General enquiries on this form should be made to:
Defra, Science Directorate, Management Support and Finance Team,
Telephone No. 020 7238 1612
E-mail: research.competitions@defra.gsi.gov.uk

SID 5 Research Project Final Report

- **Note**
  In line with the Freedom of Information Act 2000, Defra aims to place the results of its completed research projects in the public domain wherever possible. The SID 5 (Research Project Final Report) is designed to capture the information on the results and outputs of Defra-funded research in a format that is easily publishable through the Defra website. A SID 5 must be completed for all projects.
- This form is in Word format and the boxes may be expanded or reduced, as appropriate.
- **ACCESS TO INFORMATION**
  The information collected on this form will be stored electronically and may be sent to any part of Defra, or to individual researchers or organisations outside Defra for the purposes of reviewing the project. Defra may also disclose the information to any outside organisation acting as an agent authorised by Defra to process final research reports on its behalf. Defra intends to publish this form on its website, unless there are strong reasons not to, which fully comply with exemptions under the Environmental Information Regulations or the Freedom of Information Act 2000.
  Defra may be required to release information, including personal data and commercial information, on request under the Environmental Information Regulations or the Freedom of Information Act 2000. However, Defra will not permit any unwarranted breach of confidentiality or act in contravention of its obligations under the Data Protection Act 1998. Defra or its appointed agents may use the name, address or other details on your form to contact you in connection with occasional customer research aimed at improving the processes through which Defra works with its contractors.

Project identification

1. Defra Project code: PH0306
2. Project title: Aphelenchus and related taxa: molecular systematics and molecular diagnostics (Plant Health Fellowship (2))
3. Contractor organisation(s): Food and Environment Research Agency
   Sand Hutton
   York
   YO41 1LZ
4. Total Defra project costs (agreed fixed price): £ 66,000.00
5. Project: start date ..... 01 March 2005
   end date ..... 01 March 2008

6. It is Defra's intention to publish this form.
Please confirm your agreement to do so.....**YES** ☒ **NO** ☐

(a) When preparing SID 5s contractors should bear in mind that Defra intends that they be made public. They should be written in a clear and concise manner and represent a full account of the research project which someone not closely associated with the project can follow.
Defra recognises that in a small minority of cases there may be information, such as intellectual property or commercially confidential data, used in or generated by the research project, which should not be disclosed. In these cases, such information should be detailed in a separate annex (not to be published) so that the SID 5 can be placed in the public domain. Where it is impossible to complete the Final Report without including references to any sensitive or confidential data, the information should be included and section (b) completed. NB: only in exceptional circumstances will Defra expect contractors to give a "No" answer.
In all cases, reasons for withholding information must be fully in line with exemptions under the Environmental Information Regulations or the Freedom of Information Act 2000.

(b) If you have answered NO, please explain why the Final report should not be released into public domain

Executive Summary

7. The executive summary must not exceed 2 sides in total of A4 and should be understandable to the intelligent non-scientist. It should cover the main objectives, methods and findings of the research, together with any other significant events and options for new work.

In recent years there has been a desire to definitively catalogue the life on our planet. In light of the increasing extinction rates that are driven by human activities, it is unlikely that this will be achieved using traditional methods. Whilst most organisms which have a body size of more than 1cm have been described, the vast majority of animal life is smaller than this, collectively known as meiofauna, it is yet to be catalogued. Many of the plant pests of interest both in a statutory context and to agriculture belong to the meiofauna, amongst others this includes nematodes, whitefly, aphids and thrips.

Meiofaunal organisms present a range of problems for traditional taxonomy. Firstly they are microscopic, meaning that morphological features are often difficult to resolve. Secondly these creatures often exhibit cryptic diversity meaning that different species often look the same. Thirdly, it is often the case that the organisms are poorly described in the literature making it very difficult to confirm identification, assuming that someone has already described it. It is possible, however, to obtain DNA sequences from these organisms.

