshold levels of genetic divergences.

7. CONCLUSIONS
Based on data presented and reviewed herein there seems to be no convincing evidence for mitochondrial introgression and incomplete lineage sorting being much more common in amphibians than in other animals (see also Funk & Omland 2003). The cox1 gene shows high divergences within and among amphibian species, but this is due to general high mitochondrial variability rather than to a particularly fast evolutionary rate of this gene. Despite this high variability, cox1 seems to be able to correctly identify sequences from different localities to the species level. The major problems with DNA barcoding of amphibians are related to this high mitochondrial variability. First, there is a distinct overlap of intraspecific and interspecific divergence values, which complicates the establishment of threshold values to identify candidate species. Second, because high variation is also observed in the cox1 priming sites, a mix of several primers will be needed to reliably amplify this gene from all amphibian species, and the use of alternative markers with more conserved priming sites, such as 16S rRNA, should be considered, at least for some applications (Vences et al. 2005).

These conclusions may hold not only for amphibians, but also for other animal groups. In addition, mitochondrial introgression may return as a major problem in cases when identifications through DNA barcodes must have very high reliability. Depending on the required level of accuracy, nuclear barcoding markers need to be established to be able to corroborate any disputed mitochondrial species identification.

The cox1 gene has been proposed as a standard marker, and we support attempts to build up a global and complete cox1 database of eukaryotes, except plants (see Chase et al. 2005). However, there also seems to be consensus that additional markers will be helpful, and needed for certain applications. Their number will probably be limited. Besides cox1, and leaving aside the situation in plants, genes used for molecular taxonomy approaches (although usually not under the term DNA barcoding) include the mitochondrial 16S rRNA (which is being sequenced by the AmphibiaTree consortium for a large set of amphibian species at present, see http://www.amphibiatree.org),
REFERENCES


Funk, D. J. & Omland, K. E. 2003 Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Syst. 34, 397–423. (doi:10.1146/annurev.ecolsys.34.011802.132421.)


Haas, A. 2003 Phylogeny of frogs as inferred from primarily larval characters. Cladistics 19, 23–89.


Phil. Trans. R. Soc. B (2005)


The problems and promise of DNA barcodes for species diagnosis of primate biomaterials

Joseph G Lorenz, Whitney E Jackson, Jeannie C Beck and Robert Hanner

Phil. Trans. R. Soc. B 2005 360, 1869-1877

References

This article cites 31 articles, 7 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1869.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1869.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
The problems and promise of DNA barcodes for species diagnosis of primate biomaterials

Joseph G. Lorenz*, Whitney E. Jackson, Jeanne C. Beck and Robert Hanner

Coriell Institute for Medical Research, 403 Haddon Avenue, Camden, NJ 08103, USA

The Integrated Primate Biomaterials and Information Resource (www.IPBIR.org) provides essential research reagents to the scientific community by establishing, verifying, maintaining, and distributing DNA and RNA derived from primate cell cultures. The IPBIR uses mitochondrial cytochrome c oxidase subunit I sequences to verify the identity of samples for quality control purposes in the accession, cell culture, DNA extraction processes and prior to shipping to end users. As a result, IPBIR is accumulating a database of ‘DNA barcodes’ for many species of primates. However, this quality control process is complicated by taxon specific patterns of ‘universal primer’ failure, as well as the amplification or co-amplification of nuclear pseudogenes of mitochondrial origins. To overcome these difficulties, taxon specific primers have been developed, and reverse transcriptase PCR is utilized to exclude these extraneous sequences from amplification. DNA barcoding of primates has applications to conservation and law enforcement. Depositing barcode sequences in a public database, along with primer sequences, trace files and associated quality scores, makes this species identification technique widely accessible. Reference DNA barcode sequences should be derived from, and linked to, specimens of known provenance in web-accessible collections in order to validate this system of molecular diagnostics.

Keywords: primate; DNA barcoding; mtDNA

1. INTRODUCTION

The advent of molecular techniques has opened new possibilities for taxonomic research, which is important given that the vast majority of all extant species are not well characterized morphologically. Taxonomies based on morphological analyses can be problematic due to either convergence in phenotype among unrelated species or the failure to identify ‘cryptic species’ where morphologic divergence has not kept pace with genetic divergence.

In an effort to standardize the approach to species identification using molecular techniques it has been proposed that as many species as possible be characterized for the same genetic markers (Blaxter 2004). The mitochondrial gene, cytochrome c oxidase subunit I (cox1) has been proposed as a candidate locus given its conserved sequence allows for ‘universal’ primers to be used across multiple divergent taxa and its high degree of phylogenetic signal relative to other mtDNA loci that have been used for interspecific analysis (e.g. 12s or 16s ribosomal DNA). This feature is perhaps due to heavy stabilizing selection within species for mitochondrial/nuclear cytochrome protein complexes (Hebert et al. 2003a,b). Thus a sequence of several hundred nucleotides in length acts as a unique identifier for members of a given species, hence the analogy to a computerized barcode label, although the analogy is imperfect given the existence of intraspecific variation so that not all members of a species are expected to be completely identical. Nonetheless the degree of intraspecific variation compared to the degree of interspecies variation would be expected to be low enough such that sequences from polymorphic species would cluster together in a genetic distance based cluster analysis.

Early studies demonstrated the efficacy of using universal primers to amplify and sequence a variety of taxa from divergent phyla (Hebert et al. 2003a), but the use of cox1 sequences as species identifying barcodes has been limited to a study of North American birds (Hebert et al. 2004a) and one complex species of neotropical butterflies (Astraptes fulgerator) from Costa Rica (Hebert et al. 2004b); but see e.g. Janzen et al. (2005), Saunders (2005), Smith et al. (2005) and Armstrong & Ball (2005) for recent examples. To date, the efficacy of molecular barcoding has not been determined within mammalian taxa.

Primates is a mammalian order with worldwide distribution, the members of which are important in conservation, evolutionary and biomedical studies. The taxonomic classification of extant species of the primate order has been agreed upon for several decades with a few interesting exceptions (Le Gros Clark 1954). The placement of tarsiers and the demarcation of the groupings among the hominoidea (including the genus Homo) are two of those exceptions that have been the focus of extensive taxonomic reorganization. Even with the overall structure of primate taxonomy in place there remains much work to be done in understanding the relationships of closely related taxa within many of the major groupings (Ruiz-Garcia & Alvarez 2003; Singer et al. 2003).

In addition, to elucidating relationships among the lower taxonomic levels of primates, there are practical
aspects to DNA barcoding. The bush meat trade is threatening many wild populations of primates and other endangered species (Brashares et al. 2004). To effectively prosecute poachers trafficking in meat it would be beneficial to law enforcement and conservation officials to have access to a forensic database of primate cox1 samples so that positive identification of seized contraband could be made. Given that many non-human primates are important in biomedical research (Vandeberg & Williams-Blangero 1996, 1997) it has become increasingly desirable to genetically characterize the various species used in research. Thus far, there are species-specific differences (even strain differences within species) to pathological aetiology and temperamental differences that may be important in behavioural research (Champoux et al. 1994; Champoux et al. 1997). In addition biomaterial repositories would have a simple and universal means of verifying the species identity of samples submitted to them for inclusion in their collections. Finally, the ability to identify or verify the source of biomaterials from field-collected specimens may be a useful tool to conservation and range scientists as well as a means of identifying the geographical provenience of captive born animals.

Past efforts to collect, store and develop genetic resources have largely been uncoordinated efforts scattered over different institutions and countries (Savolainen & Reeves 2004). Primates are a target group for scientific and technological development due to their importance in biomedicine and conservation biology, especially given their evolutionary proximity to humans. Within this context, the collection and storage of primate resources covering all branches of their taxonomy is an urgent need to boost primate molecular biology. Such resources include living and preserved collections, tissues, DNA, frozen viable cells and cell lines. The storage of complementary information on the origin, morphology, physiology, ecology, demography or behaviour of the specimens is also crucial to explore the link between gene and function. The coordinated development of these resources will prevent repeated sampling of wild populations, reduce the number of animals used in research, and help to standardize molecular tools and protocols.

The Integrated Primate Biomaterials and Information Resource (www.IPBIR.org) provides essential research reagents to the scientific community by establishing, verifying, maintaining and distributing DNA and RNA derived from primate cell cultures. Proper quality assurance/quality control requires the ability to verify the identity of samples as they move through the accession, culture and extraction processes. At present IPBIR has 97 of the approximately 200 different species (Cheney et al. 1986) of primates representing most of the major taxonomic divisions of the order. Given the taxonomic breadth of the samples in the resource it is important to choose a molecular marker that would work in as many species as possible. IPBIR uses DNA sequences for routine identification of non-human biomaterials by simple sequence matching. The GenBank public-access database provides a working archive of available sequences, forming a valuable resource for such studies. However, the accuracy and reliability of sequences deposited in GenBank have been questioned, especially when the sequences are not linked to a voucher specimen. The identity of some DNA sequences deposited in public databases is being contested and there is a need to determine if such reports reveal a widespread phenomenon (Bridge et al. 2003).

The publicly available cox1 DNA sequences are distinct for each primate species, but represent a very incomplete data set. Moreover, sequence information can be difficult to interpret for several reasons. First, different levels of variation may occur in the same DNA region in different taxa, making generalized comparisons between taxa problematic. Second, amplification of non-target DNA from contaminants, or numts (nuclear mitochondrial DNA sequences), constitutes a danger. It has been proposed that as many as 500 copies of translocated mtDNA exist in the human nuclear genome (Richly & Leister 2004), ranging in length from 47 to 14 654 base pairs (bp) (Ricchetti et al. 2004). Hence, caution must be taken when amplifying any mitochondrial segment short of the entire mitochondrial genome. Third, sequences derived from unspecified reference materials cannot be validated.

In this study we investigate the feasibility of using cox1 sequence ‘molecular barcode’ data to verify the species designation of 225 individuals representing 56 species of primates (table 1) using both ‘universal’ cox1 primers identified in earlier molecular barcoding studies (Hebert et al. 2003a) as well as primers developed specifically for primate taxa.

2. MATERIAL AND METHODS

(a) DNA Extraction

Genomic DNA was extracted from cell pellets obtained from lymphoblastoid cell cultures, fibroblastoid cell cultures oruffy coats obtained from whole blood using standard salting-out techniques (Miller et al. 1988) or from tissue biopsies or plasma/serum using QiaAmp DNA Blood Kits (Qiagen).

(b) PCR Amplification

A region approximately 727-bp long near the 5′ terminus of the cox1 gene was amplified using one of three primer sets (table 2). PCR reactions were done in a total volume of 25 µL and consisted of 2.5 µL of 10× PCR II buffer (Applied Biosystems), 2.5 µL of 25 mM MgCl₂, 2.0 µL of 10 mM dNTP mix (2.5 mM each dNTP), 0.2 µL of each primer (25 µM stock) and 0.2 µL of 5 U/µL TaqGold DNA polymerase (Applied Biosystems), 2.0 µL of DNA template (~50 ng) and dH₂O to 25 µL. The thermocycling conditions were as follows: 96.0°C for 10 min to activate the TaqGold and then 35 cycles of 96.0°C for 1 min, 50.0 to 58.0°C for 1 min and 72.0°C for 1 min followed by a final hold of 72.0°C for 10 min. PCR products were visualized on 6% polyacrylamide minigels and the PCR product was purified using QiaQuick 96 PCR Purification Kit (Qiagen). The purified PCR product was eluted into 45 µL buffer AE (Qiagen).

(c) DNA Sequencing

Cycle sequencing reactions were carried out in 10 µL total volume. A forward and reverse reaction was performed for each sample consisting of 5.4 µL of the purified PCR

Phil. Trans. R. Soc. B (2005)
product, 4 µL of Big Dye Terminator Ready Reaction Mix, v1.1(Applied Biosystems) and 0.6 µL of 2.5 mM primer (same as PCR primer). Cycling conditions were 96.0 °C for 1 min and then 25 cycles of 96.0 °C for 10 s, 50.0 °C for 5 s, 60.0 °C for 4 min. Unincorporated fluorescent dye terminators were removed from the cycle sequencing reactions using SigmaSpin 96-well plate (Sigma Aldrich), dried at 37 °C for 20 min and resuspended in 10 µL of HiDi formamide (ABI). The cycle sequencing product was detected using an ABI 3730 DNA analyser.

(d) Sequence analysis
The forward and reverse sequence files for each sample were analysed and a consensus sequence for each sample was

<table>
<thead>
<tr>
<th>infraorder</th>
<th>suborder</th>
<th>family</th>
<th>subfamily</th>
<th>genus</th>
<th>species</th>
<th>number in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haplorhini</td>
<td>Catarrhini</td>
<td>Cercopithecidae</td>
<td>Cercopithecinae</td>
<td>Allenopithecus</td>
<td>nigroviridis</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cercopithecus</td>
<td>torquatus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cercopithecus</td>
<td>ascanius</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cercopithecus</td>
<td>cephus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cercopithecus</td>
<td>lhoesti</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cercopithecus</td>
<td>mits</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chlorocebus</td>
<td>neglectus</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lophocebus</td>
<td>petaurista</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Macaca</td>
<td>wolfi</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mandrillus</td>
<td>leucophaeus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Macaca</td>
<td>sphinx</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mandrillus</td>
<td>talapoin</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Macaca</td>
<td>anubis</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Macaca</td>
<td>cynocephalus</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Theropithecus</td>
<td>hamadryas</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Colobinae</td>
<td></td>
<td></td>
<td>Theropithecus</td>
<td>guereza</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Theropithecus</td>
<td>polykomos</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Nasalis</td>
<td>larvatus</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Semnopithecus</td>
<td>entellus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Trachypithecus</td>
<td>francoisi</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>gabriellea</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>agilis</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>lar</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>leucogenys</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>pileatus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>sindactylus</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>agilis</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>lar</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>leucogenys</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hyllobates</td>
<td>pileatus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Theropithecus</td>
<td>sindactylus</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Colobus</td>
<td>guereza</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Hominidae</td>
<td></td>
<td></td>
<td>Colobus</td>
<td>polykomos</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pan</td>
<td>paniscus</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gorilla</td>
<td>gorilla</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pongo</td>
<td>pygmaeus</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Platyrrhini</td>
<td>Callitrichidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Callimico</td>
<td>goeldii</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Leontopithecus</td>
<td>rosalia</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Saginus</td>
<td>fuscicollii</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Saginus</td>
<td>micas</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Callimico</td>
<td>goeldii</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cebidae</td>
<td>Atelinae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ateles</td>
<td>geoffroyi</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lagothrix</td>
<td>lagotricha</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Aotus</td>
<td>nancymaeae</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pithecia</td>
<td>pithecia</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Saimiri</td>
<td>boliviensis</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Saimiri</td>
<td>oerstedii</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Strepsirhini</td>
<td>Daubentoniidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Daubentonia</td>
<td>madagascariensis</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Eulemur</td>
<td>fuscus</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Eulemur</td>
<td>mozgo</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Varecia</td>
<td>variegata</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Galago</td>
<td>moholi</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Otomis</td>
<td>garnetti</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Galago</td>
<td>moholi</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Otomis</td>
<td>garnetti</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>225</td>
</tr>
</tbody>
</table>
different sequences. The paucity of whole mitochondrial products obtained using the different primers yielded previously archived (http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/9347.html) as well as pre-established and deposited in GenBank (http://www.ncbi.nlm.nih.gov/). The whole mitochondrial DNA sequence has been compared them with (Andrews & Easteal 2000; (2005) et al. 2004). In addition we constructed neighbour-joining tree (figure 2) based on Kimura 2p distances (Kimura 1980) was calculated using PAUP, bootstrap values are based on 1000 replicates.

The reverse transcription PCR experiment showed that primer sets 2 and 3 amplified \( \text{cox1} \) from RNA extracts of \( P. \text{anubis} \) lymphoblasts (figure 1). Primer set 1 did not amplify a fragment from the RNA preparation indicating that the Primer set 1 fragment amplified from genomic DNA is not from the mitochondrial genome but is possibly from a numt. The results of the sequencing analysis for each of the samples were submitted to GenBank (Accession numbers: AY544148–62, AY632376–7, AY671787–98, AY673675, AY972630–808). The sequences were aligned with \( \text{cox1} \) sequences obtained from primate whole mitochondrial genomes obtained from GenBank as well as primate \( \text{cox1} \) sequences from other studies that have been deposited in GenBank. Sequences less than 400 bp long were excluded from the analysis. A neighbour-joining tree (figure 2) based on Kimura 2p distances (Kimura 1980) was calculated using PAUP, bootstrap values are based on 1000 replicates.

The neighbour-joining tree generally agrees with the commonly accepted primate phylogeny with platyrhine sequences clustering together 99% of the time and catarrhines cluster 95%. The strepsirhines, however, do not form a cohesive cluster; this is not unexpected given the relatively short fragment used in the analysis. There are 17 \( \text{cox1} \) sequences from previous studies that were retrieved from GenBank. In ten of the cases the sequences from this study clustered with \( \text{cox1} \) sequences obtained from GenBank at the species level (table 3). In four cases the sequences from this study were not represented in GenBank by the same species, however they did cluster together with GenBank derived sequences at the generic level (Saimiri, Macaca sylvanus, Trachypithecus and Galago). In the remaining three cases the GenBank sequences did not cluster with the sequences derived from the same species in this study: i.e. Colobus polykomos and C. guereza did not cluster separately, Papio hamadyras from GenBank clustered with \( P. \text{anubis} \) in this study and Theropithecus gelada from GenBank did not cluster with \( T. \text{gelada} \) from this study nor did it cluster with any of the cercopithecine \( \text{cox1} \) sequences.

In the cases for which there are no previously reported data to which we can compare the sequences generated in this study we tallied the extent to which

---

Table 2. Primer set success by genus.

<table>
<thead>
<tr>
<th>taxa</th>
<th>primer set 1</th>
<th>primer set 2</th>
<th>primer set 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pan</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Gorilla</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Pongo</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Hylobates</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Papio</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Cercopithecus</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Macaca</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Mandrillus</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Chlorocebus</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Allenopithecus</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Leontopithecus</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Saimiri</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Pithecia</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Varecia</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Daubentonia</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

created using ‘Sequencher’ (GeneCodes). The consensus sequences were in turn aligned using ‘Sequencher’ and exported into a NEXUS file for distance analysis and cluster analysis using PAUP 4.0b10 (Swofford 1998). Mean pairwise differences were computed for all species and genera. In order to validate the \( \text{cox1} \) sequences obtained in this study we compared them with \( \text{cox1} \) sequences from primate species for which the whole mitochondrial DNA sequence has been established and deposited in GenBank (http://www.ncbi.nlm.nih.gov/GENOMES/ORGANELLES/9347.html) as well as previously archived \( \text{cox1} \) sequences (Andrews & Easteal 2000; Wu et al. 2000). In addition we constructed neighbour-joining trees using PAUP and based on Kimura 2p distances to determine whether the sequences cluster as would be expected based on overall taxonomic affinity.

(e) \( \text{rtPCR} \)

In the case of several old world monkey species, PCR products obtained using the different primers yielded different sequences. The paucity of whole mitochondrial sequences for cercopithecines limited our ability to determine which of the \( \text{cox1} \) sequences were derived from mitochondrial \( \text{cox1} \) and which were derived from numts. Reverse transcriptase PCR (rtPCR) was performed on \( P. \text{anubis} \) DNA obtained from lymphoblasts to ensure amplification of the transcribed mitochondrial \( \text{cox1} \) (Collura et al. 1996). Reverse transcription products were amplified in three separate reactions using the three primer sets. The RNA extractions were also amplified with AmpliTaq in place of reverse transcriptase to identify any DNA contamination and a positive control rtPCR reaction was also performed using primers for phosphoglycerate kinase (PGK).

3. RESULTS

All samples in this study amplified with one or more primer sets (table 2). Generally, if samples representing a given species amplified with Primer set 1 (Hebert et al. 2003a) and gave ‘phylogenetically credible’ results they were not assayed with the other two primer sets. In some cases where a species did not amplify with one primer but did with another they were tested with the third primer set in order to increase the probability the sequence was not obtained from numts (Thalmann et al. 2004 and references therein) or other spurious amplicons. In addition to checking clustering for an expected phylogenetic signal to determine whether a sequence was derived from the actual \( \text{cox1} \) and not a numt we checked to see that the fragments were identical to, or at least clustered with, the appropriate whole mitochondrial sequences obtained from GenBank. Also, sequences which yielded amino acid transcription interrupted by stop codons were not included in the analysis as they would not likely be derived from the functional mitochondrial \( \text{cox1} \) gene. This did occur in a handful of cercopithecine species but the design of primer set 3 did eliminate amplification of spurious \( \text{cox1} \) in those species.

The results of the reverse transcription PCR are listed in table 2. As expected, all species were successfully amplified with Primer set 1. In general, it was possible to amplify all species with Primer set 2, and Primer set 3 provided a useful number of additional species. The results of the reverse transcription PCR were confirmed by traditional PCR using the primers listed in table 1.

In the cases for which there are no previously reported data to which we can compare the sequences generated in this study we tallied the extent to which
our sequences cluster with members of the same species (table 4). There are 10 clusters supported in 100% of the bootstrap replicates in which all members of a given species cluster together. In fact, except for *Papio* and *Colobus* as mentioned above, there are no cases where we have multiple specimens of a species in which the sequences do not cluster together.

### 4. DISCUSSION

Using a segment of the 5′ region of *cox1* we are able to identify the appropriate species from which a biomaterial submitted to the Integrated Primate Biomaterials and Information Resource was derived. This ability to generate a ‘molecular barcode’ is useful in our case for quality control and the management of the IPBIR. It allows us to verify the identity of samples, as reported by the submitter, as they move through each stage of the accession, cell culture, DNA extraction and aliquoting processes. Since the samples comprising the IPBIR collection are from identified specimens of known species, the *cox1* sequences derived from the IPBIR resource have the potential to serve as a forensic database for the identification of primate biomaterials such as those seized in the bush meat trade.

The success of DNA barcoding depends on the amount of intraspecific variation relative to the amount of interspecific variation present among species across their range. The amount of intraspecific variation, measured as mean pairwise difference, in the present study varies (figure 3) from none (*Eulemur mongoz*, *Leontopithecus rosalia*, *Cercopithecus neglectus*) to 0.038 for *Pongo pygmaeus* (mean for all species =0.011, s.e. =0.004). The amount of intraspecific variation at *cox1* will depend on the degree to which the samples reflect the geographic diversity of widely dispersed species (e.g. *Pan troglodytes* and *Chlorocebus aethiops*), the amount of gene flow among subpopulations and also whether the species is perhaps an amalgam of multiple species. For example it has been argued that the Sumatran (*P. pygmaeus abelii*) and Bornean orangutans (*P. p. pygmaeus*) actually are distinct species *(Xu & Arnason 1996; Warren et al. 2001)*; in fact the orangutans in this study cluster robustly by subspecies (100% of bootstrap replicates) and the degree of genetic divergence is comparable to that that exists between *Pan paniscus* and *P. troglodytes*, which lends support for the case that the genus *Pongo* includes two separate species.

It is interesting to note that there are three cases where the *cox1* sequences derived from the complete mitochondrial genome sequences did not cluster with the same species from this study. In the first case GenBank sequence gi:4049475 is listed as being obtained from *P. hamadryas* but the two *P. hamadryas* from this study PR00440 and PR00559 cluster together outside of the *P. cynocephalus* / *P. anubis* group whereas gi:4049475 clusters with *P. anubis* samples. Since it is known that *P. anubis* and *P. hamadryas* do interbreed *(Szmulewicz et al. 1999)* it is possible that the GenBank sample represents a hybrid individual. In any event this study supports Newman et al.’s (2004) finding that *P. cynocephalus* and *P. anubis* are not monophyletic clades but rather cluster together.

The second case involves *C. polkomos* (gi:4239860) and *C. guereza* (gi:60392100). These two *cox1* sequences were submitted independently to GenBank; they differ at only 8 positions out of the 1545 bp that constitute the whole *cox1* sequence, in fact in the first 700 bp which constitute the region used for molecular barcoding they differ at only a single position. Clearly these two individuals fall within the range of variation of a single species. One of the samples from this study (PR00655, *C. guereza*) clusters 100% of the time with these two GenBank sequences. Two other samples (PR00597, *C. polkomos* and PR00980, *C. guereza*) also cluster together 100% of the time. The two *Colobus* samples cluster together as well 100% of the time. Thus we will need to determine whether the species named associated with these samples are indeed correct or whether *cox1* will not differentiate these species as in the *P. anubis/P. cynocephalus* case.

In the third case a specimen from GenBank (gi:12484065, listed as *T. gelada*) does not cluster with the *T. gelada* from this study nor does it even cluster with other African cercopithecids, perhaps representing a numt, contamination or a misidentification of the original sample. This sample points out the importance of having barcode sequences linked to a morphologically vouched specimen so that such anomalies can
Figure 2. Bootstrapped neighbour-joining tree calculated from Kimura 2p distances and based on 1000 replicates. The number in parentheses indicates the number of samples that share identical sequences at that position; #, indicates the position of cox1 sequences obtained from GenBank.
be resolved by returning to the original sample (Ruedas et al. 2000).

The generation of incorrect sequences appears to happen for several reasons, not the least of which being the misidentification of the original material. Other problems involve primer specificity and the amplification or co-amplification of numts. Still other problems for the repository involve contamination of the cell culture in the laboratory. Non-target DNA from contaminants or numts can easily result in the extraction or amplification of incorrect or chimeric DNA sequences.

For molecular barcoding to have forensic value, reference barcode sequences should be derived from, and linked to, voucher specimens in web-accessible collections. NCBI maintains a number of databases (including GenBank, PubMed, Taxonomy and others) that are linked together in the Entrez indexing and retrieval engine. The LinkOut program allows outside groups to maintain sets of hotlinks from objects in Entrez back to specific locations on their web sites. The Integrated Primate Biomaterials and Information Resource has indexed holdings in the taxonomy domain of Entrez and indexed barcode sequences derived from the repository specimens in the sequence domain of Entrez (GenBank). This is a simple and practical approach to the problem of linking biological specimens with the biological data and research that

Table 3. Samples from this study that cluster with cox1 sequences from fully sequenced mtDNAs and cox1 sequences from previous studies obtained from GenBank.

<table>
<thead>
<tr>
<th>GenBank acc no.</th>
<th>species</th>
<th>IPBIR sample number</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi:4239858</td>
<td>Ateles geoffroyi</td>
<td>PR00134</td>
<td>Ateles geoffroyi</td>
</tr>
<tr>
<td>gi:12484071</td>
<td>Saimiri sciureus</td>
<td>PR00741</td>
<td>Saimiri oerstedii</td>
</tr>
<tr>
<td>gi:12484069</td>
<td>Saimiri ustus</td>
<td>PR00474</td>
<td>Saimiri boliviensis</td>
</tr>
<tr>
<td>gi:4239860</td>
<td>Colobus polykomos</td>
<td>PR00655</td>
<td>Colobus guereza</td>
</tr>
<tr>
<td>gi:60392100</td>
<td>Colobus guereza</td>
<td>PR00597</td>
<td>Colobus polykomos</td>
</tr>
<tr>
<td>gi:60392086</td>
<td>Trachypithecus obscura</td>
<td>PR00980</td>
<td>Colobus guereza</td>
</tr>
<tr>
<td>gi:49146236</td>
<td>Macaca mulatta</td>
<td>PR01112</td>
<td>Macaca mulatta</td>
</tr>
<tr>
<td>gi:14010693</td>
<td>Macaca sylvanus</td>
<td>PR00408</td>
<td>with all macaques</td>
</tr>
<tr>
<td>gi:5835638</td>
<td>Papio hamadryas</td>
<td>BP00041</td>
<td>Papio anubis</td>
</tr>
<tr>
<td>gi:12484067</td>
<td>Cercopithecus aethiops</td>
<td></td>
<td>Chlorocebus</td>
</tr>
<tr>
<td>gi:5835820</td>
<td>Hylobates lar</td>
<td>PR00495</td>
<td>Hylobates lar</td>
</tr>
<tr>
<td>gi:5835834</td>
<td>Pongo pygmaeus abelii</td>
<td>PR00253</td>
<td>Pongo pygmaeus</td>
</tr>
<tr>
<td>gi:5835163</td>
<td>Pongo pygmaeus</td>
<td>PR00276</td>
<td>Sumatran</td>
</tr>
<tr>
<td>gi:5835121</td>
<td>Pan troglodytes</td>
<td>PR00744</td>
<td>Pan troglodytes</td>
</tr>
<tr>
<td>gi:5835135</td>
<td>Pan paniscus</td>
<td>PR00092</td>
<td>Pan paniscus</td>
</tr>
<tr>
<td>gi:5835149</td>
<td>Gorilla gorilla</td>
<td>PR00643</td>
<td>Gorilla gorilla</td>
</tr>
<tr>
<td>gi:12484065</td>
<td>Theropithecus gelada</td>
<td>PR00573</td>
<td>does not cluster with cercopithecines</td>
</tr>
<tr>
<td>gi:21449875</td>
<td>Lemur catta</td>
<td>PR01054</td>
<td></td>
</tr>
<tr>
<td>gi:4239864</td>
<td>Galago senegalensis</td>
<td>PR00519</td>
<td>Galago moholi</td>
</tr>
</tbody>
</table>

Phil. Trans. R. Soc. B (2005)
are derived from them. The individual barcodes validate the species identification of specimens submitted to the repository and collectively form a publicly available reference database for primate molecular diagnostics.

Although GenBank encourages the separate submission of identical sequences obtained from multiple specimens, current practice in phylogenetics research often involves reporting only the variable haplotypes. For barcoding to assume a quantitative approach to species diagnosis, barcode sequences have been submitted for all specimens in the study. Moreover, data quality is of crucial importance to barcoding if it is to develop into a forensic tool. The Consortium for the Barcode of Life (CBOL; www.barcoding.si.edu) Data-base Working Group is calling for the deposition of barcode sequences in GenBank together with the primers that were used to generate them, their trace files and associated quality scores.

With the characterization of the IPBIR collection, we have expanded the number of primate barcodes from about a dozen sequences from unvalidated source material in GenBank (derived from primate whole mtDNA sequences), to include 56 species of the approximately 200 species in the order primates. This work sheds light on the reliability of the existing data and represents a significant increase in the potential for DNA barcoding to be employed as a tool for molecular diagnostics of primates.

Prior to this study, a major practical concern for DNA barcoding was the relatively few sequences deposited for which a specimen was available for re-examination. The deposition of materials in IPBIR, a public-access collection, provides a mechanism to allow verification of potentially problematic data and the re-examination of source material as advocated by Ruedas et al. (2000). The DNA barcode data generated by IPBIR represents sequences not previously found in GenBank. Thus, we hope the quality control efforts performed on the samples provided to the IPBIR will contribute significantly to the expansion of data available for the study of primate genetic diversity.

We thank the Fannie E. Rippel Foundation for their generous support for the primate DNA barcoding initiative. We thank Patrick K. Bender, Donald Coppock, Brian Fisher and Vincent Savolainen for constructive reviews of this manuscript.

REFERENCES


Champoux, M., Higley, J. D. & Suomi, S. J. 1997 Behavioral and physiological characteristics of Indian and Chinese–


Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications

Gary W Saunders

Phil. Trans. R. Soc. B 2005 360, 1879-1888

References

This article cites 18 articles, 3 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1879.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1879.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications

Gary W. Saunders*

Centre for Environmental & Molecular Algal Research Department of Biology, University of New Brunswick Fredericton, NB, Canada E3B 6E1

Marine macroalgae, especially the Rhodophyta, can be notoriously difficult to identify owing to their relatively simple morphology and anatomy, convergence, rampant phenotypic plasticity, and alternation of heteromorphic generations. It is thus not surprising that algal systematists have come to rely heavily on genetic tools for molecular assisted alpha taxonomy. Unfortunately the number of suitable marker systems in the three available genomes is enormous and, although most workers have settled on one of three or four models, the lack of an accepted standard hinders the comparison of results between laboratories. The advantages of a standard system are obvious for practical purposes of species discovery and identification; as well, compliance with a universal marker, such as cox1 being developed under the label ‘DNA barcode’, would allow algal systematists to benefit from the rapidly emerging technologies. Novel primers were developed for red algae to PCR amplify and sequence the 5’ cox1 ‘barcode’ region and were used to assess three known species-complex questions: (i) Mazzaella species in the Northeast Pacific; (ii) species of the genera Dilsea and Neodilsea in the Northeast Pacific; and (iii) Asteromenia peltata from three oceans. These models were selected because they have all caused confusion with regards to species number, distribution, and identification in the field, and because they have all been studied with molecular tools. In all cases the DNA barcode resolved accurately and unequivocally species identities and, with the enhanced sampling here, turned up a variety of novel observations in need of further taxonomic investigation.

Keywords: cox1; cryptic species; DNA barcode; Florideophyceae; mitochondrial DNA; Rhodophyta

1. INTRODUCTION
From the student taking an introductory course in Phycology to the seasoned field biologist, there is a common, at times overwhelming, frustration when tasked with the identification of many macroalgal species. Even for the experienced systematist confronted with exclusively vegetative material—reproduction often betraying the ordinal and familial affinities of a collection—or even with reproductive material among species in a genus, accurate identification can remain elusive. This frustrating situation derives from a few commonalities of marine macroalgae that tend to confound attempts at identification viz., simple morphology and anatomy, rampant convergence (in part owing to the previous), remarkable degrees of phenotypic plasticity in response to environmental factors, and incompletely understood life histories with alternation of heteromorphic generations.

In light of the previous it is not unexpected that algal systematists have, for close to two decades, come to rely increasingly on molecular tools to resolve and identify species (see Harper & Saunders 2001). Examples include the internal transcribed spacer of the ribosomal cistron (ITS; Tai et al. 2001; Ross et al. 2003), the rubisco operon (rbcL; Hughey et al. 2001), and variable portions of the large subunit of the ribosomal cistron (LSU; Saunders & Lehmkuhl in press). The previous examples serve to highlight an unsatisfactory shortcoming in the current efforts among algal systematists—the lack of a universally applied marker has resulted in multiple, independent, and not easily comparable systems being used. Although there is ample justification for the development of multiple and divergent molecular markers for phylogenetics, agreement on a standard marker for the purposes of quick and accurate species identification would be a powerful tool for the practising taxonomist.

Genetic barcoding has championed the use of the mitochondrial marker cytochrome oxidase subunit I (cox1). In a pair of landmark publications Hebert et al. (2003a,b) established the utility of cox1 as the ‘core of a global bioidentification system for animals’. These authors reported that, for a wide variety of animal species at least, this gene could be used to assign unknown species to higher-level taxa, and where comprehensive cox1-5’ databases were established species level assignments were possible. The authors clearly articulate the power of this approach to species identification when phenotypic plasticity is a concern, morphology-based keys are only useful for particular stages in the life history or stages are unknown, or cryptic species are likely to be an issue—all of these, as noted above, considerations where macroalgae are concerned. Hebert et al. (2003a,b) justify the choice

* (gws@unb.ca).

One contribution of 18 to a ‘Theme Issue ‘DNA barcoding of life’.
Table 1. List of samples for which cox1-5' sequences were determined in this study.

<table>
<thead>
<tr>
<th>order/family</th>
<th>species</th>
<th>voucher</th>
<th>collection details</th>
<th>Genbank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceramiaceae</td>
<td><strong>Ptilota serrata</strong> Kützing</td>
<td>GWS002173</td>
<td>Vancouver I., BC, Canada. GWS</td>
<td>AY970640</td>
</tr>
<tr>
<td></td>
<td>Gigartinales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Rhodophyllis sp.</strong></td>
<td>GWS001945</td>
<td>Australia. GWS &amp; R. Withall</td>
<td>AY970627</td>
</tr>
<tr>
<td></td>
<td>Dumontiaceae</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>'Dilsea Exposed'</td>
<td>GWS002248</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970638</td>
</tr>
<tr>
<td></td>
<td><strong>Dilsea californica</strong> (J. Agardh) Kuntze</td>
<td>GWS001671/1687/1689</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970573/572/571</td>
</tr>
<tr>
<td></td>
<td><strong>Dilsea californica</strong></td>
<td>GWS002171/2238/2252</td>
<td>Vancouver I, BC, Canada. GWS &amp; C. Lane</td>
<td>AY970636/639/637</td>
</tr>
<tr>
<td></td>
<td>'Dilsea(?) Exposed'</td>
<td>GWS002283</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970582</td>
</tr>
<tr>
<td></td>
<td>Dilsea carnosa (Schmidel) Kuntze</td>
<td>GWS000746</td>
<td>Northern Ireland. C. Maggs</td>
<td>AY971151</td>
</tr>
<tr>
<td></td>
<td>Dilsea carnosa</td>
<td>GWS001216</td>
<td>France. GWS</td>
<td>AY970635</td>
</tr>
<tr>
<td></td>
<td>Dilsea integra (Kjellman Rosenvinge)</td>
<td>GWS001850</td>
<td>Nunavut, Canada. GWS</td>
<td>AY970633</td>
</tr>
<tr>
<td></td>
<td>Dilsea integra</td>
<td>GWS002334</td>
<td>Cape Breton, NS, Canada. GWS</td>
<td>AY970634</td>
</tr>
<tr>
<td></td>
<td><strong>Dumontia contorta</strong> (Gmelin) Ruprecht</td>
<td>CSM005A-C</td>
<td>Cape St Marys, NS, Canada. GWS</td>
<td>AY971154/155/156</td>
</tr>
<tr>
<td></td>
<td>Dumontia contorta</td>
<td>GWS001815</td>
<td>Ireland. GWS</td>
<td>AY970583</td>
</tr>
<tr>
<td></td>
<td>Dumontia contorta</td>
<td>GWS002137-2141</td>
<td>Plymouth, England, UK. C. Maggs</td>
<td>AY971157/158/147/148/149</td>
</tr>
<tr>
<td></td>
<td>Dumontia contorta</td>
<td>GWS002142-2146</td>
<td>Portaferry, N. Ireland, UK. C. Maggs</td>
<td>AY971150/159/160/161/152</td>
</tr>
<tr>
<td></td>
<td>Dumontia contorta</td>
<td>PC004A-C</td>
<td>Peggy's Cove, NS, Canada. GWS</td>
<td>AY970568/569/570</td>
</tr>
<tr>
<td></td>
<td>Dumontia simplex Cotton</td>
<td>GWS000209</td>
<td>Alaska, USA. S. Lindstrom</td>
<td>AY971153</td>
</tr>
<tr>
<td></td>
<td>Neodilsea borealis (Abbott) Lindstrom</td>
<td>GWS001681/1683/2176</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970614/626/625</td>
</tr>
<tr>
<td></td>
<td>'Neodilsea borealis' Exposed</td>
<td>GWS002232/2281/2282</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970617/615/616</td>
</tr>
<tr>
<td></td>
<td>Neodilsea natashae (Setchell et Gardner)</td>
<td>G0224</td>
<td>Alaska, USA. S. Lindstrom</td>
<td>AY970624</td>
</tr>
<tr>
<td></td>
<td>Lindstrom</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Chondrus crispus</strong> Stackhouse</td>
<td>PC001B</td>
<td>Peggy's Cove, NS, Canada. GWS</td>
<td>AY970567</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella affinis</strong> (Harvey) Fredericq</td>
<td>GWS001333/2259</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970577/578</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella affinis</strong></td>
<td>GWS002256</td>
<td>Vancouver I, BC, Canada. C. Bates</td>
<td>AY970576</td>
</tr>
<tr>
<td></td>
<td><em>Mazzaella flavicola</em> (Setchell et Gardner)</td>
<td>GWS002235/2245</td>
<td>Vancouver I, BC, Canada. GWS &amp; C. Lane</td>
<td>AY970575/574</td>
</tr>
<tr>
<td></td>
<td>Fredericq</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella laminariaeoides</strong> (Bory) Fredericq</td>
<td>GWS000131</td>
<td>Chacao, Chile. B. Rudolph</td>
<td>AY970593</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella linearis</strong> (Setchell et Gardner) Fredericq</td>
<td>GWS000910C/1173A–D/F–J</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY971162/592–584/593–582</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella oregona</strong> (Doty) Hughey, P.C. Silva et Hommersand</td>
<td>GWS002199/2258</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970602/603</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella parksi</strong> (Setchell et Gardner) Hughhey et al. Fredericq</td>
<td>GWS001115</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970601</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella rosea</strong> (Kylin) Fredericq</td>
<td>GWS001109</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970600</td>
</tr>
<tr>
<td></td>
<td><strong>Mazzaella sanguinea</strong> (Setchell et Gardner) Hommersand</td>
<td>GWS001146/1165</td>
<td>Vancouver I, BC, Canada. GWS</td>
<td>AY970598/599</td>
</tr>
</tbody>
</table>

(Continued.)
of a protein coding mitochondrial gene because of the relatively rapid rate of divergence in animals, the haploid mode of inheritance, the ability to design 'universal' primers at constrained portions of the gene, and the low prevalence of indels, which greatly facilitates alignment across phyla. They acknowledge that there is no \textit{a priori} reason for the selection of one mitochondrial protein gene over another, but point to two advantages of \textit{cox1}: (i) the universality of existing primers for amplification of the 5' end of this gene in a wide variety of animals; and (ii) the broad phylogenetic range covered by the gene (Hebert \textit{et al.} 2003a).

