Defining operational taxonomic units using DNA barcode data

Citation for published version:

Digital Object Identifier (DOI):
10.1098/rstb.2005.1725

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Philosophical Transactions of the Royal Society B: Biological Sciences

Publisher Rights Statement:
Freely available via PMC.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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, 2003).
unpublished). 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 taxonomic 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; Eyualem & 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 (Eyualem & 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 collemboles 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/proteinase K/phenol/chloroform protocol. The nSSU marker was amplified from individual extracts using the primers SSU_F04 (GCTTGTCTCAAGATTAGGCC) and SSU_R26 (CATCTTGCGAATGCCTTGC) (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 amplicons were amplified from a subset of tardigrade individuals (also amplified for nSSU) using the ‘universal’ primers cox1 (HC02198; TAAACCTCAAGGGTGACCCAAAAATCA) and cox1 (LC01490; GGTCAACAAATCATAAGATATTGG) (Hebert et al. 2003), yielding a ~650 bp product. Shrimp alkaline phosphatase/exonuclease I-cleaved 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/proteinase K/phenol/chloroform protocol. The nSSU marker was amplified as described above. Amplicons of nSSU generated from the bulk extract target were cloned into 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 (CTGCGCCGTGTAAACATT) and M13_R (CAGGGAAACAGCTATA), cleaned using shrimp alkaline phosphatase and exonuclease I, and sequenced using SSU_R09 (AGCTGG-AATTACCGGCGGCT) 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 trace2seq (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
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 ~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 MOTUDefine.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 metazoan-speciﬁc (we know that we can amplify environmental fungi, data not shown) and because, despite their being unicellular protozoa, ciliate macronuclei 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.

(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, nSSU PCR and sequencing was successful ~85% of the time. In contrast, the cox1 success rate was less than 40%. Indeed examination of available cox1 sequences from animals related to those expected to be found in the moss ecosystem revealed that the ‘universal’ primers employed were unlikely to be able to amplify from some phyla. We conclude that use of the cox1 target for the full diversity of animals will require additional rounds of primer pair optimization. No PCRs were successful from individually extracted rotifers, despite the nSSU primers sites being present in the available rotifer nSSU sequences. While this result could be due to the low number of cells (and thus genomes) in an individual rotifer, sequences from the bulk DNA sample were also rotifer-free (see below). 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 macronuclei 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

MOTUDefine.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 chromodorid 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

Phil. Trans. R. Soc. B (2005)
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 \textit{cox1} and \textit{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 \textit{cox1} MOTU were defined from this subset. Surprisingly, 23 \textit{nSSU} MOTU were defined, despite the overall lower level of sequence divergence, though the distance between distinct clusters was greater in the \textit{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 \textit{cox1} MOTU (with two and five members) corresponded to two \textit{nSSU} MOTU each (figure 4). The remaining 68 specimens formed two groups with complex patterns of overlap between \textit{nSSU} and \textit{cox1} MOTU (figure 4). Thus, while \textit{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.

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
matter what the addition order, this suggests that these MOTU have some congruence with biological taxa, and a distinctness from other related OTU. But if repeated clustering of a group of sequences yields discordant MOTU, this identifies a biologically interesting phenomenon, somewhere along the spectrum from population genetic processes to recently separated taxa still sharing ancestral polymorphisms. Such variability can thus alert researchers to novel features of communities not simply accessible through other means.

(b) MOTU inference using different cutoff scores

We performed 300 independent clusterings of the 295 tardigrade nSSU sequences. One hundred independent, random-addition order replicates were produced for taxon definition cutoffs of 2, 3 and 4 bp. For the 2 bp cutoff, the number of MOTU inferred ranged from 143 to 157, with a mode of 151 and a mean of 149.96 ± 2.61. The majority of the variability in MOTU number inferred was due to alternate groupings of a few clouds of sequences (not shown). The use of larger cutoff values also resulted in MOTU sets with wide ranges (~10% of the total number inferred) (figure 5). Thus increasing the fuzziness of the MOTU discriminant does not result in a simple collapse of the clouds of sequences into single taxa. 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.

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 extract the best matches (for example, all matches with a score within a small percentage of the best match), 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 a priori into more- and less-informative sites. Thus, in a protein-coding gene such as coxl, 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 studies. By comparing the barcode sequence with a database of sequences from specimens identified to Linnaean 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 Linnaean Society of London. J.M. and T.C. carried out the meiofaunal surveys as part of their major undergraduate projects.

REFERENCES