DNA barcoding, the use of short sequences of DNA to identify individuals, is now commonly used in a wide range of applications. It has been proposed that a single target gene should be sufficient to describe all organisms this way. Barcodes can be acquired from individuals or from bulk extractions from environmental samples. In the latter case, many of the sequences obtained are novel and unlikely to ever have a type specimen

associated with them. When this is the case, assessing the diversity of a sample becomes a computational exercise. However, as yet, there is no agreed standard method adopted for analyzing the barcodes produced. Indeed most methods currently employed lack objectivity.

This thesis investigates the efficiency of a range of gene targets and analysis methods for DNA barcoding, with an emphasis on meiofaunal organisms (nematodes, tardigrades and thrips). DNA barcodes were generated for up to three genes for each specimen. Sequences for each gene were analysed using two programs, MOTU_define.pl and DOTUR. These programs use different methods to assign sequences to operational taxonomic units (OTU), which were then compared. An objective method for analysing sequences such as MOTU_define.pl, which relies on only the information contained in the sequences, was found to be most suitable for designating taxa. It does not attempt to apply evolutionary models to the data, and then infer taxa from the derived data.

In addition to barcoding, some samples were pre-processed using video capture and editing (VCE). This creates a virtual slide of a specimen so that a sequence can be linked to a morphological identification. VCE proved to be an efficient method to preserve morphological data from specimens.

The DNA barcoding approaches developed were shown to be efficient methods for describing and discriminating between taxa for several groups of plant pests (nematodes and thrips). Along with VCE (to provide morphological support) the process should provide a rapid and cost effective technique for the identification of plant pests to the species level. The approach is well suited to automation which could further increase throughput and decrease costs.

Project Report to Defra

8. As a guide this report should be no longer than 20 sides of A4. This report is to provide Defra with details of the outputs of the research project for internal purposes; to meet the terms of the contract; and to allow Defra to publish details of the outputs to meet Environmental Information Regulation or Freedom of Information obligations. This short report to Defra does not preclude contractors from also seeking to publish a full, formal scientific report/paper in an appropriate scientific or other journal/publication. Indeed, Defra actively encourages such publications as part of the contract terms. The report to Defra should include:

- the scientific objectives as set out in the contract;
- the extent to which the objectives set out in the contract have been met;
- details of methods used and the results obtained, including statistical analysis (if appropriate);
- a discussion of the results and their reliability;
- the main implications of the findings;
- possible future work; and
- any action resulting from the research (e.g. IP, Knowledge Transfer).

The work presented here is taken from the thesis of Jenna Mann entitled "DNA barcodes and meiofaunal identification" submitted to the University of Edinburgh for the degree of Doctor of Philosophy. The complete thesis is available from the University of Edinburgh or the Food and Environment Research Agency.

Objective 1: Use cultured strains and identified specimens of aphelench to derive molecular sequences for a systematic phylogenetic analysis of Aphelenchida. This analysis will include isolation of sequences from related taxa such as the Bursaphelenchida, and an analysis of the relationships between Aphelenchida and the plant-parasitic Tylenchida.

Objective 2: Analyse several nuclear and mitochondrial loci for sequence divergence between taxa, and thus define which marker(s) are likely to be most useful for molecular diagnosis. The methods for isolation of these "barcode" sequences, and their analysis, will be tested in a range of field isolates