It is admittedly uncertain how well \textit{cox1}-5' will function for species discrimination in the other kingdoms of life because the mode of inheritance, rate of divergence, as well as many of the other attributes discussed above for this marker, are poorly known outside of the animals and land plants. For the latter, it was established that mitochondrial genes are generally more slowly evolving than in animals (Barkman \textit{et al.} 2000), but that the more rapidly evolving plastid genome may provide sequences that would be adequate to amplify this gene region for red algae (GazF1 5'-5' ACTT-CTGGATGTCCAAAAAAYCA 3'; GazF2 5'-5' AAAAYCARAATAATGTTGA 3'). As for the many unicellular and multicellular protists, little is known and only through exploratory research will the utility of \textit{cox1}-5' as a genetic marker be established for the various lineages. At the time of writing this manuscript only four red algal \textit{cox1} genes were accessible from GenBank and these were largely associated with mitochondrial genome projects.

There are obviously many advantages to algal systematists in adopting a standard marker for the purposes of species identification. By choosing the \textit{cox1}-5' system, should it prove suitable, there is the added advantage of being consistent with work in other kingdoms, which will facilitate universal comparisons and empower algal systematists to take advantage of emerging technologies. To facilitate this process, sequences from the four divergent red algal taxa in GenBank were acquired and used to modify the original barcoding primers published (Hebert \textit{et al.} 2003a) for animals. To date these primers have successfully amplified \textit{cox1}-5' from \textit{ca} 250 individuals spanning 15 families in six orders of the Florideophyceae. In this report the utility of \textit{cox1}-5' for species level discrimination is tested for three problematic species complexes in two orders, the Gigartinales and Rhodymeniales. These three test cases were selected because of their previous investigation with other DNA marker systems. In all cases \textit{cox1}-5' species assignments matched the earlier studies.

### 2. MATERIAL AND METHODS

Samples sequenced in the current study are identified in table 1. DNA was extracted with a protocol modified from Saunders (1993) (instead of the final agarose gel cleaning procedure, the DNA was purified with the Wizard\textsuperscript{c} DNA Clean-Up System, Promega Corp., Madison, WI). The \textit{cox1} sequences for \textit{Cyanidium caldarium} (Tilden) Geitler (Z48930), \textit{Cyanidioschyzon merolae} P. De Luca, R. Taddei et L. Varano (NC 000887), \textit{Chondrus crispus} Stackhouse (NC 001677), and \textit{Porphyra purpurea} (Roth) C. Agardh (NC 002007) were acquired from GenBank and aligned by eye in MacClade 4 (v. 4.06) for OSX (Maddison \& Maddison 2003). These sequences were used, in conjunction with the previously published \textit{cox1} barcoding primers developed for animals (Hebert \textit{et al.} 2003a), to devise specific primers to amplify this gene region for red algae (GazF1 5' TCAA-CAAACTCAATAAAAGATATTTGG 3' and GazR1 5' ACTT-CTGGATGTCCAAAAAAYCA 3'; GWS000209 used the forward primer GazF2 5' CCCACCCATAGAATCGGCG 3'); GWS002199 used the reverse primer DuniR1 5' AAAAYCARAATAATGTTGA 3'). The PCR amplification profile followed Hebert \textit{et al.} (2003a), but using an annealing temperature of 50°C. Amplified products were gel purified using a glasswool column procedure (Saunders 1993). Sequencing used the PE Applied Biosystems Big Dye (v 3.0) kit and followed the manufacturer's protocol (ABI, Foster City, CA). Forward and reverse sequence reads from the respective PCR primers were edited and aligned using Sequencher\textsuperscript{a} 4.2 (Gene Codes Corporation, Ann Arbor, MI, USA), and a multiple sequence alignment was constructed with MacClade. The final alignment included 101 taxa (table 1; plus the four taxa from GenBank) with 664 nucleotide positions. All analyses and sequence comparisons were conducted in PAUP\textsuperscript{a} 4.0b10 (Swofford 2002). Distances were corrected with a general time reversible model.
(a variety of corrections were used, but had no effect on species assignment) and neighbour-joining and UPGMA clustering algorithms were used to provide a visual display of coxl-5' variation within and between species.

3. RESULTS AND DISCUSSION
Using the primers described above ca 250 individuals spanning 15 families in six orders of the Rhodophyceae (Florideophyceae) have had coxl-5' barcodes successfully determined. The size of the amplified product is 710 base pairs (bp) of which 46 are complementary to the PCR primers and thus of no value for comparisons. In this report 97 of these sequences are presented that focus on three problematic species complexes.

Of the 101 aligned sequences, there were 87 individuals from within 16 species (from 2 to 20 isolates depending on the species; table 1) for which within species variation was between 0 and 1 (2) bp out of the 664 positions or 0–0.3% divergence. Between species comparisons within genera generally ranged between 30 and 90 changes or 4.5 and 13.6% divergence with notable exceptions for the closely related Mazzaella linearis and M. splendens (0.8–1.2%) and Dilssea carnosa and D. integra (1.1%), which are discussed below. There was thus a clear distinction in divergence within versus species observable with coxl-5' for the red alga studied—clearly a necessary attribute for a marker system to be considered useful for the task of species assignment. Below three species complexes are considered in turn.

(a) Mazzaella in the Pacific Northeast
Mazzaella is a genus of the red algal family Gigartinaceae, Gigartinales, its species common along the coast of British Columbia (BC), Canada. Species of this genus are notoriously difficult to distinguish in the Northeast Pacific (Hommersand et al. 1994; Gabrielson et al. 2000; cf. Ross et al. 2003). Mazzaella splendens from sheltered habitats (figure 1a) can be discerned from wave-exposed populations of M. linearis (figure 1b), but a continuum of morphological intermediates (e.g. figure 1c) traverse the intervening wave-exposure gradient (Shaughnessy 1996). Mazzaella flaccida is also difficult to distinguish (compare figure 1d to e, f) from M. splendens (Hommersand et al. 1994) such that the ‘Keys to the Benthic Marine Algae...of British Columbia...’ (Gabrielson et al. 2000) indicate that ‘the northern distribution limit of M. flaccida has not been established, but it may be present in southern British Columbia’. Gabrielson et al. (2000) further indicate a difficulty in distinguishing between some morphologies of M. oregona (as M. heterocarpa) (figure 1g) and M. splendens, and I have collected plants that defy identification based on morphology (e.g. figure 1h). Using sequence of the large subunit of rubisco (rbcL), Hommersand et al. (1994) were able to establish clearly genetic differences between M. flaccida and M. splendens, but not strongly between two samples of the latter and an isolate of M. linearis included in their study. A series of reciprocal transplant experiments, however, support recognition of M. linearis and M. splendens as distinct species, and indicate that plants of intermediate morphology were exposed variants of the latter (Shaughnessy 1996; Shaughnessy & DeWreede 2001).

Ross et al. (2003) outlined possible scenarios to explain the apparent morphological continuum observed in the field for the M. linearis/splendens complex: (i) a single species exists with substantial morphological plasticity in response to wave exposure; (ii) two species exist, but one or both display phenotypic plasticity such that the intermediate plants are strictly of one species or the other, or a mixture of plants from both; (iii) the intermediate plants may be tetrasporophytes of M. splendens, which was reported to have a possible heteromorphic aspect to its alternation of generations (Shaughnessy et al. 1996); (iv) the intermediates are hybrids between the two species; and (v) combinations of the previous could also explain these individuals. To evaluate the previous hypotheses, as well as the published ecological results (Shaughnessy 1996; Shaughnessy & DeWreede 2001), Ross et al. sequenced a variable region (ITS) of the nuclear genome for 17 isolates each of classic M. splendens (figure 1a) and M. linearis (figure 1b), and 20 isolates of intermediate morphology (figure 1c). They concluded that all of the intermediate plants were M. splendens, which had a broader morphological and ecological range than the strictly exposed and highly lanceolate M. linearis. The ITS varied from 0 to 4 nucleotides within a species, and was only 8–12 nucleotides different between M. linearis and M. splendens, these 32–38 nucleotides different from the next closest species, M. sanguinea. One plant field identified as M. linearis (GWS001173E; figure 1c) had ITS sequence consistent with placement in M. splendens, a result which was consistent with a morphological re-evaluation of the voucher (Ross et al. 2003); and another that was identified as M. flaccida (figure 1d), based on their interpretation of the identification keys in Gabrielson et al. (2000), also proved to be M. splendens, leaving uncertainty as to whether or not M. flaccida extends into BC. The ITS was thus a powerful tool for resolving an outstanding species issue in the flora adjacent to the Bamfield Marine Station, but it had shortcomings. The common occurrence of mononucleotide runs (poly C for example) and/or heterogeneity in the multiple copies of the ITS within an individual made it difficult to obtain sequence from both strands across the entire ITS for many of the individuals. This has serious implications for data quality and for using this marker for rapid species assignments. Additionally, the common occurrence of indels made it virtually impossible to compare the entire ITS from the M. linearis/splendens/sanguinea clade to the other species included (variable regions had to be excluded), rendering the estimation of nucleotide differences beyond the most closely related species inaccurate. Finally, results from this study could not be compared directly to the earlier study of Hommersand et al. (1994) because different marker systems were employed.

The Mazzaella linearis/splendens complex described above was used to test the coxl-5' marker for its utility in distinguishing closely related species of red algae. This gene was successfully sequenced for an isolate of
Chondrus crispus (resolves within Mazzaella in phylogenetic studies; Hommersand et al. 1999) and multiple individuals (n=42) from nine species of Mazzaella. Included were 10 M. linearis, nine M. splendens (including previously misidentified GWS001128; figure 1d), and 11 individuals of intermediate morphology (including previously misidentified GWS001173E; figure 1c), and three additional collections (figure 1 e, f, h) of Mazzaella spp. that could not be identified based on Gabrielson et al. (2000). The cox1-5′ barcode echoed exactly the results of the previous ITS study (figure 2). Mazzaella linearis and M. splendens were resolved as independent lines, and all of the intermediates were assigned to the latter. The previously misidentified samples GWS001128 and GWS001173E were unequivocally included in M. splendens. Among the 10 species sampled, there were seven with multiple isolates (between 2 and 20 per species), and the within species variation was limited to 0–2 nucleotides (0–0.3% divergence) (figure 2). Between species comparisons ranged from 35 to 91 changes (5.3–13.7% divergence) with a notable exception for the closely related Mazzaella linearis and M. splendens at only 5–8 nucleotide differences (0.8–1.2% divergence). This lower range is comparable with values obtained for the most closely related species of Lepidoptera (Hebert et al. 2003a), and presumably also represents an exceptional case among red algal species in light of the ecological and ITS studies discussed previously. Collection GWS002199 (figure 1h) was an odd carpet forming morph, midintertidal on rock from a sheltered locality on the northern end of Vancouver Island, which differed by only two nucleotides in its cox1-5′ from a typical collection of Mazzaella oregona (GWS002258; figure 1g) from a semi-exposed site—these collections were apparently morphological/ ecological extremes of a single species (figure 2). Two collections from the northern end of Vancouver Island, GWS002235 (figure 1e) and GWS002245 (figure 1f), had a novel cox1-5′ (figure 2). Based on Gabrielson et al. (2000) the collections were either M. flaccida, only questionably extending into southern BC, or M. volans (C. Agardh) Fredericq, which is not reported north of Oregon. GenBank was searched to determine

Figure 1. Gross morphology of red blades discussed in this report (Scale=centimetre ruler). Typical sheltered and exposed morphs for Mazzaella splendens (a; GWS11751) and M. linearis (b; GWS001173D), respectively, and an isolate of intermediate morphology (c; GWS001173E) (central region removed for DNA extraction). (d) Sample (GWS001128) field identified as M. flaccida, but subsequently considered M. splendens. Two collections of M. flaccida (e, GWS002235; f, GWS002245) considered as unknown during field identification. Mazzaella oregona (g; GWS002258) from the outer coast of Vancouver I., and a carpet-forming morph from a sheltered habitat (h; GWS002199). Typical habit for Dilsea californica (i; GWS002171), Dilsea integra (j; GWS001850), Dilsea carmosa (k; GWS001216), and Neodilsea borealis (l; GWS002176), as well as morphs field identified as ‘Dilsea(?) exposed’ (m; GWS002283), ‘Neodilsea exposed’ (n, GWS002282; o, GWS002281), and ‘Dilsea exposed’ (p, GWS002248). Astomomenia peltata from: North Carolina (q; NC2987K—not included in barcode analyses); Bermuda (r; GWS001252); Western Australia (s; HA703—not included in barcode analyses); and Lord Howe Island, LHA (t; GWS001062. u; GWS002072) and LHB (v; GWS002022—representative of GWS002050).
that rbcL data were available for both species, whereas ITS was reported for only the former. The rbcL was thus sequenced for GWS002235 and GWS002245 and confirmed that these isolates are M. flaccida, which clearly extends well past southern BC. This represents the first published range extension to result from the application of cox1-5′ to species identification among red algae. A standardized system for species diagnosis, as advocated here, would have obviated the need to sequence an additional genetic marker to resolve the identity of these collections.

(b) Dumontiaceae emphasizing Dilsea and Neodilsea in the Pacific Northeast

The Dumontiaceae is also a family of the Gigartinales; although low in species diversity in Canada (ca 10), it is a group with known cryptic species (Tai et al. 2001). During a variety of collecting trips I have had the
fortune of collecting *Dilsea californica* (figure 1r) along the coasts of BC, *D. integra* (figure 1j) from the Canadian Arctic, *D. carnosa* (figure 1k) from Europe, and *Neodilsea borealis* (figure 1l), again from BC. At the same time plants from exposed habitats in BC tentatively identified as 'Dilsea (?) exposed' (figure 1m), 'Neodilsea exposed' (figure 1n, o) and 'Dilsea exposed' (figure 1p) were acquired. Part of the uncertainty derives from the similarity of the species (ca 11 in total) included in the *Dilsea/Neodilsea* complex (Lindstrom & Scagel 1987), the fact that only *D. californica* and *N. borealis* are recorded from BC (Lindstrom & Scagel 1987; Gabrielson et al. 2000), and the recovery by Tai et al. (2001) of divergent ITS sequences for *D. californica* from Alaska and Oregon indicating cryptic species.

As a first step toward a *cox1*-5' analysis of the *Dilsea/Neodilsea* complex in BC, 17 individuals of a related taxon, *Dumontia contorta*, were collected from the Maritime Provinces of Canada (n = 6) and Europe (n = 11). The three Peggys Cove isolates were identical and differed from the other collections at only one nucleotide difference. Voucher numbers correspond to records in table 1 (central zeros omitted in the matrix).
It was expected that these two would represent closely related species (Tai et al. 2001), along with Dilsea socialis (Postels et Ruprecht) Perestenko from Alaska and the Northwest Pacific, which has not been included in the cox1-5'0 alignment to date. Six isolates of Dilsea californica (figure 1i) had identical cox1-5'0 sequences, as did 'Dilsea exposed' (GWS002248; figure 1p), which clearly represents the exposed morphology of this species (figure 3). Neodilsea borealis (figure 1l) formed a distinct lineage in the cox1-5' analyses with only 0–1 nucleotide differences noted among the three isolates, but these differed substantially (51–52 nucleotide differences) from collections tentatively identified as exposed morphs ('Neodilsea exposed'; figure 1, o) of this species (figure 3). 'Neodilsea exposed' resolved as sister (36 nucleotides divergent) to Dilsea californica and represents a new species, or a range extension of one of the four Northwest Pacific species (cf. Lindstrom 1994). Tai et al. (2001) also resolved two divergent entities in their ITS investigation (identified as D. californica from Oregon and Alaska), which may correspond to Dilsea californica and Neodilsea exposed uncovered here. Neodilsea natashae, reported from Alaskan waters as far south as the Sumner Strait, Alexander Archipelago, can be excluded as a possible designation for Neodilsea exposed because it was included in the cox1-5'0 analyses (figure 3). Consistent with the previous, a subsequent anatomical investigation revealed it to be

![Figure 4](image_url)

Figure 4. Phylogram (neighbour joining) displaying clustering of the included species of the genera Leptofauchea and Asteromenia, Rhodymeniales, and a matrix of actual nucleotide differences. Voucher numbers correspond to records in table 1. Isolate 6268 in bold represents an anomalous 'Bermuda morph', which is only distantly related to Asteromenia spp.
the exposed morph of another dumbiotean species, *Farlowia mollis* (Harvey et Bailey) Farlow et Setchell (cf. Lindstrom & Scagel 1987). A comprehensive barcode alignment would have provided an accurate identification for this collection.

In summary, all six of the species (figure 3) for which multiple isolates were studied in the Dumontiaceae displayed intraspecific *cox1*-5′ divergence of 0–1 nucleotide changes, with the closely related *D. carnosa* and *D. integra* differing at seven sites, but with most interspecific divergences > 30 changes.

(c) The genus *Asteromenia* from three oceans

*Asteromenia* was recently erected in the red algal order Rhodymeniales to include the single species *Fauchea peltata* W. R. Taylor, which was reportedly widely distributed throughout tropical and warm temperate waters (Huisman & Millar 1996). However, Saunders *et al.* (unpublished) have used LSU sequence data (GenBank DQ068294–DQ068301) in combination with anatomical analyses to argue that at least five cryptic species are included in this complex (figure 1q–r). The first (figure 1q) is distributed throughout the Caribbean, to Bermuda and North Carolina, and is anatomically the most divergent supporting the molecular evidence in resolving it as a separate species. A second species is common in Bermuda, but extends throughout the Caribbean including Puerto Rico, which had the characteristic anatomy of the *Bermuda morph* and identical LSU sequences, clustered together in the *cox1*-5′ analysis with 0–2 nucleotide changes (figure 4). Again consistent with the LSU and anatomical observations, a second species encompassed plants assigned to LHA; the *cox1*-5′ sequences identical among the three included samples (figure 4). Lord Howe Island LHB resolved as a third species, again consistent with the earlier observations, and was 31 nucleotides different in *cox1*-5′ from its sister species LHA (figure 4). Isolate 6268 from Puerto Rico was considered *Asteromenia peltata* Bermuda morph based on its gross morphology and was not included in the LSU investigation. The *cox1*-5′ sequence determined from this collection, however, was the most divergent of all the members of the Rhodymeniales studied here (figure 4). A preliminary assessment of the internal anatomy confirmed that this collection is not related to *Asteromenia* (Ballantine & Saunders, unpublished) and further investigation is required to determine its identity. Of importance here, the *cox1*-5′ barcode clearly established the novelty of this collection despite its gross morphological similarity to the Bermuda morph of *Asteromenia*.

4. CONCLUSION

The case studies presented here indicate that *cox1*-5′ barcoding will be a powerful ally in the identification of red algal species. In all cases intraspecific divergence values ranged from 0 to 2 bp, whereas interspecific divergences > 30 bp were usually observed. Exceptions were noted for the closely related *Mazzaella linearis* and *M. splendens* (5–8 differences) and *Dilsea carnosa* and *D. integra* (7 differences), but in both cases the marker successfully assigned collections to the correct species. Indeed the *cox1*-5′ system without fail matched the previous anatomical and molecular results. Considering the low number of taxa studied in this report (table 1), the number of new records is astonishing and indicates that considerable taxonomic and biogeographical work remains among the Rhodoplantae. Genetic barcoding will thus not signal the end of taxonomy for phycologists, but will initiate a revolution of molecular-assisted alpha taxonomy that will greatly change the number and distribution of species that are recognized in this lineage. It is not, however, a tool that should be used in isolation, particularly during the development stages when it will be desirable to accompany the molecular results with thorough anatomical observations, and in the case of closely related species where it will be prudent to assess the mitochondrial data with nuclear markers to search for introgression, hybridization and incomplete species boundaries. Further, the utility of *cox1*-5′ for species identification in asexual lines, a common manifestation in protists including the red algae, has not been adequately assessed and requires detailed investigation.

I am indebted to Melissa Brooks, Chris Lane and Tanya Mossman for generating the sequence data used in this study. I also thank the collectors and dive partners listed in table 1, and additionally acknowledge Drs Gerry Kraft, John Huisman and Craig Schneider for organizing substantial collecting trips to the more exotic locations, and the last two for the generous loan of herbarium collections used in figure 1 (HA703 and...
REFERENCES


Land plants and DNA barcodes: short-term and long-term goals

Mark W Chase, Nicolas Salamin, Mike Wilkinson, James M Dunwell, Rao Prasad Kesakurthi, Nadia Haidar and Vincent Savolainen

*Phil. Trans. R. Soc. B* 2005 *360*, 1889-1895
Land plants and DNA barcodes: short-term and long-term goals

Mark W. Chase\textsuperscript{1,*,}, Nicolas Salamin\textsuperscript{2}, Mike Wilkinson\textsuperscript{3}, James M. Dunwell\textsuperscript{3}, Rao Prasad Kesana\textsubscript{kurthi}\textsuperscript{3}, Nadia Haidar\textsuperscript{3} and Vincent Savolainen\textsuperscript{1}

\textsuperscript{1}Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3DS, UK
\textsuperscript{2}Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
\textsuperscript{3}School of Plant Sciences, University of Reading, Reading RG6 6AS, UK

Land plants have had the reputation of being problematic for DNA barcoding for two general reasons: (i) the standard DNA regions used in algae, animals and fungi have exceedingly low levels of variability and (ii) the typically used land plant plastid phylogenetic markers (e.g. \textit{rbcL}, \textit{trnL-F}, etc.) appear to have too little variation. However, no one has assessed how well current phylogenetic resources might work in the context of identification (versus phylogeny reconstruction). In this paper, we make such an assessment, particularly with two of the markers commonly sequenced in land plant phylogenetic studies, plastid \textit{rbcL} and internal transcribed spacers of the large subunits of nuclear ribosomal DNA (ITS), and find that both of these DNA regions perform well even though the data currently available in GenBank/EBI were not produced to be used as barcodes and BLAST searches are not an ideal tool for this purpose. These results bode well for the use of even more variable regions of plastid DNA (such as, for example, \textit{psbA-trnH}) as barcodes, once they have been widely sequenced. In the short term, efforts to bring land plant barcoding up to the standards being used now in other organisms should make swift progress. There are two categories of DNA barcode users, scientists in fields other than taxonomy and taxonomists. For the former, the use of mitochondrial and plastid DNA, the two most easily assessed genomes, is at least in the short term a useful tool that permits them to get on with their studies, which depend on knowing roughly which species or species groups they are dealing with, but these same DNA regions have important drawbacks for use in taxonomic studies (i.e. studies designed to elucidate species limits). For these purposes, DNA markers from uniparentally (usually maternally) inherited genomes can only provide half of the story required to improve taxonomic standards being used in DNA barcoding. In the long term, we will need to develop more sophisticated barcoding tools, which would be multiple, low-copy nuclear markers with sufficient genetic variability and PCR-reliability; these would permit the detection of hybrids and permit researchers to identify the ‘genetic gaps’ that are useful in assessing species limits.

\textbf{Keywords:} BLAST; molecular taxonomy; plant DNA barcoding; phylogenetics; population genetics

1. INTRODUCTION

Efforts to produce DNA barcodes (the so-called ‘DNA-taxonomy’ of Tautz \textit{et al.} 2003) are proceeding apace for animals and fungi using a standard DNA region, the mitochondrial \textit{cox1} gene (which codes for subunit 1 of cytochrome oxidase). It appears that this region can also be used in at least some groups of ‘algae’ (Saunders 2005; note that algae, even those that are multi-cellular, belong to various distantly related clades), but in land plants \textit{cox1} sequences are highly invariant and therefore unsuitable for use as DNA barcodes. Alternative single regions in the plastid genome have seen wide use in phylogenetic studies (e.g. exons such as \textit{rbcL}, \textit{atpB}, \textit{ndhF} and \textit{matK} and non-coding regions such as the \textit{trnL} intron and \textit{trnL-F} intergenic spacer), but these have ‘appeared’ not to be variable enough to be useful as barcodes because in phylogenetic studies results from individual loci have been highly unresolved due to too few phylogenetically informative sites. Another commonly sequenced region for land plant phylogenetic studies is nuclear ribosomal ITS (the internal transcribed spacers of the large subunit of ribosomal DNA), but this region does not work well in at least some groups of plants due to problems of paralogy and other factors associated with the complex concerted evolution of this highly repeated part of the nuclear genome. Thus, to at least some workers, barcoding in land plants with an approach similar to that being used in other organisms (employing \textit{cox1}) has appeared to be ‘on hold’ until more appropriate plastid markers have been developed (but see Kress \textit{et al.} 2005). However, lack of utility in a phylogenetic context does not necessarily mean that standard phylogenetic markers could not function well as identity codes; in the latter case, autapomorphies (unique single substitutions) are important whereas in 1889 © 2005 The Royal Society

\* Author for correspondence. (m.chase@kew.org).

One contribution of 18 to a ‘Theme Issue ‘DNA barcoding of life’.

Published online 12 September 2005

\textit{Phil. Trans. R. Soc. B} (2005) 360, 1889–1895
the former these are uninformative. Therefore, the first task is to make an attempt to see how such markers would perform in the altered context of DNA barcoding. We provide here an assessment of autapomorphies in several plastid DNA regions for two large South African genera, *Moraea* (approximately 200 species; *Goldblatt et al.* 2003) and *Protea* (approximately 85 species; *Reeves et al.* in press). Then, in a second phase, we provide an estimate of the utility of the two most abundant plant sequences in GenBank/EBI, the plastid gene *rbcL* and nuclear ribosomal ITS, using the BLAST procedure (*Altschul et al.* 1990) to make the ‘identification’. These DNA sequences and BLAST were never intended to be used in this manner, but they should provide some insights into how well we can expect perhaps more appropriate plastid DNA regions to perform as barcodes.

### 2. USERS OF PLANT BARCODES

There are two general categories of potential users of DNA barcodes: plant taxonomists/systematists, who wish to use these methods/markers as tools to elucidate species limits, and scientists in other fields, who are ‘end users’ of taxonomic concepts developed by taxonomists/systematists. For the latter category, there is an urgent need to establish at least a crude system of barcoding, and for this purpose plastid DNA regions are perfectly suited. By ‘crude’, we mean an easily developed but sometimes coarse system that is based on a uniparentally inherited marker, which makes it a less than perfect system. In some, perhaps many instances, this sort of marker will not provide an accurate identification, but there is still a great deal of utility in developing such a system. Many applications require DNA markers that can be easily amplified from degraded DNA samples, particularly in forensics and economic uses, such as traditional-drug authentication efforts. However, the incidence of hybridization, introgression and (allo)polyploidy in land plants is well documented, and to improve the taxonomic base upon which DNA barcoding efforts rest there is also a need to assess variation in multiple nuclear DNA regions. This also applies to algae, animals and fungi, although perhaps to a lesser degree due to a lower incidence of hybridization in such groups compared to higher plants.

Population genetic studies typically have large numbers of freshly collected specimens at their disposal, so DNA quality is a lesser concern, and it is upon such high-quality DNA samples that more accurate barcoding techniques would depend. To improve species concepts, we need to develop a more sophisticated approach to barcoding, which would ideally include sequences from multiple (perhaps six to eight) independent markers, a multi-locus barcode, and specific inference tools that could be used to explore species limits and identify genetic ‘gaps’. This second type of barcode would improve the information base upon which the cruder plastid and mitochondrial DNA barcodes depend. We will in the last part of this paper propose and describe in more detail our vision of how such a system could be developed.

#### 3. AN ASSESSMENT OF UNIQUENESS OF CURRENTLY USED PHYLOGENETIC MARKERS

To be used in phylogenetic studies, markers must exhibit sufficient variability to link species and groups of species by possessing shared (synapomorphic) substitutions; unique substitutions or autapomorphies are not used in assessing phylogenetic relationships of species and other taxa (but note that they are used in dating of phylogenetic trees, i.e. in molecular clock studies, and establishing overall genetic distances between species). Therefore it is not appropriate to determine utility of such markers for use in DNA barcoding efforts by comparing how well individual loci perform in phylogenetic studies; lack of resolution (the production of many equally optimal trees) caused by low levels of informative characters is not a useful measure when evaluating markers for use as barcodes, which requires unique substitutions that provide ‘species markers’. Therefore, we provide here an assessment of uniqueness in several markers commonly used in phylogenetic studies. We selected for this purpose two large genera emblematic of the flora of South Africa (two of the 34 global diversity hotspots, Cape and Succulent Karoo): (i) *Moraea* (peacock irises, approximately 250 species; Iridaceae) studied by *Goldblatt et al.* (2003) using three plastid DNA regions, the *rbcL* gene, the *trnL*-F intron/intergenic spacer and the *rps16* intron; and (ii) *Protea* (proteas, approximately 85 species; Proteaceae) studied by *Reeves et al.* (in press) using three plastid genes and a low-copy nuclear region, glutamine synthetase (the copy expressed in plastids, *ncpGS*; *Emshwiller & Doyle* 2002). We determined the degree to which species could be separated by unique changes (table 1) and found that in *Moraea*, which is approximately 25 million years old (*Goldblatt et al.* 2003), even the ‘relatively conserved’ *rbcL* exon exhibited sufficient numbers of autapomorphies to separate more than 99% of the 170 species in our DNA matrices. In contrast, for *Protea*, which is of similar age (*Reeves et al.* in press), these same sorts of standard plastid markers did not fare so well (table 1): the best plastid region exhibited enough unique variation to separate only 65% of the 82 species in our matrices. However, the fragment of *ncpGS* sequenced, which contains three introns (about 80% of its length), separated >99% of the species. There are at least two factors that could make these statistics less meaningful than they appear: some of the unique changes could be sequencing errors (which by some estimates, *Kristensen et al.* 1992, are as high as 4%, a figure that we would dispute in this case) and bad taxonomy, such that variants of the same biological entity have been given two or more names. These two phenomena would to a degree compensate for each other. These studies did not include many accessions of the same species, so we cannot assess the degree of intraspecific variation for these markers. Thus it appears that in the case of *Moraea* a single plastid marker would be highly useful as a tool for barcoding whereas in *Protea* two or more plastid markers would be required; *ncpGS*, however, could be successful on its own for *Protea*. In any case, this result demonstrates that although not sufficient as phylogenetic markers these DNA regions contain...
Table 1. Probability ($\rho$) of identifying the correct species based on DNA sequences. (Pair-wise distance matrices of absolute numbers of differences were computed using PAUP* 4.0b10 (Swofford 2001). Note that the probability of identifying the correct species was calculated as the proportion of comparisons in which at least one nucleotide difference was found between species pairs (in practice we would aim at targeting genes that have more than just one nucleotide difference between species).)

<table>
<thead>
<tr>
<th>taxa</th>
<th>DNA regions (number of base pairs, bp)</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moraea (n=170)</td>
<td>trnL-F intron/spacer (1229 bp)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td></td>
<td>rbcL (1354 bp)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td></td>
<td>rps16 intron (992 bp)</td>
<td>&gt;0.99</td>
</tr>
<tr>
<td>Protea (n=88)</td>
<td>trnL-F (1074 bp)</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>atpB-rbcL spacer (842 bp)</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>rps16 intron (832 bp)</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>ncpGS (854 bp)</td>
<td>&gt;0.99</td>
</tr>
</tbody>
</table>

unique changes that could serve as DNA barcodes (provided that as follow-up studies intraspecific variation would be assessed to determine species limits).

In a second case study, we utilized the two most abundant DNA sequences for plants in GenBank/EBI, plastid rbcL (6 741 sequences) and nuclear ribosomal ITS (total 33 508, some treating ITS1, 5.8S and ITS2 as separate entries). All available entries for these two DNA regions were extracted from the Euphyllophyta dataset in GenBank release 144. In turn, each ITS and rbcL entry was used as a query for a BLAST search against the entire Euphyllophyta GenBank dataset. First, the percentages of identity, as returned by the BLAST algorithm (Altschul et al. 1990), were calculated between and within several taxonomic levels (species, genus, family and order). The genera were assigned to families and orders following the APGII classification (APG 2003; list of genera available from MWC and VS), but those genera present in GenBank but not recognized by APGII were removed from the BLAST analyses. Second, for each query sequence, the percentages of incorrect assignment were calculated at the genus and species level. These proportions represent the number of BLAST hits not identical to the query at the genus or species level but with a better hit score than the first and last correct hit for the query. At the species level, we examined the number of best hits before finding with BLAST the same sequence used as input for the search, plus any other entries of the same species. At the genus level, we looked at all entries for a target genus and recorded the number of incorrect hits (i.e. species belonging to other genera) that appeared above the lowest ranked entry of a species correctly assigned to the genus. GenBank/EBI accessions for ITS contain either the complete ITS region or part of it. The results for this DNA region were therefore split according to which part represented the query sequence (table 2).

One can imagine several reasons why this procedure is a less than ideal measure of the barcode potential of these markers, ranging from incorrect name assignments in GenBank/EBI to the use of BLAST as a tool for which it was never designed, and we did not expect this procedure to work well. Hence, evaluating the feasibility of barcoding with the available data and tools is likely to produce an unfavourable outcome; such methods should perform much better when the reference database is more complete, when intraspecific levels of variation have been determined and included in the procedure and when a more appropriate search tool is employed (e.g. string barcoding, the use of particular motifs or combinations of particular bases; Rash & Gusfield 2002; DeSalle et al. 2005). For ITS, the two more variable parts, ITS1 and ITS2, performed better than the more conserved 5.8S region (table 2). On average, with an ITS1 probe, the cluster of best BLAST hits formed by all sequences from the same species contained 6.79% of sequences from other species (table 2). At the genus level, the cluster of genus-specific sequences contained ca 40% of other genera (table 2), which is some cases could be the result of current generic limits being unnatural rather than BLAST not being able to discriminate between natural (i.e. monophyletic) genera. For ITS2, these percentages were of 33.70 and 51.68% for the species and genus level, respectively (table 2). However, ITS1 had a higher percentage identity at all taxonomic levels compared to ITS2. In our comparisons, the intraspecific levels of variation were not an important factor for correct assignment of species. Although rbcL sequences had higher levels of taxonomic fidelity (i.e. getting an appropriate match), the proportions of erroneous assignment were ca 17 and 68% for species and genus level, respectively (table 2). We are encouraged by the relatively high levels of the target species being in the highest BLAST categories. This bodes well for using markers similar to those already widely sequenced as phylogenetic markers as barcodes for land plants.

4. DNA BARCODING AS A TOOL FOR GENETIC DELIMITATION OF BIOLOGICAL SPECIES

Most potential users of DNA barcoding are not taxonomists (or systematists, which we here consider synonymous, although many would distinguish between these two, with the former being equivalent to nomenclaturists and the latter population and/or evolutionary biologists). These users in other fields need a quick, easily used and accurate system of identification, and in many cases a relatively crude diagnosis would be acceptable. The need for such a system is immediate and cannot wait for something more sophisticated to be developed. Taxonomists have worried that DNA barcoding will be less successful than it could be because species limits in many groups of organisms are merely statements of what we think rather than what we know, and therefore the old adage applies: ‘rubbish in, rubbish out.’ We think that taxonomists are overly critical of their work and focus on the gaps in their knowledge, which is in many ways admirable (users of taxonomic data would prefer to be told of a lacuna in knowledge rather than that it is perfect when it is not). Many of the situations in which barcodes would be applied can accept the application of a broad species concept (i.e. identification to an aggregated species, a group of species for which limits

Phil. Trans. R. Soc. B (2005)
are not clear, or to one of a closely related set of species, a species complex, rather than a single species) and identification of a hybrid or introgressed plant as its maternal parent (because plastid DNA are maternally inherited in most plants) would not be hugely problematic on a practical level. Moreover, an ecologist trying to identify sterile plants, perhaps seedlings, in his plots or the forensic scientist trying to tie a vehicle to a particular location where a rare plant grows will not be overly concerned about the effects of hybridization, introgression or polyploidy on the results of the barcoding effort. This is not a denial that such phenomena are factors in the identification of species, but rather that a high degree of precision is not always required; just getting the field of possibilities narrowed to this extent provides such immense benefits that a degree of imprecision can be easily tolerated.

In contrast, a specialist working on a particular group of organisms (perhaps one for which the specialist is the world-recognised expert) worries a great deal about all the phenomena that can make identification of a particular accession problematic. The working taxonomist is concerned about the effects of hybridization, introgression and (allo)polyploidy, and some groups are notoriously difficult in these respects (e.g. Nicotiana in Solanaceae, Chase et al. 2003; Clarkson et al. 2004; and some orchids, e.g. Dactylorhiza, Pillon et al. in press), so it is not surprising that many taxonomists have viewed the issue of DNA barcoding with a great deal of suspicion and scepticism. Even if they were willing to recognize that it could be done, they are simultaneously critical of how good the results would be simply due to the ambiguities of species limits that they know exist within their groups of particular interest and hence by extrapolation to all others as well. DNA barcodes based on uniparentally inherited markers can never reflect the complexity that exists in nature, and many taxonomists have by and large ignored or been highly critical of the barcoding movement as a waste of time and money with at best the prospect of dubious results.

However, we can easily imagine a more sophisticated barcoding technique that would provide taxonomists with a new tool to investigate species limits and identify the genetic gaps that result when gene flow has become negligible. An example of these gaps is presented by Richardson et al. (2003) for a group of Phylica species (Rhamnaceae) using AFLPs (a genetic fingerprinting technique not suitable for barcoding because of difficulties in inferring the homology of amplicons on the basis of shared fragment size and to a lesser extent the need to find PCR primers appropriate

<table>
<thead>
<tr>
<th>DNA regions</th>
<th>taxonomic levels</th>
<th>identity (%)</th>
<th>wrong assignment (%)</th>
<th>first occurrence</th>
<th>last occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITS1 + 5.8S + ITS2</td>
<td>species</td>
<td>95.75/97.20</td>
<td>6.59</td>
<td>45.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>94.95/95.65</td>
<td>0.30</td>
<td>43.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>94.35/95.24</td>
<td>92.98/93.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS1 + 5.8S</td>
<td>species</td>
<td>97.24/97.43</td>
<td>0.84</td>
<td>43.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>94.69/95.76</td>
<td>0</td>
<td>40.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>93.02/94.27</td>
<td>92.07/92.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.8S + ITS2</td>
<td>species</td>
<td>95.48/97.21</td>
<td>3.57</td>
<td>40.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>93.50/93.72</td>
<td>0.90</td>
<td>34.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>91.38/93.21</td>
<td>90.75/91.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS1</td>
<td>species</td>
<td>95.16/96.08</td>
<td>0.47</td>
<td>49.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>93.98/94.42</td>
<td>1.13</td>
<td>46.28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>93.20/92.03</td>
<td>91.86/92.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS2</td>
<td>species</td>
<td>93.98/95.28</td>
<td>7.08</td>
<td>33.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>93.78/93.10</td>
<td>0.10</td>
<td>51.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>91.79/89.46</td>
<td>88.89/89.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.8S</td>
<td>species</td>
<td>97.18/99.80</td>
<td>32.76</td>
<td>62.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>96.30/98.45</td>
<td>36.04</td>
<td>64.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>95.32/97.25</td>
<td>93.25/95.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rbcL</td>
<td>species</td>
<td>97.62/99.73</td>
<td>3.69</td>
<td>16.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>genus</td>
<td>95.75/97.13</td>
<td>0.23</td>
<td>67.71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>family</td>
<td>91.43/95.92</td>
<td>88.81/92.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
for each group, i.e. because it cannot be made universal; Vos et al. 1995). Genetic gaps do not necessarily reflect species limits and hence cannot always be used as a guide to the application of names (i.e. there is not a 1 : 1 relationship of genetic gaps to taxonomic names), but knowing where such gaps exist is extremely useful information in the quest to make meaningful taxonomic decisions. Such a more accurate barcode would have to be based on highly variable markers in the nuclear genome, and development of such a method would have to utilize multiple loci (hybridization and introgression, for example, cannot routinely be diagnosed by sequencing only a single locus).

The need for this more sophisticated tool ('gold standard') should not be seen as an argument to delay implementation of single-locus barcoding ('silver standard') until a time when we have developed such methods. We advocate the immediate development in land plants of a silver standard system based on one or two plastid DNA regions plus perhaps ITS (in those groups for which it has been demonstrated to work well, e.g. orchids). This system would serve the needs of the wider scientific community that needs rapid and reasonably accurate identification of unknowns. While this phase is being implemented, another effort should be made to develop this more sophisticated gold standard method (see below).

We can imagine that in fact a two-step barcoding system with a traffic light approach might eventually be routinely used (we thank Kenneth Cameron, New York Botanical Garden for this idea). For many groups of organisms and many applications, the first step would be production of the uniparentally inherited barcode, and then the process would end with this crude answer. If the result was a name for which we know the taxonomy is simple and robust, then the name would be produced with a ‘green light’ beside it, meaning that this identification is uncomplicated and clear, whereas if a ‘yellow light’ appears then the identification is from a group that has some problems, and the user can decide if a more detailed investigation is necessary to satisfy the level of precision required. If the third situation is encountered, then a ‘red light’ appears, and the user would be informed that the identification is highly likely to be inaccurate because it belongs to a species complex in which phenomena such as hybridization/introgression are commonplace. In such a situation, the user could still stop there if there might be no need for a more accurate identification, but in many cases greater precision would be desirable, in which case the more diagnostic procedure could take place, provided that high quality DNA is available. Of course, for most taxonomists, the only reasonable barcode would be the one based on multiple nuclear DNA loci, the gold standard barcode.

5. DEVELOPMENT OF A NUCLEAR MULTI-LOCUS BARCODE AS A TOOL FOR INVESTIGATING SPECIES LIMITS
To develop this multi-locus barcode (MBC) system, there would need first to be an effort to identify conserved sites flanking regions containing variable sequences, most notably introns of appropriate size. These conserved sites would serve as universal PCR priming locations in all land plants and could enable amplification of these variable regions in products of an appropriate size for single-pass sequencing reactions. This phase of the project would take advantage of the completely sequenced genomes of Arabidopsis, Populus and Oryza and the EST libraries of a diverse range of plants that are now populating GenBank. An algorithmic approach to the identification of potential sites could be automated to develop over 1000 such sites that will then be evaluated on a set of DNA samples representing all major clades of land plants. For instance, Kozik & Michelmore (2003) identified a provisional list of over 3000 putative single copy genes for Arabidopsis. Although an unknown proportion of these candidates will inevitably prove to be incorrectly classified, we are in the process of refining this list to assemble a first set of primers for potential use on all higher plants. Testing these on DNA samples representing pairs of closely related species that have previously been studied with some of the standard phylogenetic markers, such as the Moraea and Protea studies described above, would provide evidence of the relative levels of sequence variability in the newly developed loci and permit selection of those appropriate for further testing. To be practical, these putatively low-copy markers would need to satisfy a set of requirements similar to those already widely discussed, particularly with respect to length, which is even more critical if such nuclear loci are to be routinely amplified from DNA samples of highly variable quality.

The development and subsequent application of the MBC as a tool to investigate species limits will utilize the large numbers of intact DNA samples that are routinely associated with the field of population genetics. Unlike plastid and nuclear rDNA regions that can be amplified from highly degraded DNA samples due to their highly iterative nature, the MBC method will be much more subject to failure to amplify from even moderately degraded DNA samples, but this should not be seen as a fatal flaw. This system would not be a replacement for the simpler, more robust, but less sophisticated methods. If six to eight appropriately variable regions are identified, then the MBC might consist of 50–100 base pairs of DNA sequence from each region, perhaps then sequenced with some of the newer techniques, such as pyrosequencing (Pacey-Miller & Henry 2003; Mochida et al. 2004) that do not produce such long sequences as do standard methods but do so with much greater speed.

The reason why the MBC has to be multi-locus is because detection of hybridization/introgression cannot be reliably accomplished by examination of a single nuclear DNA region. An F1 hybrid on the other hand can be detected with DNA sequences from a single locus because it will be heterozygous at most if not all loci, but F2 hybrids or backcrossed progeny (to one or the other parent) will be homozygous at many loci, so dependence on a single locus could be misleading.

It has been suggested that once barcodes have been established for most species a quick way to identify these taxa would be array-based rather than sequence
based. If one were studying seed germination and seedling establishment in set of previously studied forest plots in which all adult trees had been barcoded, then an array could be developed with dots for each species, and one could expect then to be able to see which species were present as bright spots on the array. If an ecologist were instead working on many previously unstudied plots, e.g. in the tropics, then it would be important to have a barcoding system from which it could be expected that species previously unknown from those plots could be identified. Ultimately, sequence-based barcodes will have greater power and more applications than array-based technologies, but we acknowledge that there may be specific high-throughput applications where an array-based approach offers a quick and practical solution.

6. A PLEA FOR DNA BANKING

To make as rapid progress as possible, nearly all early efforts to establish a set of reference standards have focused their attention on using herbarium or museum specimens (Kress et al. 2005, for example). It is impractical to expect, particularly in the short run, that a representative of each species on Earth can be newly collected. Such efforts would have to be global in scale, and there are many political as well as practical difficulties that make such efforts unrealistic. Therefore, reliance upon already collected material (often with highly degraded DNA) has been viewed as an imperative. This is one of the reasons why plastid and mitochondrial DNA have been viewed as the most appropriate regions to sequence: these highly repeated genomes are the most likely to survive reasonably intact within herbarium and museum specimens. In the rush to get a barcoding system up and running quickly, procurement of more intact materials has been ignored, but this is something to which more attention needs to be paid. In the course of work on barcoding, it should be possible to assemble DNA banks from those samples that are of higher quality. These high quality samples would be the basis upon which the MBC can be established. Given the high rates of global extinction, it is imperative that DNA banks of high-quality DNAs be established (Savolainen & Reeves 2004; Savolainen et al. in press). There are a number of large DNA banks already in operation, such as the one at the Royal Botanic Gardens, Kew, which currently holds 23,000 DNAs (www.rbgkew.org.uk/data/dnaBank/homepage.html). These DNAs are available to researchers worldwide, and many of these samples could become the standards used in the development of the MBC system for plants. The Royal Botanic Gardens, Kew, is willing to hold tissue and extract DNAs from plants collected anywhere in the world, with no charge to the collectors, and handling of these samples is done in accordance with the international Convention of Biological Diversity, so that the rights of countries of origin to profits derived from exploitation of their genetic resources are taken into account. Such partnerships have been already successfully put in place between the Royal Botanic Gardens, Kew, and South Africa under the umbrella of the UK Darwin Initiative for the Survival of Species. Before it is too late, DNA samples from as wide a range of organisms as possible should be assembled and curated. If, during the rush to barcode every organism, we lose sight of the need to document their genetic makeup, it would be most tragic. At the least, we seem to be well on our way to banking DNA from a large percentage of land plants, and this effort must be expanded as quickly as possible.

REFERENCES


Microcoding: the second step in DNA barcoding

R.C Summerbell, C.A Lévesque, K.A Seifert, M Bovers, J.W Fell, M.R Diaz, T Boekhout, G.S de Hoog, J Stalpers and P.W Crous

Phil. Trans. R. Soc. B 2005 360, 1897-1903

References

This article cites 30 articles, 8 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1897.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1897.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here
Microcoding: the second step in DNA barcoding

R. C. Summerbell1,*, C. A. Lévesque2, K. A. Seifert2, M. Bovers1, J. W. Fell3, M. R. Diaz3, T. Boekhout1, G. S. de Hoog1, J. Stalpers1 and P. W. Crous1

1CBS Fungal Biodiversity Centre, Utrecht, The Netherlands
2Biodiversity (Mycology and Botany) Agriculture and Agri-Food Canada, Ottawa, Canada
3Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA

After the process of DNA barcoding has become well advanced in a group of organisms, as has in the economically important fungi, the question then arises as to whether shorter and literally more barcode-like DNA segments should be utilized to facilitate rapid identification and, where applicable, detection. Through appropriate software analysis of typical full-length barcodes (generally over 500 base pairs long), uniquely distinctive oligonucleotide ‘microcodes’ of less than 25 bp can be found that allow rapid identification of circa 100–200 species on various array-like platforms. Microarrays can in principle fulfill the function of microcode-based species identification but, because of their high cost and low level of reusability, they tend to be less cost-effective. Two alternative platforms in current use in fungal identification are reusable nylon-based macroarrays and the Luminex system of specific, colour-coded DNA detection beads analysed by means of a flow cytometer. When the most efficient means of rapid barcode-based species identification is sought, a choice can be made either for one of these methodologies or for basic high-throughput sequencing, depending on the strategic outlook of the investigator and on current costs. Arrays and functionally similar platforms may have a particular advantage when a biologically complex material such as soil or a human respiratory secretion sample is analysed to give a census of relevant species present.

Keywords: macroarrays; species identification; barcoding; fungi

1. INTRODUCTION

The concept of DNA barcoding emerged from a background of molecular phylogenetic analysis (Hebert et al. 2003). Scientists working with fungi, like those working with other microorganisms, were attracted to molecular phylogenetics very early in its history. This was partly because the organisms they dealt with were morphologically enigmatic, and thus biosystematically intractable, and partly because most of the organisms could readily be grown in culture, facilitating centralized DNA sequence comparison of strains sampled all over the world (e.g. O’Donnell et al. 1998, Kurtzman 1994, Scorzetti et al. 2002). Comparative biosystematics by nature demands the comparison of homologies—folk wisdom about comparing apples to apples and not to oranges applies here—and therefore there was a rapid emergence of a very small number of easily sequenced gene loci that could be roughly used as comparison standards at different taxonomic levels. For the most part, the nuclear ribosomal internal transcribed spacer (ITS) regions and the variable D1/D2 domains within the 28S ribosomal subunit were used at the species and generic levels. The 28S subunit and the 18S ribosomal subunit were utilized for taxonomic levels above the genus. In a few fungal groups where these ribosomal regions were clearly shown to provide inadequate resolution, one or two additional housekeeping gene loci such as translation elongation factor 1α (EF-1) and β-tubulin were also brought into play. Though the appropriateness of this strong concentration on a few individual genes was occasionally questioned—there were legitimate concerns, for example, about individual gene trees being mistaken for species phylogenetic trees (Taylor et al. 2000)—what this intuitive comparative standardization accomplished was to introduce basic DNA barcoding to fungal microbiology long before the expression was coined.

It should be noted that there was no tradition in mycology of using the mitochondrial cytochrome oxidase subunit I gene (cox1) used for most DNA barcoding of animals (Hebert et al. 2003). Preliminary results with Penicillium, however, indicate that it may be of interest and should be investigated further (K. Seifert, A. Lévesque, unpublished data).

For several years now, many thousands of fungal ribosomal sequences as well as a significant representation of EF-1 and other sequence types have been available in public repositories. Most fungal groups of any economic or practical importance are already represented to some degree. A significant proportion of the fungi represented in sequence databases are misidentified (de Hoog & Horré 2002, Hawksworth 2004, Lévesque & de Cock 2004, Kopchinskiy et al. 2005), and many more are unvouchered (Crous 2002) and thus stripped of their biological context. Nonetheless, careful sifting of the data for sequences connected to accessible, nomenclaturally significant isolates (ex-type strains, biosystematically well studied strains, etc.) allows assembly of a useful preliminary
barcode set for numerous fungal taxonomic and ecological groups. The question then becomes how most efficiently to press this information, plus any comparable new sequences investigators wish to contribute, into service to accomplish one of the major aims of DNA barcoding, that is, facilitating identification. It should be noted that most twentieth century molecular identification approaches that relied on comparison of nucleic acid or protein band migration rates were vulnerable to coincidence (non-homologous bands migrating at essentially the same rate) and other factors limiting their resolution. The very sharp specificity of sequences has thus become the unquestioned molecular identification gold standard in recent years. The application of sequencing remains preliminary in insufficiently studied fungal groups where sequence types are not yet known to correspond to well-delimited species or other biosystematically or ecologically relevant units, but in many groups, sequences can now be straightforwardly used for species identification, especially when specialized databases are consulted (e.g. the Fusarium Database, \texttt{http://fusarium.cbio.psu.edu/}). Investigators trying to devise effective routine species identification systems for fungi that have already been barcoded soon find that they are at 'a fork in the road', in that one of two strategies must be followed in sequence-based identification. Since both strategies can be conceived of as subtypes of DNA barcoding, let us call them basic barcode identification—the use of whole gene or large partial gene amplicons in sequence identification—and microcoding, the use in identification of oligonucleotide sequences that are often not much longer than PCR primers. Both strategies are becoming increasingly rapid and cost-effective over time. Some strengths and weaknesses of these approaches are discussed here. Though these approaches are mainly discussed in the context of current examples involving economically important, culturable fungi, parallel considerations are expected to become salient in work with many other groups of organisms, including other fungal groups, as these groups reach the state where a significant amount of primary barcoding work has already been accomplished.

2. BASIC BARCODE IDENTIFICATION

Rapid development in sequencer and DNA handling technologies has made it ever more practical to routinely sequence whole-gene or large partial-gene amplicons as an identification technique. A front-running and well-known example in the field of medical mycology was provided by Pryce et al. (2003), who showed that all but a very small number of medically important fungal species growing in culture could be identified by full-length ITS sequences obtained within 24 h and costing less than AU$ 10 (~US$ 7.50 or € 6.25) per sequence inclusive of materials and labour. Unlike the many PCR-restriction fragment length polymorphism techniques developed in the immediately preceding period of technological development, this allowed a large number of species from a very broad taxonomic range to be handled in a maximally uniform way. Some years prior to this publication, important US reference centres for basidiomycetous (Fell et al. 2000) and ascomycetous yeasts (Kurtzman 1994) as well as ascomycetous species (O’Donnell et al. 1998) had already altered procedures to identify (or, in the case of poorly resolved taxa, cluster) essentially all incoming isolates by means of well-chosen sequence regions. This was made possible in part by the use of high-throughput sequencing apparatus and robotics.

The main obstruction to using a uniform basic barcoding procedure for rapid identification of fungi is that it is a major challenge to find DNA extraction and PCR procedures, as well as primer pairs, that will reliably yield suitable high-quality, long sequence reads in rapid cycle sequencing with a wide range of fungi. This tends to be true even for loci such as ITS with highly conserved flanking regions, theoretically reliably amplified by ‘universal primers.’ Buried in the seemingly minor variations in many mycological papers’ materials and methods sections are the results of months of struggle with isolates that at first seemed completely resistant to ‘universal’ sequencing or even to PCR. Currently, it is seldom possible to predict which fungi will prove difficult to sequence, and this means that there is no way as yet to make \textit{a priori} adjustments to streamline basic barcode identification in such groups. There are a few clues available. Anecdotally, it can be suggested that slow-growing filamentous fungi with dense colonies, especially when these colonies are heavily pigmented, appear to be especially likely to generate cycle sequencing problems. Recently, for example, repeated, concerted and technically varied attempts by two very competent molecular mycology laboratories to obtain ribosomal sequences for the dark, slow-growing new species \textit{Oidiodendron fimicola} (Rice & Currah in press) were unsuccessful, and the species ultimately had to be described based on its phenotype. Other groups of fungi where sequencing is difficult may not offer any obvious clue to this state of affairs. Even when isolates have phylogenetic affinity with easily sequenced groups, this may not be predictive of success in normal barcoding procedures. For example, in the present first author’s laboratory, we have recently dealt with members of a group of \textit{Acremonium} species, including \textit{A. ochraceum} and \textit{A. bacillisporum}, that largely withstood ITS and 28S sequencing attempts involving various primers and purification techniques (ranging from high-cost affinity columns to primitive cetyltrimethylammonium bromide procedures). It was ultimately found that they could be sequenced once 3% dimethylsulfoxide was added to the amplification step of PCR (Demeke & Adams 1992). Meanwhile, successful 18S sequences done over 2 years earlier without problem had already revealed that the recalcitrant species were a close sister group, or perhaps even an internal clade, related to the uniformly easily sequenced clade containing \textit{Acremonium strictum} and \textit{A. kilianse} (Bills et al. 2004). The members of the easily sequenced group and the difficult sister group were very similar in phenotype except that the latter produced conidia not in the mucoid clumps typical of the former group, but rather in chains, signalling a switch to airborne conidial dissemination. This switch likely involved at least one major change in fungal

\textit{Phil. Trans. R. Soc. B} (2005)
biochemistry, namely the additional production of hydrophobic anti-desiccation substances, but at present, we cannot directly attribute difficulty in sequencing to this factor. Perhaps in future, however, links may be discovered between cycle sequencing difficulties and certain readily discernible phenotypic, ecological or chemical properties. Authors are advised to briefly document such difficulties in their own studies, and to look for any patterns that could be used in predicting the occurrence of similar problems in other fungi.

In some cases, isolates that are problematic in cycle sequencing can be sequenced via cloning procedures. In the case of our A. ochraceum-complex isolates, however, the problem interfering with cycle sequencing appeared also to obstruct cloning; such findings are not uncommon in our experience. In any case, cloning requires a specially licensed laboratory in many countries (e.g. The Netherlands) and may be held in low favour because of its relative inconvenience and, in certain circumstances, its vulnerability to random selection of paralogous gene forms that may be present at low copy number.

Fungal identification by basic barcoding is also vulnerable to the common molecular phylogenetic problem, seen in many groups of organisms, that ‘one size does not fit all’ in molecular sequence identification. Some predominantly phytopathogenic groups like Gibberella (anamorphs: Fusarium subgenus Liseola), perhaps influenced by the constant generation of variation that is of advantage in pathogenicity (Brasier 2000), and, partially in consequence of this, by relatively rapid speciation, show poor resolution at the species level in concertedly evolving multicycloc loci such as ITS; in some cases they have also developed paralogous forms that complicate both cycle sequencing and cloning procedures (O’Donnell & Cigelnik 1997). In such cases, recourse must be made to single-copy genes such as EF-1. These in turn then must occasionally be optimized for primer choice and PCR techniques, as flanking regions and topologies may not be as highly conserved as in the ribosomal coding regions. Single-copy genes that work well in some fungal groups may be troublesome in others.

Development of a uniform barcode standard for species identification in fungi is thus hindered by the variation in the evolutionary ages of species in different groups. Given that there can be no rules in biology for how quickly or slowly functional species may evolve, construction of a uniform barcode identification procedure may be regarded as a theoretical impossibility. In some fungi, for example, minor changes in the mating type loci, followed by inbreeding, may be sufficient to initiate a newly separated sexual lineage intersterile with its parental forerunner. This lineage may rapidly emerge as a new biological species (Aanen et al. 2000) that will inexorably diverge evolutionarily from its forerunner even though, early in its evolutionary history, the sequences of its housekeeping genes are mainly unchanged from the ancestral type even in the introns and spacers. The real possibility of such events occurring merely entails that barcode identification may need to follow nested protocols to deal with known rapidly evolving species groups, or that investigators may need to do a preliminary glance at morphology in order to make the correct selection of barcode genes to examine for members of particular fungal groups. Much like the image of a barcode that is over 500 characters long, however, these contingencies do tend to detract from the simple vision of a tricorder-like reading of fungal biodiversity (Godfray & Knapp 2004; If the term ‘tricorder’ is unfamiliar, it is the fictional device in the science fiction drama series Star Trek that can be pointed at life forms to immediately read their species identity and physiological properties).

To a certain extent, the PCR and cloning inhibition problems discussed above also affect techniques based on use of microcodes, and the inability of any single gene region to serve as an identifying standard for all fungi applies a fortiori when gene regions are represented by small subregions. With microcode procedures, however, there is a built-in expectation that organisms will be batched in advance into particular groups or that analytical platforms will be tailored to specific taxonomic or ecological circumstances, so this is not generally perceived as a disadvantage. If the difficulties in using basic barcoding in rapid identification can be optimistically summarized with the phrase, ‘some customization may be necessary for certain species groups’, then it is clear that microcoding procedures, by nature, cannot make a better offer. For the most part, they can only become relatively advantageous by being faster, cheaper or more convenient to handle than full barcoding techniques. The circumstances in which they might achieve these advantages are not yet well resolved. Nonetheless, these techniques appear to offer a potential for very rapid handling, and there has been considerable interest in their development despite the increasing convenience of routine whole-gene sequencing.

### 3. MICROCODING (OLIGONUCLEOTIDE BARCODING)

In recent years, as indicated above, an increasing number of techniques have been deployed that are predicated on the use of distinctive short nucleic acid segments—that is, short DNA barcodes—in rapid identification procedures. Among these are silicon-based microarrays, nylon-membrane-based macrosays, and the patented Luminex system of DNA-tagged polystyrene beads sorted by flow cytometry. These systems share the feature that fully sequenced genes are studied to find small, thermodynamically stable and non-hairpin-forming areas of high sequence uniqueness that can then be used as fixed single-stranded oligonucleotides on an identification platform. The bound oligonucleotides then anneal specifically to matching, complementary DNA regions in a test solution bearing labelled single-stranded amplicons from one or more unidentified strains. These systems can only be developed as a second step in DNA barcoding procedures for any given group of organisms, because all the relevant basic barcodes need to be assembled before the unique short segments useable as oligonucleotides can be discerned. Because of practical considerations related to how many oligonucleotides...

*Phil. Trans. R. Soc. B (2005)*
can be proof-tested as unique in relation to one another and then attached together in a common test platform, oligonucleotide-based systems so far have been limited to groups not larger than circa 200 test species. For example, it was found that macroarrays spotted with oligonucleotides for soilborne Oomycota in the genera *Pythium* and *Phytophthora* (as per Lévesque et al. (1998) but with an updated version of the array) facilitated the early detection of the biocontrol agent *Pythium unn* in soybean roots early in the growing season (figure 2), when this species was present only at inoculum levels lower than those of several other oomycetous species (Anonymous 2002). Coupled with appropriate techniques for unbiased, proportionate extraction and amplification of DNA of all relevant organisms from substrate material such as soils, this technique does offer a rapid census of populations present in otherwise inscrutable material, giving a genuine foretaste of tricorder-like biodetection.

The selection of appropriate microcodes can be greatly facilitated by the use of software that can go through a large, multi-species alignment (or even an unaligned collection) of full-length barcodes and determine the best signature microcode regions that also are free of thermodynamic and conformational problems. Success in such automated microcode selection has been reported by Seifert & Lévesque (2004) with the software package Signature Oligo, which has been used, for example, to select useable diagnostic oligonucleotides from sequences of mycotoxigenic *Aspergillus*, *Fusarium* and *Penicillium* species. Oligonucleotides can be selected specific to any clustering level in sequence dendograms, not only allowing the detection of variant organisms not represented by individual species-level microcodes, but also providing a level of redundancy serving as a control against false-positive results (e.g. if the *Fusarium thapsinum* microcode gives a positive signal in macroarray analysis, the overarching general *Fusarium* subgenus *Liseola* microcode should also be positive). A second software package, Array Designer 1.1, is used to ensure that the oligonucleotides selected have a suitable melting temperature (55 °C at 6×SSC) and lack dimers and hairpins. It can also be used to BLAST (Altschul et al. 1990) a batch file of microcodes against GenBank to monitor their uniqueness. As Seifert & Lévesque (2004) state, ‘it is surprising how often 18–22 bp oligonucleotides have no BLAST matches apart from their own sequences’. Occasional random matches that are found tend to have no practical consequence, Seifert and Lévesque explain, because such coincidences have an extremely high likelihood of involving parts of completely different gene regions from organisms considered unrelated to the groups under consideration in the arrays. Even if
the fortuitously coinciding DNA were present in a sample being analysed by a microcode-based technique, it would not give a positive result because the gene containing it would not amplify with the PCR primers used.

A similar principle for the selection and use of unique microcodes is embodied in the Luminex system (Spiro et al. 2000, Diaz & Fell 2004). In Luminex, however, sensitivity is enhanced by attaching the synthesized oligonucleotide segments to polystyrene microspheres that can be read, one by one, in a flow cytometer. Each oligonucleotide is partnered with one of 100 types of differently coloured microspheres; each of the 100 colours available is identifiable when probed with a red laser. The amplicons from the test material are tagged with a fluorescent marker readable with a green laser. In using the system to identify, for example, a fungal culture, the 100 types of differently coloured beads partnered with 100 different oligonucleotides are exposed to the test solution of single-stranded, fluorescently tagged amplicons derived from the culture. The amplicon then, at the correct temperature, anneals with the unique complementary species-specific oligonucleotide that is attached to just one of the microsphere types. It may anneal with only that oligonucleotide or, depending on the design of the test, it may also anneal with another, hierarchically supervening oligonucleotide such as that specific for the relevant fungal genus. As the microspheres are channelled through the flow cytometer, they are simultaneously read by two lasers, the red one classifying the microspheres and the green one showing which microspheres bear the annealed DNA tagged with the fluorescent marker. Putting these two data sets together, the identification of the species involved becomes clear.

The use of microcode-based oligonucleotides for species identification in the Luminex system requires the same selection for uniqueness and thermodynamic and conformational appropriateness that must be used for microarray and macroarray technologies. One disadvantage of all such systems is that this process must be carefully done, and then the identification platforms must be thoroughly tested for sensitivity and specificity using all species and other genetic types that are intended to be detected. In particular, cross-reactions among oligonucleotides for closely related species may be difficult to preclude based on computer analysis alone, and occasionally, an oligonucleotide will need to be replaced with a more robustly specific one, or the distinction of two closely related taxa may need to await further testing (Diaz & Fell 2004).

The limited number of species that can be identified by any given array or sphere set may be problematical in some contexts, e.g. in general fungal soil ecology, where more than 100–200 organisms may be potentially significant in a given microhabitat. Of course, use of a second array or sphere set doubles the number of species that can be identified. Similarly, if some species are identified based on ITS sequences and others must be analysed with another locus such as EF-1 in order to establish specificity, either two arrays or sphere sets need to be analysed, or the platforms must be configured with the oligonucleotides corresponding to

Figure 2. Macroarray exposed to *Pythium*- and *Phytophthora*-specific nuclear ribosomal internal transcribed spacer-region amplicons obtained from a DNA sample extracted from field-grown soybean roots (T. Barasubiye, Agriculture and Agri-Food Canada, unpublished). Dark spots appear where specific oligonucleotides have captured labelled amplicons corresponding to various species present on the roots. The array is covered by a mask indicating the identity of the various spots seen. ‘X’ symbols are placed over known cross-reacting oligonucleotides.
signature microcode regions in both amplicons types. In connection with latter option, both amplicon types will need to be simultaneously or serially exposed to the detection platforms, increasing the amount of work that needs to be done to obtain an identification. Optimization of multiplex PCR for the two gene regions tested may minimize the extra work involved in such identifications. Microarrays can be used with labelled whole-cell DNA, so strictly speaking, amplification is not necessary (Trad et al. 2004). Clearly, however, the use of amplicons would improve sensitivity by ensuring that a relatively high quantity of compatible DNA was present for binding with the appropriate complementary oligonucleotide.

It is unclear if the use of microcodes would alleviate the problem, mentioned above in the context of basic barcode identification, of some species being difficult to analyse because difficult-to-remove inhibitors in their cellular chemistry obstruct one or more of the PCR-related steps involved in cycle sequencing. Such organisms often seem to show adequate DNA amplification, and they frequently even yield short sequence runs of varying quality. It is thus possible that microcode platforms may successfully detect and identify them even though the amplicons involved are of inadequate quality for full barcode reading. The matter has not, to our knowledge, been tested.

One feature of microcodes is that, being lower in information content than the full-length barcodes that served as their sources, they tend by nature to be a less sensitive taxonomic indicator than the full barcodes. This may, however, be an advantage in some cases. In a relatively genetically complex species, there may be considerable minor variation within the entire barcode region, but the statistical probability that this variation will have affected the particular small region designated as the microcode is relatively small. Thus the use of microcodes in identification may minimize the potential influence of distracting mutational ‘noise’ and provide users with clear, unambiguous results. By the same token, however, there is always the chance that an unknown and unbarcoded sibling species will be misidentified as its already described sibling based on 100% microcode similarity. With full-length barcodes, the chance is much greater that information distinguishing the unknown species from the known would come to light. Recently, a specialized BLAST search program for Trichoderma species was made available (Kopchinskiy et al. 2005; www.isth.info) based on a sophisticated use of microcodes to increase identification accuracy beyond the level obtainable with standard whole-gene BLAST searching. In part, this was to diminish the influence of long, highly homologous regions that strongly influence sorting in common BLAST procedures (Kopchinskiy et al. 2005). The full potential of microcodes as a diagnostic tool has not yet been circumscribed.

4. MICROCODING FOR ORGANISMS OTHER THAN FUNGI

Microcoding has mainly been used with culturable organisms, partly because many were sequenced at signature barcoding regions long ago and partly because plenty of material is always available for extraction of the DNA used in microcode test development. Microcodes may be very advantageous, however, in the identification of organisms that are inconvenient for artificial cultivation or husbandry. In application to living organisms, the strengths and weaknesses of these techniques do not differ significantly from those mentioned above in connection with culturable fungi and Oomycetes. Where microcoding methods may have unique value is in the DNA-based identification of preserved specimens, e.g. plant herbarium specimens, with degraded DNA (but see Chase et al. 2005). Presumably in most such degraded DNA, the less-than-25-bp segments used in microcoding are much more likely to remain intact than whole barcodes over 500 bp long. Double-ended ligation of restriction-enzyme-pretreated DNA fragments to DNA linkers serving as templates for PCR amplification, similar to the process used in amplified fragment length polymorphism (AFLP) analysis, could speed the process of identifying preserved specimens. Various related techniques are possible. Proof of principle for the convenience of short signature DNA regions in such situations is provided by the successful identification and strain typing of Mycobacterium spp. strains from Egyptian mummies using the spoligotyping (spacer oligotyping) technique (Zink et al. 2003). This technique is based on membrane array detection of amplicons derived from a series of 34–41 bp spacer regions in the chromosomal DR (direct repeat) locus. Each spacer is separated from neighbouring spacers by a conserved direct repeat region that can serve as a primer binding site for amplicons spanning single short spacers or pairs or triads of adjacent spacers (Kamerbeek et al. 1997). The spoligotyping technique itself is not directly applicable to most non-mycobacterial organisms, as far as is known, but, as one of numerous possible alternative strategies based on similar principles, conserved primers closely flanking signature microcode regions within closely related species groups could be designed.

The microcodes used for identification would first, naturally, need to be demonstrated as sensitive and specific in prior testing with living material of the same species, if this is at all possible (i.e., if the species in question is not extinct). The use of microcoding techniques, however, offers to redeem many DNA-degraded taxonomic specimens that belong, for example, to closely interrelated species complexes where species identification can now only be accomplished using molecular or other in vitro techniques. Such techniques may also allow forensic DNA analysis of materials that contain DNA that has been damaged but not completely destroyed, e.g. lightly burned or dried, heavily pulverized material. Ultimately, with microcode platforms prepared with various selections from complete genome sequence databases, extensive genomic analysis may be possible for materials bearing only fragmented DNA. Such prospects have been widely publicized in connection with microarrays, but recent biosystematic work has shown, as summarized above, that more tractable and cost-effective alternatives may exist.

Phil. Trans. R. Soc. B (2005)
REFERENCES


Anonymous 2002 Early season diseases take their toll. Dec.


Anonymous 2002 Early season diseases take their toll. Dec.


Crous, P. W. 2002 Adhering to good cultural practice (GCP).

Kamerbeek, J., Schouls, L., van Agterveld, M., van

Kopchinskiy, A., Komoni, M., Kubicek, C. P. & Druzhinina, I. S. 2005 TrichoBLAST, a multi locus database for


Rice, A. V. & Currah, R. S. In press. Profiles from Biolog FF plates and morphological characteristics support the recognition of Oidiodendron fimicola sp. nov. Stud. Mycol. 53.


The unholy trinity: taxonomy, species delimitation and DNA barcoding

Rob DeSalle, Mary G Egan and Mark Siddall

Phil. Trans. R. Soc. B 2005 360, 1905-1916

References

This article cites 64 articles, 10 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1905.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1905.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
The unholy trinity: taxonomy, species delimitation and DNA barcoding

Rob DeSalle*, Mary G. Egan and Mark Siddall

Division of Invertebrate Zoology, American Museum of Natural History, 79th Street at Central Park West, New York, NY 10024, USA

Recent excitement over the development of an initiative to generate DNA sequences for all named species on the planet has in our opinion generated two major areas of contention as to how this ‘DNA barcoding’ initiative should proceed. It is critical that these two issues are clarified and resolved, before the use of DNA as a tool for taxonomy and species delimitation can be universalized. The first issue concerns how DNA data are to be used in the context of this initiative; this is the DNA barcode reader problem (or barcoder problem). Currently, many of the published studies under this initiative have used tree building methods and more precisely distance approaches to the construction of the trees that are used to place certain DNA sequences into a taxonomic context. The second problem involves the reaction of the taxonomic community to the directives of the ‘DNA barcoding’ initiative. This issue is extremely important in that the classical taxonomic approach and the DNA approach will need to be reconciled in order for the ‘DNA barcoding’ initiative to proceed with any kind of community acceptance. In fact, we feel that DNA barcoding is a misnomer. Our preference is for the title of the London meetings—Barcoding Life. In this paper we discuss these two concerns generated around the DNA barcoding initiative and attempt to present a phylogenetic systematic framework for an improved barcoder as well as a taxonomic framework for interweaving classical taxonomy with the goals of ‘DNA barcoding’.

Keywords: DNA barcoding; taxonomy; species delimitation; muntjac; leeches; sturgeon

1. INTRODUCTION: BUILDING A BETTER DNA BARCODER

One of the major issues concerning the inclusion of molecular information into taxonomic aspects of biology that has yet to be discussed in detail in the commentaries on this subject is concerning the best way to read the barcodes. There are two separate tasks to which DNA barcodes are currently being applied. The first is the use of DNA data to distinguish between species (equivalent to species identification or species diagnosis) and the second is the use of DNA data to discover new species (equivalent to species delimitation, species description). These two activities differ in the types and amount of data required. Below we highlight some of the issues that may limit the utility of current DNA barcoding endeavours (especially those used for species discovery) and suggest a framework for the development of a barcoder that addresses these issues.

(a) The barcoder engine: distances or characters?

A major issue that needs to be resolved is how to read the organisal barcode once it is generated. Most recently published approaches to DNA barcoding have utilized distance measures to make the inference as to species designation (Hebert et al. 2003a,b, 2004a,b). Distances are used in two major approaches; the first is a simple BLAST (Altschul et al. 1990) approach where a raw similarity score is used to determine the nearest neighbour to the query sequence. The second approach utilizes distances in tree building (Hebert et al. 2003a,b). We point out the following shortcomings with these approaches and further suggest that character based approaches are more appropriate for DNA barcoding both for theoretical and for practical reasons.

A major shortcoming of using distances in DNA barcoding is that all classical studies and taxonomic schemes that accomplish the same thing that barcodes are meant to accomplish are character based, making the union of classical and DNA barcoding a difficult process if the use of distances is continued in barcoding studies (see below). This shortcoming also is related to the need for diagnostic characters that classical studies use to validate the existence of a species. A second shortcoming is that similarity scores often do not give the nearest neighbour as the closest relative (Koski & Golding 2001). Nevertheless, similarity scores will always give a nearest neighbour. Character based methods have the logical advantage that when diagnostic character data are lacking, they will fail to diagnose, allowing for a degree of hypothesis testing not available when using distances. A third shortcoming involves the lack of an objective set of criteria to delineate taxa when using distances. For example, a universal similarity cut-off to determine species status will simply not exist, because of the broad overlap of inter- and intra-specific distances (Goldstein et al. 2000). Researchers will have to constantly revise their similarity cut-offs from group to group. We suspect that

* Author for correspondence (desalle@amnh.org).

One contribution of 18 to a Theme Issue ‘DNA barcoding of life’.

2005 The Royal Society
distance-based criteria for different species groups within genera often will have different parameters, making the delineation of species using distances fairly subjective.

We suggest that an alternative approach including character based phylogenetic analysis is more appropriate for establishing or ‘printing’ barcodes. The character based approach is compatible with classical approaches allowing the combination of classical morphological and behavioural information. Character based approaches sidestep the nearest neighbour problems of distances because they can reconstruct hierarchical relationships where common ancestry is inferred when two entities share derived characters. Neither BLAST (Altschul et al. 1990) nor neighbour joining (NJ; Saitou & Nei 1987) tree building approaches allow for character-by-character diagnoses on branches of trees. Any such diagnosis would need to be Parsimony or Maximum Likelihood based. Furthermore, the diagnosis of two separate entities in nature can be accomplished by the existence of a single character shared by a group of organisms to the exclusion of others, whether it be a DNA character or a morphological character.

(b) The barcoder engine: to tree or not to tree, that is the question

Given that character based information is a viable alternative to the distance-based approaches already implemented in barcoding studies, the question arises which approach to analysis of characters is more appropriate for barcoding—the non-tree based population aggregation analysis approach (PAA; Davis & Nixon 1992) or a tree based approach. There are several drawbacks to the use of tree-building approaches to species identification. The first relates to the use of distances (described above) to construct trees. But problems with tree building are not limited to trees constructed with distance data. Rather, the second drawback is in the use of single gene trees as evidence of phylogenetic relationships. Several studies have demonstrated both theoretically (Kluge 1989) and empirically (Rokas et al. 2003; Gatesy et al. 2004) that combined analyses of multiple data partitions yield better representations of evolutionary history than single gene trees. The combined analysis approach has the additional advantage that it allows for the exploration of the character contribution of data partitions to the combined tree and can reveal character support for the combined tree that was not evident in separate analyses of individual gene partitions (Baker & DeSalle 1997; Baker et al. 1998; Gatesy et al. 1999, 2002, 2003). From these advantages it could be argued that a corroborated total evidence tree could be used as a guide tree for identifying the phylogenetic affinities of an unknown individual’s sequence, assuming that the query sequence is one of the gene regions used to construct the total evidence tree. There are bioinformatics tools to aid in the placement of a query sequence based on the presence of shared characters that are diagnostic for nodes on the tree (Sarkar et al. 2002). However, there is a third drawback to the use of a tree building approach to species identification. This relates to the use of hierarchical methods (tree building) and terminology (monophyly as a criterion for species delimitation) when the underlying system (of individuals and populations) is not a hierarchical system of ancestor–descendant relations, redefinitions of monophyly as reciprocal monophyly (Avise et al. 1987; Avise 1989; Avise & Ball 1990) or exclusivity (Baum 1992; Baum & Donoghue 1995; Baum & Shaw 1995) notwithstanding (reviewed in Goldstein & DeSalle 2000). A more practical alternative is the exploration of character diagnostics in the sequences themselves without reference to trees. This mirrors the two-step procedure of traditional taxonomic studies in which relationships among species are assessed only after the terminals in the analysis (in this case, species) are first identified by diagnostic characters. In this approach as formulated by Davis & Nixon (1992), sequences are examined using PAA (Davis & Nixon 1992). Diagnostics are accepted if they are fixed and different from aggregate to aggregate of organisms—such diagnostics are termed ‘pure’ (Sarkar et al. 2002). This approach and its relevance to species delimitation has been discussed at length (Davis & Nixon 1992; Goldstein & DeSalle 2000; Goldstein & DeSalle 2003; Goldstein et al. 2000; Nixon & Wheeler 1990) and its relevance to diagnosing entities in nature has been discussed both from the technical and theoretical standpoints (Cracraft 1983, 1989; Frost & Kluge 1994; McKitrick & Zink 1988).

Some tree based methods attempt to aggregate terminals on the basis of character distribution (Brower 1999) or on tree topology (Wiens & Penkrot 2002) and these are an improvement over distance based tree methods; however, for the DNA barcoding with multiple individuals within a species we feel it inappropriate to use a tree based approach (Davis & Nixon 1992; Goldstein & DeSalle 2000).

(c) The barcoder database: is cox1 enough?

A controversial aspect of the DNA barcoding initiative has been which molecular tool to use to generate the DNA barcodes (Prendini 2005). The published efforts so far in animal systems have used the cytochrome $c$ oxidase subunit I gene ($cox1$) of the animal mitochondrion. One of the major criticisms of this approach is that a single molecular probe such as $cox1$ will not necessarily provide sufficient information to deliver the resolution needed to diagnose the large number of species targeted by the initiative. In arguing for the sufficiency of $cox1$ (or any other single molecular marker), Hebert et al. (2003b) pointed out that just 15 variable sites in the $cox1$ gene offers 1 billion different combinations of bases giving more than enough possible barcode ‘patterns’ at the DNA level. Missing from that assertion is a recognition that relatively few of those combinations could ever result in a viable translated protein observable in an extant species. Based on a study of birds (Hebert et al. 2004b) it was suggested that $cox1$ might have broader utility across the animal kingdom and that a universal distance cut-off of 10 times the distance within species could be used to distinguish between species. However, even among the birds surveyed for $cox1$, there were anomalous taxa that showed greater than expected within species divergences. In addition, studies of copepods...
Figure 1. Hypothetical example of character based diagnosis (Davis & Nixon 1992) in action. The twelve sequences represent two populations of six individuals each. The solid line through the middle of the matrix represents a geographical barrier between the two populations. A. DNA sequence attributes in these columns are purely diagnostic characters (sensu Davis & Nixon 1992). B. DNA sequence attributes in this column are not purely diagnostic, but rather the G in the three individuals in the top population are ‘private’ to that population. C. The DNA sequence attributes in the two columns by themselves constitute two private DNA positions. However, in combination these two columns provide a ‘pure’ diagnostic combination (AA versus AG or GA; ‘compound pure’ character in the terminology of Sarkar et al. 2002). D. The four columns marked by the shading for D are neither diagnostic nor private. Yet in combination the four columns provide a diagnostic system for the top population versus the bottom. The top population is diagnosed by GA, AG/GA, AG for the four columns.

(Edmands 2001; Goetze 2003) have found high levels of cox1 variation (up to 20%) even among conspecifics. Having to leave aside these outliers argues against the sufficiency or the universality of the gene region. While the fact that these taxa showed great divergence in genetic distance is suggestive that there may be unrecognized taxonomic diversity present, to test that hypothesis would require more than one line of evidence.

So we examine here the character based approach to diagnosis and the power of character-based approaches. Sarkar et al. (2002) recognized that combinations of attributes that are not ‘pure’ diagnostics could indeed be used to develop compound ‘pure’ diagnostics. The simplest compound diagnostic is when two attributes that are ‘private’ for aggregates (found only in one aggregate but not fixed; e.g. if aggregate 1 is fixed for all individuals at position 1 of a sequence with a G and aggregate 2 has 5 individuals with a G in position 1 and 5 individuals with an A in position 1, the A in aggregate 2 is a private diagnostic for aggregate 2) are combined to produce a ‘pure’ diagnostic. Even more complex combinations can be found if two or more aggregates are defined. For instance, figure 1 shows four positions in a hypothetical sequence that are polymorphic (i.e. neither fixed nor private for the alternative character states), which when combined together create a pure diagnostic. This approach has been used to evaluate character diagnostics in sturgeon species. The system examined was generated from comparisons of over 150 Acipenseridae individuals in two species, Acipenser gueldenstaedtii and A. baerii (Doukakis et al. 1999). While the molecular probe used was 700 base pairs (bp) of cytochrome B gene (cytB) of the mitochondrion, this example will suffice to demonstrate the power of finding diagnostics using this approach. Between these two species for the 700 bp region of cytB, we observed 36 variable sites, of which three were ‘pure’ diagnostics for the two species. Nearly half of the sites were ‘private’ to one species and over 1000 combinations of two sites produced compound ‘private’ sites (i.e. the two sites together were ‘private’ to one of the species. More interestingly, there were seven combinations of these 15 singly ‘private’ sites and the 1000 two position compounds that produced ‘pure’ diagnostics.

The answer to ‘Is cox1 enough?’ is then yes and no. Cox1 is certainly not enough to delineate phylogenetic relationships of organisms. However, it may be enough to generate suites of characters that can and will diagnose aggregates of organisms as entities in nature.