Nematode DNA Barcoding: Comparisons of the performance of different markers

Introduction

Molecular diagnosis of taxonomic affinities of specimens is now a common practice across biological research disciplines. The idea of using molecular barcodes (Floyd *et al.*, 2002; Hebert *et al.*, 2003a, 2003b) for designation of taxa is now widely accepted although the definition of a taxon (or what constitutes a species) is still a matter of discussion. It is now common practice with high-throughput projects routinely producing hundreds if not thousands of barcodes. Initiatives such as the Consortium for the Barcoding of Life (CBOL, http://www.barcoding.si.edu/) produce much data and have also generated numbers of daughter projects such as FISH-BOL (http://www.fishbol.org/) and the All Birds Barcoding Initiative (ABBI, http://www.barcodingbirds.org/). Barcodes have been used for various purposes from forensic identifications (Lorenz *et al.*, 2005) to large-scale environmental diversity surveys (Venter *et al.*, 2004; Sogin *et al.*, 2006). Broadly speaking, two current uses exist for the application of barcodes. Firstly, they can be used for confirmation. Whether regulatory, forensic or investigatory, a barcode is associated with a particular taxonomic scheme. So a novel sequence can be assigned by asking whether it is the same (or within certain limits can be considered the same) to a barcode previously generated. Secondly, barcodes can be inferential. They can be used in specimen independent environmental surveys to assess the molecular diversity of organisms present (Blaxter *et al.*, 2005). Further questions can be asked of these well-differentiated groups, for example, are they of any biological significance i.e. do they represent putative species, or are they populations within species? Failure to delineate species can be a result of a lack of variation among their barcode sequences. On occasion, this can be overcome by more intensive sampling, by increasing the number of individuals sampled from a taxon to *Chapter 3. Nematode barcoding* 68 reveal the intraspecific variation. Given intraspecific variation, it is important to generate sequence from many specimens per taxon so divergent specimens can be correctly allocated. However, there are situations where increased sampling will not aid resolution (e.g. (Trewick), 2008). Geographically and recently speciated taxa may not have accumulated sufficient differences to allow resolution from barcodes (Elias *et al.*, 2007; Wiemers and Fiedler, 2007) and may retain ancestral polymorphism. This situation may be compounded if the species definition used is incorrect (e.g. morphological traits which are unable to differentiate cryptic species). In this case, increased sampling will still result in unresolved taxa as the molecular taxa will not be congruent with the morphological taxa. Meiofauna, organisms with a maximal body axis of less than 1 mm, are hyper-abundant and include nematodes, tardigrades and rotifers (Lawton *et al.*, 1998). Traditional morphological species definitions are difficult to apply to meiofauna due to their microscopic size. Further, they are ubiquitously distributed and some taxa exhibit high levels of molecular diversity (Blaxter *et al.*, 2004). These factors will challenge the efficiency of