(d) The barcoder database: how many individuals are enough?

Another controversial aspect of the DNA barcoding initiative relates to the number of individuals of each putative species to include in the analysis. Classical taxonomic endeavours screen numerous individuals from multiple localities across the range of a given species to distinguish variation within a species from variation between species in order to identify those characters that are uniquely shared among all members of a species. One or only a few individuals may not be representative of the species as a whole, especially for taxa with widespread distributions (Davis & Nixon 1992; Goldstein et al. 2000; Walsh 2000). The necessity for adequate numbers of individuals applies to both distance and character based methods and it is not likely that there will be a universal sample size that will be appropriate for all species. Neither is a universal geographic distance likely to provide a reasonable proxy for determining the appropriate sampling strategy. As with gene region choice, sampling sufficient numbers of individuals to capture representative within-species variation will require pilot studies and the use of background information on life history, dispersal ability and mating patterns, among other information.

2. A CHARACTER BASED BARCODER PROPOSED

Recent work and commentary on the barcoding initiative in the literature has stimulated concern and excitement both from taxonomists and from those who are based in molecular approaches. Concerns range from the philosophical to the technical. These commentaries could loosely be separated into the taxonomic perspective (Agosti 2003; Dunn 2003; Lipscomb et al. 2003; Proudlowe & Wood 2003; Seberg et al. 2003), the molecular perspective (Baker et al. 2003; Blaxter & Floyd 2003; Ronquist & Gardenfors 2003; Tautz et al. 2003) and commentaries sympathetic to both (Mallet & Willmott 2003; Wilson 2003). The pro-taxonomy commentaries strongly deride the lack of consideration of the intellectual content of classical taxonomy and can be summarized as in the following quote from Lipscomb et al. (2003) p. 65: ‘advocates of
DNA taxonomy seem not to understand the peerless intellectual content of taxonomy based on all available information, or the hypothesis-driven basis of modern revisionary work.'

It is clear to us that genomic information should be an active component of modern taxonomy, but DNA should not be the sole source of information retrieval. ‘Fashionable DNA bar-coding methods are a breakthrough for identification, but they will not supplant the need to formulate and rigorously test species hypotheses.' (Wheeler et al. 2004, p. 285). A barcode should incorporate diagnostic characters both from the classical morphological approach and from the newer molecular approaches; one without the other misses the synergy that an integrated taxonomy is capable of attaining (Godfray 2002). We see a major strength to an integrated approach in that descriptive taxonomy and phylogenetic taxonomy together produce a synergy of resolution that neither can attain in the current fragmented 'tower of babel' (Mallet & Willmott 2003). It should also be clear that integration of the ‘fashionable’ molecular approaches with the classical taxonomic approach is a critical component of reconciling both camps and to move towards the use of barcodes in modern biology. Consequently we present an operational, integrative approach to taxonomy that attempts to reconcile molecular information with other sources of characters.

(a) The taxonomic circle; breaking out

We offer figure 2 as a heuristic for how modern taxonomy can be viewed. While any diagram describing the workings of taxonomy would suffer from oversimplification of the intellectual process that taxonomists use in plying their trade, we feel figure 2 captures many elements of modern taxonomy—hypothesis testing, corroboration, reciprocal illumination and revision. The main problem that needs to be addressed in any attempt to determine the boundary of a species and hence raise the entity to species status is to avoid circular or tautological reasoning. Breaking out of the circle of inference (figure 2, central diagram) in species delineation work is one descriptive way to describe the job of the taxonomist and hence the role of DNA sequence information (and barcoding) in taxonomy.

Figure 2 shows a highly simplified version of several taxonomic problems that have faced systematists and DNA barcoders. The classical process of using morphology in taxonomy is shown first (figure 2, panel A). In this diagram the data points on the ‘taxonomy’ circle consist of geographical, morphological, ecological, reproductive and behavioural information. In most morphological taxonomic studies an initial hypothesis based on geography is made. The taxonomist then crosses over the interior of the circle to either ecological characters or to morphological characters to test the geographical hypothesis. If morphological, behavioural, reproductive or ecological information relevant to the geographical hypothesis assist in rejecting the null hypothesis that there is no differentiation of the two geographical entities, then the taxonomist can ‘break out’ of the circle.

Cryptic species detected by DNA approaches is shown next (figure 2, panel B). In this case we add DNA sequence information to the circle. Initially a geographical hypothesis is formulated, a null hypothesis established and tested with the classical tools of the taxonomist. In this case, none of the classical tools—reproductive biology, morphological, behavioural or ecological characters—can reject the null hypothesis. The taxonomist can turn to DNA sequences where the null hypothesis based on geography is rejected because of fixed DNA differences among the aggregates hypothesized by geography. In essence, the aggregates contain morphologically cryptic species that are only detected at the DNA sequence level, which allows the taxonomist to break out of the circle.

The third panel in figure 2 (panel C) represents the power of integrating novel methods into this operational scheme. In this case, several individuals within a single geographic area show morphological differences. Because these individuals are considered to reside in the same geographic region, a geographical hypothesis cannot be made. But in this case the morphologically different entities can be aggregated and tested for fixed differences with other sources of data. In the case in the diagram, we imply that DNA sequence information can be used and if fixed DNA differences corroborate the morphological hypothesis then the conclusion of the analysis is that two species exist in sympathy and can be delineated by morphological differences.

The converse situation is also possible—a researcher could examine a ‘population’ of organisms with morphology and see no morphological differences. When the genomes of the organism are examined, the researcher might discover a DNA sequence polymorphism that clearly separates the single population into individuals with one haplotype and individuals with a distinct second haplotype. The only way to break out of the circle here would be to re-examine morphology or to move on to some other source of information. If no corroboration of the molecular aggregation can be found then the conclusion should be that a single population with two clearly distinct haplotypes exist. If corroboration is attained, then two distinct entities should be concluded to exist.

3. THE CHARACTER BASED (BREAKOUT) BARCODER IMPLEMENTED

We feel that a formalized method for inclusion of molecular information into taxonomy will clarify the intellectual content of taxonomy from a molecular perspective, but it will also clarify how DNA sequence information can most efficiently be used in the DNA barcoding initiative. The following section uses a mammalian case study (the genus Muntiacus or barking deer), a fish case study (the Acipenseridae or sturgeons—the source of caviar) and invertebrate examples (leeches)—that have manageable numbers.
of taxa and unique taxonomic problems to examine the incorporation of molecular data into taxonomical issues. Specifically, we first examine the use of type specimens in barcoding of muntjac and the impact such type specimens will have on future DNA taxonomic efforts. Second we examine the use of DNA barcoding in the leech genus *Hirudo* to demonstrate the importance of broad scale sampling of groups in taxonomic surveys. *Hirudo* also serves as a strong example of the effect of hybridization on how taxonomy is done. Finally, we use the commercially important fish family Acipenseridae to examine a cryptic species problem. Each of these examples will demonstrate that while DNA is an important factor in all three, the interaction of DNA sequence data with other kinds of characters produces a more precise taxonomical framework.

(a) **Muntjac barcoding: Muntiacus rooseveltorum example**

The muntjac study (Amato *et al.* 1999) used DNA barcoding for species discovery in a framework that is
compatible with classical taxonomy. The study highlights some of the issues related to barcoding such as the use of independently identified vouchers, sample size and gene region choice.

It began with field reports of what may have been representatives of a new species of muntjac in Laos for which the only material available at the time consisted of dried tissue samples. In order to devise a method to explore the question of the species status of these muntjacs using DNA, the study looked to established practices in taxonomy for guidelines. Sample size and gene region choice are factors that can affect the ability to discern true diagnostic data. In order for DNA data to have the potential to be used in a species discovery process, all species in the group must be included in the analysis. In addition, the numbers of individuals sampled for each species must be large enough to be representative of the variation in a given gene region for the species as a whole and the gene region must be variable enough to detect true differences between species. Given a large enough and representative sample size, the problem of a too variable gene region can be minimized.

The number of available samples for the putative new species was low (n = 10). Given the low number of individuals, a pilot study was undertaken that explored several gene regions for representatives of each species in the group and chose a gene region (in this case 16S mt rDNA) to balance the potential for two types of errors: (1) mistaking individual variation for species level variation by using too few individuals and a highly variable gene region; or (2) failing to identify true species differences, by using a conserved gene region sequenced for too few individuals to recover sufficient variation.

A diagnosis matrix of 114 individuals representing all species in the genus plus outgroup taxa was constructed. For the more widespread species, this included larger sample sizes from several localities across the range of those species as in the case of Muntiacus muntjak (n = 49). In addition, in order to be able to associate newly collected material with species names, for each species DNA sequence was obtained from museum collections and included in the diagnostic matrix. The sequences of the putative new species compared to other species in the group were unique. However, the inclusion of sequence data from museum specimens proved critical to an accurate assessment of the species status. In the literature there had been a description of a species (M. rooseveltorum Osgood 1932) based on a single individual collected in Laos ~70 years ago and never collected again. The DNA from the Type specimen of M. rooseveltorum was shown to share diagnostic sites with all the newly collected specimens of the putative new species (figure 3). This led to the conclusion that the newly collected specimens represented a rediscovery of M. rooseveltorum. The use of DNA alone, without the inclusion of all species in the group (and the type specimen in this case), would have led to the incorrect conclusion that the M. rooseveltorum specimens represented a new species. This highlights the importance of complete taxonomic sampling, literature review and corroboration with a second line of evidence. It also highlights the need to obtain voucher specimens that would provide not only DNA but also provide the means to examine a second line of evidence such as morphology.

(b) Leech barcoding

For a variety of reasons, leeches provide a unique framework for examination of the utility of DNA barcoding methods. The group (Hirudinida) is well circumscribed with approximately 750 known species. There is, nonetheless, a diversity of habitat preferences and life-history strategies represented across the clade including some extremes of parental care, various trophic modes ranging from blood feeding to predation as well as life in marine, freshwater or terrestrial environments. Furthermore, leeches already are well characterized for the cox1 locus (e.g. Siddall & Burreson 1998; Siddall et al. 2001; Borda & Siddall 2004) that is typically advocated for barcoding studies (Hebert et al. 2003a,b, 2004a,b). However, the use of cox1 alone for barcoding leech diversity may warrant some caution in that this locus has a highly biased base composition. Among the New World medicinal leeches, for example, adenosine and thymidine represent up to 72% of the nucleotide composition in cox1 (and up to 96% at third positions; Siddall & Burreson 1998). As a result, approximately 24% of the variable sites in cox1 are rendered binary among leeches as opposed to having all four nucleotides available.

(i) Glossiphonidae example

One of the best-represented families of leeches in freshwater environments is Glossiphoniidae. The clade comprises taxa that are dorsoventrally flattened freshwater species normally found feeding on anuran or chelonian hosts, though a few are fish parasites and one (Placobdelloides jaegerskioeldi) is even specific to the rectal tissues of hippos (Hippopotamus amphibius). However, many species in the family, in particular those in the unrelated genera Glossiphonia and Helobdella have abandoned sanquivity in favour of a predatory lifestyle on mollusks and oligochaetes, respectively. Species in the genus Helobdella have their greatest diversity in South America and less so in North America. Since being described by Linnaeus, Helobdella stagnalis was the only species known from Europe, and none were known to occur either in Africa or Australia. As such, recent descriptions of new species like Helobdella europea from an urban pond in Berlin (Kutschera 1985, 1987) and Helobdella papillornata from streams in Australia (Govedich & Davies 1998) were entirely unexpected, as was the surreptitious discovery of undescribed representatives of Helobdella in each of South Africa, Hawaii and New Zealand, all in a span of three years. Siddall and Budinoff (2005) employed DNA barcoding to assess this distribution of leeches, which at first presented a historical biogeographic conundrum. Their results (reproduced here in figure 4) clarified the fact that in each case the leeches were genetically indistinguishable both at the cox1 and ND1 mitochondrial loci and represented a single species of Helobdella nested in a South American
Figure 3. A DNA barcoding example for barking deer (genus *Muntiacus*). The table at the top of the figure shows variable nucleotide positions including several diagnostic sites in the 16s mt rDNA of multiple individuals of muntjac species. DNA from the type specimen of *Muntiacus rooseveltorum* was compared to recently collected putative *M. rooseveltorum* specimens to clarify their nomenclature (Amato et al. 1999). The word Type after the binomial indicates the sequence obtained from the type specimen of *M. rooseveltorum*. Shaded area indicates nucleotide position diagnostic for *M. rooseveltorum*. Dots (.) indicate sequence identity to the reference sequence on the first line. Colons (:) indicate missing data. Position 1 in the region sequenced corresponds to position 2305 in the *Bos taurus* mitochondrial DNA, complete genome (GenBank Accession Number: AB074962). Photograph of the skull of the Type specimen of *M. rooseveltorum* courtesy of the Field Museum of Natural History (Field Museum negative number Z82184: *Muntiacus rooseveltorum* Zoology specimen 31783), is used with permission. The graphic in the centre shows the multiple gene region barcode for *M. rooseveltorum* separated by right brackets; reading from left to right, it shows the diagnostic nucleotides and position numbers found in the mitochondrial gene regions: 16s, cytochrome b, 12s and D loop.

*Phil. Trans. R. Soc. B* (2005)
group known as the *triserialis* complex (Ringuelet 1943). Accidental introductions to each locality could easily have been coincident with introductions of common aquatic invasive plant species like *Pistia stratiotes* and *Salvinia molesta*, known to have happened in each of Germany, Australia, New Zealand and Hawaii. The genetic determination alone, however, was not accomplished in isolation from taxonomic considerations. Rather, through dissections and comparison to described taxa, Siddall & Budinoff (2005) asserted that this leech species corresponds exactly to Ringuelet’s (1943) *Helobdella triserialis* var *lineata*; an unfortunate result since Verrill’s (1872) North American *Helobdella lineata* preoccupied that appropriate specific epithet. The globally invasive leech species, thus, is now known as *Helobdella europaea* notwithstanding its suspected South American origin. Significant to the successful barcoding result in the foregoing was a broad geographic coverage of the species of *Helobdella* from the known range of the genus. Without that global coverage, and by, for example, only focusing collections on a restricted geographic area in which the suspect leech was found (e.g. Hebert et al. 2004a) probably would have abrogated discovery of its true identity and ultimate origin.

(ii) *Hirudo* example

The European medicinal leech remains a valuable tool available to the biomedical sciences notwithstanding its having historically been used for some rather dubious purposes like the treatment of obesity and Stalin’s fatal strokes. In fact, just this past year the US Food and
Drug Administration formally approved the European medicinal leech (*Hirudo medicinalis*) as a ‘medical device’ and several companies like LeechesUSA, BioPharm and Ricarimpex specialize in the global distribution of leeches for use in microsurgery and related procedures. One would think that a species of annelid that is so broadly used in medicine, neurobiology, developmental biology and genomics, and for which various genomic libraries are being developed would have been better characterized in terms of its species limits. The first phylogenetic analyses to incorporate the European medicinal leech as a taxon using *cox1* were Black *et al.* (1997) and Siddall & Burreson (1998), though neither of those analyses considered multiple representatives of the species. More recently, Trontelj & Utevsky (2005) demonstrated several unusual findings on the basis of *cox1*: specifically, that so-called European medicinal leeches group into four distinct lineages and that what Black *et al.* (1997) and Siddall & Burreson (1998) each sequenced bears little resemblance to the *cox1* gene found in wild-caught European medicinal leech populations. Revisiting their analysis here, we have reanalysed the available data with some wild-caught and commercially available material in a broader taxonomic scope for the Hirudinidae (figure 5). Notably, the results corroborate the findings of Trontelj & Utevsky (2005) in that the European medicinal leech species complex seems to include at least four species, three of which previously had been synonymized with *Hirudo medicinalis*. If DNA barcoding results are accepted as is, then we would need to resurrect each of *Hirudo verbana* Carena 1820, *Hirudo troctina* Johnson 1816, and establish a new species for the Persian medicinal leech denoted ‘*Hirudo sp.*’ in figure 5. Conveniently, each of these species may not be as ‘cryptic’ as previously (Sawyer 1986) thought insofar as they appear to be readily distinguishable on the basis of external colour patterns. More alarmingly, though, is the status of those leeches that previously have been called ‘*Hirudo medicinalis*’. For example, a leech obtained from Ward’s Biological for this study and shipped under the name of ‘*Hirudo medicinalis*’ unequivocally groups with *Hirudo verbana* (figure 5). Also, both Black *et al.* (1997) and Siddall & Burreson (1998) obtained their representatives of *Hirudo medicinalis* from Carolina Biological supply, and both of those sequences group with the Asian *Hirudinaria manilensis* notwithstanding the fact that the specimen used by Siddall & Burreson (1998) is morphologically indistinguishable from *H. verbana*. The latter suggests a remarkable ability for introgression that is as yet not well understood for these leeches, and yet which should cause some concern for the overall utility of DNA barcoding methods based on a single locus.

(c) **Sturgeon barcoding**

There are three species of fish in the family Acipenseridae—*Hucho huso*, *Acipenser gueldenstadii* and *A. stellatus*, all of them listed as endangered by CITES—that are the source for the grand majority of the world’s commercial caviar trade. One of these species, *A. gueldenstadii*, has been an enigma with respect to the surveillance of imported caviar since DNA sequence methods were introduced to monitor importation of caviar from the three highly endangered fish (DeSalle & Birstein 1996; Birstein *et al.* 2000). One of the major problems with the diagnosis of *A. gueldenstadii* caviar has been the occurrence at high frequency of caviar purportedly from *A. baerii* (a close relative to *A. gueldenstadii*), or the Siberian sturgeon (Birstein *et al.* 2000). More detailed examination of the problem using larger numbers of individuals from both the *A. gueldenstadii* and *A. baerii* clades of sturgeons now indicates the presence of a cryptic species identical to *A. gueldenstadii* in morphology, but also similar (but not identical to) *A. baerii*. In fact, several DNA sequence changes exist that diagnose this aggregate of fish that are morphologically identical to *A. gueldenstadii*, as distinct from *A. baerii*. In this case the DNA diagnostics indicate that this second confused form of *A. gueldenstadii* is a separate entity (Birstein *et al.* 2005). This case is an excellent example of cryptic species and how DNA sequence information can reveal the crypticism. More importantly, this case exemplifies two important technical aspects of DNA barcoding. First, the need for large sample sizes and continual revision using larger sample sizes is highlighted by this example. When small sample sizes are used, the second cryptic *A. gueldenstadii*-like species is improperly diagnosed as *A. baerii*. Second, the case emphasizes the importance of precision in species delimitation in the practical application of any DNA barcoding system. Since animal forensics is a major positive outcome of DNA barcoding, this example reinforces the notion that large sample sizes and comprehensive databases coupled with classical techniques (as in this case meristics) be incorporated to implement the barcoding approach.

4. CONCLUSIONS

We conclude for the following reasons that the non-tree based approaches are more appropriate for the construction of a barcode reader. First, tree based approaches will produce phylogenies based on a single poorly chosen (for phylogenetics) molecule. While the trees will often times make sense, the support for hypotheses from such trees is almost always low, limiting the robustness of any phylogenetic hypothesis from such trees. Related to this issue is the well known widely acknowledged data matrices to produce phylogenies (Kluge 1989; Gatesy *et al.* 2003; Rokas *et al.* 2003). To base any inferences of relationship of species on a phylogenetic tree generated from a single molecular marker would be in conflict with the current approaches to modern systematics. Second, current taxonomic approaches use diagnostics discovery independent of trees to establish taxonomic systems. Using DNA characters in a diagnostic context would be entirely compatible with the process of current taxonomic research. Third, our proposed framework, requiring corroborations from more than one line of evidence, is also consistent with current taxonomic practices, would serve as a bridge between morphological and molecular approaches and provides sufficient rigor for species identification and discovery. We readily admit that certain barcoding problems such as environmental microbial species
identification will be problematic due to the lack of geographical and morphological information for corroboration. However, we suggest that in these problematic cases additional gene regions and ecological information might also be used to support or refute hypotheses of species cohesion.

Finally, when thinking about the possible formats for an actual field usable DNA barcoder, having a diagnostic system would be most appropriate for a small device. The coding of the diagnostics can be included in the design of a microarray format or in a rapid single nucleotide polymorphism detection format. These highly technical molecular approaches utilize character based detection methods, and would bring the development of a small field usable DNA barcoder closer to reality.

REFERENCES


Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences

Melanie Markmann and Diethard Tautz

Phil. Trans. R. Soc. B 2005 360, 1917-1924

Supplementary data
"Data Supplement"
http://rstb.royalsocietypublishing.org/content/suppl/2009/02/12/360.1462.1917.DC1.html

References
This article cites 22 articles, 8 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1917.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1917.full.html#related-urls

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences

Melanie Markmann¹,† and Diethard Tautz²,*

¹Zoological Institute, University of Munich, Luisenstrasse 14, 80333 Munich, Germany
²Department of Genetics, University of Cologne, Zülpicherstrasse 47, 50674 Cologne, Germany

Organisms living in or on the sediment layer of water bodies constitute the benthos fauna, which is known to harbour a large number of species of diverse taxonomic groups. The benthos plays a significant role in the nutrient cycle and it is, therefore, of high ecological relevance. Here, we have explored a DNA-taxonomic approach to access the meiobenthic organismic diversity, by focusing on obtaining signature sequences from a part of the large ribosomal subunit rRNA (28S), the D3–D5 region. To obtain a broad representation of taxa, benthos samples were taken from 12 lakes in Germany, representing different ecological conditions. In a first approach, we have extracted whole DNA from these samples, amplified the respective fragment by PCR, cloned the fragments and sequenced individual clones. However, we found a relatively large number of recombinant clones that must be considered PCR artefacts. In a second approach we have, therefore, directly sequenced PCR fragments that were obtained from DNA extracts of randomly picked individual organisms. In total, we have obtained 264 new unique sequences, which can be readily placed into taxon groups, based on phylogenetic comparison with currently available database sequences. The group with the highest taxon abundance were nematodes and protozoa, followed by chironomids. However, we find also that we have by far not exhausted the diversity of organisms in the samples. Still, our data provide a framework within which a meiobenthos DNA signature sequence database can be constructed, that will allow to develop the necessary techniques for studying taxon diversity in the context of ecological analysis. Since many taxa in our analysis are initially only identified via their signature sequences, but not yet their morphology, we propose to call this approach ‘reverse taxonomy’.

Keywords: DNA taxonomy; large subunit D3–D5; fresh water lakes; benthos fauna

1. INTRODUCTION
The benthos harbours a community of organisms including micro-organisms, animals and plants. The term meiobenthos fauna relates usually to multicellular animals with a size between 50 and 500 μm (Giere 1993). This includes, for example, nematodes, rotifers, mites, tardigrades, annelides, crustaceans, as well as larval stages of organisms that become larger as adults, such as chironomids. The meiobenthos has so far mainly been studied in the context of the formation of sediments and ecotoxicology in marine environments and freshwater lakes (McIntyre 1969; Traunspurger & Drews 1996, Soltwedel 2000). However, it should also be a particularly interesting subject of food web studies, since it represents a significant part of the biomass in water. However, even for taxonomic experts, the fauna is too complex and varied to get a complete picture of all species on a routine basis. Thus, in spite of its undoubted importance, the ecological analysis of the meiobenthos fauna remains superficial. The principal goal of our study is, therefore, to develop an assay for automatic taxon determination in complex samples to aid ecological research. The use of DNA signature sequences to distinguish taxa (Floyd et al. 2002; Hebert et al. 2003; Tautz et al. 2003; Blaxter 2004) is a potential solution for achieving this goal. For prokaryotic species this is often the only means to identify them, because of the lack of sufficient morphological markers. However, even for organisms where morphological differentiation is possible in principle, it has advantages over traditional approaches, because it can be automated. For example, microarray techniques based on DNA signature sequences allow the parallel determination of thousands of different sequences in a single experiment, making them particularly suitable for the qualitative and quantitative analysis of complex samples.

As a target molecule for analysis we have chosen here the nuclear encoded large ribosomal RNA subunit. Ribosomal RNA genes are universally present and have a very conservative organization (figure 1). Small subunit (SSU—often called 18S) and large subunit (LSU—often called 28S) rRNAs are always transcribed together and then processed into individual molecules. The SSU/LSU unit is tandemly repeated and present in
sieved material (see above) was mixed in a 1:5 ratio with this.

2. MATERIAL AND METHODS

(a) Sampling and sample purification

Samples were taken from the sediment within three metres of the shoreline of the respective lakes by taking the upper 5 to 10 cm of the sediment layer. About 200 mL sediment slurry were filled in a two litre measuring cylinder and topped with lake water. The mixture was shaken and the rough parts of the sediment were left to settle for 30 s. The supernatant was decanted into a series of graded mesh sieves. All materials larger than 250 μm and smaller than 30 μm were discarded.

For a further removal of anorganic material, we used centrifugation in polysilicate buffer (Burgess 2001). The polysilicate (LudoxTM 50, Dupont de Nemours, Antwerpen) was diluted to 1.14 g cm\(^{-3}\) with water (approximately 30 vol.% Ludox) and adjusted to pH 7.0 with HCl. The sieved material (see above) was mixed in a 1:5 ratio with this solution and centrifuged for 5 min at 800g. This leads mainly to sedimentation of the remaining inorganic material, while the organisms remain in the supernatant. The supernatant was then again concentrated on a 20 μm sieve and washed with water.

A further purification step was used for the DNA extraction in the batch approach. This made use of a step gradient of Ludox with a density of 1.4 g cm\(^{-3}\) as the cushion and the organism fraction from the above step (in water) as upper layer. Centrifugation at 800g for 5 min lead to the concentration of the organisms at the interface between the Ludox cushion and the water. From there they were retrieved with a pipette and washed again over a 20 μm sieve with water.

(b) DNA extraction

For the DNA extraction in the batch approach, we used the organism fraction from the step gradient. This was transferred into HOM-buffer (20 mM Tris–HCl, 100 mM EDTA pH 7.5) and homogenized with a glass pestle homogenizer. Sodiumdodecylsulfate was then added to a final concentration of 1% and proteinase K to a final concentration of 500 μg mL\(^{-1}\). Protein digestion was carried out over night at 50 °C. The following steps are based on the protocol given by Porteus et al. (1997), which is designed for soil samples. Per millilitre lysate, 125 μL 5 M potassium acetate and 420 μL 40% polyethylene glycol 8000 (Sigma) were added. The tube was then incubated for 1 h at −20 °C to precipitate the DNA. After centrifugation for 15 min at 13 000g the supernatant was discarded and the pellet was dissolved in CTAB buffer (2% hexadecyltrimethylammonium bromide, 1.4 M NaCl, 100 mM EDTA pH 7.5) and extracted with one volume of chloroform. After centrifugation for 10 min at 13 000g the supernatant was transferred into a new tube and precipitated by adding 1.15 volume of isopropanol (15 min at −20 °C). Centrifugation was as above and the pellet was then dissolved in 2.5 M ammoniumacetate and precipitated again with 2.2 volumes of ethanol (15 min, −20 °C, centrifugation as above). The final precipitate was dissolved in TE (10 mM Tris–HCl, 0.1 mM EDTA pH 7.5) and further purified on a Microcon-100 centrifugation filter (Millipore) washed with TE to remove remaining impurities and degraded DNA fragments. Single organisms were transferred into a buffer containing 10 mM Tris–HCl pH 7.5 and 140 μg mL\(^{-1}\) proteinase K. Larger organisms were squashed with a pipette tip to allow the buffer to penetrate the tissue. Digestion was for 4 h (or over night) at 50 °C.
The proteinase K was then denatured for 10 min at 100 °C and the solution was further treated with Gene releaser (Bioventures Inc.) before amplification (Schizas et al. 1997).

(c) PCR amplification and sequencing
The primers used for amplification are depicted in figure 1. The primers were both used for PCR and for sequencing. The sequences of the primers are (always 5′–3′ direction): 1274: GACCCGTCTTTGAAACCGGA; 1480: TAGGGGCG AAAGACTCG; 1275: TCGGAAGGAAAACGCTACTA; 706: CGCCAGTTCTGTTTACC; 689: ACACACTCC TTGGCGGA. Two microlitre DNA template was used in different dilutions (up to 1:10 000 in water) in 20 μL reaction volumes. PCR cycles were 2 min denaturation at 96 °C, then 20 cycles with 45 s 96 °C, 60 s 48 °C, 60 s 72 °C followed by 20 cycles with 45 s 90 °C, 60 s 48 °C, 60 s 72 °C and final elongation at 72 °C for 10 min. The resulting fragments were either cloned into pZERO vector (Invitrogen), or directly sequenced by cycle sequencing, following the protocols of the supplier of the respective kits. For the clones we used the standard sequencing primers that flank the inserts for sequencing, for the PCR fragments we used primers 1480 and 706. Sequencing reactions were run on an ABI377 sequencer. All resulting sequences were manually inspected and only clear sequence reads were retained. Ambiguous base callings were manually inspected and edited, if necessary, including the information from the opposite strand when this was available (note that only partial information from the opposite strand was available for the PCR fragments from individual organisms, due to the internal localization of the sequencing primers; see figure 1). Our experience with this sequencing strategy suggests that the upper bound of the error is less than one wrongly assigned nucleotide in a given sequence (i.e. < 0.15%).

(d) Computational analysis
To place new sequences among known sequences, we have retrieved approximately 360 LSU sequences from the EBI database and built a local database. All of our sequences were then compared to the sequences in this database using the FASTA algorithm (Pearson & Lipman 1988). New sequences were initially assigned to taxon groups on the basis of the best similarity scores obtained from FASTA, whereby a minimum of 70% similarity was required to assign sequences to a major taxon.

The FASTA algorithm was also used to screen for potential recombinant clones. Clones were considered possible recombinants if the FASTA score of the first half and the second half of the sequence was very different, using a subjective cut-off, based on the further manual inspection of the FASTA alignment. This procedure is, therefore, only a first approximation to identify possible molecular chimaeras.

The further analysis was mainly done with the help of the Arb program (Ludwig et al. 2004), which includes alignment and tree building features. The alignment was optimized taken stem-loop structure criteria into account, as described in Friedrich & Tautz (1997). Additional analyses were done with Phylogenetic analysis using parsimony (PAUP; Swofford 1993).

The Arb neighbour-joining (NJ) tree building feature was used to obtain a phylogenetic tree of all sequences. This tree served then to reassess all initial taxon assignments obtained from the FASTA analysis, which led to some minor corrections. The tree allowed a tentative placement of all sequences that were not already assigned by the FASTA analysis.

3. RESULTS
Meiobenthos samples were obtained from 11 different lakes around Munich (figure 2). One further sample was taken from a site near Braunschweig (ca. 600 km north of Munich). Since our main objective was to sample a large diversity of taxa, we have chosen lakes from rather different ecological settings (table 1). The organisms were retrieved from the sediments (see §2) and a size fraction of 30–250 μm was selected via appropriate sieves.

(a) Sequence signatures
In a first approach to obtain taxon specific signature sequences, we have simply pooled all organisms extracted from the meio benthos fraction of a given lake sample and prepared DNA from the pool. From this DNA we made two types of amplification, one encompassing the D3 region alone and one encompassing the whole D3–D5 region using the universal primers depicted in figure 1. The resulting fragments were cloned and approximately 900 randomly picked clones were sequenced. All sequences were checked against a database of available LSU sequences (see §2) to assess whether they can be associated with a known sequence or at least placed close to a known taxon. This analysis showed for some fragments similarity with more than one taxon group. This is apparently due to the presence of hybrid sequences that were most likely caused by ‘jumping PCR’ (Meyerhans et al. 1990; Pääbo et al. 1990). We, therefore, tested for all sequences whether they yielded different results when the first versus the second half of the sequence was compared with the database sequences. This showed that approximately one third of the sequences had to be considered as possible hybrid sequences from at least two different organisms. These sequences were removed from the
were submitted to GenBank (DQ086498–DQ086776).

Thirty-five of these were identical to sequences obtained from the batch approach. The new unique sequences involved. With this approach, we successfully obtained mutations is highly reduced, because no cloning step is obtained wrong sequences caused by PCR induced sequences can be excluded; and third, the chance of stereo microscope; second, the generation of hybrid possibilities based on the visual identification under the sequenced without cloning. This approach has three advantages. First, a rough taxon assignment is already possible based on the visual identification under the stereo microscope, and DNA was extracted from them individually.

Among the 600 clones, we identified 159 unique sequences (124 for the whole D3–D5 fragment and 35 for the D3 fragment).

As was to be expected, the different parts of the D3–D5 region show different degrees of sequence divergence. Using the distance measure implemented in ARB, we find that the average similarity between all further analysis. This left about 600 useable clones, which were not obviously the product of artificial recombination of two very distinct sequences, although we cannot rule out that some of them might still be due to recombination between two similar sequences. Among the 600 clones, we identified 159 unique sequences (124 for the whole D3–D5 fragment and 35 for the D3 fragment).

Although the effect of jumping PCR and recombination is well known in principle, it was nonetheless surprising that such a high fraction of artificial clones was generated from the batch PCR approach. Since the fraction of hybrid sequences obtained was somewhat different between the lake samples, it seems possible that different mixtures of sequences, or different DNA preparations (e.g. degradation status) are more or less prone to jumping PCR artefacts. Still, from these initial results we have to conclude that the batch approach is not the best method to reliably obtain signature sequences that represent single taxa. In a second approach we used, therefore, individual animals that were randomly picked under a stereo microscope and DNA was extracted from them individually. The amplified fragments were then directly sequenced without cloning. This approach has three advantages. First, a rough taxon assignment is already possible based on the visual identification under the stereo microscope; second, the generation of hybrid sequences can be excluded; and third, the chance of obtaining wrong sequences caused by PCR induced mutations is highly reduced, because no cloning step is involved. With this approach, we successfully obtained approximately 400 sequences, of which 140 were unique. Thirty-five of these were identical to sequences obtained from the batch approach. The new unique sequences were submitted to GenBank (DQ086498–DQ086776).

Table 1. List of lakes sampled with rough description of ecological context. (The column ‘sediment structure’ refers to the coarseness of the sediment ranging from one star (*, very fine sediment) to five stars (*****), coarse gravel). This classification refers of course only to the respective sampling sites.)

<table>
<thead>
<tr>
<th>acronym</th>
<th>name</th>
<th>description</th>
<th>sediment structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am</td>
<td>Ammersee</td>
<td>large mesotrophic lake, with several adjacent towns,</td>
<td>****</td>
</tr>
<tr>
<td></td>
<td></td>
<td>broadly used for fishing, water sports and shipping</td>
<td></td>
</tr>
<tr>
<td>Di</td>
<td>Dietlhofer Weiher</td>
<td>small lake next to a town, bathing lake in summer,</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>occasionally exceeds bacteriological pollution limits</td>
<td>**</td>
</tr>
<tr>
<td>Lu</td>
<td>Lussee</td>
<td>small mesotrophic lake under environmental protection,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>with broad reed girdle</td>
<td>**</td>
</tr>
<tr>
<td>Ma</td>
<td>Maisinger See</td>
<td>small mesotrophic lake, bathing lake in summer,</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>low pollution level</td>
<td></td>
</tr>
<tr>
<td>Nu</td>
<td>Nussberger Weiher</td>
<td>small lake used mainly for pisciculture</td>
<td>****</td>
</tr>
<tr>
<td>Os</td>
<td>Ostersee</td>
<td>midsized mesotrophic lake, next to Lussee, but without</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>special protection status, bathing lake in summer</td>
<td></td>
</tr>
<tr>
<td>Pi</td>
<td>Pilsense</td>
<td>midsized eutrophic lake, broadly used for water sports,</td>
<td>******</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bathing and fishing</td>
<td></td>
</tr>
<tr>
<td>St</td>
<td>Starnbergersee</td>
<td>very large mesotrophic lake, with several adjacent towns,</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>broadly used for fishing, water sports and shipping</td>
<td></td>
</tr>
<tr>
<td>Tk</td>
<td>Tonkuhle</td>
<td>artificial lake in a former clay pit near Braunschweig</td>
<td>*</td>
</tr>
<tr>
<td>To</td>
<td>water reservoir near</td>
<td>mesotrophic storage lake of the Isar, fluctuating water</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Bad Tölz</td>
<td>levels, can fall dry during summer</td>
<td></td>
</tr>
<tr>
<td>We</td>
<td>Weißlinger See</td>
<td>small polytrophic lake within a town (Weißling), artificially</td>
<td>******</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aerated</td>
<td></td>
</tr>
<tr>
<td>Zo</td>
<td>—</td>
<td>very small artificial pool in the garden of the Zoological</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Institute in Munich</td>
<td></td>
</tr>
</tbody>
</table>

**b** Sequence variation

About a third of all unique sequences differed at less than 10 nucleotide positions from the next closest sequence. This raises the question whether they might represent variations within species, rather than different species. Unfortunately, unequivocal species identification is difficult for the taxa that we look at and it is, therefore, not easy to sequence several representative of the same species to assess within species variance. However, for 12 taxa (five insects, three annelids, two crustaceans and two molluscs) we have been able to sequence between two and four individuals from the same species (assigned by morphological criteria). In all cases we found identical sequences for the respective species, including where the second sample was obtained from a lake from northern Germany. This suggests that the intra-species variance cannot be very high on average, although this issue will need to be further studied in the future.

As was to be expected, the different parts of the D3–D5 region show different degrees of sequence divergence. Using the distance measure implemented in ARB, we find that the average similarity between all D3 and D5 sequences in the database is 74%, i.e. the region is relatively well conserved. The pattern of conservation and divergence can be plotted onto the secondary structure predictions of the region (figure 3).

The most divergent parts are helices 30 and 31a, with adjacent loops. They show also major length differences between the taxa. Helix 39a is very variable with respect to nucleotide exchanges, but less variable in length. It is, thus, evident that the most variable regions are not necessarily confined to loops, but can also form stem regions. This implies that compensatory changes should often be found in these regions, which is indeed the case. Thus, although these regions show a high
divergence, they are clearly not free of selective constraints. This raises the question of whether they are actually sufficiently fast evolving to distinguish closely related species. Again, we have only preliminary data on this question so far. For all identified species from the same genus (including database sequences), we found at least six and usually more than 10 nucleotide differences. However, only 11 such comparisons were possible in our dataset, indicating that this is also an issue for further research.

**Taxon assignments**

To place our new sequences from meiobenthos organisms within a phylogenetic framework, we have combined them with about 400 sequences from the database. The latter included also bacterial sequences and vertebrate sequences. The sequences were aligned within the ARB program (Ludwig et al. 2004), taking secondary structure constraints into account. The NJ option in ARB was used to build a tree of all sequences. The subtrees of the monophyletic groups that are relevant for the meiobenthos fauna are shown in the Electronic Appendix. Although our mode of tree reconstruction must be considered as only a first approximation, it is nonetheless clear that almost all anonymous sequences were assigned to a known taxon group. Thus, although D3–D5 rDNA sequences are still somewhat underrepresented in the database, it is already possible to place almost any unknown sequence into a phylogenetic framework that allows taxon assignment on a rough scale.

Figure 3. Divergence profile of the sequences in our analysis correlated with secondary structure. The divergence measures are based on similarity values provided by the ARB program, the naming of the helices and loops follows Hassouna et al. (1984). The depiction does not include indel differences, which can be substantial.

**Figure 4a** provides an overview of the number of different sequence signatures in the taxon groups that

---

*Phil. Trans. R. Soc. B* (2005)
are represented in our samples. Nematodes are most abundant, followed by Protozoa, Chironomids and Cyclopoda. Protozoa would have been expected to be absent from the meiobenthos fraction, because most are too small. However, they were abundantly represented among the batch sequences, suggesting that they are in some way co-extracted with larger organisms.

Approximately 18% of signature sequences were found in more than one lake, although this differed for the different taxon classes (figure 4). Similarly, the number of signature sequences in each taxon class differed between the lakes, giving each of them a more or less unique representation of taxa (figure 5). However, we cannot expect to have fully saturated the possible types of sequences from any of these lakes. Because the sequences are from single collections, and neither the clones nor the single animals have been sampled exhaustively, the picture should be considered only a snapshot at a given time.

To assess how far away we are from saturation, we have plotted the new unique sequences that were found per lake versus the total number of unique sequences. Such a plot should approach a plateau, once saturation is reached. However, this is clearly not the case in our study (figure 6). We find that there are on average about 75% new unique sequences with each lake sample added.

4. DISCUSSION

Although the meiobenthos fauna plays undoubtedly a significant role in the ecology of water bodies, it remains poorly studied. The main reason for this is that most species of the fauna can only be identified by expert taxonomists, who are specialists for the respective groups. Routine surveys of the whole fauna are, therefore, very difficult, if not impossible. Our approach of using DNA signature sequences may be a solution for this dilemma. The initial results from our study are very promising in this respect. In the following we want to discuss the various aspects of our results that need to be considered, if a broad application to ecological studies is envisaged.

(a) Choice of marker sequence

Our study has focused on a fragment of the LSU ribosomal RNA as a basis of obtaining taxon specific sequence signatures. It was previously known that this fragment can be aligned between very diverse taxon groups and can be used for phylogeny reconstruction (Friedrich & Tautz 1995, 1997). It was less clear...
whether this would also be useful for distinguishing closely related species. With a 74% overall sequence conservation across the phyla, the chances for this might have seemed low. However, there are a few highly variable parts in the region that apparently provide enough information for distinguishing closely related taxa with good discriminatory power. For those cases, where we have species pairs from the same genus, we always find a sufficient number of nucleotide differences. Although the full discriminatory power of the D3–D5 LSU region will only become clear when a sufficiently large number of sequences exists from closely related species, it seems that we deal with a highly suitable signature sequence region, at least for the taxon groups analysed here.

The fact that the region is not free of constraints may also be the reason why we have found no sequence polymorphisms within species, at least in the cases where we could test this. This may even be advantageous, since neutral sequence polymorphisms can be a potential problem because they require to sequence a large number of samples from each species to assess the divergence within the group. Conversely, it is not to be expected that the most closely related species can be easily discriminated on the basis of D3–D5 sequences alone.

Ribosomal RNA genes offer an additional advantage for DNA-taxonomy schemes, because they are pre-amplified in the nucleus and because their products are very abundant in any living cell. This will allow devising detection schemes that do not need to rely on PCR amplification, if one focuses directly on the transcribed RNA.

We conclude that the D3–D5 LSU region may be a very good compromise between conservation and divergence across a large range of taxa. In particular, we should like to emphasize that the primers that we have used appear to be universally applicable for all eukaryotic taxa. However, there are still other regions of the LSU that might be even more suitable, in particular for discriminating very closely related taxa. In a parallel study we are currently exploring the D1–D2 region, which appears to show even greater discriminatory power (Nolte, Sonnenberg and Tautz, unpublished).

(b) Taxon assignment

The fact that we have obtained identical signatures from different lakes suggests that the taxon diversity is not infinite. However, our sampling was certainly not yet exhaustive either. Our sampling strategy was designed to obtain an overview on the total diversity of organisms, i.e. we have intentionally sampled lakes from very different ecological contexts.

In our batch cloning approach we have detected a significant number of sequences that were artificially generated through PCR-induced recombination. Such an approach can, therefore, lead to an overestimate of taxon diversity, even if one corrects for obvious recombination artefacts. Another problem with the batch approach is that there can be biases with respect to the ‘amplifiability’ of certain fragments, which results in wrong conclusions with respect to taxon representation. We have observed this for one case, a sequence from a harpacticoid species, which turned out in high numbers among the clones, but never among the individual sequences. Thus, the approach to pick individuals and to directly sequence the PCR fragments from them is clearly the better strategy to obtain reliable sequences and appropriate representations. Unfortunately, this strategy is also much more laborious and less easy to automate. Thus, batch approaches may still be warranted, as long as the shortcomings are fully considered.

Most taxa appear to harbour only one ribosomal sequence cluster. The sequence variants in this cluster are subject to concerted evolution (Elder & Turner 1995), i.e. intra-cluster divergence is rather low. On the other hand, some organisms appear to harbour two clusters with rather different sequences. This has for example been described for Plathelmintes (Carranza et al. 1996) and we have also found this when we sequenced several individuals of Dugesia polychroa from the Ammersee. Two sequence types were found which differed at 27 positions, which is as much as one can find for differences at the family level. It remains unclear how such very separate clusters evolve and what their function might be. There might be stage specific differences in expression but this has still to be further explored. Although such separate clusters can complicate the analysis it should be noted that both sequence variants are nonetheless specific for the respective species.

(c) Reverse taxonomy

Although the majority of the sequence classes determined in a pure sequencing approach originate from anonymous taxa, this information will nonetheless be extremely useful (see also Blaxter 2004, Blaxter et al. 2005). The reason is that even the anonymous sequences can be assigned to taxon groups that represent different trophic levels and will thus allow studying foodweb structure. For those sequence types that appear to play a particularly important role, it will then be warranted to actually identify and properly describe the species that harbours it. Such an approach may be called ‘reverse taxonomy’ akin to ‘reverse genetics’ where one identifies the sequence of a new gene first and its function later. Given that sequence determination and re-identification has become so
highly efficient nowadays, we expect that ‘reverse taxonomy’ will become a fruitful approach towards all those cases where taxon diversity cannot be handled with traditional approaches.

**d) Outlook**

Once a comprehensive database for meiothons organisms exists, one would have to develop further techniques to make it useful for ecological studies. The challenge is to devise a tool that allows to re-identify the sequence classes in temporal and spatial samples and to correlate this with ecological parameters. Technically this would best be achieved via a microarray approach. Here, one selects short oligo-nucleotides that represent the different sequence classes (Pozhitkov & Tautz 2002) and fixes them onto the surface of a microarray. These can then be directly hybridized with the RNA extracted from a sediment sample, providing a qualitative and quantitative measure of the sequence classes present. It should be emphasized that the sensitivity of microarray techniques is sufficient to identify e.g. single nematodes in a given sample (Markmann 2000). Furthermore, it will be technically possible to develop hybridization schemes that can be performed with low cost chemicals and equipment. Thus, a broad scale application of this technique for ecological research is within short reach.

We thank Walter Traunspurger for discussion and his support in sampling and extraction of the meiothons fauna. Gerhard Mayr and Sabine Giessler have provided rotifers and daphnid cultures, Florian Bernhard has helped with sampling, sample processing and sequencing, Wolfgang Ludwig and Konny Rassmann with setting up the ARA program and Arne Nolte has provided comments on the manuscript. This work was done at the Zoological Institute in Munich and partially funded through the DFG.

**REFERENCES**


The supplementary Electronic Appendix is available at http://dx.doi.org/10.1098/rstb.2005.1723 or via http://www.journals.royalsoc.ac.uk.

Phil. Trans. R. Soc. B (2005)
DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers

Michael T Monaghan, Michael Balke, T. Ryan Gregory and Alfried P Vogler

*Phil. Trans. R. Soc. B* 2005 *360*, 1925-1933

**References**

This article cites 33 articles, 6 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1925.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1925.full.html#related-urls

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](http://rstb.royalsocietypublishing.org/subscriptions)

To subscribe to *Phil. Trans. R. Soc. B* go to: [http://rstb.royalsocietypublishing.org/subscriptions](http://rstb.royalsocietypublishing.org/subscriptions)
DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers

Michael T. Monaghan1,2, Michael Balke1,†, T. Ryan Gregory3 and Alfried P. Vogler1,2,*

1Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK
2Division of Biology, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK
3Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada N1G 2W1

DNA barcoding has been successfully implemented in the identification of previously described species, and in the process has revealed several cryptic species. It has been noted that such methods could also greatly assist in the discovery and delineation of undescribed species in poorly studied groups, although to date the feasibility of such an approach has not been examined explicitly. Here, we investigate the possibility of using short mitochondrial and nuclear DNA sequences to delimit putative species in groups lacking an existing taxonomic framework. We focussed on poorly known tropical water beetles (Coleoptera: Dytiscidae, Hydrophilidae) from Madagascar and dung beetles (Scarabaeidae) in the genus Canthon from the Neotropics. Mitochondrial DNA sequence variation proved to be highly structured, with >95% of the observed variation existing between discrete sets of very closely related genotypes. Sequence variation in nuclear 28S rRNA among the same individuals was lower by at least an order of magnitude, but 16 different genotypes were found in water beetles and 12 genotypes in Canthon, differing from each other by a minimum of two base pairs. The distribution of these 28S rRNA genotypes in individuals exactly matched the distribution of mtDNA clusters, suggesting that mtDNA patterns were not misleading because of introgression. Moreover, in a few cases where sequence information was available in GenBank for morphologically defined species of Canthon, these matched some of the DNA-based clusters. These findings demonstrate that clusters of close relatives can be identified readily in the sequence variation obtained in field collected samples, and that these clusters are likely to correspond to either previously described or unknown species. The results suggest that DNA-assisted taxonomy will not require more than a short fragment of mtDNA to provide a largely accurate picture of species boundaries in these groups. Applied on a large scale, this DNA-based approach could greatly improve the rate of species discovery in the large assemblages of insects that remain undescribed.

Keywords: taxonomy; large subunit ribosomal RNA; Madagascar; DNA barcoding; cox1; COI

1. INTRODUCTION

A reliable and accessible classification of species is fundamental to research in ecology, evolutionary biology, biodiversity and conservation biology. While ca 1.5 million species have been described to date, this represents only a fraction of the actual diversity on Earth (Tudge 2000; Wilson 2003). Owing to the constant threat of biodiversity loss, there is an increasingly urgent need to accelerate the pace of species discovery and taxonomic databasing (Godfray 2002). Even the routine identification of known species can be difficult, often requiring highly specialized knowledge and representing a limiting factor in ecological studies and biodiversity inventories. In response, recent proposals have called for a more prominent role of efficient DNA-based methods in the delineation and identification of species (Blaxter 2004; Floyd et al. 2002; Hebert et al. 2003a; Tautz et al. 2003). Reactions to such proposals have ranged widely, from strongly supportive (Janzen 2004; Proudlove & Wood 2003; Stoeckle 2003) to vigorously opposed (Lipscomb et al. 2003; Seberg et al. 2003; Wheeler 2004; Will & Rubinoff 2004). The use of DNA-based methods for the delineation and discovery of new species, and thus their broader role in taxonomy, represents an especially contentious issue in this regard. Unfortunately, much of this debate has remained rhetorical, with limited empirical assessment of the benefits and limitations of a DNA-assisted programme of species discovery.

The objective of any method of species delineation, including DNA-based approaches, is to identify reproductively isolated groups of organisms that warrant classification as distinct species. It is widely acknowledged, and reflected in the Linnaean taxonomic system, that living organisms fall into largely discrete groupings recognizable by differences in morphology or other traits. It is then the role of taxonomy to define and name these groupings.
Instead of the identification of pre-defined species (e.g. Hebert et al., 1992), conclusions about population separation (Davis & Nixon 1992) depend on the degree of sampling, as local variation may affect the accuracy of species delineation. Further, the accuracy of species delineation depends on the degree of sampling, as local variation may affect the accuracy of species delineation. However, where more than a single individual per species has been sequenced, a minimum threshold of approximately one-tenth of the average $p$-distance found between well-established species in a lineage has been interpreted as intra-specific variation, while greater divergences are thought to indicate misidentifications of specimens or overlooked cryptic species (Hebert et al., 2004a). These cut-off values roughly correspond to maximum intra-specific divergences in mtDNA of 1–2%, and at the upper bound of this range may include several geographically defined ‘phylogroups’ (Avise & Walker 1999).

The discovery of new, cryptic species from existing, morphologically indiscriminate groups using DNA is neither controversial nor novel (Knowlton 1993), and potential taxonomic revisions inspired by DNA barcoding results have been typically left to experts to resolve on the basis of morphology, behaviour and other features (Hebert et al., 2004b). It is to be expected that a successful global DNA barcoding program would provide a comprehensive barcoding inventory for a majority of described taxa in the foreseeable future, facilitating the systematic discovery of cryptic species. However, with 85% or more of species still unknown to science, a much greater challenge lies in the potential application of DNA-based methods to the discovery and delineation of new species in poorly characterized taxa.

To explore the utility of DNA-based approaches to species recognition in poorly known groups, we evaluated patterns of variation in both mitochondrial and nuclear genes in a broad sample of water beetles collected from Madagascar. These samples exhibited an unknown level of species diversity within the families Dytiscidae and Hydrophilidae. Using the same methods, we also examined specimens from a single genus of Neotropical dung beetles (Canthon). Specimens were collected from various localities in the Neotropics, comprising an unknown number of species in a group that is acknowledged to present difficulties for morphological discrimination. The analysis shows that sequences cluster into cohesive, well-differentiated groups, and identical groups are recovered by both nuclear and mitochondrial markers. Based on multiple lines of evidence, these DNA-based clusters are taken to represent putative species boundaries and could assist with the assembly of a framework for the taxonomy of poorly studied lineages.

**2. METHODS**

(a) **Field sampling, selection of specimens and DNA sequencing**

Water beetles (Dytiscidae and Hydrophilidae) were collected at five sites in the North and central parts of Madagascar as part of a survey of insect biodiversity in 2004 (Monaghan et al., unpublished). Specimens were collected by sieving through small stream pools, ponds, and packs of leaf litter, and were sorted under a dissecting microscope (10×) into externally distinct morphological groups. Between two and five individuals from each group and each locality were selected for DNA analysis, as a representative sample of the variation of this group. This initial morphological treatment was superficial and was intended to maximize the disparity included in the subset of samples used for sequencing. Specimens of Canthon were obtained using baited pitfall traps from locations in Madagascar, French Guiana, Ecuador and Costa Rica between 1997 and 2001 (Inward 2003). Additional samples were collected from Belize in 2004 (L. Powell, MSc, Imperial College London, 2004). They were assigned to the genus Canthon based on a phylogenetic analysis combining them with unpublished sequences for most major groups of Scarabaeinae that also included five species of Canthon: C. doosburgi, C. indigaceus, C. luteoilis, C. smaragdulus and C. viridis, plus the closely related Scybalocanthon pygidialis (GenBank accessions: AY131633-7 for 28S, AY131814-7 for cox1, plus AY131673 and AY131849 for Scybalocanthon).

Genomic DNA extraction was performed using Wizard SV 96-well plates (Promega, UK). For both groups, a ca 700 bp fragment of 28S rRNA was amplified using primers FF and DD (Inward 2003). Fragments of cox1 were amplified with primers Pat and Jerry (Simons et al. 1994) for Canthon (800 bp) or with LCO1490 and HCO2198 (Folmer et al. 1994) for water beetles (600 bp). Sequencing was performed in both directions using a BigDye v. 2.1 terminator reaction with the same primers used for PCR. Sequences were analysed on an ABI3730 automated sequencer and forward and reverse strands were assembled in SEQUENCHER software. The 28S fragment was length-variable and was aligned separately for the two datasets using BLAST Alignment (Beshaw & Katzourakis 2005). Cox1 was not length-variable for either group.

(b) **Tree construction**

 Parsimony trees were obtained with PAUP v. 4b10 (Swofford 2002), with gap characters treated as a ‘fifth character state’, and branch length optimized under accelerated transformation. Heuristic searches were performed using TBR branch swapping and 100 replicates. We performed 1000 random addition replicates saving only a single tree in each case. Because the dataset contained many identical or very similar haplotypes, a large number of trees were found, one of which was selected arbitrarily for further analysis. To calculate Bremer Support (Bremer 1994), constraint files for parsimony searches enforcing the absence of the focal nodes were produced with TREEROT v. 2.0a. Bremer Support values of 0 indicate unresolved nodes which would be collapsed in a strict consensus of all shortest trees. Trees were rooted with sequences from related Carabidae taken from GenBank in the case of water beetles, and using a sequence from the Canthon dataset generated here for rooting the 28S tree. The single species of Scybalocanthon was used as the outgroup to root the tree of Canthon.

(c) **Variation in cox1**

We examined cox1 variation within and among clusters of sequences (see §3) using analysis of molecular variance.

Phil. Trans. R. Soc. B (2005)
3. RESULTS

A total of 75 and 71 individuals were included in the analysis of water beetles and _Canthon_, respectively. Sequencing of 28S rRNA was successful for 63 and 62 specimens, and aligned matrices contained 699 and 746 characters in the respective groups. In the water beetle dataset, we detected 16 different 28S genotypes. Four genotypes were present in only a single individual, and the remaining occurred in groups ranging in size from two to nine individuals (figure 1a). Sequences differed from one another by a minimum of two nucleotides, e.g. an AC insertion separated WB-4 and WB-5 (figure 1a). DNA sequencing for _cox1_ revealed 42 haplotypes. Parsimony tree searches uncovered 14 clusters of similar _cox1_ sequences, plus five isolated sequences without close relatives (‘singletons’) (figure 1b). Results for _Canthon_ were very similar. There were 12 different 28S genotypes and all but two were represented by > 1 individual (figure 2a). The 46 _cox1_ haplotypes grouped into 12 clusters, with two singletons (figure 2b).

The clustering of _cox1_ sequences in the parsimony trees showed complete congruence with the 28S genotypes for both the water beetle and _Canthon_ datasets. Closely related _cox1_ haplotypes all exhibited the same 28S genotype and none of the groups defined by 28S genotypes were polyphyletic in the _cox1_ tree (figures 1 and 2). It was not possible to judge incongruence in the 12 water beetles for which the 28S sequencing had failed (figure 1b). Equally, the _Canthon_ dataset included missing sequences in both the 28S and _cox1_ datasets, although there was perfect congruence for the 50 individuals sequenced for both genes (called ‘core terminals’, below).

Sequences making up the _cox1_ clusters were very similar to each other, but very different from other clusters. Based on pairwise differences in AMOVA, within-group variation accounted for only 4.1% of the total variation in the dataset in _Canthon_, and only 2.5% in water beetles (table 1). Absolute divergence
within clusters, and basal nodes had low support.

For water beetles, 80% for defining the clusters (88% of total Bremer Support was derived from nodes immediately below (i.e. between the clusters), the majority of total tree support a cluster, and nodes defining basal relationships nodes within a cluster, nodes immediately subtending separately for three categories of node levels (tip supported. When Bremer Support was considered more diverse sample of water beetles.

between clusters was 10 and 19% for water beetles, respectively (table 2). These patterns of related to one another than were the clusters in the other barcoding studies (Hebert et al. 2003b). The findings also appear to indicate that clusters in Canthon were more closely related to one another than were the clusters in the more diverse sample of water beetles.

The monophyly of the cox1 clusters was highly supported. When Bremer Support was considered separately for three categories of node levels (tip nodes within a cluster, nodes immediately subtending a cluster, and nodes defining basal relationships between the clusters), the majority of total tree support was derived from nodes immediately below (i.e. defining) the clusters (88% of total Bremer Support for water beetles, 80% for Canthon; table 3). Tip nodes within clusters, and basal nodes had low support.

Figure 2. Parsimony trees for Canthon, as in figure 1.

Table 1. cox1 variation among and within clusters (figures 1 and 2) measured with AMOVA (Excoffier et al. 1994). (*p<0.001.)

<table>
<thead>
<tr>
<th>source of variation</th>
<th>among variation</th>
<th>within variation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>df</td>
<td>Var</td>
</tr>
<tr>
<td>Canthon water beetles</td>
<td>9</td>
<td>1632.451</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>3130.381</td>
</tr>
</tbody>
</table>

(uncorrected p-distance) of sequences within clusters ranged from 0 to 2%, whereas the mean divergence between clusters was 10 and 19% for Canthon and water beetles, respectively (table 2). These patterns of divergence are similar to those reported for intraspecies comparisons in other barcoding studies (Hebert et al. 2003b). The findings also appear to indicate that clusters in Canthon were more closely related to one another than were the clusters in the more diverse sample of water beetles.

The monophyly of the cox1 clusters was highly supported. When Bremer Support was considered separately for three categories of node levels (tip nodes within a cluster, nodes immediately subtending a cluster, and nodes defining basal relationships between the clusters), the majority of total tree support was derived from nodes immediately below (i.e. defining) the clusters (88% of total Bremer Support for water beetles, 80% for Canthon; table 3). Tip nodes within clusters, and basal nodes had low support.

(uncorrected p-distance) of sequences within clusters ranged from 0 to 2%, whereas the mean divergence between clusters was 10 and 19% for Canthon and water beetles, respectively (table 2). These patterns of divergence are similar to those reported for intraspecies comparisons in other barcoding studies (Hebert et al. 2003b). The findings also appear to indicate that clusters in Canthon were more closely related to one another than were the clusters in the more diverse sample of water beetles.

The monophyly of the cox1 clusters was highly supported. When Bremer Support was considered separately for three categories of node levels (tip nodes within a cluster, nodes immediately subtending a cluster, and nodes defining basal relationships between the clusters), the majority of total tree support was derived from nodes immediately below (i.e. defining) the clusters (88% of total Bremer Support for water beetles, 80% for Canthon; table 3). Tip nodes within clusters, and basal nodes had low support.

Table 2. Mean uncorrected p-distances for cox1 within and among clusters.

<table>
<thead>
<tr>
<th></th>
<th>among clusters</th>
<th>within clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>range</td>
</tr>
<tr>
<td>Canthon water beetles</td>
<td>0.115</td>
<td>0.099–0.125</td>
</tr>
<tr>
<td></td>
<td>0.162</td>
<td>0.140–0.190</td>
</tr>
</tbody>
</table>

(uncorrected p-distance) of sequences within clusters ranged from 0 to 2%, whereas the mean divergence between clusters was 10 and 19% for Canthon and water beetles, respectively (table 2). These patterns of divergence are similar to those reported for intraspecies comparisons in other barcoding studies (Hebert et al. 2003b). The findings also appear to indicate that clusters in Canthon were more closely related to one another than were the clusters in the more diverse sample of water beetles.

The monophyly of the cox1 clusters was highly supported. When Bremer Support was considered separately for three categories of node levels (tip nodes within a cluster, nodes immediately subtending a cluster, and nodes defining basal relationships between the clusters), the majority of total tree support was derived from nodes immediately below (i.e. defining) the clusters (88% of total Bremer Support for water beetles, 80% for Canthon; table 3). Tip nodes within clusters, and basal nodes had low support.

(uncorrected p-distance) of sequences within clusters ranged from 0 to 2%, whereas the mean divergence between clusters was 10 and 19% for Canthon and water beetles, respectively (table 2). These patterns of divergence are similar to those reported for intraspecies comparisons in other barcoding studies (Hebert et al. 2003b). The findings also appear to indicate that clusters in Canthon were more closely related to one another than were the clusters in the more diverse sample of water beetles.
Table 3. Parsimony analysis on the four datasets produced trees with minimal length and homoplasy values as indicated. (Bremer Support was calculated for three categories of nodes, corresponding to those near the tips within a cluster (tips), the nodes immediately below a cluster (sub-cluster) and the nodes defining basal relationships between the clusters (basal). The number given is the sum of the nodal Bremer Support values for this category in the entire tree, and numbers in parentheses give the number of nodes assigned to each category (many of them collapsed near the tips because sequences are identical or very similar). Total Bremer Support refers to the sum of the values for the entire tree. For Canthon, a combined analysis of cox1 and 28S datasets was conducted, either combining all terminals in a single ‘supermatrix’ (all terminals), or removing all terminals which were not complete for either one of the two gene partitions (core terminals, n = 50).)

<table>
<thead>
<tr>
<th>Data</th>
<th>Bremer Support (no. nodes)</th>
<th>Bremer support (no. nodes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>tips</td>
</tr>
<tr>
<td>water beetles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox1</td>
<td>284</td>
<td>17</td>
</tr>
<tr>
<td>28S</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>Canthon (all terminals)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox1</td>
<td>296</td>
<td>13</td>
</tr>
<tr>
<td>28S</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>cox1 + 28S</td>
<td>92</td>
<td>7</td>
</tr>
<tr>
<td>Canthon (core terminals)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox1</td>
<td>260</td>
<td>10</td>
</tr>
<tr>
<td>28S</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>cox1 + 28S</td>
<td>277</td>
<td>12</td>
</tr>
</tbody>
</table>

n = 50; table 3). Total tree support was much higher in the analysis of core terminals as compared to when all individuals were included in the combined analysis (figure 3, table 3). However, the incongruence length difference was minimal, indicating that the drop is not due to conflict between both markers but the reduced discriminatory power of the dataset once a large number of missing entries is included in the data matrix. In all cases, total tree support was higher for cox1 than 28S for all analyses, regardless of whether all Canthon or only core individuals were used in the calculation (table 3), presumably due to the larger number of character changes in the former.

4. DISCUSSION

(a) The partitioning of genetic variation

The most striking result of the DNA analysis was the strong clustering of the sequence variation, with comparably large distances between groups of closely related sequences. In addition, these clusters showed remarkably high levels of nodal support for their monophyly according to cox1 and 28S genes. Support in the combined analysis was even higher and is essentially the sum of the individual partitions, showing a high degree of congruence for the two markers. Nodes defining the clusters included >80% of the total Bremer Support provided by the datasets, although they specify only one-third or less of the total number of nodes in the tree. Variation in the 28S nuclear gene, while showing overall fewer character changes, was also strongly clustered. Genotypes were shared by many individuals and were separated by a minimum of a single base pair (bp) in Canthon and a minimum 2 bp insertion differentiating two water beetle genotypes (e.g. WB-4 and WB-5). For the cox1 variation, >95% of variation occurred among these different clusters, with only ca 2.5–4.5% of the variation within these groups. Remarkably, the observed pattern of clustering was very similar in the two groups of beetles, even though they are composed of members of two different suborders Adephega and Polyphaga, and obtained from different parts of the world (Madagascar and the Neotropics).

A further key finding of this study is that the nuclear and mitochondrial gene data were completely congruent for both samples of beetles. Individuals were grouped into clusters in the exact same way whether based on the cox1 or 28S genotypes. Cox1 sequences were more variable than 28S, but multiple cox1 haplotypes in a cluster were monophyletic with respect to a single 28S genotype. For the water beetles this may be biased by the fact that these were field-samples from a given locality, raising uncertainty as to whether sister or even closely related taxa co-occur and were collected. For Canthon, by contrast, a single lineage was deliberately chosen from a larger sample of dung beetle communities (unpublished), and a wider sampling range covered, in order to increase the probability of sampling sister taxa. Notably, even in the case of the very closely related Can-1 and Can-2 clusters, where only a 2 bp insertion segregated 28S genotypes, the cox1 phylogenetic analysis was completely congruent with separation into two distinct groups.

(b) What is the nature of the clusters?

Several lines of evidence suggest that the clusters identified in this study represent distinct species, rather than any other level of hierarchical organization. Phenetic sequence divergence in mtDNA within these groups never exceeded 2% and usually was much lower, whereas divergence between the clusters was often greater by more than an order of magnitude. This is in general agreement with empirical levels of divergence found between species in phylogeographic analyses (Avise & Walker 1999) and barcoding studies (Hebert et al. 2003b). For Canthon, the existing molecular phylogenetic and taxonomic framework...
also supports this conclusion, as clusters from our study necessarily represent subgroups below the genus level, and GenBank database entries of various species of Canthon correspond to different clusters in our analysis. Beyond simple comparisons of phenetic divergence, inspection of the phylogenetic trees revealed a striking shift in branch length, long branches leading to subtending nodes and short branches within tip clusters, as seen in other studies of closely related species (Barraclough unpublished; Pons et al. unpublished). In addition, the strong support for the sub-cluster nodes (and no other node level) also confirms the unique status of this particular level of

Figure 3. Tree from the combined data matrix (28S-cox1 all terminals; table 2) for Canthon. Bremer Support values are reported above branches for the tree pictured. Values in parentheses are from the analysis of only the core terminals (50 individuals; table 2).
hierarchy in the trees. Whether or not this represents the species boundary remains to be investigated further. Ultimately, additional information, such as field studies of the sampled populations and broader genetic surveys including sister species, is required to confirm that the groups defined by these nodes are defining the species. However, the species category does take up a special place in the taxonomic hierarchy as the only ‘natural’ level of organization of the classificatory system, in contrast to the higher levels, such as genera and families (Cracraft 1983). It is our hypothesis that the transition in branching patterns, and the shift from strong to negligible branch support, represents a genetic signature of this unique level of organization.

(c) Methodological issues of species delineation from sequence data

Existing approaches to species delineation from sequence variation alone have been applied mainly to very small organisms, such as prokaryotes or soil nematodes, in which morphological discrimination is difficult or impossible (Floyd et al. 2002; Gregory & DeSalle 2005). In the case of nematodes, Molecular Operational Taxonomic Units (MOTUs) have been assigned based solely on sequence divergence (Blaxter 2004). While there may be no better way to classify these organisms to date, it remains unclear how these MOTUs correspond to evolutionarily differentiated groups, and how meaningful they are with respect to species cohesion. While the observation of large inter- and low intra-species variation promises easy identification of described species and the discovery of many cryptic species (Hebert et al. 2003b; Hebert et al. 2004b), there is concern regarding variability in the threshold values both between individual sister species pairs and among major lineages (DeSalle et al. 2005; Moritz & Cicer 2004).

In part, the problem of quantitative species delimitation could be overcome by searching for diagnostic character combination (Cracraft 1983), or complex character combinations (DeSalle et al. 2005) to define the species limits based on quantitative methods (Sites & Marshall 2003). These tests, which are rooted in the phylogenetic species concept, establish whether a priori populations can be ‘aggregated’ into a single species based on the distribution of characters or tree topology. These methods may not be practical when applied to large-scale species discovery and barcoding studies, where the cohesion of populations is unknown and broad sampling across species’ geographic ranges may not be possible.

A possible alternative to aggregation methods is to interpret branch length itself as being suggestive of species boundaries, assuming that the long branches defining the clusters could only have arisen if populations diverged longer than around Ne (effective population size) generations ago (Hudson & Coyne 2002). Appropriate methods for estimating these shifts include Templeton’s statistical parsimony analysis that partitions the variation into homoplastic (i.e. long branches) and non-homoplastic (short branches) variation (Templeton 2001). Similarly, it may be possible to statistically differentiate the shifts of lineage branching from interspecific, long branches to intra-specific, short branches using maximum likelihood methods (Barraclough, unpublished; Pons et al. submitted). A further approach could be based on population genetics analyses. It is possible to interpret the AMOVA results used to calculate intra- versus inter-cluster variation in a way analogous to F-statistics (Wright 1978). In this scenario, FST > 0.95 for both water beetles and Canthon datasets, meaning that > 95% of the total genetic variation in the dataset arises from differences among groups. A threshold of < 5% within-group variation seems a reasonable means of minimizing the chance of overlooking distinct taxa. As an example, combining Can-1 and Can-2 into a single group (‘Can1-2’) and recalculating AMOVA statistics results in a 15.6% value for within-group variation, as opposed to the 2.5% value when these two groups are considered as separate entities (table 1). This demonstrates the stringent clustering of the data, and provides a simple procedure to identify groups that have been grouped incorrectly.

(d) Conclusions and prospects

The aim of the present study was to investigate the efficacy of short sequence fragments for use in the discovery, delineation and routine identification of species. The analysis neither strictly constitutes a test of whether DNA can delimit pre-defined species, nor was it an analysis of Type I or Type II errors of species assignment (Quicke 2004). Instead, we used parsimony analysis to simultaneously examine a large number of sequences to assess patterns of variation in nature, and enquired whether this conforms to expectations of clustering at the species level of the biological hierarchy. The results were striking: sequence variation clustered very strongly for both nuclear and mitochondrial markers, and nodes defining these clusters were well supported whereas tip nodes, connecting closely related individuals, were not. Whether all the clusters we identified correspond to pre-existing, named species remains to be tested, and would require the input of specialists experienced in these taxonomic groups. Notably, comparisons with GenBank sequences indicated that at least three of the clusters identified here do indeed correspond to named species of Canthon.

The analytical approach will allow these clusters to be delineated objectively and repeatedly by anyone using the sequence data matrix. As a result, DNA data can form the basis of testable taxonomic hypotheses that could be examined with additional types of data in the future. The benefits of this approach are manifold: it provides a rapid division into probable groups of reproductively isolated individuals and generates more direct links to their evolutionary past; it will facilitate the determination of distinctive morphological features (i.e. through a focussed comparison of pre-delineated groups); it would allow the study of these putative species to proceed even while formal description is pending; it would link individuals from the same species collected in different localities or in different studies in a way that arbitrary designations (e.g. ‘Canthon sp.1’) do not; and it would immediately
provide the data needed for future DNA barcoding identification.

The results of the present study are based on a relatively small number of species, but nonetheless demonstrate the general feasibility of using DNA-based methods in the large-scale delineation and discovery of new species, even in poorly characterized groups. While more research is needed to establish the best approach for species delineation using DNA (e.g. through phylogenetic or coalescent methods, with phenetic barcode ‘thresholds’, or some combination thereof), it is becoming evident that DNA methods present a promising new means of assessing and identifying biological diversity in some of the most species rich taxa and environments on Earth. There is reason for optimism that, if fully developed and implemented on a broad scale, DNA-based tools such as those examined here may provide the first opportunity for creating a comprehensive inventory of life.

We are grateful to Daegan Inward, Richard Davies and Liz Powell for collecting *Canthon* dung beetles; to David Lees, Raymoniarana Ranaivosolo, Pierre Razafindraire, Roger Andriamampy and Doug Otuke for assistance with water beetle collection; to Ruth Wild and Miranda Elliot for laboratory analysis; and to Silvia Fabrizi for mounting specimens. In Madagascar, we thank York Pareik at King de la Piste and Madame Liva and Benjamin Andriamihaja at MICET (Madagascar Institut pour la Conservation des Ecosystèmes Tropicaux). Tim Barraclough and an anonymous referee provided helpful comments on the manuscript.

REFERENCES


Wright, S. 1978 *Evolution and the genetics of populations*. Chicago: University of Chicago.
Defining operational taxonomic units using DNA barcode data

Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd and Eyualem Abebe

*Phil. Trans. R. Soc. B* 2005 360, 1935-1943

References

This article cites 29 articles, 9 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1935.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1935.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to *Phil. Trans. R. Soc. B* go to: http://rspb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
Defining operational taxonomic units using DNA barcode data

Mark Blaxter*, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd† and Eyualem Abebe‡

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, King’s Buildings, Edinburgh EH9 3JT, UK

The scale of diversity of life on this planet is a significant challenge for any scientific programme hoping to produce a complete catalogue, whatever means is used. For DNA barcoding studies, this difficulty is compounded by the realization that any chosen barcode sequence is not the gene ‘for’ speciation and that taxa have evolutionary histories. How are we to disentangle the confounding effects of reticulate population genetic processes? Using the DNA barcode data from meiofaunal surveys, here we discuss the benefits of treating the taxa defined by barcodes without reference to their correspondence to ‘species’, and suggest that using this non-idealist approach facilitates access to taxon groups that are not accessible to other methods of enumeration and classification. Major issues remain, in particular the methodologies for taxon discrimination in DNA barcode data.

Keywords: DNA barcodes; molecular operational taxonomic units; tardigrades; nematodes; meiofauna; small subunit ribosomal RNA

1. INTRODUCTION: THE UNSEEABLE ANIMAL

The total number of unique taxa described to the species level is circa 1.5 million, but the total number of ‘species’ is likely to be in the region of 10 million (May 1988). The overall ‘taxonomic deficit’ (the ratio of expected taxa to named taxa) is thus approximately sixfold. However this deficit, like all phylogenetic things, is not immune to systematic bias (Blaxter 2003). For vertebrates, the current described species total is likely to be relatively close to the ‘true’ total: we have described most of these relatively large organisms. The same is true of most groups whose members have body sizes greater than 10 mm. However, the vast majority of organisms on the Earth have body sizes less than 1 mm, and for these groups the taxonomic deficit is likely to be several fold worse than for land plants and vertebrates (Lambshead 1993; Platt 1994; Lambshead & Boucher 2003). These meio- and micro-fauna and flora are, however, key to the functioning of ecosystems and are the productive and saprophytic base upon which the macro-organisms rely. Their size precludes facile visual identification, and indeed much of their important morphology may be at scales that are beyond the resolution of light microscopy (De Ley & Bert 2001; De Ley et al. 2005). Wendell Berry quotes from his daughter in his poem ‘To the unseeable animal’: ‘I hope there’s an animal somewhere that nobody has ever seen. / And I hope nobody ever sees it.’ (Berry 1970). We suggest that DNA barcoding may permit rational access to these animals.

DNA barcoding, the use of a specified DNA sequence to provide taxonomic identification for a specimen, is a technique that should be applicable to all cellular (and much viral) life (Floyd et al. 2002; Hebert et al. 2003; Tautz et al. 2003; Blaxter et al. 2004). Theoretically, this should allow rapid and high-throughput identification, either of individual organisms or of sequences isolated from an environmental DNA sample. Specimen-independent DNA surveys are already used for microbial (Giovannoni et al. 1990) and protozoal communities (Diez et al. 2001; Lopez-Garcia et al. 2001; Moreira & Lopez-Garcia 2002; Amaral-Zettler et al. 2002), and have revealed a wealth of hidden diversity. Meiofauna would appear to be an ideal group in which a molecular identification system could be used (Lambshead 1993; Lawton et al. 1998; Blaxter 2004).

2. BARCODING MEIOFAUNA: CHALLENGES

The number of meiofaunal taxa, animals with a body size ~1 mm (or less), can only be guessed at. Thus, the number of described species of nematodes is quoted as between 26 000 and 40 000, but the real total estimated to be above one million (Lambshead 1993; Platt 1994; Lambshead & Boucher 2003). The deficit may be put into perspective by considering that the number of described species of soil dwelling nematodes for the UK is approximately 400, a figure surprisingly close to the inventory of UK breeding birds. Is the UK nematode fauna really that depauperate? Our surveys of nematodes in soils in relatively degraded habitats (upland farm grassland) suggest that taxon numbers identifiable from even a small area may be remarkably high (R. Floyd, A. Eyualem and M. Blaxter, 2005).
unextracted. Similarly, for tardigrades, the described UK fauna is ~100 species (Maucci 1986; Kinchin 1994), but we have identified over 50 taxa from one restricted set of sample sites (Blaxter et al. 2003). While some authors have argued for a relatively low number of meio-taxa matched by a near-ubiquitous distribution (Finlay 2002), we have found that different sites, though close geographically, can have very different taxon assemblages (Blaxter et al. 2003). If organisms with a body size <1 mm really do have no biogeographical structure, and are all essentially ubiquitous, the sampling we have carried out suggests at least that relative abundances must vary greatly between sites. Meiofaunal barcoding must fall into the purview of the third community identified above: experimental investigation of biodiversity.

We have been generating DNA barcode datasets for meiofaunal specimens (mostly nematodes and tardigrades) for several years (Floyd et al. 2002; Blaxter & Floyd 2003; Blaxter et al. 2003; Eyuaelem & Blaxter 2003; Blaxter 2004; Blaxter et al. 2004). We are agnostic as to whether the taxa we can define using these barcode sequences (which we call ‘molecular operational taxonomic units’ or MOTU) are ‘species’ or not, though in the case where we have compared and contrasted MOTU, morphological species hypotheses and breeding-based biological species, MOTU and biological species hypotheses were congruent while morphological analyses disagreed internally, and with the other modes of taxon definition (Eyuaelem & Blaxter 2003). We have traditionally used the nuclear small subunit (nSSU) as a marker, but have also tested nSSU alongside cytochrome oxidase subunit I (cox1), with equivalent resolution. Here, we use a new dataset of meiofaunal barcodes to discuss what we feel are very interesting and important features of DNA barcode data: they can be used not only to define taxa, but also to identify sets of specimens for which robust taxonomic hypotheses are difficult to construct. These clouds of related specimens are immediately of interest for further study: is this evidence for recent, rapid radiation of distinct taxa or is it evidence for a highly variable single taxon?

3. METHODS: OBTAINING MEIOFAUNAL BARCODE SEQUENCES

(a) Sampling of moss ecosystems
Moss samples for this study were collected from dry stone walls surrounding Ettrick Old Church, in Glen Ettrick in Southern Scotland (Blaxter et al. 2003). Meiofauna were isolated by modified Baermann funnel separation through milk filters into sterile tap water. Larger fauna (such as collembolans and mites; body sizes >2 mm) were excluded from the separation by the pore size of the filter: some of these arthropods were picked from moss individually. Relative numbers of animals from each phylum were counted from a subsample of the filtrate, and a few of each phylum picked individually: the remainder was processed for DNA extraction.

(b) Individual specimen barcoding
Individual animals were extracted using the NaOH direct lysis procedure: this yields ~40 μl of stable extract per specimen from which over a dozen PCRs can be performed (Floyd et al. 2002). Bulk filtrate animals were concentrated by centrifugation and extracted using a snap-freezing/protease K/phenol/chloroform protocol. The nSSU marker was amplified from individual extracts using the primers SSU_F04 (GCTTTGCTCAAGGATTAAGCC) and SSU_R26 (CATTCTTGCGAAATGCTTCCG) (Blaxter et al. 1998), yielding a ~900 base pair (bp) product. These primers were designed to be metazoan-specific (Blaxter et al. 1998). The cox1 amplifications were amplified from a subset of tardigrade individuals (also amplified for nSSU) using the ‘universal’ primers cox1 (HC02198; TAAAATTCAGGGTGACCAAAAAATCA) and cox1 (LC01490; GGTCAACAAATCATAAAGATATTGG) (Hebert et al. 2003), yielding a ~650 bp product. Shrimp alkaline phosphatase/exonuclease I-cleaned PCR products from single specimens were sequenced directly using SSU_R09 or cox1.

(c) Barcodes from bulk ecosystem DNA
Bulk filtrate animals were concentrated by centrifugation and extracted using a snap-freezing/protease K/phenol/chloroform protocol. The nSSU marker was amplified as described above. Amplicons of nSSU generated from the bulk extract target were cleaned using a Montage gel extraction kit and cloned into pTOPO2.1 (Invitrogen). After growth on LB/kanamycin IPTG/Xgal, recombinant colonies were picked to 200 μl of LB broth with kanamycin in microtitre plates and grown overnight. Inserts in the recombinant plasmids were amplified from ~1 μl of overnight liquid culture using the primers M13_F (CTGCGCCGTGTTTAC) and M13_R (CAGGAAACACGTATA), cleaned using shrimp alkaline phosphatase/exonuclease I, and sequenced using SSU_R09 (AGCTGG-AATTACCGCGGCCGTG) and ABI BigDye3.0 reagents to produce ~500 bp of sequence.

(d) Molecular operational taxonomic unit definition
The sequencing was carried out on an ABI3730 capillary sequencer, and sequencing chromatograms were post processed with tracseq (a perl program that uses phred to identify high-quality base calls and crossmatch to identify vector sequence; A. Anthony and M. Blaxter, unpublished). All sequences have been deposited in EMBL/GenBank/DDBJ. The perl program ‘MOTU_define.pl’ (R. Floyd and M. Blaxter, unpublished; based on CLOBB (Parkinson et al. 2002)) was used to allocate the resulting high-quality sequences to MOTU, based on pairwise identity scores and a user-defined cutoff.