molecular barcoding as a tool for meiofaunal identification. Consideration must be given to the gene selected for barcoding. There need to be conserved regions so that universal primers can be used, and variable regions are also required to give taxonomic information. Placing all inferences solely on the results from one gene is unlikely to reflect much more than the gene history. Results from some recent studies proposed that at least two genes should be used, preferably independent of each other (i.e. one nuclear and one mitochondrial) (Elias *et al.*, 2007). Hebert *et al.* (2003a; 2003b) have proposed the mitochondrial marker cytochrome oxidase I (CO1) and this has been adopted by CBOL as the barcoding standard. However its 'universality' is under scrutiny (Lorenz *et al.*, 2005) and it has been suggested that is unreliable for inferring phylogenetic relationships (Moritz and Cicero, 2004; Hurst and Jiggins, 2005; Sonnenberg *et al.*, 2007). As the third base position is less constrained than *Chapter 3. Nematode barcoding* 69 the first and second, variable site saturation can obscure phylogenetic signal. Whilst deep phylogenetic resolution may not be necessary for barcoding, using barcodes to generate trees, which are then used to define species, should be approached with caution. Likewise, relying solely on a nuclear marker (such as large or small ribosomal subunit LSU and SSU respectively) may not provide enough resolution between taxa. When COI was proposed as a standard universal marker, previous work (Johns and Avise, 1998) suggested that a 2% difference was sufficient to discriminate closely related vertebrate species based on results from mitochondrial cytochrome b (*cytb*). Hebert *et al*. (2003b) investigated the COI divergences across species pairs representing 11 animal phyla and found divergences for congeneric species pairs up to 53.7% for COI. Whilst most taxa showed interspecific differences of more than 8% (Hebert *et al.*, 2003b) there were some taxa that exhibited much lower COI differences (e.g. cnidarians showed less than 2% difference between species (Hebert *et al.*, 2003b)); and others that have much greater differences (e.g. amphibians; 10-14% for mantellid frogs (Vences *et al.*, 2005)). The idea of a barcoding gap underpins all molecular barcoding studies. Molecular barcodes rely on differences or similarities of DNA sequences to separate or cluster sequences into taxa. Small differences (less than 1%) are assumed to represent individual differences (intraspecific variation), while larger differences (approximately 2%) are assumed to reflect distinctions between different species (interspecific divergence). In order for barcoding to work, there should be no overlap between the intraspecific variation and interspecific divergence (Meyer and Paulay, 2005). So when two sequences are compared, the amount of genetic distance between them would indicate if they belonged to the same species. For COI data from cowries (Meyer and Paulay, 2005), there was an overlap between intra- and interspecific variation and a threshold could not easily be chosen to define taxa without the risk of over-splitting or lumping taxa. Other taxa also fail to show a barcoding gap (Elias *et al.*, 2007; Wiemers and Fiedler, 2007). Once a dataset has been generated (preferably using multiple nuclear and mitochondrial genes for a complementary dataset), there is no one standard way of assigning sequences to taxa. There is a suite of different analytical methods to develop taxa from. One is to classify sequences into taxa by inferring a phylogenetic tree and choosing taxa using various parameters. Ultimately clades are defined by visual inspection of trees by the investigator. This is a highly subjective method as one person's well-defined clade is another's nested subpopulation. An objective method would designate taxa purely on the information contained within the sequence data. One method used is the basic local alignment search tool (BLAST) to search for most similar named sequences. This method is dependent upon a broad ranging and correctly identified database from which to search from (e.g. GenBank). An issue remains regarding the cut-off to use. Alternatively, sequences can be algorithmically clustered into operational taxonomic units (OTUs) independent of any species designation. In this study we investigated the behaviour of one mitochondrial and two nuclear genes (mtCO1, nLSU and nSSU) for definition of molecular operational taxonomic units (MOTUs) in a sample of terrestrial nematodes. We used these genes to assess the use of two algorithms for clustering the sequence data, MOTU_define.pl and DOTUR. The data were also interrogated for the presence of a barcoding gap.

Results

PCR and Sequencing From the collection of 92 lysates, 53 COI, 82 LSU and 73 SSU nematode sequences were generated before lysates were exhausted (57.6%, 89.1% and 79.3% success respectively). Sequences were only used if the top BLAST hits were nematode sequences. Most SSU sequences (52) had been previously generated by Dr Cutter before the samples were used in this study. These were compared with newly generated sequences (21) to ensure that specimen-order followed the order used by Dr Cutter. The COI primer pair LCO1490 and HCO219 was initially used for PCR, producing 43 sequences. Samples that did not amplify were then tried with C1-J-1718 and C1-J-2191 (generating 10 positive results) however there were 35 samples that failed to amplify with either COI primer pair (other positive PCRs generated non nematode sequences). Despite the support for COI as the universal barcode (Hebert *et al.*, 2003a), these results suggest that COI is not always as readily available for barcoding as previously thought. SSU primer pairs also had variable success, the most successful primer pair was SSU_F_07 and SSU_R_09, generating 69 out of 72 sequences in the complete data set. The three gene sets were filtered to generate a congruent data set where sequences for all three genes had been recovered for the same sample. This consisted of 48 sequences for each gene. The mean sequence length and standard deviation for congruent sequences for COI, LSU and SSU was 533.9 (± 95.7) bp, 558.7 (± 68.1) bp and 430.3 (± 30.2) bp respectively.