The MOTU_define.pl program adds sequences one at a time to a growing database of barcode sequences (figure 1). It is a very simple procedure, internally consistent, and has the benefit of allocating stable MOTU identifiers to the dataset. As more sequences are generated, they can be added incrementally to the existing MOTU sets and thus continuity between experiments is attained. Indeed, sequence data can be acquired from other sources (such as GenBank/EMBL) and added to the dataset without compromising or changing the MOTU
 assignment of local data. This sort of process is ideal for building up a shared database of MOTU assignments and sequences. It is relatively rapid and reasonably scaleable (a variant of the program using the megaBLAST algorithm can cluster 100 000 expressed sequence tag sequences in $\sim 20$ h on a desktop computer (Parkinson et al. 2002); we do not yet have barcode datasets of this magnitude to test). The MOTU_define.pl is freely available from M. Blaxter, and requires only perl and a local copy of the NCBI BLAST suite (it is thus installable on UNIX, MacOSX and Windows systems).

---

**Figure 1.** The MOTU_define.pl system. (a) A schematic of the process by which MOTU_define.pl allocates sequences to MOTU. The process can be run any number of times with different sequence addition order to assess MOTU stability. (b) The effect of addition order on MOTU definition. Three sequences, A, B and C, are clustered into MOTU. A differs from B, and B from C by less than the MOTU discriminant cutoff, but C differs from A by more than the cutoff. Depending on the order of analysis of the sequences, either one or two MOTU will be defined.
Following our previous analyses of similar data, and our measured error rate in sequencing (~1 base in 3500) (Floyd et al. 2002), we standardly use a cutoff of 2 base differences in ~500 bp of sequence to discriminate MOTU: this can be varied. The program can also be rerun multiple times over the same set of sequences, randomizing the input order each time, and thus can be used to identify sequences and MOTU that do not behave simply under the cutoff statistic used. The use of single linkage clustering in MOTU_define.pl (where each sequence is clustered based on its identity to a single comparator) avoids issues of ambiguous alignment across a wide range of distantly related sequences. The high-quality sequences were aligned to each other and to a set of relevant control sequences from named taxa derived from GenBank or our previous studies and the alignment analysed using Maximum Parsimony in PAUP* 4.0b10 (Swofford 1999).

4. RESULTS: ETTRICK MOSS MEIOFAUNA

The moss fauna included animals from five animal phyla: Arthropoda (mites and collembolans), Tardigrada, Annelida (enchytraeids), Nematoda and Rotifera. The filtrates also included many protozoa (ciliates and amoebas) and some plant material. There was doubtless also a thriving unicellular fungal and algal, and prokaryote presence. Nematodes were most abundant, followed by rotifers and tardigrades (a ratio of 132±20.8 nematodes to 6±0.6 rotifers to 3±0.9 tardigrades; mean and standard error of four samples corresponding to 0.5% of the extract from ~1 g dry weight of moss ecosystem); collembolans, mites and enchytraeids were rare in the moss, and excluded by their size from the filtrate.

(a) Barcode sequence generation from single specimens

Barcode sequences were derived from single specimens of nematodes, mites, collembolans, and enchytraeids. A total of 121 cox1 sequences were generated from over 270 tardigrade specimens. For all taxa except rotifers, cox1 PCR and sequencing was successful (Floyd 1938 M. Blaxter and others) Defining taxonomic units nSSU PCR and sequencing was successful in ~500 bp tardigrades. For all taxa except rotifers, cox1 nSSU PCR and sequencing was successful. We conclude that we will have to improve our extractions specifically to enhance rotifer DNA recovery.

(b) Barcode sequences from nSSU libraries from bulk DNA

A total of 145 sequences were generated from the bulk nSSU PCR library. Comparison to database sequences and single-specimen sequences from the same collection site (Blaxter et al. 2003) indicated that most derived from nematodes (123 or 85%) and four from tardigrades (3%). This ratio corresponds to that derived from the visual survey, excepting that no rotifer nSSU was recovered. In addition to these animal sequences, we isolated 18 nSSU sequences that clearly derived from ciliate protozoa, though none had an exact match in the public databases. We presume that these DNA segments were amplified because our primer set is not strictly metazoan-specific (we know that we can amplify environmental fungi, data not shown) and because, despite their being unicellular protozoa, ciliate macro-nuclei contain a many thousand fold amplification of the genes archived in the micronucleus, including the ribosomal RNA operons. No enchytraeid or arthropod sequences were recovered because the filtration excludes these larger meiofauna. Chimaeric amplicons are the bane of environmental sampling PCR. They arise from mispriming by amplification products during PCR, and result in DNA sequences that match one taxon at the 5’ end and another, unrelated one at the 3’ end. No chimaeric amplicons were identified, based on finding no discrepant BLAST matches for the first 250 compared to the last 250 bases of each.

(c) Comparing single specimen and bulk nSSU MOTU

MOTU_define.pl was used to infer MOTU from the nSSU datasets using a 2 bp difference cutoff. Data from the bulk sample and the single specimen sequences were clustered independently. For each nSSU MOTU, we derived a consensus sequence to represent that cluster for subsequent phylogenetic analysis (figure 2; but note that the definition of membership of a MOTU is not based on phylogenetic analysis). The use of a consensus sequence does not imply that this sequence correctly represents some ideal version of the true sequence, but rather is used to represent the diversity of the constituent sequences. The most abundant nSSU MOTU, derived from the bulk dataset, has 106 representatives, and is most similar to the chromadorid nematode Plectus aquitilis. Two of 16 single-specimen nSSU MOTU were also found in the bulk sample data (the P. aquitilis-like MOTU and a Clarkus (nematode)-like MOTU; figure 2). The bulk sequence dataset reflects the expected distribution of animals observed, excepting the Rotifera, and comparison with other more extensive datasets from soils and moss environments affirms that within the phyla that were amplified there is no apparent phylogenetic bias. Examination of this dataset suggests that the rate of identification of novel taxa using the barcode is not yet at saturation, despite the presence of the hyperabundant P. aquitilis-like Bulk_2bp_MOTU0001/Sin_2bp__MOTU0005 (58% of all sequences, and 73% of the bulk sample sequences). Presumably, the rate of new MOTU identification could now be enhanced by prescreening for P. aquitilis-like sequences.

(d) Comparison of cox1 and nSSU barcode analyses

A representative MOTU definition set for the cox1 sequences is shown in figure 3. Twenty-two MOTU were defined, containing from 1 to 65 sequences...
The distribution of abundances of taxa implies one abundant taxon (~50% of the sample) and a larger number of taxa with low abundance.

For 82 tardigrade specimens, we obtained sequences of both cox1 and nSSU with >490 bp of high-quality data. The two markers were used to infer independent clusterings, using a 2 bp cutoff, and the resultant clusters compared (figure 4). Seventeen cox1 MOTU were defined from this subset. Surprisingly, 23 nSSU MOTU were defined, despite the overall lower level of sequence divergence, though the distance between distinct clusters was greater in the cox1 dataset (as would be expected from the known higher substitution rate in animal mitochondrial genes). Seven MOTU with single members were found in both datasets, and two cox1 MOTU (with two and five members) corresponded to two nSSU MOTU each (figure 4). The remaining 68 specimens formed two groups with complex patterns of overlap between nSSU and cox1 MOTU (figure 4). Thus, while cox1
and nSSU are both effective at defining MOTU, and there was a general agreement between the two cluster sets, there were also significant disagreements. Whether these disagreements are due to the population history and hybridization patterns of the specimens sampled or are indicators of real incongruence between the markers is not clear. The two clouds of taxa (marked in figure 4) may correspond to particularly variable single taxa, or perhaps diverging radiations of taxa. Many tardigrades can reproduce asexually, or have sex only very rarely (Kinchin 1994), and thus this pattern may reflect divergence of clonal or matrilineal lines.

Figure 3. Tardigrade MOTU defined using cox1 sequences. (a) A consensus sequence was derived for each MOTU, and these were aligned. The branch lengths are proportional to the number of discrete changes mapped to each. The number of sequences assigned to each MOTU is given in brackets after the MOTU name. (b) Histogram of MOTU abundance in the 121-sequence cox1 dataset.

5. RESULTS: PROPERTIES OF EXACT SCORE MOTU DEFINITION

(a) Variability due to single linkage clustering
Assignment of any single sequence to a MOTU depends critically on what sequences have been added previously (figure 1b). If one takes three sequences, where only two differ by more than the chosen cutoff, the order of addition changes the number and membership of MOTU inferred. Rather than being a failing of this procedure, we regard this as being a feature: it permits exploration of the ‘clouds’ of taxa that are closely related. If a set of specimens robustly clusters into a particular set of MOTU, no
MOTU_define.pl: the same issue must arise in all other meiofaunal datasets. We have also observed this pattern in other meiofaunal datasets (Blaxter et al. 2003) (Floyd, Blaxter et al. unpublished).

The variability of attribution observed between independent clusterings is not a unique feature of MOTU_define.pl: the same issue must arise in all other methods, as the data we are using to infer taxa is essentially quantal.

![Figure 4. Comparison of MOTU definition using nSSU and cox1 markers. This Venn diagram shows cox1 MOTU sets (solid circles) and nSSU MOTU sets (dotted squares). The numbers within each partition indicate the number of individual specimens (out of 82) placed there.](image)

![Figure 5. Variability in the number of MOTU defined by replicate analyses. The histogram shows the frequency distribution of total numbers of clusters inferred from 295 tardigrade nSSU sequences from the Glen Ettrick study site using MOTU_define.pl at three different cut off values: 2 bp (black), 3 bp (hatched) and 4 bp (open). The mean and standard deviation of each set of analyses is given.](image)

**6. DISCUSSION: TURNING SEQUENCES INTO MOTU**

The MOTU-define.pl program is but one, obvious way of inferring MOTU. Other methods could also be applied. One common choice is to use a BLAST algorithm (usually BLASTn) (Altschul et al. 1997) to identify the best match in a reference database, and to assign the identifier of the best match to the barcoded specimen if the similarity is judged to be good enough. This method has many pitfalls, not least its reliance on a well-populated (and correctly named) database of barcodes. In meiofaunal surveys such as presented here, the lack of close relatives in the database can make this approach less-than-rewarding. More importantly, the BLAST algorithm (Altschul et al. 1990; Altschul et al. 1997) was not designed for barcode identity assignment, and simply taking the top-scoring match, with some predefined quality score cutoff, may miss issues of, for example, equal top scoring matches. A variation on the BLAST approach would be to perform a complete alignment with the barcode query sequence, and then subject this alignment to model-driven phylogenetic analysis to ask if the barcoded specimen is a credible member of a monophyletic clade with any of the references.

Because much DNA barcode sequence is derived from single sequencing reads on only one strand of the DNA, the quality of the sequences may not be as good as those in the databases. The sequencing chromatogram can be analysed to yield a quality score for each base (Ewing & Green 1998; Ewing et al. 1998), and these could be incorporated into a BLAST-and-align method for MOTU definition that down-weights any differences associated with low quality scores and pays more attention to high-quality scores. A variation on this method might also include partitioning the aligned sequences.
sequences *a priori* into more- and less-informative sites. Thus, in a protein-coding gene such as *cox1*, one might give first and second base changes more weight than those observed in fourfold degenerate sites. In a RNA gene such as nSSU, one could differentially weight residues by their involvement in secondary structure, and their observed conservation in large aligned datasets.

As barcoding is applied somewhere on the span between population genetics and taxon phylogenetics, the use of network-based algorithms may also assist. Templeton network analysis is much used in population studies to examine patterns of haplotype distribution and relatedness (Clement et al. 2000). For DNA barcode data, such network analysis, with different cutoffs for the breaking of ties between subnetworks, can assist in understanding the patterns of diversity in the sequences and thus the likely status of the MOTU defined. In genomics, definition of protein families has been achieved using multiple cluster linkage methods, where complex networks of similarity between sequences can be examined at different levels of granularity to identify coherent clusters (Enright et al. 2002). A similar approach applied to DNA barcode data might be doubly informative of not only final MOTU but also the interrelationships of MOTU clouds.

Ultimately, we might want to use rigorous phylogenetic methods to affirm the monophyly of our newly defined MOTU, and to place them in the context of named sequence diversity. However, we must be aware of the issues of partial sorting of haplotypes between lineages as they diverge. Wide-ranging studies on several taxa have clearly shown that while rapidly evolving sequences are very well suited to generation and testing of taxon hypotheses at local scales, they are often very much unsuited to deeper phylogenetic analysis. Processes such as base substitution bias and variable site saturation can rapidly obscure real phylogenetic signals and generate spurious trends in data. The barcode data will be rather unsuitable for reconstructing the deeper branches of the tree of life, including perhaps all those below the generic level (Vogler et al. 2005). Simply using trees to infer taxa from barcode data can be positively misleading: we should rather define the taxa and then examine their relationships through rigorous phylogenetics.

Taxa defined by MOTU methods can be used for standard taxonomic and ecological surveys. By comparing the barcode sequence with a database of sequences from specimens identified to Linnean taxa before sequencing, the anonymous survey specimens can be placed within the known taxonomic framework, and the organismal biology of the organisms from which they derived inferred (Floyd et al. 2002; Blaxter & Floyd 2003; Blaxter 2004). By this method we can move from anonymous sequence to ecosystem biology.

This work was carried out as part of ongoing investigations into meiofaunal diversity in our laboratory, and was funded by the UK Natural Environment Research Council and the Linnean Society of London. J.M. and T.C. carried out the meiofaunal surveys as part of their major undergraduate projects.

**REFERENCES**


Defining taxonomic units  M. Blaxter and others 1943


---


An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding


Phil. Trans. R. Soc. B 2005 360, 1945-1958

References

This article cites 48 articles, 8 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1945.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1945.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding

Paul De Ley¹,*, Irma Tandingan De Ley¹, Krystalynne Morris², Eyualem Abebe², Manuel Mundo-Ocampo¹, Melissa Yoder¹, Joseph Heras¹, Dora Waumann³, Axayácatl Rocha-Olivares⁴, A. H. Jay Burr⁵, James G. Baldwin¹ and W. Kelley Thomas²

¹Department of Nematology, University of California, Riverside, CA 92521, USA
²Hubbard Center for Genome Studies, University of New Hampshire, 35 Colvos Road, Durham, NH 03824, USA
³Facultad de Ciencias Marinas, Universidad Autónoma de Baja California, Apdo. Postal 453, Ensenada, Baja California, CP 22800, México
⁴Centro de Investigación Científica y de Educación Superior de Ensenada, Apdo. Postal 2732, Ensenada, Baja California, CP 22860, México
⁵Department of Biological Sciences, Simon Fraser University, Burnaby, Canada BC V5A 1S6.

Molecular surveys of meiofaunal diversity face some interesting methodological challenges when it comes to interstitial nematodes from soils and sediments. Morphology-based surveys are greatly limited in processing speed, while barcoding approaches for nematodes are hampered by difficulties of matching sequence data with traditional taxonomy. Intermediate technology is needed to bridge the gap between both approaches. An example of such technology is video capture and editing microscopy, which consists of the recording of taxonomically informative multifocal series of microscopy images as digital video clips. The integration of multifocal imaging with sequence analysis of the D2D3 region of large subunit (LSU) rDNA is illustrated here in the context of a combined morphological and barcode sequencing survey of marine nematodes from Baja California and California. The resulting video clips and sequence data are made available online in the database NemATOL (http://nematol.unh.edu/). Analyses of 37 barcoded nematodes suggest that these represent at least 32 species, none of which matches available D2D3 sequences in public databases. The recorded multifocal vouchers allowed us to identify most specimens to genus, and will be used to match specimens with subsequent species identifications and descriptions of preserved specimens. Like molecular barcodes, multifocal voucher archives are part of a wider effort at structuring and changing the process of biodiversity discovery. We argue that data-rich surveys and phylogenetic tools for analysis of barcode sequences are an essential component of the exploration of phyla with a high fraction of undiscovered species. Our methods are also directly applicable to other meiofauna such as for example gastrotrichs and tardigrades.

Keywords: identification; taxonomy; Nematoda; meiofauna; Gulf of California

1. CAPTURING NEMATODE DIVERSITY, IDENTITY AND MORPHOLOGY

Nematodes are estimated to be the most abundant metazoans on earth (Lambshead 2004). Even though they probably constitute one of the most diverse metazoan phyla in terms of species richness, the overwhelming majority of these species remains unknown (Hugot et al. 2001; Coomans 2002). Contrary to common assumption, many nematode taxa are morphologically highly diverse (De Ley in press), but this diversity remains under-appreciated as it often requires high-resolution light microscopy and electron microscopy to be observed, as well as knowledge of the wide range of forms described in specialized taxonomic literature. Nematode identification with morphological characters is often difficult and laborious because of these same two requirements. To make matters worse, nematode species can be variable in morphology, while the differences between valid species can be obscured by cryptic diagnostic differences (e.g. De Ley et al. 1999). As a result, ecological studies and surveys of nematode diversity are usually restricted to identifications at genus level, since most genera can be identified at first sight, while taxonomic surveys hardly ever approach completeness in identifying all species isolated from all but a few samples.

Aside from the challenges of identification, vouchering of microscopic nematode specimens presents some...
interesting problems in its own right. A number of well-established protocols exist for preserving and mounting nematodes in permanent slides, and in our experience type specimens may survive over a hundred years of storage. Unfortunately, in practice it turns out that within 5–15 years most vouchers deteriorate substantially, are shattered accidentally or go missing, not least because of the steady erosion of long-term investment in taxonomic collections. Obtaining detailed images of their appearance could mitigate the loss of type specimens. A fundamental limitation in the documentation and vouchering of microscopic invertebrates has always been the critical need for observation with high-resolution microscopy, before any level of accuracy can be attained in interpretation of characters and identification of species. Nematode specimens, as a particular case in point, are not only thin and transparent but are also three-dimensional objects that must be examined de rigueur with multiple magnifications and at multiple focal planes. This multifocal aspect explains why in most cases single photographs are not sufficient to document nematode morphology, and why re-examination of type specimens is often really necessary (though rarely possible) if published drawings and descriptions do not suffice to confirm or reject conspecificity. In order to fully capture the multifocal nature of nematode microscopy, a new approach for recording nematode morphology was therefore proposed by De Ley & Bert (2002). Inspired by the complex software and equipment used to record nematode embryonic development with a technique known as ‘four-dimensional microscopy’ (Thomas et al. 1996; Schnabel et al. 1997), this new approach consists of capturing a multifocal series of images to computer hard disk in the format of a single video clip. This clip can then be edited with any consumer-type video editing program to optimize it for distribution via the Internet or on CD. The resulting technique was dubbed ‘video capture and editing’ (VCE) and examples of some of its many different applications can be seen at http://faculty.ucr.edu/~pdeley/vce.html on the World Wide Web.

The process of capturing morphology of transparent microscopy specimens as multifocal clips is highly adaptable and scaleable. It can be applied at minimal expense in any laboratory equipped with at least one good research microscope, one video camera and one personal computer. In this case, nematode images can be recorded directly as video clips while simply focusing by hand through parts of the specimen. However, with sufficient resources the process can be qualitatively optimized and standardized using an automated computer-controlled microscope and digital camera. A first example of the latter approach was published by Eyualem et al. (2004) to document the findings of an exploratory marine nematode survey in the Gulf of Maine. In this case, nematode morphology was recorded by stepwise capturing of successive focal plans as image stacks, which were subsequently converted into digital video format to produce digital multifocal images (DMI) that are viewable in any multimedia player software. Public access to the resulting clips is available through a versatile online database called NemATOL (http://nematol.unh.edu/).

2. LINKING BARCODED GENOTYPES WITH PHENOTYPES

Molecular methods have opened up a wide range of new approaches to nematode surveys, particularly in the context of characterizing short sequence stretches amenable to species identification. Two recent developments are especially significant for barcoding applications. The first of these is the development of a phylogeny-based species concept for nematodes that is both theoretically sound and applicable in diagnosis (Adams 1998, 2001; Nadler 2002). The second is the development of nematode barcoding techniques that attain maximal processivity by foregoing any attempt at morphological vouchering or species identification, focusing instead on assignment to molecular operational taxonomic units (MOTU; Floyd et al. 2002; Blaxter 2004; Blaxter et al. 2005). It has been argued in the context of such surveys that it is impossible to describe biodiversity with traditional approaches (Blaxter 2003), that DNA sequences should become the basis of a new taxonomic reference system (Tautz et al. 2003; Blaxter 2004) and that cytochrome oxidase subunit I (coxI) is an especially appropriate choice for species discrimination in triploblastic Metazoa (Hebert et al. 2003). In the heat of the debate over these exciting new directions and their implied priorities for funding and collaboration, others have strongly cautioned against the dangers of mistaking the huge potential of new approaches for a justification to effectively sideline established methodologies and expertise (Lipscomb et al. 2003; Wheeler 2004).

Despite explicit statements to the contrary by proponents of new DNA-based standards for identification and classification (Hebert et al. 2003; Tautz et al. 2003; Blaxter 2004), there is a clear suspicion among other members of the taxonomic community that such approaches will yet again draw away from the already very meagre resources for morphological studies. This suspicion probably partly derives from comparisons made with current practice in microbial taxonomy, where genotypic criteria for species identification are now seemingly given absolute primacy over older phenotypic standards. In reality, microbiologists approach the thorny issue of prokaryote identification and classification with levels of ambiguity and circumspection (e.g. Torsvik et al. 2002) that are perhaps even greater than those found among eukaryote taxonomists, and it would be erroneous to assume that the former have altogether given up on expressed characters in general and on morphology in particular. Much like prokaryotes, nematodes have the paradoxical reputation of being biologically diverse yet morphologically uniform, to the point where it is tempting to conclude that nematode taxonomy should essentially reset itself and switch to similar principles as those of microbiology. In reality, nematodes harbour endless morphological diversity and the genuine problems with ‘traditional’ nematode identification are instead rooted in quite different obstacles. These obstacles include (but are not limited to) the particular constraints of observing, interpreting and representing nematode morphology through various microscopy techniques. Scanning or transmission electron microscopy often reveals clear diagnostic characters,
Table 1. Origin of nematodes and other meiofauna from which D2D3 sequences were obtained, along with their best BLAST hits in GenBank.

<table>
<thead>
<tr>
<th>specimen ID</th>
<th>provisional taxon ID</th>
<th>sample origin</th>
<th>BLAST % ID; GenBank accession number(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1P6K2</td>
<td>Ascolaimus</td>
<td>Newport beach, CA, USA</td>
<td>91.6; AF210406</td>
</tr>
<tr>
<td>2P6K2</td>
<td>Enoplolaimus</td>
<td>Newport beach, CA, USA</td>
<td>96.2; AF210421</td>
</tr>
<tr>
<td>3P6K2</td>
<td>Viscioa</td>
<td>Newport beach, CA, USA</td>
<td>97.4; AF210413</td>
</tr>
<tr>
<td>4P6K2</td>
<td>Metachromadora</td>
<td>Newport beach, CA, USA</td>
<td>88.9; AF210410</td>
</tr>
<tr>
<td>2I6K2</td>
<td>Enoplolaimus</td>
<td>Newport beach, CA, USA</td>
<td>97.4; AF210413</td>
</tr>
<tr>
<td>5P9K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>96.3; AF210421</td>
</tr>
<tr>
<td>1I11K2</td>
<td>Chromadorida</td>
<td>San Felipe, BC, Mexico</td>
<td>92.2; AF210421</td>
</tr>
<tr>
<td>2I11K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>96.7; AF210421</td>
</tr>
<tr>
<td>3I11K2</td>
<td>Enoploides/Enoplolaimus</td>
<td>San Felipe, BC, Mexico</td>
<td>99.1; AF210409, AF210407</td>
</tr>
<tr>
<td>4I11K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>96.6; AF210421</td>
</tr>
<tr>
<td>5I9K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>96.3; AF210421</td>
</tr>
<tr>
<td>1I12K2</td>
<td>Chromadorida</td>
<td>San Felipe, BC, Mexico</td>
<td>97.3; U94755, AF226573-89,</td>
</tr>
<tr>
<td>2I12K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>96.7; AF210421</td>
</tr>
<tr>
<td>3I12K2</td>
<td>Enoploides/Enoplolaimus</td>
<td>San Felipe, BC, Mexico</td>
<td>99.1; AF210409, AF210407</td>
</tr>
<tr>
<td>4I12K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>96.6; AF210421</td>
</tr>
<tr>
<td>5I12K2</td>
<td>Odontophora</td>
<td>San Felipe, BC, Mexico</td>
<td>94.7; AF210409</td>
</tr>
<tr>
<td>7I12K2</td>
<td>Linhomeus</td>
<td>San Felipe, BC, Mexico</td>
<td>98.1; U61758, AY652779,</td>
</tr>
<tr>
<td>8I12K2</td>
<td>Ceramonematidae</td>
<td>San Felipe, BC, Mexico</td>
<td>95.5; AY377676</td>
</tr>
<tr>
<td>11I4K2</td>
<td>Viscoa</td>
<td>San Felipe, BC, Mexico</td>
<td>92.6; AF210406</td>
</tr>
<tr>
<td>21I4K2</td>
<td>Metachromadora</td>
<td>San Felipe, BC, Mexico</td>
<td>93.8; AF210410</td>
</tr>
<tr>
<td>2P15K2</td>
<td>Laronema</td>
<td>San Felipe, BC, Mexico</td>
<td>95.7; AF210400</td>
</tr>
<tr>
<td>3P15K2</td>
<td>Tricoma</td>
<td>San Felipe, BC, Mexico</td>
<td>86.1; AF210421</td>
</tr>
<tr>
<td>11I2K3</td>
<td>Rhabdoca</td>
<td>Sta. Clara, BC, Mexico</td>
<td>96.9; AF210418</td>
</tr>
<tr>
<td>1P10K3</td>
<td>Laronema?</td>
<td>Sta. Clara, BC, Mexico</td>
<td>95.7; AF210400</td>
</tr>
<tr>
<td>21I2K3</td>
<td>Cyatholaimidae</td>
<td>Sta. Clara, BC, Mexico</td>
<td>97.2; AF210400</td>
</tr>
<tr>
<td>3I23B4</td>
<td>Phanoderma</td>
<td>Laguna beach, CA, USA</td>
<td>90.9; AF210409</td>
</tr>
<tr>
<td>5I23B4</td>
<td>Phanoderma</td>
<td>Laguna beach, CA, USA</td>
<td>91.3; AF210399</td>
</tr>
<tr>
<td>6I23B4</td>
<td>Pontonema</td>
<td>Laguna beach, CA, USA</td>
<td>89.8; AF210409</td>
</tr>
<tr>
<td>1I22G4</td>
<td>Odontopharynx</td>
<td>near Pomponio State Beach, CA, USA</td>
<td>93.7; AB189984</td>
</tr>
<tr>
<td>1M21G4</td>
<td>Chromadorina?</td>
<td>near Pomponio State Beach, CA, USA</td>
<td>87.5; AY300843</td>
</tr>
<tr>
<td>7I16G4</td>
<td>Haliplectus</td>
<td>near Pomponio State Beach, CA, USA</td>
<td>95.9; AY210845, AY364867,</td>
</tr>
<tr>
<td>8M21G4</td>
<td>Choanolaimus</td>
<td>near Pomponio State Beach, CA, USA</td>
<td>98.1; AF210406</td>
</tr>
<tr>
<td>4I22G4</td>
<td>Helicotylenchus</td>
<td>near Pomponio State Beach, CA, USA</td>
<td>92.4; AY592987-93</td>
</tr>
</tbody>
</table>

Reference nematodes (no multifocal imaging)

- **Hirschmanniella Hirschmanniella pomponiensis** near Pomponio State Beach, CA, USA 93.5; U47556
- **Mesocricosoma Mesocricosoma xenoplax** Parlier, Fresno, CA, USA 99.3; AF133304
- **Apor BDGW Aporcelaimellus** Big Sur, CA, USA 98.5; AY061632
- **Brewbucca SB261 Brewbucca saprophaga** SB 261 petri plate culture 95.8; AB189984, AY210806
- **Brewbucca SB117 Brewbucca punctata** SB 117 petri plate culture 96.4; AY210894, AY210806
- **Basiria A GPhi Basiria** UCR campus, CA, USA 87.6; U47557, AF545025, AF435797-801
- **Bunonema Bunonema** UCR campus, CA, USA 97.9; AF151919, AY220628
- **Diphtherophora Diphtherophora** UCR campus, CA, USA 94.6; AY377672
- **Alaimus Alaimus** Botanical Garden, UCR, CA, USA 96.9; AB087186
- **Tyllocephalus PP2 'Tyllocephalus'** Sweeney Granite Mountain, CA, USA/petri plate culture 99.9; AF147068

(Continued.)
but is not applicable on the scale required for routine surveys. Light microscopy is much more limited in attainable magnification and resolution, and furthermore imposes specific constraints on the nature of the visual information it generates.

In this paper, we present the next logical development in the use of multifocal recordings: an integrative approach for morphological vouchering of individual nematodes, as well as other meiofauna, in the context of data-rich nematode surveys combining microscopy with molecular barcoding. Although our strategy does not maximize specimen processivity to the extent of barcoding-only approaches (e.g. Floyd et al. 2002), we argue that barcoding combined with morphological vouchering is as much of an essential component of the exploration of biodiversity in nematodes as it is in other organisms. Such a combination not only maintains a direct link with established taxonomy and ecology, but it also provides a set of reference morphologies and sequences that are essential for interpreting the data obtained from barcoding-only surveys. We strongly believe that such an approach is essential for sustained growth of taxonomy as a holistic discipline employing an ever-widening array of tools and benefiting from an ever-increasing range of data sources.

3. MATERIAL AND METHODS

(a) Sample preparation and image capture

In order to maximize taxonomic diversity and optimize our methods, we sought to include as wide a range of nematode species as possible. Specifically, we focused on marine nematodes as an excellent test bed for establishing our approach, since they include some of the most neglected major nematode taxa, they are uniquely challenging in terms of their potentially enormous diversity (Lambshead 1993), their biogeographical record is largely limited to northwestern Europe, and they are currently extremely underrepresented in sequence databases—especially so when compared to terrestrial or parasitic nematode taxa. To test for applicability outside Nematoda, we also included microscopic individuals from four other invertebrate phyla. Specimens were collected mainly in the context of our ongoing surveys in the Gulf of California (Baja California, Mexico) and various coastal locations in California. Additional species and specimens were obtained from various other localities (table 1; see http://nematol.unh.edu for detailed locality data).

Nematodes were extracted alive from sediment, soil or organic litter samples through decanting and sieving, tray extraction or mist chamber extraction. Live specimens were individually mounted and immobilized in various ways under cover slip on ringed fluorescence slides, a Taylor microcompressor Mk II (Taylor 1993), or ordinary glass slides. If the specimen was not immobilized by the pressure of the cover slip alone, temporary paralysis was obtained through a heat shock of 5–10 s at 60–70°C (depending on the type of sample and size of nematode), or chemical anaesthesia with 7.5% MgCl2 for marine meiofauna. The temporary mounts were placed on a microscope equipped with differential interference contrast optics, and the diagnostically most important body parts were captured at various magnifications as multifocal images. Our respective laboratories avail of three different systems for multifocal imaging. The one used for this study is a more budget-conscious configuration that is manually controlled and writes S-VHS quality video clips directly to hard disk (VCE procedure as described in De Ley & Bert 2002). By comparison, the other two are more sophisticated systems with motorized microscope controls and capture as digital image stacks that are then saved as high definition video clips (cf. Eyualem et al. 2004). Multifocal vouchers of the barcoded specimens included in the analyses below were deposited in NemATOL (http://nematol.unh.edu/) with the specimen ID codes listed in table 1.

<table>
<thead>
<tr>
<th>Other Meiofauna subjected to multifocal imaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>specimen ID</td>
</tr>
<tr>
<td>Tardi oak</td>
</tr>
<tr>
<td>1PolyA3</td>
</tr>
<tr>
<td>Macro 1</td>
</tr>
<tr>
<td>1FlatA3</td>
</tr>
<tr>
<td>1GasIIA3</td>
</tr>
<tr>
<td>Tetra 1</td>
</tr>
</tbody>
</table>

(b) Molecular analysis of captured specimens

Immediately after image capture, each specimen was recovered alive from its temporary mount, cut into 2–4 pieces (depending on body size) in 20 μl Worm Lysis Buffer (50 mM KCl, 10 mM Tris-Cl pH 8.3, 2.5 mM MgCl2, 0.45% NP40, and 0.45% Tween 20 as described in Williams et al. 1992) transferred to 1–5 microcentrifuge tubes, digested with 2 μl (60 μg ml⁻¹ stock) Proteinase K and stored at −80°C. In addition to 37 video captured nematodes, the same methods were applied to six animals from four other phyla. Furthermore, 11 individuals from various other nematode genera and species were prepared for freezing and PCR without video capturing, in the course of protocol testing or various soil surveys. These were included as reference taxa in the analysis in order to cover some of the major nematode orders not encountered in our marine samples.
Series of frozen (fragments of) specimens were subsequently thawed and subjected to PCR amplification of the D2 and D3 expansion segments of the LSU rDNA gene. Three of the 11 reference nematodes were first genome-amplified using GenomiPhi™ DNA amplification kit (Amersham-GE Health Care, Sunnyvale, CA, USA) following the manufacturer's protocol, before performing D2D3 amplification. Depending on the size or type of the organism, genome-amplified products are used as undiluted or diluted (10×–1000×) template for D2D3 amplification. PCR amplifications were performed using 2.5 μl genomic DNA as template in 25 μl reaction volume containing 2.5 μl of 10× reaction buffer with MgCl2, dNTP-mix at 0.2 mM each, 1 μM each of primers D2A (5′-ACAAGTACCGTGGAAGGAAATTTG-3′) and D3B (5′-TCGGAG-GAACAGCCTACTA-3′) D3b and 1 unit of DyNAzyme EXT DNA polymerase (MJ Research, Waltham, MA 02451, USA) with 40 cycles each involving denaturation at 94 °C for 30 s, annealing at 55 °C for 1 min, and extension at 72 °C for 2 minutes; followed by a 7-minute polymerization at 72 °C. Amplified PCR products were separated in 1% agarose (FMC, Rockland, ME, USA), excised and purified with exonuclease I and shrimp alkaline phosphatase (USB, Cleveland, OH, USA) following the manufacturer's protocol, or separated on 1.5% SeaPlaque agarose (FMC, Rockland, ME, USA), excised and purified using Qiaquick gel extraction kits (Qiagen, Los Angeles, CA, USA). Sequencing reactions were performed at the UCR Core Instrumentation Facility using Big Dye Deoxy Terminator Cycle Sequencing kits following the manufacturer's protocol prior to fragment analysis using an ABI 3100 automated sequencer (Applied Biosystems Inc., Foster City, CA, USA); and at the Hubbard Center for Genome Studies using the DTCS Quickstart kit and a CEQ-8000 capillary automated sequencer (Beckman-Coulter, Fullerton, CA, USA) with 40 cycles each involving denaturation at 94 °C for 30 s, annealing at 55 °C for 1 min, and extension at 72 °C for 2 minutes; followed by a 7-minute polymerization at 72 °C. Amplified PCR products were separated in 1% agarose (Shelton Scientific, Inc.) stained with 0.003% ethidium bromide (0.02 μg ml−1) with 1 Kbp DNA ladder (Promega Corp., Madison, WI, USA) as size markers. PCR products were purified with exonuclease I and shrimp alkaline phosphatase (USB, Cleveland, OH, USA) following the manufacturer's protocol, or separated on 1.5% SeaPlaque agarose (FMC, Rockland, ME, USA), excised and purified using Qiaquick gel extraction kits (Qiagen, Los Angeles, CA, USA). Sequencing reactions were performed at the UCR Core Instrumentation Facility using Big Dye Deoxy Terminator Cycle Sequencing kits following the manufacturer's protocol prior to fragment analysis using an ABI 3100 automated sequencer (Applied Biosystems Inc., Foster City, CA, USA); and at the Hubbard Center for Genome Studies using the DTCS Quickstart kit and a CEQ-8000 capillary sequencer (Beckman-Coulter, Fullerton, CA, USA).

Sequences were assembled in GeneTool 2.0 (BioTools Inc., Edmonton, AB, Canada) and compared with published data deposited in GenBank by means of a BLAST search (Altschul et al. 1997), and by phylogenetic analysis of all sequences over 500 bp. As our marine nematode survey was still in its early stages and coverage of the taxonomic spectrum was still very fragmentary, we expected substantial problems with alignment of variable regions in D2 and D3 across highly divergent nematode orders, and especially across different phyla. The purpose of our analyses was therefore neither to establish a definitive assignment of barcodes to species, nor to provide a credible topology of deeper phylogenetic relationships. Rather, we set out to explore methods for providing a first assessment of the performance of D2D3, in the context of phylogeny-based barcoding and cross-verification with multifocal recordings of morphology, but also with an eye on applicability and automation in future analyses of much larger datasets with hundreds of barcode sequences. Additional sequences were obtained from GenBank and included to further broaden taxonomic representation and to benchmark the ability of our analyses to match known species identities across different taxa and taxonomic situations. An alignment was created using the server version of MAFFT 5.6 (Katoh et al. 2005), an alignment algorithm based on fast Fourier transformation that was developed specifically for large datasets and which is claimed to be more accurate than other automated alignment packages. MAFFT was left at the default settings and the iterative refinement chosen was the option ‘E-INS-i’. The result was trimmed at both ends in MacClade 4.0 (Maddison & Maddison 2000) to an alignment width of 1538 characters and then subjected to neighbour-joining and heuristic maximum parsimony analysis (respectively, with 3000 and 1000 bootstrap values) in PAUP* 4.0b10 (Phylogenetic analysis using parsimony and other methods; Swoford 1999).

Variable regions were clearly arbitrarily aligned, due to the often great divergence between sequences. Alignment ambiguities are well known to be a major source of error and uncertainty in phylogenetic analyses, and one approach to avoiding such errors is to remove ambiguous positions. In order to examine the usefulness of ambiguity removal algorithms in the context of barcoding applications, we applied the program Cutter (http://goodey.unh.edu) to our manually edited alignment and generated two alignments of reduced length. The first one will henceforth be referred to as Cutter alignment 1 and was created with column threshold setting and wall threshold setting both at 0.10, resulting in trimming down of the original alignment to 570 positions. The second alignment, hereafter referred to as alignment 2, was created with a column threshold setting of 0.05 and a wall threshold setting of 0.65. This alignment retained 662 positions. Mean differences (corrected for missing data) were calculated between all sequence pairs for each alignment, using PAUP*. In order to estimate possible matches of our barcoded specimens with published D3 sequences of nematode species sequenced by Litvaitis et al. (2000), we also created an alignment of just the D3 expansion segment. This alignment only included our barcoded specimens and the sequences deposited in GenBank by Litvaitis et al. (2000). We again analysed this with neighbour-joining and heuristic maximum parsimony. All unpublished sequences used in the above analyses were deposited in GenBank with accession numbers DQ077749–DQ077803.

4. RESULTS

(a) Quality and practicality of multifocal images

At 720 by 480 pixels, the video clips obtained with our VCE system were clearly lower in resolution than the level of detail that could be observed by eye directly through the microscope. Nevertheless, the clips generally captured all the detail that matters to identification. Significantly, the ability to scroll through successive focal planes greatly facilitates interpretation of the shape, location and orientation of structures, especially when compared to single still images. For example, sensory organs like amphids and setae are some of the most important diagnostic structures that are difficult to adequately capture in a single photograph, particularly if a specimen is not oriented at a perfect angle to fit the structure in question in a single optical section (see figures 1 and 2 for examples). By providing continuity across adjacent focal planes, we find that multifocal clips typically make it much easier to recognize such structures—in a way that cannot be adequately reproduced or demonstrated on a printed page.

In terms of storage, the file size of one uncompressed video clip at S-VHS resolution is usually around 4 to 20 megabytes, an order of magnitude roughly comparable to that of a single uncompressed still image captured at resolutions around 2 million pixels. This is well within the handling capacity of current desktop and laptop computers. If necessary it is relatively easy to optimize file sizes with further...
manipulations such as batch compression. Our multifocal imaging procedures as described above (with manual focusing and analogue video camera) typically added about 15–30 min to the processing time of each specimen prior to DNA extraction, thus allowing one person to capture the morphology and prepare frozen lysates of between 10 and 20 nematodes during an average working day.

(b) First results with barcoding of nematodes from the Gulf of California

Multifocal capturing followed directly by PCR and direct sequencing yielded good D2D3 sequences for over 80% of all specimens. Genomic amplification was successful and the resulting DNA amplicons easily allowed PCR and direct sequencing of the D2D3 locus. By comparison, PCR and/or sequencing of other loci were less consistently successful and clearly need more optimizing (data not shown). Of the 43 barcode sequences obtained (including six from other phyla), only one matched GenBank data with a BLAST percent identity score just above 99% (table 1). Trees produced from all alignments suggested six pairs or groups of putative matches within our set of nematode barcodes (marked by horizontal blue bars in figure 3). A number of other interesting sequence pairs also occurred among the sequences downloaded from GenBank or determined without prior video capturing (marked by grey bars in figure 3). We will discuss these in light of strength and weaknesses of our approach as confronted with the complications of actual identification. Four sets of sequences were scored with 0% mean divergence in all three alignments. Two of these consisted of pairs of identical sequences known to correspond to either one species (Acrobeloides bodenheimeri AF147064-5) or one single individual (2I6K2 Oncholaimidae was sequenced twice for quality control purposes). A third identical pair included Acrobeloides thornei and a cultured nematode originally presumed to belong in the genus Tylocephalus. Apparently A. thornei was accidentally carried over into the latter culture (a mistake that could easily have been avoided had we first video imaged the nematode in question). All three identical pairs show that at least the MAFFT alignment correctly matched up identical sequences.

The fourth set with 0% mean differences included not two but three sequences: two of these came from individuals identified as Odontophora (figure 2) while the third nematode was identified as Pomponema (figure 1). These two genera are morphologically quite different from one another, traditionally being classified in the orders Monhysterida and Chromadorida respectively. The obtained multifocal images strongly suggest that in this case the chosen locus and/or species-matching algorithm did not perform optimally. The sequences themselves were not identical but the only differences consisted of a few unresolved nucleotides.

Figure 1. Five frames (images) extracted from a multifocal video clip of the lip region of barcoded nematode specimen 2P12K2, identified as the genus Pomponema. Note the various setae revealed in different focal planes, as well as the different aspect of the faint but tightly coiled amphid in successive frames (a–c) and the presence of a subventral tooth (e). The placement and shape of the amphids and most of the setae would be impossible to record in a single photograph in this specimen. The position of the tooth would be impossible to ascertain without determining the angle of view, which information must be gleaned from other focal planes (e.g. from the relative position of the amphid, which is located laterally).
Among the video captured nematodes, two sequence pairs appear to constitute bona fide matches: *Pontonema* specimens 3I24B4 and 6I23B4 derived from the same sample and had a mean divergence of 0.2% in all three alignments, while *Latronema* 1P10K3 and 2P15K2 came from different sites but had a mean divergence of 0.6% or less. Two other sequence pairs were more problematical, each clustering together as

Figure 2. Four successive frames from a multifocal video clip of the lip region of barcoded specimens 2I11K2 and 5P9K2, both identified as belonging in the genus *Odontophora*. Note the extremely long setae pressing against the coverslip (a, e) as well as the horseshoe-shaped amphid (b, f) and the radially arranged teeth (c, d, g, h). The placement of the setae, shape of the amphid and arrangement of the teeth is not recordable with a single (or even a few) still photograph(s). Both these specimens cluster together with *Pomponema* 2P12K2 (shown in figure 1) in phylogenetic analyses (see figure 3) but they are clearly not conspecific, illustrating the ability of multifocal images to pinpoint potential errors in barcode-based species delineations.

Among the video captured nematodes, two sequence pairs appear to constitute bona fide matches: *Pontonema* specimens 3I24B4 and 6I23B4 derived from the same sample and had a mean divergence of 0.2% in all three alignments, while *Latronema* 1P10K3 and 2P15K2 came from different sites but had a mean divergence of 0.6% or less. Two other sequence pairs were more problematical, each clustering together as
sister taxa with maximal bootstrap values in the trees (data not shown) but nevertheless differing by higher mean percentages. *Richtersia* 5P12K2 was an adult male, while 4P11K2 was a juvenile and could not be morphologically matched with certainty to the same species or even genus; their sequences differed by 2.8–5.3% and it seems possible that these could actually represent two different species. *Phanoderma* 3I23B4 and 5I23B4 matched more clearly to the same species morphologically, for example in sharing presence of eyespots, but nevertheless had a mean difference in D2D3 sequence of 1.0–4.7% in our three alignments.

Among the nematode sequences that did not derive from video captured specimens, a number of other groupings are noteworthy with respect to species identity versus mean sequence differences. The included species in the entomopathogenic genus *Steinernema* and the plant parasitic genera *Globodera* and *Meloidogyne* are morphologically confounding among their respective congeners, but nevertheless considered valid species based on life history characters among others. The respective mean difference values between D2D3 sequences are small to very small (especially for *Meloidogyne*), especially when compared to sequences from putative conspecific individuals such as our two *Latronema* individuals, or two sequences of *Discolaimus major* as downloaded from GenBank.

Clearly, D2D3 mean differences do not correlate consistently with conspecificity across the phylum, and external data (such as morphological vouchers) are needed to corroborate putative matches. Overall, we cautiously estimate that our 37 video captured and barcoded nematode specimens represent 32–34 different species, none of which matches any previously sequenced species. Trees derived from the D3-only alignment suggested not a single match with any of the sequences of Litvaitis *et al.* (2000), and placed AF210426 *Tricoma* among Ceramonematidae, well away from our 3P15K2 *Tricoma* (data not shown), the former sequence presumably deriving from a misidentified specimen.

(c) Phylogenetic signal or noise

Phylogenetic reliability of the obtained trees was clearly not very high. The untrimmed Mafft alignment as well as Cutter alignment 1 yielded trees in which non-nematode sequences were inserted in at least three separate places in between nematode taxa. Only alignment 2 yielded a relatively coherent grouping of the other phyla (figure 3). Various other dubious placements occurred, e.g. with respect to separation of *Enoplolaimus* 2P6K2 from other Enoplidae, placement of Oncholaimidae (including *Viscosia* and *Pontonema*) in two separate clades, as was also the case for 3P12K2 (unidentified member of Monhysterida) versus 3P11K2 *Xyala*. Heuristic maximum parsimony analysis yielded respectively 61, 57 and 84 best trees for the untrimmed alignment, Cutter alignment 1 and Cutter alignment 2. In view of all this evidence for poor resolution of deeper nodes, which is undoubtedly connected with the fragmentary taxonomic representation in our data, we did not at this stage attempt species assignment by autapomorphy analysis as proposed by Adams (1998).

5. DISCUSSION

(a) Are multifocal images useful for barcoding studies?

Nematodes and many other microscopic Metazoa are optically unusual as three-dimensional objects in that they are both minute and largely transparent. As a result, nearly all internal features are visible without requiring dissections or special staining techniques—but these features are also transparent in their own right and must be observed at the highest magnifications in optical conditions where focal depth is reduced to a single optical section. In order to understand any particular anatomical structure, one must therefore both detect subtle differences in transparency as well as examine multiple two-dimensional focal planes, before one can mentally visualize the three-dimensional shape of the feature in question. This visualization problem also applies to any images that are created to record and represent nematode morphology: unlike for example a photograph of an insect wing, no single photograph of a nematode body part can accurately embed and reflect its three-dimensional contents and context. On the other hand, line drawings can represent structures at multiple levels of focus, but require substantial personal skill and are incapable of accurately representing subtle contrasts or shades of transparency. Video and digital camera technology have opened up new possibilities for morphological vouchering through analogue or DMI. Many cameras used in scientific applications are optimized for applications that require the highest resolution and/or the highest sensitivity in low light conditions. The question therefore arises whether such cameras are also suitable for capturing the subtle nuances of transparency that reveal the outline and internal organization of nematode structures. Based on our material, we feel this is definitely the case—even for relatively inexpensive video cameras of limited resolution. PCR can be optimized for small fragments and/or for formalin-fixed material (Thomas *et al.* 1997; Dorris *et al.* 2002), but high-throughput DNA extraction of large numbers of taxonomically diverse individual animals usually requires the destruction of entire specimens, thereby obliterating all morphological characters. Molecular barcoding without morphological vouchering results in the loss of all other biological data, creating at best a first record that will not be available for further study until future rediscovery of specimens in nature. Morphological vouchering through multifocal imaging cuts through this Gordian knot, and is actually easier to implement with specimens that are small enough to require complete destruction during DNA extraction. Furthermore, it reverses the traditional problem of time-consuming species identifications or descriptions in taxonomically challenging phyla such as nematodes. Identifications can now be deferred until molecular barcodes are obtained, and identification effort can be prioritized based on both the outcome of barcoding analysis and
any apparent morphological novelty, or particular significance to other studies. Also, vouchering through multifocal imaging provides a direct link to existing keys, classifications and hypotheses based on previously documented morphologies and ecologies. The virtual vouchers are not likely to suffice in and of themselves for descriptions of new species, but they do provide an effective way of linking molecular data from destroyed specimens with unsequenced (and generally unsequenceable) type material in permanent slides, through comparison of their respective morphologies.

(b) Do barcoding studies need named species?
In a number of cases, the advantages listed above will clearly outweigh the benefits of maximal processing speed that could be attained with a voucher-less MOTU approach. Although maximal processing speed is excellent for applying mathematical biodiversity analyses, the sacrificing of morphology disallows comparisons with species-based alpha diversity surveys. Combined acquisition of molecular and morphological data greatly enhances the strengths of each while bypassing many of their respective weaknesses. For example, juveniles can be detected through microscopy but are rarely accurately identifiable, since most morphological keys are based only on adults (cf. our two Richtersia specimens). Conversely, barcoding works with both adults and juveniles, but does not reveal developmental stage unless morphological data are collected as well.

Floyd et al. (2002) have noted that biology of sequenced but voucher-less MOTUs can be inferred from proximity to known species. However, this process can easily lead to a significant fraction of erroneous conclusions in poorly sampled phyla, while phylogenetic analyses are themselves volatile and typically generate multiple topologies with at least some unresolved areas. Furthermore, in order to be able to make any inferences on biology, MOTU approaches must necessarily rely on a reasonable representative set of barcodes (the ‘stepping-stone sequences’ of Blaxter 2004) from species identified with independent criteria. Multifocal imaging combined with barcoding of specimens provides a structured approach to establishing such a set of reference points. This selected set must necessarily grow at a slower pace than MOTU studies that maximize on high throughput of specimens. Nevertheless, a combined strategy such as the one outlined in this paper will dramatically expand the coverage and accuracy of biological inferences by phylogenetic proxy, even in habitats where the large majority of species are unknown and unlikely ever to be formally named and described.

Similar to the spectacular diversity currently being discovered in prokaryotes (see e.g. Torsvik et al. 2002) and protists (Countway et al. 2005), MOTU surveys of nematodes can play a very useful role in generating first-line data on ‘operational’ biodiversity in different sites or habitats, thereby providing an objective basis for e.g. selecting samples that need to be prioritized for identification and description of species. However, the fundamental problem of nematode surveys is the identification of species, and this problem clearly will not be tackled by sidestepping the theoretical and diagnostic challenges of species concepts in nematodes. Instead, a genuine solution to the nematode corner of the taxonomic impediment will have to squarely face the challenges of diagnosing species, as opposed to diagnosing operational units. In a recent review of the operational aspects of species diagnosis, Sites & Marshall (2004) noted that no method would invariably succeed in delimiting species boundaries properly. They explicitly caution against reliance on any single data set or method. Similarly, in a barcoding study of cave-dwelling eyeless spiders, Paquin & Hedin (2004) demonstrated the need for external information to allow for the discovery of incongruence between different data types, in order to make sense of otherwise one-dimensional molecular taxonomies.

(c) Do barcoding studies need phylogenies?
Phylogenetic analysis has become essential to the delineation of species in nematodes. The practical application of species concepts used to be a major stumbling block to nematode taxonomy, and it was not unusual for leading experts to declare the problem to be basically insurmountable (e.g. Heyns 1983; see review in Coomans 2002). This conceptual paralysis has changed substantially in recent years, with the proposal of a tree- and autopomorphy-based species concept (Adams 1998) that is rapidly gaining ground in nematode taxonomy (see e.g. De Ley et al. 1999; Stanton et al. 2001, Coomans 2002; Nadler 2002; Troell et al. 2003). This has in many ways opened the doors for more objective approaches to nematode surveys. Even though occasional failures to properly diagnose species are to be expected (Sites & Marshall 2004), the current state of the art in nematode taxonomy explicitly requires species allocations to incorporate phylogenetic criteria. However, as shown by our data the application of autopomorphy analysis in barcoding surveys may not be all that easy, at least until a more comprehensive set of sequences is available for constructing more robust alignments.

The use of short DNA sequences as identification tags of species is a unifying concept with multiple uses. Existing or projected applications include: development of portable identification technology for walkabout surveys in the field; conservation inventory management of listed species (which are by definition known species); comparative biodiversity analyses of alpha and beta diversity; supporting taxonomic discovery by generating a first baseline for quantifying and selecting novel taxa for description. It is important to note that these different applications do not all present the exact same requirements in terms of optimal choice of barcoding locus. For example, the former two would benefit from a single uniform standard barcoding locus across all organisms, which may well require compromises in terms of attainable resolution and representation for some particular groups. In contrast, the latter two applications typically involve more methods development along the way, to determine which loci (if any) are truly optimal for species discrimination among the particular organisms studied.

Barcoding of taxa with a high ratio of unknown to known species is unlikely to be worth the effort if no

Phil. Trans. R. Soc. B (2005)
Figure 3. Neighbour-joining phylogram based on alignment 2 of the D2D3 expansion segment of barcoded and reference organisms (aligned with MAFFT 5.6 using setting E-INS-i followed by ambiguity trimming with CUTTER using column setting $Z = 5$, wall setting $Z = 65$). Taxon names are coloured as follows: red for downloaded sequences of non-nematode taxa, purple for newly video vouched and sequenced non-nematode taxa, blue for newly video vouched and sequenced nematode taxa, black for nematode taxa of which sequences were downloaded or newly determined without video capturing. Deeper relationships are not necessarily meaningfully resolved as the D2D3 locus is too short and too divergent for that purpose. Interesting benchmark sequences as discussed in the text are highlighted with horizontal bars (blue for video vouchered nematodes, grey for others). Percentage mean difference values for the relevant sequence pairs are listed on the right for the original MAFFT alignment without any removal of ambiguities, for alignment 1 (MAFFT followed by CUTTER using column setting $Z = 10$, wall setting $Z = 10$) and for alignment 2.
space of undiscovered taxa (De Ley 2000). In the absence of any phylogenetic context, the information conveyed by the barcode alone is purely archival and will only be relevant to future studies if the exact same taxon happens to be encountered again. With a phylogenetic context and a virtual voucher, morphology and DNA sequences can be reciprocally verified, reducing errors and maximizing the chances of discovery of biologically interesting patterns of diversity. Even more importantly, a barcode that carries meaningful phylogenetic signal will ensure relevance to other surveys in the present and in future, because related taxa can be recognized even if they are not identical, while morphological vouchers can establish links with past studies through comparison with type specimens.

(d) Choices of barcoding loci for nematodes and other meiofauna

Our experience with nematodes alone already suggests that no single barcoding locus will perform optimally across all forms of life, and it is likely that different loci should be used for barcoding different types of organisms. For example, using the same primer pair for macroscopic as for microscopic organisms is likely to be counter-effective because it could well lead to mixed PCR products due to the ubiquity of symbionts. Nuclear genes are likely to be more diagnostically useful than mitochondrial loci because they are not subject to lineage sorting. Compared to rDNA, protein genes offer an additional layer of information with respect to different evolutionary rates according to codon positions, and translation into amino acids, but they also present potential problems due to occurrence and variability of introns. Furthermore, the choices for barcoding of nematodes and other very poorly covered taxa are not only constrained by methodological considerations, but also by the availability of research interest and funding. Nematodes are not likely to show up on any lists of endangered species. They are therefore also unlikely to receive sufficiently focused and sustained support to allow an expansive database to be compiled for the specific use in barcoding of any single locus. Thus, the most appropriate loci from the point of view of funding priorities will also be those that can serve multiple tasks, e.g. by being useful in phylogenetics as well as in diagnostics. In the longer term, however, the situation will undoubtedly change quite drastically as sequencing technology improves. The barcoding gene may rapidly become superseded by the advent of ‘genomic barcodes’ and ‘Genomic Operational Taxonomic Units’ constructed by instantaneous sequencing of numerous species-diagnostic markers sampled across entire genomes. For the time being, the following loci appear to be most relevant to barcoding efforts in nematodes and other microscopic invertebrates:

(i) small subunit (SSU) rDNA has received the greatest attention as a barcoding locus in recent literature (Floyd et al. 2002; Blaxter 2004; Powers 2004). SSU targets usually have a high success rate with PCR, but may require primer optimization for different nematode taxa, and for use in other phyla. The gene is not known to be subject to significant amounts of polymorphism; in nematodes it has a wide range of diagnostic resolution but it tends to work better for separating species in some groups than in others. The locus has a high phylogenetic information content and often works well for resolving relationships, especially at family and order levels (e.g. Blaxter et al. 1998; Félix et al. 2000; Rusin et al. 2003). The public record of partial and complete SSU sequences is the single most abundant component of all known nematode sequences, and it is likely to remain the best sampled gene in nematodes, due to its widespread application in both phylogenetic studies and molecular surveys.

(ii) The LSU rDNA gene has been in use for almost ten years as a source of diagnostic sequences in nematodes, particularly the region that spans the D2 and D3 expansion segments (Thomas et al. 1997). In nematodes it covers about 600–1000 bp, fairly close to the 5′ end of the gene. As with the other expansion segments of LSU, sequence divergence of D2 and D3 between related species is often high. By contrast, the conserved regions alternating with D2 and D3 are highly constant, even across phyla, and provide very robust primer sites. In our experience, the D2D3 primer pair has the highest success rate when applying PCR amplification to a phylum-wide selection of nematodes, and based on our limited testing it also works well in other phyla of microscopic metazoans. The locus is not known to be subject to significant amounts of intraspecific polymorphism, and provides very good separation of cryptic species in some groups (De Ley et al. 1999). Previous studies have included phylogenetic applications of the D2 or D3 alone (cf. Litvaitis et al. 2000). The combination of both expansion segments generally provides more robust signal, at a level corresponding roughly to species, genus and family in local taxonomic context, although it is probably overall too variable for meaningful analysis at deeper levels in classification and phylogeny. Our data presented here also indicate that mean differences in D2D3 do not always correspond reliably with accepted species boundaries. Published sequences of partial or complete LSU are more limited in number than SSU, but this part of the overall nematode dataset is likely to expand considerably due to increasing interest in LSU sequence data for phylogenetic studies.

(iii) The internal transcribed spacer region (ITS) initially received great attention as a potentially diagnostic tool in nematodes (Powers et al. 1997). It consists of two variable regions (the spacers) alternating with the much more conserved SSU, 5.8S and LSU rDNA genes, in an arrangement rather similar to the alternation of conserved and
variable sequence stretches of the D2/D3 region. However, in our experience no single primer pair seems capable of allowing robust PCR success across all major nematode groups. To make matters worse, several recent studies have revealed occurrence of ITS haplotypes, sometimes with such levels of polymorphism that species boundaries are confounded and direct sequencing becomes impossible (Hugall et al. 1999). As a result, in some nematode groups the ITS region may prove to be a much more interesting marker for population genetics than for species diagnostics. Nematode ITS data in public databases are less numerous than SSU sequences but more so than LSU data. The sequence pool for this region will undoubtedly continue to grow, due to its frequent application in phylogenetic studies of selected taxa (e.g. Chilton et al. 2001; Subbotin et al. 2001), in studies of potential hybridizations (Hugall et al. 1999) or in analyses of intraspecific diversity (Kaplan et al. 2000; Elbadri et al. 2002; Ye et al. 2004).

(iv) No phylum-wide primers are available for the mitochondrial cytochrome oxidase I gene in nematodes, in our experience PCR success rates are well below 50% across various taxa within the phylum. Reasons for these problems may relate to the emerging evidence that nematode mitochondrial genomes are highly diverse, displaying unusual properties such as recombination (Lunt & Hyman 1997), insertional editing (Vanfleteren & Vierstraete 1999) and multipartitioning (Armstrong et al. 2000). Thus, designing nematode primers for mitochondrial protein genes may present many complications if these are to be useful for barcoding across the phylum. At present, little is known about the applicability of nematode coxI in phylogenetic analyses and very little is available in public databases in terms of sequences.

(e) Conclusions and wider perspectives
Multifocal imaging opens up possibilities for vouchersing that can bring molecular surveys of nematodes closer to the protocols of barcoding studies of larger organisms. However, the flavour of barcoding surveys of nematodes is likely to remain quite different from those focusing on e.g. insects, because it is also determined by the high preponderance of undescribed species, by greater debate about the relevance of morphology, and by differences in theoretical developments with respect to species concepts. We are still in the early stages of exploring the biodiversity of meiofaunal, interstitial and other microscopic Metazoa. Although we can make certain predictions about the most appropriate search strategy for charting this vast taxonomic wilderness, it is unlikely that such predictions will be very accurate on the basis of known species, which might easily constitute less than 1% of all existing microscopic Metazoa. As with geographic exploration of uncharted continents and planets, it therefore makes sense to try multiple strategies of exploration and to maximize data exchange between these different approaches. We consider a data-rich, somewhat less processive approach to be an essential component of this overall process, documenting morphological and molecular diversity in formats that are easily archived and distributed electronically. We are therefore also concentrating our efforts on setting up the necessary infrastructure to facilitate data exchange and to foster new collaborations and linkages. Communication and integration between different surveys will be mediated by websites such as NemATOL, and these will in turn integrate through higher-level web resources. The probable outcome will be an emergent network of Internet resources dedicated to the discovery and documentation of biodiversity.

This research was supported by a UCMEXUS/CONACYT joint project award and by the National Science Foundation through award no. DEB-0315829. We are indebted to Matt Bemus for development of CUTTER. We also wish to thank Nicola Debenham, John Lambshed and Ian King for help with provisional genus identifications, and are grateful to Ian and the anonymous reviewers for many constructive suggestions during the writing of this paper.

REFERENCES


Critical factors for assembling a high volume of DNA barcodes

Mehrdad Hajibabaei, Jeremy R deWard, Natalia V Ivanova, Sujeevan Ratnasingham, Robert T Dooh, Stephanie L Kirk, Paula M Mackie and Paul D.N Hebert

Phil. Trans. R. Soc. B 2005 360, 1959-1967
**Critical factors for assembling a high volume of DNA barcodes**

Mehrdad Hajibabaei*, Jeremy R. deWaard, Natalia V. Ivanova, Sujeevan Ratnasingham, Robert T. Dooh, Stephanie L. Kirk, Paula M. Mackie and Paul D. N. Hebert

*Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, ON, Canada N1G 2W1

Large-scale DNA barcoding projects are now moving toward activation while the creation of a comprehensive barcode library for eukaryotes will ultimately require the acquisition of some 100 million barcodes. To satisfy this need, analytical facilities must adopt protocols that can support the rapid, cost-effective assembly of barcodes. In this paper we discuss the prospects for establishing high volume DNA barcoding facilities by evaluating key steps in the analytical chain from specimens to barcodes. Alliances with members of the taxonomic community represent the most effective strategy for provisioning the analytical chain with specimens. The optimal protocols for DNA extraction and subsequent PCR amplification of the barcode region depend strongly on their condition, but production targets of 100K barcode records per year are now feasible for facilities working with compliant specimens. The analysis of museum collections is currently challenging, but PCR cocktails that combine polymerases with repair enzyme(s) promise future success. Barcode analysis is already a cost-effective option for species identification in some situations and this will increasingly be the case as reference libraries are assembled and analytical protocols are simplified.

Keywords: museum specimens; DNA isolation; PCR; species identification; taxonomy; cox1

1. ASSEMBLING DNA BARCODES: THE CHALLENGE

DNA barcoding promises fast, accurate species identifications by focusing analysis on a short standardized segment of the genome (Hebert et al. 2003). Several studies have now established that sequence diversity in a ~650 bp region near the 5' end of the mitochondrial cytochrome oxidase subunit I (cox1; also referred to as COI) gene provides strong species-level resolution for varied animal groups including birds (Hebert et al. 2004b), fishes (Ward et al. 2005), springtails (Hogg & Hebert 2005), spiders (Barrett & Hebert 2005) and moths (Hebert et al. 2003; Janzen et al. 2005). These early results have provoked larger-scale barcoding efforts and global projects for fishes and birds have now been initiated (Marshall 2005). These projects represent the first wave in a series of initiatives which will demand the capability to assemble barcodes rapidly and cost-effectively. As one looks further to the future, the need for substantial analytical capacity looms. For example, an effort to barcode the 1.7 million described species (Hawkesworth 1993) would require the assembly of some 20 million barcodes, given a target of about 10 barcodes per species. This total will rise fivefold if barcode coverage is desired for all 10 million eukaryote species (e.g. Hammond 1992), producing a sequence library of 65 billion base pairs, approximately twice the current size of GenBank (April 2005). This task could be completed within a decade by establishing 50 core laboratories, each producing 200 000 barcode records per year. When viewed from the perspective of major genomic facilities, some of which generate more than 50 million sequences a year, the production goals for barcode facilities may seem modest. However the business of generating barcodes is complex; each record represents a sequence derived from a specimen that had to be collected, archived and databased.

In the balance of this paper, we direct most of our attention to an evaluation of how the primary steps in the analytical chain extending from specimens to barcode records can be optimized, scaled up and economized. Because the single most critical step to achieve high production involves a move from protocols based on single specimens to those compatible with 96-well format, we only consider methods compliant with this approach.

2. SPECIMENS

(a) Sourcing specimens

Specimens are the raw material for any barcode facility. This need can be met most easily by sequencing all specimens encountered, but because of the lognormal distribution of species abundance (May 1975), most of the resultant sequences will derive from a few common species. Collaborations with taxonomists represent a far more effective strategy for provisioning the analytical chain with specimens (Janzen et al. 2005; Smith et al. 2005; Ward et al. 2005). With this approach, it is...
feasible to assemble a library of sequences that provides both broad species coverage and similar sampling intensity across species (e.g. 10 barcodes each). Moreover, sample sizes can be increased in cases where complexities, such as cryptic species, are encountered in the first pass (Hebert et al. 2004a; Janzen et al. 2005).

We have adopted the TrakMates micro-plate system (Matrix Technologies, Hudson, NH, USA) to force the organization of specimen shipments into the blocks of 96 needed for the later stages of analysis. One micro-plate holds 96 vials (94 specimens, two controls), each uniquely barcoded on the bottom of the vial. These barcoded vials can be rapidly scanned, aiding the tracking of specimens as they enter the analytical chain. Aside from an organized flow of specimens to the barcode facility, there is a critical need for the firm connection of specimens to their collaterals. To facilitate this, we have developed a spreadsheet that organizes key specimen information. We have, as well, developed web-based software to both organize the specimen information and to connect each barcode sequence with its source specimen (see below).

Preservation/handling
Whenever possible, animal specimens should be killed and preserved in a DNA-friendly fashion (freezing, cyanide and ethanol). Even brief exposure to agents that damage DNA, such as ethyl acetate or formaldehyde, should be avoided (Prendini et al. 2002). While fresh or freshly frozen tissues are ideal for analysis, DNA in dried specimens ordinarily remains easily analysed for 5 years, although degradation rises as time passes. Specimens preserved in absolute ethanol are easily analysed when young, but acidification soon degrades their DNA unless it is regularly replaced or buffered. As a general principle, barcode analysis should follow collection as soon as possible, but delays of a few months will cause little problem.

To minimize external or cross-contamination, all tissue samples should be handled on a clean working surface and all instruments should be acid or flame sterilized before handling a new specimen. When using 96-well plates for tissue assembly, particular care must be taken when adding samples to avoid cross-contamination between wells.

The importance of archival specimens
Natural history museums and herbaria maintain most of the world’s known biodiversity within their collections. In some groups, species coverage may be nearly complete. For example, museums hold nearly 10 million bird specimens (Roselaar 2003), assuring deep coverage for most of the 10 000 known species. The analysis of museum specimens could enable rapid growth in barcode coverage (Janzen et al. 2005).

Unfortunately, they are generally poor targets for analysis because of DNA degradation due to hydrolysis and oxidation (Lindahl 1993), exposure to ultraviolet light (Eglinton & Logan 1991) and preservation agents such as formaldehyde (Schander & Halanych 2003). Methods used to retrieve DNA from museum specimens typically aim to isolate DNA with high efficiency (Junqueira et al. 2002). Because many copies of the mitochondrial genome are present in each cell, its component genes, such as cox1, represent optimal targets for analysis in archival specimens. However, because the template DNA is degraded, few amplicons longer than 300–400 bp can be obtained from specimens more than a decade old (Su et al. 1999; Junqueira et al. 2002; Rohland et al. 2004). When degradation is particularly severe, one common strategy involves the amplification of less than 100 bp DNA fragments (Goldstein & Desalle 2003). In such cases, obtaining a DNA barcode will require the concatenation of several short sequences (i.e. Su et al. 1999).

3. DNA isolation
(a) Different strategies
Methods for DNA isolation fall into two broad categories: DNA release and DNA extraction. DNA release protocols aim to rapidly release DNA into solution, making it accessible for downstream applications such as PCR. Release-based methods also enable DNA isolation from samples without their physical disruption. In this case, the entire specimen can be removed after DNA isolation, allowing the retention of a voucher in cases where this would otherwise be possible. Release methods are, however, not very sensitive and do not produce high purity DNA suitable for long-term storage (e.g. more than 1 year). By contrast, DNA extraction methods aim to purify DNA, often by binding it to a membrane (e.g. silica) or by chemical fractionation. Some classical methods, such as phenol/chloroform extractions (Sambrook et al. 1989), are not attractive because they are time consuming and involve toxic materials. The type and condition of specimens is a key factor in selecting a DNA isolation method. For fresh or recently collected tissue, a release-based DNA extraction usually provides sufficient DNA for barcoding. However, for archival material, more sensitive approaches should be used. Because little DNA is needed for barcode analysis, the amount of tissue used in DNA isolation is usually minute. Figure 1 shows four typical tissue samples for barcode analysis.

(b) Comparing DNA isolation techniques
In order to determine an optimal procedure for high volume barcoding, we compared five DNA isolation methods on four sets of specimens (birds, fish, recent and archival moth—see Electronic Appendix part 1A for details). The major criterion for the inclusion of methods in this performance test was their capacity for high-throughput analysis, but we also considered cost and sensitivity. These methods included an artisanal (=homemade) DNA release method, called DryRelease, which employs Chelex resin as a DNA release agent (Walsh et al. 1991). We also examined three DNA extraction methods that use silica to bind DNA: Silitom, an artisanal method based on the protocols of Elphinstone et al. (2003) and Boom et al. (1990), NucleoSpin96 tissue kit (Machery-Nagel, Düren, Germany) and DNeasy96 tissue kit (Qiagen, Hilden, Germany). Finally, we tested a DNA extraction method that uses magnetic beads to bind DNA: ChargeSwitch Forensic kit (Invitrogen, Carlsbad,
The effectiveness of these five DNA isolation methods was compared by testing their success in PCR amplification (using visual inspection on an agarose gel; see Electronic Appendix part 1F for details) of the full-length (~650 bp) \( cox1 \) barcode with primer sets specific for each taxonomic group (see Electronic Appendix part 1E for details) (figure 2; table 1). Overall, the NucleoSpin96 kit was most effective, producing more than 75% success for three of the groups of specimens, and 31% for the most difficult group (archival moths). However, it was not always the best: the Silitom and ChargeSwitch methods produced higher success for bird samples. Interestingly, the DNeasy96 kit was less effective than the NucleoSpin96 kit, despite their very similar methodologies. This difference was particularly striking for fishes where the NucleoSpin96 kit delivered three times as many successful amplifications. The ChargeSwitch method produced the most variable results with 90% PCR successful amplifications. The ChargeSwitch method, where the DNA is attached to a silica gel membrane inside a column offer more security. All PCR reactions were sequenced to ascertain their performance in delivering both a clean \( cox1 \) sequence and one that derived from the presumptive source specimen (Electronic Appendix part 1G provides sequencing protocol). In most cases, a small percentage of the visible PCR products failed to generate a clean sequence, but the differences between extraction methods were small (figure 2). The only exception involved the ChargeSwitch method for fishes where the number of sequences obtained was higher than the number of visible PCR products (figure 2).

The strong performance of silica-based approaches makes them appropriate for high-throughput barcoding, especially when work is focused on the analysis of small tissue samples. Substantial cost savings (80%) can be realized by the use of an artisanal protocol such as Silitom rather than commercial kits. Moreover, when samples are young or well-preserved a simple release-based artisanal protocol such as DryRelease could represent the optimal approach in many cases (table 1). We note that the amount of tissue used in the DryRelease protocol, where potential PCR inhibitors in the tissue are not removed, can greatly influence the results. As a consequence, success rates can often be substantially increased by optimizing the amount of tissue at the initiation of a study and we have done this for many large-scale projects.

None of the negative controls (six in each set of 96 samples for a total of 24 per method) produced an amplification product, except the ChargeSwitch analysis on birds, where four of six negative controls showed an amplicon (figure 2). However, we did observe more cases where the PCR product did not derive from the correct specimen. The ChargeSwitch kit showed the highest number of such contaminated sequences (five and three in recent moths and birds, respectively), but four sequence contaminations were observed in Silitom isolations (two in both recent moths and birds) and two in the DNeasy96 amplifications, both from birds. These results suggest that some protocols are more sensitive to contamination than others, and that the risk of its occurrence is greatest when one is working with tissues that yield relatively large amounts of DNA. Our results further suggest that the ChargeSwitch method, where the DNA is attached to magnetic beads in liquid phase, is particularly sensitive to handling errors leading to contamination, especially when the protocol is performed in 96-well format. By contrast, the two kits (NucleoSpin96, DNeasy96) where the DNA is bound to a silica gel membrane inside a column offer more security.

(c) Which DNA isolation method is best for archival specimens?

In our studies, the NucleoSpin96 kit produced the highest amplification success for the full-length barcode region of \( cox1 \) in archival moths (31%) followed by Silitom (20%), DNeasy96 kit (18%), DryRelease (8%) and ChargeSwitch kit (1%; figure 2). These results make it clear that silica-based methods should be used for DNA isolation from archival specimens.
primer sequences can have a large impact on barcode recovery. Primer reconfiguration begins by aligning all available sequences for the target taxonomic group. Subsequent adjustments in sequence to maximize matches have enabled the development of effective primer sets (more than 95% amplification across species) for large taxonomic assemblages, such as Lepidoptera (Janzen et al. 2005), birds (Hebert et al. 2004b) and fish (Ward et al. 2005). In most cases, effectively complete barcode recovery for all species in a group can be achieved with two sets of non-degenerate primers. Primers with degenerate positions or modified bases such as inosine (which can form base pairs with all four nucleotides) can help with recalcitrant groups where variable nucleotide positions across taxa compromise amplification (Candrian et al. 1991). Using primers with degenerate positions may also reduce the chance of preferential amplification of nuclear pseudogenes (Sorenson et al. 1999). Many software packages are available to aid primer design, but we recommend PRIMER3 (Rozen & Skaletsky 2000) for designing non-degenerate primers and CODEHOP (Rose et al. 2003) for degenerate primers.

(b) PCR optimization
An optimized PCR for the barcode region of cox1 should yield a single sharp amplicon, with no more than minor sub-banding when examined on an agarose gel. This can often be achieved by optimizing cycling conditions, especially the annealing temperature, and by altering the concentration of PCR reagents such as magnesium, dNTPs and primers through pilot studies on a few taxonomically divergent members of the target assemblage. Optimization often also dramatically increases amplification success and can eliminate the need for PCR cleanup prior to the sequencing reaction. PCR amplification can also be enhanced with additives such as bovine serum albumin, betaine, DMSO (Abu Al-Soud & Radstrom 2000) and trehalose. Trehalose is especially useful because it acts as a potent PCR enhancer by both lowering the DNA melting temperature and stabilizing Taq polymerase (Spiess et al. 2004). Trehalose can also overcome the effect of PCR inhibitors that are often present in crude DNA extracts (e.g. DNA release methods). Minimalization of the volume of each PCR reaction is also important to reduce reagent use and cost; 10 μl reactions should be employed.

(c) Evaluation of different polymerases
Taq DNA polymerase from Thermus aquaticus (Saiki et al. 1988) is standard for PCR, but a wide variety of other polymerases have higher fidelity or processivity (e.g. Cline et al. 1996). As well, more complex PCR cocktails that include one or more repair enzymes offer new hope for the amplification of degraded DNA (Di Bernardo et al. 2002; Mitchell et al. 2005). Restorase (Sigma-Aldrich, St. Louis, MO, USA) represents one recently introduced commercial enzyme cocktail that couples AccuTaq, a high accuracy polymerase, with a repair enzyme.
We evaluated the effectiveness of four polymerases on DNA isolated using the NucleoSpin96 kit from two sets of specimens: recent moths (90 samples, six negative controls) and archival moths (84 samples, 12 negative controls). The recent moths were all less than 1 year old, whereas the archival moths included 14 specimens from each of six age groups (2, 4, 8, 16, ~32, ~64 years; See Electronic Appendix part 1A for details). We tested amplification of the DNA extract from each specimen using: Taq polymerase, Diamond DNA polymerase (Bioline, Randolph, MA, USA), AccuTaq (Sigma-Aldrich, St. Louis, MO, USA) and Restorase. Each enzyme was used according to the manufacturer’s instructions, but the amount of template DNA was constant across all four enzymes (See Electronic Appendix part 1D for details). We tested these enzymes for their ability to amplify the full-length *cox1* barcode (658 bp), as well as partial barcode sequences of 407 and 155 bp (see Electronic Appendix part 1E for sequences of primers).

As expected, positive PCR results were much higher for recent than archival specimens (figure 3). For recent moths, the highest overall success was obtained with Taq polymerase (86%, 93% and 87% success for 658, 407 and 155 bp amplicons, respectively). For archival moths, Restorase performed best overall (44%, 50% and 26% success for 658, 407 and 155 bp amplicons, respectively), but the Diamond and Taq polymerases outperformed it for the smallest amplicon. This latter result was not wholly surprising as Restorase is not recommended for the amplification of small targets. Our results indicate that standard Taq polymerase provides both high performance and low cost for specimens whose DNA has not been degraded, while the use of Restorase merits consideration in archival specimens.

We further compared PCR and sequencing results for different age groups of the archival moths (figure 4). For all four enzymes, success in both PCR and sequencing declined with specimen age for all three amplicons. Restorase delivered the highest PCR success for the full-length product, but none of the enzymes produced 658 bp amplicons from samples older than 8 years (figure 4). For the 407 bp amplicon, all four enzymes performed well, amplifying almost 100% of the samples 8 years and younger. In samples older than 8 years, Taq polymerase showed lower success compared to the other three enzymes (figure 4). Restorase, AccuTaq and Diamond polymerase performed similarly and produced 407 bp amplicons from about 70% of the samples as old as 32 years (figure 4). However, results with Diamond polymerase were inflated by two sequence contaminations. Surprisingly, for the smallest amplicon (155 bp), all four enzymes performed poorly for samples older than 8 years (figure 4).

Direct sequencing of all PCR reactions revealed an interesting result: sequences were sometimes recovered from samples where no PCR product was evident on the agarose gel. This was particularly the case for samples 8 years and older. For example, agarose gels revealed only two of 14 positive PCR products in Restorase amplification of 32 year old moths, but we obtained 11 sequences (with no sign of contamination) from the same reactions (figure 4). Visualization of PCR products on agarose requires a product concentration of at least 1–2 ng µl⁻¹ (Sambrook et al. 1989; White & Wu 2001), while capillary sequencers are known to be more sensitive. This fact suggests that all PCR products from archival specimens should be sequenced.

(d) Archival specimens and DNA repair

AccuTaq is the polymerase present in the Restorase enzyme blend so a comparison of results using Restorase versus those using AccuTaq can indicate if the repair mechanism in Restorase aids barcode recovery from archival specimens. We found that Restorase produced more PCR positives on agarose gels for both the 658 and 407 bp amplicons than AccuTaq (figure 4). However, we found no difference once these samples were sequenced (figure 4). This result suggests that Restorase aids PCR yield, perhaps by repairing template damage, but that the effect is small. In earlier experiments with Restorase, we were able to amplify full-length 658 bp *cox1* barcodes from moths up to 70 years old. However, this success required extensive optimizations that are not time- or cost-effective when the goal is high production rates. However, in the case of extremely rare or extinct species, this capacity could be valuable.

5. SCREENING PCR PRODUCTS

When working on a new taxonomic group or on specimens where PCR success is uncertain, it is helpful to screen PCR reactions for product. This has traditionally been a laborious task involving gel casting and the loading of individual reactions onto the gel. We have explored two options to accelerate this process: microfluidic devices and pre-cast agarose gels. Microfluidic devices ‘sip’ small volumes of the PCR reaction from each of the 96 wells on a plate and then run electrophoresis on a very small scale to determine both the size and concentration of the PCR
Unfortunately, current devices have several limitations for DNA barcoding; they are expensive (more than $100K, €80K), have high operating costs and are relatively slow. These disadvantages are not offset by any gain in sensitivity: the detection limits for these systems mirror those that can be achieved through agarose gels. Pre-cast agarose gels represent a second option: they are fast, require little capital investment (less than $1K, €0.8K) and have modest operating cost. We regularly employ the E-Gel 96 system (Invitrogen, Carlsbad, CA, USA) to screen PCR products, but similar gels are manufactured by several other suppliers.

6. SEQUENCING

(a) Sequencing reaction optimization

Sequencing reactions employ standard chemistry, but reactions can be run in low volume format with diluted sequencing mix (i.e. BigDye; Applied Biosystems, Foster City, CA, USA) without compromising sequencing success or quality. By employing a 10 μl reaction volume containing 0.25 μl BigDye (1/16 of standard reaction), the cost of each sequencing reaction can be substantially reduced. Before the reaction product is submitted for sequencing, it must be cleaned up. There are a variety of solutions for this step and several are scalable to very large production rates. Ethanol precipitation and magnetic bead protocols are widely used by major genomic facilities, but column-based approaches are also effective. Any high volume DNA barcoding facility requires access to one or more capillary sequencers, such as the ABI 3730 DNA Analyser (Applied Biosystems, Foster City, CA, USA). Based on a bidirectional sequencing each barcode record represents two ‘reads’ (see below). Operating seven days a week, a single ABI 3730 can generate just 200K reads or 100K barcode records per year, setting a production threshold for the facility.