MOTU results MOTU_define.pl was used to analyse each congruent gene set separately. The number of MOTUs and clustering behaviour was examined over the range of cut-offs used (Figure 1). For COI, the number of MOTUs defined decreased as the cut-off (number of base pairs) was increased. The mean number of unique sequences, i.e. the number of MOTUs using 0 bp as a cut-off value, was 44.06 ± 0.5. Initially there was a sharp decrease in the number of MOTU defined as the cut-off value was increased to 3 bp (Figure 1). The rate of MOTU definition then decreased as the cut-off value approached 10bp. There was then little change in the number of MOTU defined as the line plateaus. Between 15 and 20 bp cut-offs, the mean number of MOTUs defined was stable at 13 MOTUs (15 bp cut-off defined 13.24 ±0.48, 20 bp cut-off defined 13.00 ±0.48) and 25 bp cutoff defined 12.62 ±0.73 MOTUs. After 30 bp there was another drop in the number of MOTU defined as the threshold allowed less similar sequences to be assigned to the same MOTU. At a cut-off of 60 bp (11.2% of mean sequence length), the number of MOTU defined was 5.75±1.03. At the plateau phase the number of MOTU defined were stable. If this is to be used as an indicator of a barcoding gap, then the members of MOTUs should also be stable. The members of MOTUs were investigated for all resamples at 15 bp cut-off. From the re-sample data, eleven MOTUs were equivalent to the 13 MOTUs defined in the primary run. MOTU0001 only differed once in the re-samples where sequence COI_48 split to form a singleton. MOTU0004 differed in 19 of the re-samples. In eleven of these cases, this was a result of sequence COI_57 forming a singleton MOTU, although sequences COI_84 and COI_93 also formed singletons once and twice respectively. In five of the re-samples, MOTU0004 split to form two MOTUs with differing members.

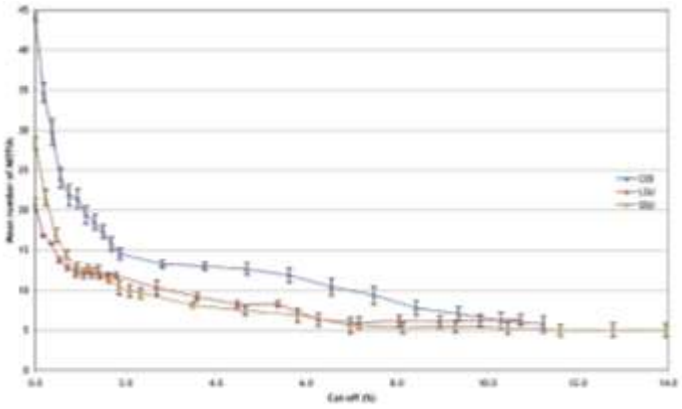


Figure 1 Mean number of MOTUs and standard deviation calculated from 100 re-samples at each cut-off for congruent COI, LSU and SSU gene sets (*n* = 48).

MOTUs defined using a 15 bp cut-off were also strongly supported when a 20 bp cut-off was used. The 20 bp cut-off equivalent of MOTU0004 differed in 14 of the 100 re-samples. The MOTU split simply in 11 of the re-samples and merged with two other MOTUs in three of the re-samples. There was one case where the MOTU had split and one of the MOTUs was joined with four other sequences. The cut-off values of 15 and 20 bp represent the range of thresholds at which the designation of sequences to and the definition of COI MOTUs was stable. This is analogous to the barcoding gap (Meyer and Paulay, 2005). Both the LSU and SSU datasets showed a similar pattern of MOTU definition. That is both genes followed the pattern of COI, where initially there was a sharp decrease, followed by a levelling off of the number of MOTU defined, followed by another decrease. For the LSU, the unique number of sequences (i.e. MOTUs defined at 0 bp cut-off) was 20.69±0.76 and there were 16.85±0.1 MOTUs that differed by 1 bp. This dropped to 6.31±0.66 MOTUs at 60 bp cut-off. The number of unique sequences in the COI dataset was more than double those in the LSU dataset, suggesting that COI has more variation than LSU. As the number of MOTUs at 60 bp was approximately the same (5.75±1.03 for COI), then the variation was due to small changes such as single point mutations rather than largescale insertions and deletions. Like the COI data, the LSU exhibited a range of stable MOTU definition (Figure 1) between the cut-off values of 5 and 10 bp. The mean number of MOTUs fell from 12.19±0.6 to 11.83±0.36 over 6 bp. At 8, 9 and 10 bp, where the mean number of MOTUs was 11.8 with standard deviations of 0.37, 0.35 and 0.36 respectively, the members of the MOTUs were investigated. The majority of the re-sample MOTUs were equivalent to those designated in the primary LSU run. Differences were due to MOTUs joining. MOTU0002 merged with MOTU0008 in four of the re-samples and with MOTU0006 in 12 of the re-samples.