(b) Sequence assembly and edit

A bidirectional sequencing strategy has the advantage of enabling the use of automated sequence assembly software to both assign quality values like PHRED scores (Ewing et al. 1998) for each base position and produce a consensus barcode sequence from the reads. It also enhances the quality of the final barcode and ensures its compliance with the minimum read length (i.e. 550 bp) needed to gain barcode designation (by avoiding signal deterioration that often occurs at the end of the reads). Manual inspection and editing of the barcode sequence at the electropherogram level are still required to validate sequence quality and to check for

Figure 4. Effectiveness of four PCR enzymes for the amplification of cox1 in archived specimens of varied ages. Taq polymerase (Taq), Restorase (Res), AccuTaq (Acc) and Diamond polymerase (Dia) were compared for the amplification of full and partial cox1 barcodes (658, 407, 155 bp) in archival moth specimens from six age groups (2, 4, 8, 16, ~32, ~64 years). Results are shown as (a), % PCR success and (b), % sequencing success.
possible polymorphic sites. Their presence, which is often overlooked by sequence assembly software, can indicate the co-amplification of nuclear pseudogenes (Bensasson et al. 2001) along with the authentic mitochondrial sequence. Several software packages are available for visualizing, editing and assembling sequences. SEQUENCHER (Gene Codes Corporation, Ann Arbor, MI, USA) and SEQSCAPE (Applied Biosystems, Foster City, CA, USA) are the most popular commercial software options and include features such as internal basecallers, automatic alignment, contig assembly and trimming of sequences.

7. BARCODE OF LIFE DATA SYSTEMS

Large-scale DNA barcoding projects will create a substantial number of sequence records that must each link to a voucher specimen, as well as to its collateral data. These records need to be organized and analyzed. In addition, for the barcode database to be useful for species identification, it must be searchable by sequence, as well as by species name and higher taxonomic categories. The Barcode of Life Data Systems (BOLD), see http://www.barcodinglife.org, provides varied support for a large-scale barcode project. It is an online repository for \( \text{cox1} \) sequences as well as a workbench for barcode analysis that includes three components: a laboratory information management system (LIMS), a data management and analysis system (DMAS), and a species identification engine.

(a) Laboratory information management system

The assembly and storage of hundreds of thousands of barcode records requires a LIMS to ensure the accurate tracking of all specimens passing through the multi-step analytical chain. Commercial LIMS are available, but they typically cost $50–100K ($60–$80K) per site license, and they lack some key functionality required to support the DNA barcoding initiative.

(b) Data management and analysis system

The DMAS of BOLD provides support for both the warehousing and analysis of barcode records. It includes a simple interface enabling the upload of new sequences to password-protected projects. Its web-based delivery allows work to proceed simultaneously in different labs while being managed from a centralized location, improving communication and preventing data loss or duplication. Moreover, because the DMAS includes information on where each specimen was collected, where it is currently deposited, copies of sequence traces, and high-resolution photographs of each specimen, it allows straightforward traceability of the data stream back to the source. The DMAS was designed to operate at a global scale, ideally supported by mirror sites based at facilities active in barcode analysis.

The DMAS also includes a suite of analysis tools that allow processing or visualization of data. Sequence records, which can be submitted via a simple interface, are automatically aligned. Specimen pages are created automatically from the user-defined data, including an automated plot of GIS coordinates on high-resolution, multi-scale geographic maps. BOLD also includes programs for assembling and exporting neighbor-joining trees (Saitou & Nei 1987), which include colour coding to indicate taxonomic affiliation or other user-defined parameters, as well as tools for specimen display. Finally the DMAS includes an interface that allows the bulk submission of barcode records to GenBank.

(c) Species identification engine

The first step in creating a DNA-based species identification system lies in the assembly of a comprehensive barcode sequence library. The second step involves the development of an effective engine for the comparison and matching of sequences from new specimens to the barcode library. The species identification engine, BOLD-ID, includes a simple user interface to allow \( \text{cox1} \) sequences to be entered into a search field and automatically compared against the existing dataset. BOLD-ID makes use of a combination of Local Alignment Search Tool (BLAST; Altschul et al. 1990) and hidden Markov models based on a global protein alignment for the \( \text{cox1} \) gene, which increase both the speed and accuracy of the matching procedure. Using this algorithm, BOLD returns a probability-based match profile indicating the likely identity of the source species. Links to the species page provide additional information about it (e.g. photographs) that can be useful in confirming the identification. Aside from identifying single specimens, BOLD-ID also performs batch identifications on 96-well plates of samples. The current version of BOLD-ID is optimized for \( \text{cox1} \) gene. However, future versions will include the capability to analyse barcode data from other genes or non-coding regions because barcoding systems in some groups (i.e. plants) will use such data (Chase et al. 2005; Kress et al. 2005).

---

Table 2. DNA barcodes generated in various projects.

<table>
<thead>
<tr>
<th>Project</th>
<th>Sample</th>
<th>Storage condition</th>
<th>Specimens age range (year)</th>
<th>DNA isolation</th>
<th>Barcodes generated&lt;sup&gt;a&lt;/sup&gt;</th>
<th>% success&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lepidoptera of North America</td>
<td>leg</td>
<td>dried</td>
<td>1–3</td>
<td>DryRelease</td>
<td>6510</td>
<td>99</td>
</tr>
<tr>
<td>Lepidoptera of the ACG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>leg</td>
<td>dried</td>
<td>1–28</td>
<td>NucleoSpin96</td>
<td>2419</td>
<td>80</td>
</tr>
<tr>
<td>birds of North America</td>
<td>muscle, liver, feather</td>
<td>ethanol, dried, DMSO</td>
<td>0–38</td>
<td>DryRelease</td>
<td>1782</td>
<td>74</td>
</tr>
<tr>
<td>fishes of Australia</td>
<td>muscle</td>
<td>ethanol</td>
<td>1–15</td>
<td>DryRelease</td>
<td>913</td>
<td>96</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of barcode sequences generated as of April 7, 2005.

<sup>b</sup> Percentage successful DNA barcoding from total number of specimens tested.

<sup>c</sup> Area de Conservacion Guanacaste in northwestern Costa Rica.

---

Phil. Trans. R. Soc. B (2005)
8. PROSPECTS FOR HIGH VOLUME DNA BARCODING

Few molecular taxonomy and evolution laboratories process more than a few thousand specimens a year, but the assembly of a comprehensive barcode library will require, as noted earlier, 100-fold higher production rates. In one sense, the protocols described in this paper are unproven because no barcoding facility has yet achieved this production target. However, we are confident, based on our own experience (table 2), that these protocols will allow the 100K goal to be achieved (see Electronic Appendix part 2 for routine protocols). We emphasize that there is no single optimized protocol if varied types of specimens are being analysed. For example, our work on recently collected North American Lepidoptera employed the DryRelease protocol for DNA isolation, followed by PCR recovery of the full-length barcode. By contrast, results on Costa Rican Lepidoptera, aged from 1 to 28 years, were greatly improved by using the NucleoSpin96 kit for DNA isolation. Moreover, when a full-length cox1 barcode could not be recovered (mainly in samples more than 10 years old), additional PCRs were used to obtain 400 and 350 bp barcode sequences that were concatenated to produce the full-length sequence. These two examples provide a sense of the methodological flexibility that is critical to achieve high success while minimizing costs. While our work has been mainly focused on animals, we expect that barcode analysis of other organisms, such as plants, will require substantial protocol changes, particularly in the isolation of DNA and in the choice of a target genomic barcode region (Kress et al. 2005).

Although barcoding can be executed in a decentralized fashion, economies of scale are gained by establishing core facilities. The capital costs involved in creating a facility capable of generating 100K barcodes a year will range from US$0.5–0.8M (€0.4–0.6M) with the higher figure allowing the emplacement of two capillary sequencers. However, much smaller capital investments ($50K, €40K) will allow the creation of facilities capable of generating 100K PCR products that might then be submitted to any sequencing facility for analysis. The generation of 100K barcode records based on bidirectional sequencing will require an operating budget of approximately $0.3M (€0.2M) before salaries. Although such investments will allow an impressive advance on past production levels, it may be insufficient as work moves from the construction of barcode libraries to the routine application of DNA barcodes for rapid, large-scale assessments of biodiversity in conservation biology and other ecological contexts (DeSalle & Amato 2004). Fortunately there are prospects for both further reductions in cost and increases in production. Costs will drop as reaction volumes shrink and microfluidic devices, which employ nanolitre reaction volumes for PCR and sequencing, are under development. There is also the potential for robotic intervention, which when coupled with unidirectional short reads for identifications (as opposed to reference barcodes) could drive production levels to more than 500 000 specimens per year from a single sequencer. In short, the prospects for both the assembly and use of barcode libraries appear bright enough to expect illumination of many key problems in biodiversity science.

Funding for this study was provided by the Gordon and Betty Moore Foundation, the Canada Foundation for Innovation, the Ontario Innovation Trust, the Canada Research Chairs Program and NSERC. We thank Daniel Janzen, Scott Miller, Jean-François Landry, John Burns, Tyler Zemlak and Kevin Kerr for providing specimens and Andrey Poltaraus for aid with analytical protocols. We also thank the Sigma-Aldrich R & D group, especially Kevin Kayser, for providing Restorase and Accutaq enzymes and for helpful discussions. Alex Borisenko aided the assembly of graphics for this paper, while Janet Topan oversaw the sequencing. We thank the Consortium for Barcoding Life for facilitating the DNA barcoding initiative.

REFERENCES


Greiner, M., Carter, P., Korn, B. & Zink, D. 2004 New approach to complete automation in sizing and quantitation of DNA and proteins by the Automated Lab-on-a-Chip Platform from Agilent Technologies. *Nat. Methods* 1, 87–89. (doi:10.1038/nmeth0487.)


Lindahl, T. 1993 Instability and decay of the primary structure of DNA. *Nature* 362, 709–715. (doi:10.1038/362709a0.)


A likelihood ratio test for species membership based on DNA sequence data

Mikhail V Matz and Rasmus Nielsen

*Phil. Trans. R. Soc. B* 2005 360, 1969-1974

References

This article cites 31 articles, 9 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1969.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1969.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to *Phil. Trans. R. Soc. B* go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2005 The Royal Society
A likelihood ratio test for species membership based on DNA sequence data

Mikhail V. Matz¹ and Rasmus Nielsen²,*

¹Whitney Laboratory for Marine Bioscience, Department of Molecular Genetics and Microbiology, University of Florida, 9505 Ocean Shore Blvd, Saint Augustine, FL 32080, USA
²Center for Bioinformatics, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark

DNA barcoding as an approach for species identification is rapidly increasing in popularity. However, it remains unclear which statistical procedures should accompany the technique to provide a measure of uncertainty. Here we describe a likelihood ratio test which can be used to test if a sampled sequence is a member of an a priori specified species. We investigate the performance of the test using coalescence simulations, as well as using the real data from butterflies and frogs representing two kinds of challenge for DNA barcoding: extremely low and extremely high levels of sequence variability.

Keywords: DNA barcoding; likelihood ratio tests; assignment of individuals; coalescent simulations

1. INTRODUCTION
DNA barcoding is a technique for assigning specimens to species or other taxonomic units based on their DNA sequence in a standard marker, the DNA barcode (Hebert et al. 2003a; Tautz et al. 2003). Despite growing popularity of the approach, it remains unclear which statistical procedures should be used to assign an unknown specimen to a species, and how to associate the assignments with measures of statistical uncertainty. The major problem arises because within-species variability and between-species variability can easily be confounded (Lipscomb et al. 2003; Seberg et al. 2003). The extent of these variations may be notably different across the animal phylogeny (Hebert et al. 2003b). Statistical methods in DNA barcoding must, therefore, be aimed primarily at discriminating intra- from inter-specific variability on the basis of actual sequence data.

There are several population genetic approaches for assignment of individuals, such as the methods of Paetkau et al. (1995), Rannala & Mountain (1997), Cornuet et al. (1996), Cornuet et al. (1999), and Pritchard et al. (2000). All these methods consider multi-locus data, e.g. restriction fragment length polymorphisms (RFLP), single nucleotide polymorphisms (SNP) or microsatellite data. The situation in DNA barcoding is slightly different because assignment is based on a single DNA sequence, usually not subject to recombination. New statistical and population genetic methods are needed to deal with this special situation.

In this paper we consider the problem of testing membership of a particular, a priori specified, species. This situation may arise in many cases, for example in screens of invasive species, pathogenic bacterial strains, toxic algae or products from protected species (e.g. Tautz et al. 2002, 2003; Armstrong & Ball 2005; Lorenz et al. 2005; Markmann & Tautz 2005). We will assume that more than one sequence from the focal species is known. A new sequence has been obtained from a sample and it is then of interest to the hypothesis that the new sequence was obtained from an individual from the focal species. We propose a new likelihood ratio test to address this problem, based on existing population genetic Markov Chain Monte Carlo methods (MCMC).

2. METHODS AND THEORY
In the following, we will describe how a likelihood ratio test can be constructed to test the hypothesis of species membership. The basic idea is to use a model with two populations, one population containing all existing database sequences from the species, and another population containing the new sampled sequence. The null hypothesis is then specified as $H_0: T=0$, where $T$ is the divergence time between the two species.

The likelihood function is defined as a function proportional to the sampling probability of the data, conditional on the parameters, e.g. $Pr(X|\Psi)$, where $X$ is a vector of DNA sequences from multiple individuals, and $\Psi$ is a vector of parameters. A fundamental problem in the analysis of DNA sequence data in population genetics is that the likelihood function cannot be calculated analytically. However, the likelihood conditional on the underlying gene tree (or coalescent tree) can be easily calculated using standard methods of phylogenetics. The following representation of the likelihood is often used:

$$Pr(X|\Psi) = \int_{\Omega} Pr(X|G, \Psi)f(G|\Psi)dG,$$

where $G$ is the gene tree, $\Omega$ is the set of all possible gene trees, and $f(G|\Psi)$ is the density of gene trees. The integral can be thought of as a sum over all possible gene tree topologies and a multi-dimensional integral.
over all possible coalescent times (lineage divergence times in the gene tree). This integral cannot be evaluated analytically, or using numerical integration, for realistically large sets of DNA sequences, but must be evaluated using stochastic methods. For example, Griffiths & Tavaré (1994), as well as Stephens & Donnelly (2000), used a simulation technique known as sequential importance sampling (e.g. Liu 2001), while Kuhner et al. (1995) and Nielsen & Wakeley (2001) used MCMC. In the MCMC methodology of Nielsen & Wakeley (2001), a Markov chain is defined with state space on $O \times \Psi$, where $\Psi$ is the parameter space, i.e. $O \times \Psi$ for all values of $\Psi$. A prior distribution is assigned to $\Psi$ and a Markov chain is then constructed with stationary distribution (equilibrium distribution) given by the joint posterior distribution of $O$ and $\Psi$, i.e. $f(O, \Psi|X)$. When simulating this Markov chain, parameter values sampled from it (at stationarity) follow the marginal posterior for $\Psi$, $f(\Psi|X)$. Using a uniform prior for $\Psi$, the integrated likelihood function for $\Psi$ is simply given by $f(\Psi|X)$. Inferences regarding the parameters can then either proceed in a Bayesian or a frequentist framework.

Nielsen & Wakeley (2001) considered a model with migration and divergence between two populations. We will here use a simplification that assumes no ongoing gene flow between the two populations (i.e. zero migration rate). In tests of $T=0$, including migration in the model would add an additional parameter, leading to an undesirable increase in degrees of freedom in the test. Also, since only one sequence has been assigned to one of the populations, the effective population size and $m$ for the parameters can then either proceed in a Bayesian or a frequentist framework.

The simulation of sequences under the coalescent model was performed using the MS program (Hudson 2002) under an infinite sites model, assuming $\theta=3$, 12 and 30 (per locus); $\theta=12$ represented more or less common situations in animal barcoding (Hebert et al. 2003b), while $\theta=3$ and 30 modelled low-divergence and high-divergence situations, respectively. For each of the three values of $\theta$, we produced 100 datasets composed of either 4, 11 or 26 sequences, one of which was then randomly chosen as the query sequence, resulting in $n=3$, 10 or 25 ‘database’ sequences. For each simulated data set, 500,000 updates of the Markov chain were performed using a burn-in period of 50,000. Although more updates might be recommendable for real data analysis, using 500,000 updates provides reasonably accurate estimates of the likelihood ratio while allowing multiple simulations to be performed in a short amount of time.

The frequency distribution of likelihood ratios are shown in figure 1. For small sample sizes or small values of $\theta$, the type I error rate (rejection of the correct species) for the critical value of 2.7 was slightly lower than 5%. However, for large sample sizes a test based on a critical value of 2.7 would reject the correct species much more often than in the expected 5% of cases. The most extreme case is $n=25$ and $\theta=30$, where the critical value of 2.7 for likelihood ratio statistic resulted in 15.1/100 rejections. From the simulations it appeared that a critical value of 4 would ensure a 5% significance level test of the right size (figure 1b). Still, the coalescent simulations did not address the magnitude of type II error rate for this critical value. This issue was further investigated in the simulations involving real data (see below).

The major reason why the rejection rate increases with $\theta$ and $n$ appears to be that likelihood ratio is harder to estimate for larger sample sizes. The increased Monte Carlo variance for these parameter values inflates the rejection rate. This problem can be rectified by increasing the number of cycles in the MCMC procedure. In real data analysis it may also be desirable to obtain critical values directly by parametric bootstrapping, although such procedures may be computationally expensive. Future research may also be focused on more computationally efficient methods for estimating the likelihood ratio.

3. ANALYSIS OF SIMULATED DATA

The frequency distribution of likelihood ratios are shown in figure 1. For small sample sizes or small values of $\theta$, the type I error rate (rejection of the correct species) for the critical value of 2.7 was slightly lower than 5%. However, for large sample sizes a test based on a critical value of 2.7 would reject the correct species much more often than in the expected 5% of cases. The most extreme case is $n=25$ and $\theta=30$, where the critical value of 2.7 for likelihood ratio statistic resulted in 15.1/100 rejections. From the simulations it appeared that a critical value of 4 would ensure a 5% significance level test of the right size (figure 1b). Still, the coalescent simulations did not address the magnitude of type II error rate for this critical value. This issue was further investigated in the simulations involving real data (see below).

The major reason why the rejection rate increases with $\theta$ and $n$ appears to be that likelihood ratio is harder to estimate for larger sample sizes. The increased Monte Carlo variance for these parameter values inflates the rejection rate. This problem can be rectified by increasing the number of cycles in the MCMC procedure. In real data analysis it may also be desirable to obtain critical values directly by parametric bootstrapping, although such procedures may be computationally expensive. Future research may also be focused on more computationally efficient methods for estimating the likelihood ratio.

Phil. Trans. R. Soc. B (2005)
We examined two real data sets representing two marginal cases of extremely low and extremely high sequence variability, both at the intra- and inter-specific levels. The first data set contains sequences from the skipper butterfly *Astraptes fulgerator*, which recently has been proposed to be a complex of perhaps as many as 12 separate species (Hebert et al. 2004a). Genetic differentiation between these species was originally identified by the phylogeny of cytochrome oxidase I (*cox1*) sequences, and was corroborated by the presence of morphological difference in caterpillars and the species of plants preferred by them as food. Still, both the degree of divergence within and between these species is very small (figure 2a), average $\theta$ estimate per species being only 0.65 for the whole locus. In sharp contrast to the butterflies, our second example—four species of the Australian rainforest frogs of the genus *Litoria*—displayed intra-specific *cox1* sequence often exceeding 10% pairwise difference (figure 2b) (Schneider et al. 1998), with average $\theta=26$. Such a high level of *cox1* variability appears to be common in amphibians (Vences et al. 2005a,b).

In the simulations described here, for each of the two datasets 100 sequences were drawn with replacement from the data pool to represent every species with equal probability (i.e. in each draw, first the species name was randomly chosen and then a sequence from this species was randomly selected). These queries were then tested using the likelihood ratio test for affiliation with their true species, to estimate the type I error rate, as well as for affiliation with the most similar ‘sister species’ identified on the basis of the *blastn* score (Altschul et al. 1997), to evaluate type II error rate. The use of the word sister species here reflects, therefore, sequence similarity and does not reflect any phylogenetic assessment of taxonomic status. From the sequences originating from each of these two species, several ‘database sequences’ were randomly chosen to represent them when applying the likelihood ratio test. We performed simulations either with three or ten representative database sequences. In each case 100 000 updates of the Markov chain under a finite sites model was performed.

The frequencies of the likelihood ratios obtained in the simulations are shown in figure 2c,e (skipper butterfly) and figure 2d,f (tree frogs). Notice, that the distributions in most cases do not show much overlap when comparing the true species and the sister species. This shows that, if appropriate critical values have been obtained, powerful tests for rejecting species membership can be constructed. For 10 sequences, assuming a critical value of 2.7 for the likelihood ratio test results in a conservative test with high power to reject the wrong species, the type I error was less than 5% and the power (1—type I error) was larger or around 95% (figure 3). Note that in both these cases, using a critical value of 2.7 results in a test with good properties (figure 3), whereas re-setting the critical value to 4 would result in a dramatic increase in type II error rate (loss of power). However, when analysing only three representative database sequences, the power was significantly reduced (figure 3). Additionally, a critical value of 2.7 would result in an anti-conservative test in the case of *Litoria*.

**5. DISCUSSION**

The likelihood ratio test presented here is one of the first attempts to introduce statistical rigor into DNA barcoding. It is designed to explicitly evaluate the possible boundaries of intra-specific variation on the basis of the available sequence data, using population genetic inference based on coalescent theory. It is essentially a test of population subdivision between a
query sequence and a set of database sequences. This procedure represents a significant step towards realistic species modelling in comparison to the previously suggested limiting threshold approach, which simply assumes that the intra-specific sequence variation cannot exceed certain pairwise distance (Floyd et al. 2002; Hebert et al. 2003a,b).

The taxonomic affiliation of the query sequence is sometimes inferred using tree-building methods on the basis of phylogenetic grouping of the query (Hebert et al. 2004a,b). Our method is in fact a tree-based one as well, but it takes into account phylogenetic uncertainty and uses population genetic theory to determine cut-offs in ambiguous cases (e.g. cases with...
lineage sorting). For example, based on purely topological arguments, at least 12 sequences would be needed to have a theoretical chance of rejecting species membership at the 5\% significance level, assuming all topologies are equally likely. The population genetic approach essentially makes use of the distribution of branch-lengths to determine if species membership is plausible.

Throughout this work, we assumed the simplest model of intra-specific variability based on a single panmictic population of constant size. Caution should be exercised, therefore, in direct applications of the test to cases where these assumptions may be seriously violated, for example, when the query sequence has been obtained from a distinct geographic area than the database sequences. Although we expect that the concept of effective population size may mitigate most of the complications, the extent to which our test is applicable to non-ideal populations remains to be investigated. In the future, it may be feasible to include additional assumptions regarding $\theta$. Otherwise, the implementation of the DNA barcoding technology would be delayed, or even worse, would lead to artefacts resulting from confusion between intra- and inter-specific genetic variation.

Finally, we would like to point out that the DNA barcoding technology at the present moment sacrifices a great deal of potential power by relying on just a single molecular marker. The data acquisition for a multi-locus DNA barcoding database may be more feasible than it seems, since the stages associated with sample collection, DNA isolation and (partially) PCR set up would not require additional investments in comparison to single-locus data acquisition (e.g. see Chase et al. 2005; Summerbell et al. 2005). The current statistical method could also be modified to analyse data from multiple loci (e.g. Hey & Nielsen 2004; Hey et al. 2004).

Funding for this project was provided by grants from NIH (GM066243) and US Department of Defense (SERDP program) to MVM, and HFSP Grant RGY0055/2001-M, NSF Grant DEB-0089487 and NSF/NIH Grant DMS/NIGMS—0201037 to RN.

REFERENCES


\textit{A test for species membership} M. V. Matz & R. Nielsen 1973

The likelihood ratio test can be easily adapted for use in routine DNA barcoding. However, it is obvious that the power of the test depends greatly on the number of samples in the database. Although tests based on only three sequences have some power, it is clear that more sequences are desirable. Databases containing only a single sequence from each species cannot be used to test for species membership without making additional assumptions regarding $\theta$. Unfortunately, at the current stage of development of DNA barcoding databases, in which the depth of individual species sampling is usually sacrificed in favour of greater taxonomic coverage, it will in most cases not be possible to test for species membership using our procedure. For example, the database of DNA barcodes for the North American birds (Hebert et al. 2004) includes just one or two sequences for the majority of species. We believe that the priorities in the data accumulation for DNA barcode databases should be adjusted to balance the two essential informational components, intra-specific and inter-specific sampling. Otherwise, the implementation of the DNA barcoding technology would be delayed, or even worse, would lead to artefacts resulting from confusion between intra- and inter-specific genetic variation.

Figure 3. Summary of error rates obtained for \textit{Astraptes} (a) and \textit{Litoria} (b) datasets with different number of sequences per species in the database, assuming the critical value of 2.7. Open bars, type I error rate; filled bars, type II error rate.


TaxI: a software tool for DNA barcoding using distance methods
Dirk Steinke, Miguel Vences, Walter Salzburger and Axel Meyer

Phil. Trans. R. Soc. B 2005 360, 1975-1980

References

This article cites 27 articles, 9 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/360/1462/1975.full.html#ref-list-1

Article cited in:
http://rstb.royalsocietypublishing.org/content/360/1462/1975.full.html#related-urls

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to Phil. Trans. R. Soc. B go to: http://rstb.royalsocietypublishing.org/subscriptions
TaxI: a software tool for DNA barcoding using distance methods

Dirk Steinke1, Miguel Vences2, Walter Salzburger1 and Axel Meyer1,*

1Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, 78457 Konstanz, Germany
2Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Zoological Museum, Mauritskade 61, 1092 AD Amsterdam, The Netherlands

DNA barcoding is a promising approach to the diagnosis of biological diversity in which DNA sequences serve as the primary key for information retrieval. Most existing software for evolutionary analysis of DNA sequences was designed for phylogenetic analyses and, hence, those algorithms do not offer appropriate solutions for the rapid, but precise analyses needed for DNA barcoding, and are also unable to process the often large comparative datasets. We developed a flexible software tool for DNA taxonomy, named TaxI. This program calculates sequence divergences between a query sequence (taxon to be barcoded) and each sequence of a dataset of reference sequences defined by the user. Because the analysis is based on separate pairwise alignments this software is also able to work with sequences characterized by multiple insertions and deletions that are difficult to align in large sequence sets (i.e. thousands of sequences) by multiple alignment algorithms because of computational restrictions. Here, we demonstrate the utility of this approach with two datasets of fish larvae and juveniles from Lake Constance and juvenile land snails under different models of sequence evolution. Sets of ribosomal 16S rRNA sequences, characterized by multiple indels, performed as good as or better than cox1 sequence sets in assigning sequences to species, demonstrating the suitability of rRNA genes for DNA barcoding.

Keywords: DNA barcoding; molecular taxonomy; cox1; 16S rRNA; species recognition

1. INTRODUCTION

Recent work suggested that a DNA-based identification system can aid the resolution of the vast diversity of life with its millions of species (Tautz et al. 2003). Hebert et al. (2003a,b) proposed that a DNA barcoding system for animal life could best be based upon sequence diversity in the 5' section of the mitochondrial gene cytochrome oxidase subunit I (cox1). Although DNA variation has long been used successfully for the identification and classification of microorganisms (Rosello-Mora & Amann 2001), scepticism against this approach for more complex taxa has been expressed. Two primary objections have focused on (i) the concern that DNA sequence differences among closely related species will often be too small to allow their discrimination (Mallet & Willmot 2003) and (ii) the fact that present strategies and programs suffer from difficulties of aligning sequences of different lengths (Lipscomb et al. 2003), especially in automated large-scale analyses. Hebert et al. (2003b) argued against the mitochondrial 12S and 16S rRNA genes as standard DNA barcoding markers because the presence of multiple insertions and deletions in these genes pose potential problems due to difficulties and ambiguities in their alignment. This problem would apply as well to the nuclear 28S rRNA and internal transcribed spacer genes (ITS). On the other hand, a variety of arguments have been voiced and evidence has been brought forth to suggest that 28S (Tautz et al. 2003; Markmann & Tautz 2005), ITS (Blaxter 2003) and 16S rRNA (Vences et al. 2005b) could be valuable taxonomic markers in the framework of a large-scale DNA barcoding system.

One of the most promising applications of DNA barcoding is the molecular identification of often phenotypically disparate life-history stages in taxonomic, ecological, behavioural and conservation studies. Animal juvenile or larval morphology is often distressingly uniform among different species such as in fish larvae. Larvae might, on the other hand, be radically different in morphology from the adult (e.g. in frogs or holometabolous insects), and some species such as parasites might even display a complex variety of larval, semi-adult and adult stages. Reliable field identification in these taxa requires extensive expertise, and even then can be impossible. Furthermore, any classical taxonomic method to estimate population numbers and monitor trends is costly and time-intensive. Recent developments in non-invasive genetic sampling techniques in combination with DNA barcoding provide a practical alternative for such research.

In this study we present a flexible software tool for DNA taxonomy named TaxI (available at http://www.evolutionsbiologie.uni-konstanz.de/Software). This tool is based on pairwise sequence divergences between query and reference sequences that can be defined by the user. We test the suitability of this program for the identification of larval and juvenile stages of organisms with two datasets, fish larvae from Lake Constance, on August 19, 2011
Central Europe, and juvenile land snails, and compare the performance of two proposed markers, cox1 and 16S rRNA, in DNA barcoding.

2. MATERIAL AND METHODS

(a) Molecular datasets and methods

Thirty 16S rRNA fish reference sequences were used in this study representing most fish species that occur in Lake Constance (Eckmann & Ro¨sch 1998). These sequences were obtained from GenBank (accession numbers given in table 1).

Snail reference sequences representing the Mediterranean family Hygromiidae (23 species) and the Helicidae s.l. (39 species including the previous 23 Hygromiidae species) were taken from a previous study (Steinke et al. 2004; GenBank accession numbers AY546342–546381 for 16S rRNA and AY546262–AY546301 for COI = cox1).

The fish larvae and juveniles (seven putative species) and juvenile snail samples (five putative species) were identified by classical morphological methods. Following a proof of principle approach we only used samples of those specimens that could be reliably identified. Sequences were obtained by preparing total DNA extracts from tissue samples with a proteinase K digestion followed by sodium chloride extractions and ethanol precipitations. Published ‘universal’ primers were subsequently used to amplify a 420 bp fragment of the 16S rRNA gene (16Sar-L and 16Sbr-H by Palumbi et al. 1991) in fish and snail specimens and a 520 bp fragment of the cox1 gene (LCO1490 and HCO2198 by Folmer et al. 1994) in juvenile snails. Polymerase chain reaction (PCR)-amplifications were performed according to standard protocols on a GeneAmp 9700 thermocycler (Applied Biosystems). The PCR products were purified using the QiaQuick spin columns extraction kit (Qiagen), sequenced in both directions with the BigDye termination reaction chemistry, and determined on an ABI 3100 Automatic Capillary Sequencer (Applied Biosystems).

(b) Data analysis using TaxI

TaxI is a program for Windows platforms to compute pairwise distances among sequences after their pairwise alignment. Data input is in fastA format, a simple and widely used file format (Pearson & Lipman 1988). TaxI can process pairwise distances among sequences after their pairwise alignment. Data input is in fastA format, a simple and widely used file format (Pearson & Lipman 1988). TaxI can process multiple files containing a single sequence each or single files with multiple sequences (aligned or unaligned). Hundreds of query sequences can be analysed in one program run. As depicted in the process flowchart in figure 1, the program considers all possible pairs among query and reference files. All these pairs are then aligned using the T-Coffee algorithm (Notredame et al. 2000). T-Coffee performs with high accuracy even if long internal deletions require a method that is able to deal with local similarity. Following the alignment, sequence divergence is determined for each pair by dividing the number of different nucleotides by the total number of nucleotides examined in the alignment. All alignment positions with gaps are excluded from distance computation (complete deletion).

The evolutionary distances that are computed from DNA sequence data are primarily estimates of the number of nucleotide substitutions per site (d) between two sequences. There are many methods for estimating evolutionary distances, depending on the pattern of nucleotide substitutions, which might best account for back mutations among more distantly related sequences (see Nei 1987; Gojobori & Nakamura 1986).

Table 1. Fish reference sequences used in this study and obtained from GenBank given by species and accession number.

<table>
<thead>
<tr>
<th>Order</th>
<th>family</th>
<th>species</th>
<th>accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anguilliformes</td>
<td>Anguillidae</td>
<td>Anguilla anguilla</td>
<td>AJ244828.1</td>
</tr>
<tr>
<td>Cypriniformes</td>
<td>Balitoridae</td>
<td>Barbatula barbatula</td>
<td>DQ077974</td>
</tr>
<tr>
<td></td>
<td>Cobitidae</td>
<td>Cobitis taenia</td>
<td>AJ247080.1</td>
</tr>
<tr>
<td></td>
<td>Cyprinidae</td>
<td>Abramis brama</td>
<td>AJ247067.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alburnus alburnus</td>
<td>AJ247063.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Barbus barbus</td>
<td>AJ247065.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blica hjortkna</td>
<td>AJ247064.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Carassius carassius</td>
<td>AJ247070.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chondrostoma nasus</td>
<td>AJ247047.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cyprinus carpio</td>
<td>NC001606.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Golio gobio</td>
<td>AJ247068.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leuciscus cephalus</td>
<td>AJ247054.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leuciscus leuciscus</td>
<td>AJ247052.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leuciscus souffia agassizi</td>
<td>AJ247079.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phoxinus phoxinus</td>
<td>AJ247062.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rhodetus sericus</td>
<td>AJ247086.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rutilus rutilus</td>
<td>AJ247045.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scardinius erythrophtalmus</td>
<td>AJ215479.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tinca tinca</td>
<td>AJ247083.1</td>
</tr>
<tr>
<td>Esociformes</td>
<td>Esocidae</td>
<td>Esox lucius</td>
<td>AP006446.1</td>
</tr>
<tr>
<td>Gadiformes</td>
<td>Lotidae</td>
<td>Lota lota</td>
<td>AP004412.1</td>
</tr>
<tr>
<td>Gasterosteiiformes</td>
<td>Gasterostidae</td>
<td>Gasterosteus aculeatus</td>
<td>AF355030.1</td>
</tr>
<tr>
<td>Perciformes</td>
<td>Percidae</td>
<td>Gymnocephalus cernuus</td>
<td>AY141443.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perca fluviatilis</td>
<td>AY141442.1</td>
</tr>
<tr>
<td>Salmoniformes</td>
<td>Salmonidae</td>
<td>Coregonus lavaretus</td>
<td>NC002646.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oncorhynchus mykiss</td>
<td>AF312573.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Salmo trutta</td>
<td>X77564.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Salvelinus alpinus</td>
<td>AJ319820.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thymallus thymallus</td>
<td>AY430237.1</td>
</tr>
<tr>
<td>Scopaeiformes</td>
<td>Cottidae</td>
<td>Gobius gobio</td>
<td>AP004442.1</td>
</tr>
</tbody>
</table>
et al. 1990). In TaxI, the user can choose between six methods: Jukes-Cantor distance (Jukes & Cantor 1969), Kimura 2-parameter distance (Kimura 1980), Tajima-Nei distance (Tajima & Nei 1984), Tamura distance (Tamura 1992), Tamura-Nei distance (Tamura & Nei 1993), LogDet (Lockhart et al. 1994). All possible pairwise distances are saved in a matrix and the lowest $d$ is calculated with a simple sort routine for every multiple hit correct method chosen. The output files are in basic ASCII format and contain evolutionary distances and other information related to the data and delimited by tabs. Most spreadsheet programs (e.g. Microsoft Excel) allow editing those files.

Phil. Trans. R. Soc. B (2005)
Snail sequences

<table>
<thead>
<tr>
<th></th>
<th>fish sequences</th>
<th>16S rDNA</th>
<th>cox1</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Barbatula barbatula (juvenile)</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Rutilus rutilus (juvenile)</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Perca fluviatilis (juvenile)</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Phoxinus phoxinus (juvenile)</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

In our test datasets, all juvenile sample sequences (table 2) were tested against the reference sequences with all possible distance methods. The percentage of correct matches was calculated to test the accuracy of the method. Twenty-eight newly determined 16S rRNA fish sequences (GenBank accession numbers DQ077946–DQ077973) were tested against the Lake Constance reference data set containing fish species occurring in that lake (Eckmann & Rösch 1998).

Twenty 16S rRNA and cox1 juvenile snail sequences were determined morphologically to belong to five different species and were tested against two data sets respectively. In the latter case sequences were tested against a second dataset containing 16 more species at a more-inclusive taxonomic level to test the performance of TaxI in a situation with considerable amounts of sequence divergence.

3. RESULTS

In the 16S rRNA fish dataset all 28 query sequences were correctly assigned to seven species under all possible model regimes. The divergence value of the query sequences to the most similar lineage was between 0 and 2.7% and the mean divergence to the complete reference dataset lies between 15.5 and 26.9%. Query sequences of the snail dataset, using uncorrected distances, were in all cases assigned to the correct genus and in 90% to the right species. Exceptions were Candidula unifasciata juveniles that grouped most closely with a morphological distinct sister species (Candidula spadac). In all cases, there was 0–16.8% sequence divergence between the test taxon and the lineage in the profile that was most similar to it (table 3), which was in any case lower than the mean divergence to the complete reference dataset (19.6–25.9%). Eighteen out of the 20 juvenile snail specimens were successfully classified using 16S rRNA sequences and 16 were assigned to the correct species using cox1. The use of Tamura–Tamura–Nei- and LogDet-distances in calculating snail classifications resulted in 100% success of species recognition (figure 2) whereas other substitution models performed worse, especially with cox1 sequences. The identification using cox1 sequences was more reliable in the taxonomically less inclusive dataset, i.e. when available sequences for the comparison were a priori

4. DISCUSSION

In this study we developed and tested a new software tool for a DNA identification system based on pairwise sequence divergence under different model regimes. As highlighted by Moritz & Cicero (2004), DNA barcoding or DNA taxonomy focuses on phenetic identification rather than phylogenetic reconstruction, and the software requirements are different in the two cases. Although trees are used for the visualization of results (e.g. Hebert et al. 2003a,b), the tree topology (especially at inclusive phylogenetic levels) is of less concern in DNA barcoding as long as tips are correctly grouped at the relevant taxonomic level (usually species or genera). In the case of gastropod cox1 sequences, recent studies (Medina & Walsh 2000) indicated a high variability of third codon positions and therefore these are suspected to be of limited utility for phylogenetic analyses among closely related species. However, as long as these positions contain information to identify and distinguish among closely related species, they are valuable for DNA barcoding.

In phylogenetic analyses, a correct alignment of sequences is of paramount importance but is highly complicated in DNA fragments characterized by multiple insertions and deletions (e.g. Morrison & Ellis 1997). Most methods for maximum parsimony, maximum likelihood and Bayesian analyses require a multiple alignment, which is rather time consuming if hundreds of sequences are used. Furthermore, algorithms for multiple alignments often produce errors in such large datasets (Wheeler et al. 1994). For DNA
barcoding purposes, distances need to be compared among sequences. Homologous nucleotide positions must be identified and compared between pairs of sequences, but homology between the separate alignments is less relevant because mutations are not analyzed in a phylogenetic context and evolution at particular nucleotide positions needs not to be reconstructed. Although different pairwise alignments may yield different homology statements for one locus, the pairwise alignment approach followed in TaxI is faster than large multiple alignments, and produces an output that is easier to interpret than those of the faster BLAST algorithm (Altschul et al. 1990), particularly when more than one query sequences are subjected to analysis. A major drawback of BLAST is that this algorithm often partitions highly divergent sequences into separate fragments that are then compared separately, and therefore it is not straightforward to obtain a standard measure of overall sequence divergence. TaxI also provides information on alignment length, number of transitions and transversions and the number of introduced gaps for each pairwise comparison to interpret an assignment and to resolve ambiguous findings. Due to the fact that a user can define the reference dataset TaxI is flexible and reduces computational time. We suggest that this program will facilitate the application and wider acceptance of DNA barcoding, and increased practical use will help to identify further tools required for such applications, to be incorporated in this or other software packages.

In ribosomal DNA indels are relatively common because the excision or insertion of a few nucleotides often has little impact on rRNA function. Their alignment and analysis can therefore be particularly difficult. However, this study provides further evidence that even partial 16S rRNA gene fragments are useful in DNA barcoding. In our fish dataset, 16S performed successfully, and in the snail dataset, species identification was more successful using 16S compared to cox1. We have furthermore successfully used TaxI to calculate distances among large numbers of amphibian sequences of the 16S rRNA and cox1 genes, and to identify conspecific and closely related taxa from these datasets (Köhler et al. in press; Vences et al. 2005a).

Helicid land snails pose a challenge for DNA barcoding because they are characterized by extremely high levels of intraspecific mitochondrial DNA divergence (e.g. Thomaz et al. 1996). Strong differences among the reference and query sequences explain why our snail dataset species identification was not in all cases successful. However, a correct identification of all query sequences was achieved using particular models of sequence evolution (figure 2), namely the LogDet, Tamura, and Tamura and Nei models. The latter two correct for multiple hits taking into account substitutional rate differences between nucleotides and inequality of nucleotide frequencies. These methods also distinguish between transversional and transitional substitution rates, which is also the case in the Kimura-2-parameter model misplacing a sequence only in one case. Given the fact that there is a high transition/transversion bias in mitochondrial DNA this indicates that models that consider these two classes of substitutions or calculate additive distances with variable base composition (like LogDet) are most suitable for DNA barcoding in groups of large intraspecific divergences of mitochondrial haplotypes.

We would like to thank Elke Hespeler for technical assistance. Support from the Deutsche Forschungsgemeinschaft (DFG) to A.M. and M.V. and from the European Union to W.S. is gratefully acknowledged.

REFERENCES

Figure 2. Bar graph of the percentage success in classifying snail species to the correct taxon. Positive 16S rRNA identifications are shown in grey, positive cox1 identifications using the Hygromiidae profile group (23 species) are shown in black, and positive cox1 identifications using the Helicidae s.l. (39 species) profile group are shown in white.

DNA taxonomy using distance methods D. Steinke and others 1979

Phil. Trans. R. Soc. B (2005)