The LSU sequences also clustered in the same way as the COI sequences. For example, LSU_8bp_MOTU0003 was equivalent to COI_15bp_MOTU0010. The difference in the number of MOTUs designated between COI_15bp and LSU_8bp was due to LSU equivalents of COI_15bp_MOTU0007 (sequences 02, 03 and 04), and COI_15bp_MOTU0003 (05, 06 and 11) forming LSU_8bp_MOTU0006 containing all six sequences. A cut-off value of 30 bp was required before these six COI sequences were defined as a robust MOTU. At first glance, the results for the SSU data appeared similar to COI and LSU. There were 28.34±0.75 unique MOTUs and the number of MOTUs decreased as the cut-off value was increased in a similar fashion to COI and LSU, with the same sharp initial decreases levelling off (Figure 1). At the plateau phase, between 4 and 6 bp, the mean number of MOTUs ranged from 12.75±0.61 to 12.60±0.47. As with COI and LSU, most SSU MOTUs were stable at. Only MOTU0009 and MOTU0003 were weakly supported. The re-samples showed that the three sequences were just as likely to form one MOTU or a MOTU of SSU_51 and SSU_52 or SSU_52 and SSU_53 and a singleton of SSU_53 or SSU_51. However, although the number of MOTUs defined at the plateau phase was the same for COI and SSU, the membership of the MOTUs was not equivalent. The COI data (and to a certain extent the LSU data) defined MOTU0001, MOTU0006 and MOTU0009 at 15 bp cut-off. In the SSU results, the member sequences of

these MOTU were clustered very differently. This was not due to experimental error as for some specimens, where independently derived sequences were available, the same clustering was observed.

DOTUR results

DOTUR (Schloss and Handelsman, 2005) was used to analyse the three gene sets for comparison with MOTU_define.pl results (Figures 2-4). There were only minor differences in the number of DOTUs defined using either Kimura “2-paramter” or Jukes-Cantor distances. DOTUs generated with Jukes-Cantor distances were compared with MOTU results. For the COI dataset, FN clustering defined more OTUs than NN clustering as expected (Figure 2). At 0% cut-off, there were 45 furthest neighbour DOTUs (fn-DOTUs) and 44 nearest neighbour DOTUs (nn-DOTUs). As with MOTU results, the number of fn-DOTUs and nn-DOTUs decreased rapidly as the cut-off increased. Unlike MOTU, DOTUR does not generate a plateau phase. Cut-offs were only reported where there was a change in the number of DOTUs. The decrease in the number of DOTUs slowed after 3% up to 12% cut-off where NN clustering had been used and 15% for FN clustering. The number of nn-DOTUs and fn-DOTUs deceased by three over 1% of the mean sequence length at the respective cut-offs (Figure 2). There were more COI fn-DOTUs than COI MOTUs. At higher cut-offs (above 5.5%), there were more nn-DOTUs than MOTUs. Below this threshold, NN clustering defined similar numbers of DOTUs to MOTUs. In comparison with COI DOTUR results, there were fewer LSU DOTUs found (Figure 3). There were 24 fn-DOTUs and 23 nn-DOTUs at 0%. The number of DOTUs fell rapidly as the cut-off was increased to 2%. Unlike the COI DOTUs (and SSU), the number of fn-DOTUs and nn-DOTUs were very similar (Figure 3.6). Again, there was no plateau phase. However, between 3.2-8.5% and 4.1-8.9% (for nn-DOTUs and fn-DOTUs respectively) the number of DOTUs for both clustering methods decreased by one.

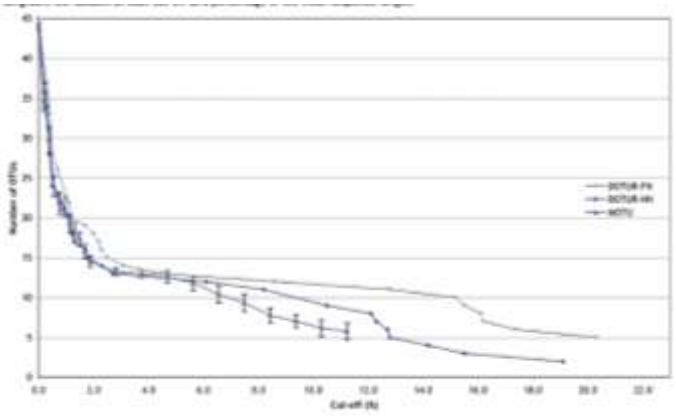


Figure 2: The number of COI OTUs defined using furthest (FN) and nearest (NN) neighbour clustering in DOTUR and MOTU_define.pl for the congruent COI dataset at each cut-off as a percentage of the mean sequence length.

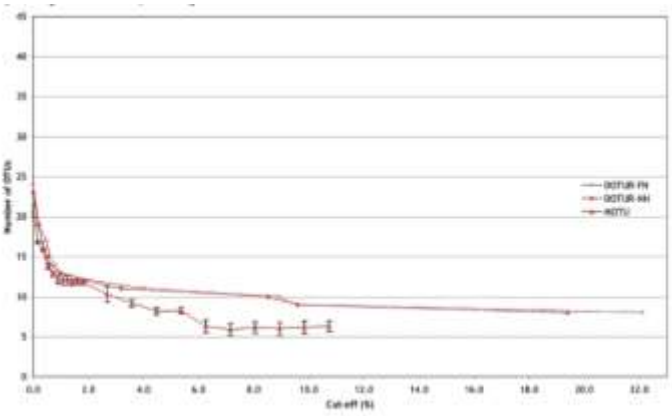


Figure 3: Number of LSU OTUs defined using FN and NN clustering in DOTUR and MOTU_define.pl for the congruent LSU dataset at each cut-off as a percentage of the mean sequence length.

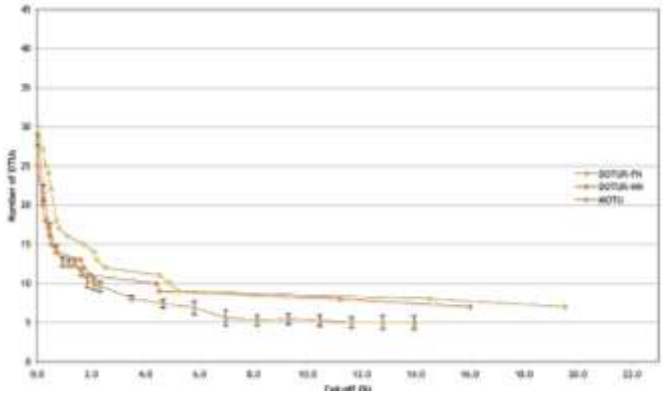


Figure 4: Number of SSU OTUs defined using FN and NN clustering in DOTUR and MOTU_define.pl for the congruent SSU dataset at each cut-off as a percentage of the mean sequence length.

At 0% cut-off, there were 30 fn-DOTUs and 25 nn-DOTUs (Figure 4). This was the only dataset where there were more SSU MOTUs found at 0% cut-off than nn-DOTUs. There was a rapid reduction in the number of DOTUs until 1% cut-off. This initial decrease phase was much shorter in the SSU results compared with the COI and LSU results. Unlike COI and LSU, the SSU data exhibited two phases where the proportional decrease (i.e. where a small change in the number of DOTUs occurred over a large range of cut-off values) in the number of DOTUs over the range of cut-offs was small for both fn-DOTUs and nn-DOTUs (Figure 4). Membership of sequences to DOTUs was investigated where the proportional decrease in the number of DOTUs was smallest for FN and NN clustering. For COI and LSU DOTUs, NN and FN clustering had only one phase where the number of DOTUs decreased by one over a range of cut-off values. The SSU data indicated two phases, so membership of sequences to DOTUs at both phases was investigated. As with MOTUs, COI and LSU DOTUs are congruent at the beginning of the phase but cluster differently. Not only are SSU DOTUs different to COI and LSU DOTUs, but NN and FN clustering also result in different orders and defined different DOTUs.

MOTUs versus DOTUs The clustering behaviour of COI sequences were broadly similar between MOTU_define.pl and DOTUR. The main difference was DOTUR grouped sequences from MOTU0008 and MOTU0009 at the phase of smallest proportional change. This did not occur in the MOTU results until a 35 bp cut-off was used, and then the sequences were members of a MOTU containing many more sequences. The first difference between LSU MOTUs and DOTUs, was the definition of one DOTU in comparison with two MOTUs. This was where the number of OTUs was most stable. The merged MOTU was not supported by MOTU_define.pl until a cut-off of 20 bp (3.6%) was used. The second difference in LSU DOTUs was the joining of sequences LSU_41 and LSU_54 with LSU_64. MOTU_define.pl did not support this join. Sequence LSU_64 joined with LSU_08, LSU_24 and LSU_32 before LSU_41 and LSU_54 at 35 bp cut-off. Overall, the SSU MOTUs defined were similar to either SSU nn-DOTUs or fn-DOTUs. Again, SSU OTUs did not form equivalent OTUs in comparison with the COI or LSU results.

Tree comparisons of all sequences

Unrooted NJ phylograms were constructed for the complete gene sets. Branch tips were compared with mean MOTU designation from the plateau phase of each gene.

The addition of the extra sequences to each gene set did not alter the COI and LSU MOTUs and there was similarity in branching patterns. There were 48 sequences in the congruent data set for each gene, so the addition of five COI sequences may not be expected to dramatically change the COI MOTU designations. However, the LSU data set was increased from 48 to 82 sequences, yet the clusters did not change. As seen in the MOTU and DOTUR results, SSU clusters differed from COI and LSU clusters. COI and LSU clusters separate sequences 70 and 72, where as for the SSU results show that the sequences are the same. Most SSU pairs

of duplicated sequences were found to cluster close together (grey sequence names. There are two exceptions. The first is JM_SSU_41 and AC_4-a-4 which were generated from the same specimen, but are separated on the tree (boxed sequence names. JM_SSU_41 is 316 bases compared to 457 bp for AC_4-a-4 and 459 bases of AC_4-a-2 (Appendix 3.2). Since the tree is based on absolute differences including end gaps, significant differences in length are likely to separate sequences on the phylogram. The second, even more striking exception, were sequences JM_SSU_32 and AC_8-a-4. These sequences were generated using the same primers, sequenced using the same primer, are 416 and 382 bases respectively, yet are on opposite branches. The plausible explanation for this surprising pattern is due to mislabelling of this particular specimen. Great care was taken to ensure correct specimen order during amplification and sequencing. However there was no monitoring of specimens before they were received as lysis samples in a 96-well plate. Without being able to trace the complete history of individual samples, this needs to be considered as the most likely cause of this result.

Discussion

PCR and Sequencing success

The results clearly indicate that LSU is more reliably recovered from this sample set than SSU and especially COI sequences. Differences among the number of positives were not due to a lack of DNA as PCRs were successful for the other genes. There was no systematic bias of the recovery COI sequences. Comparing LSU taxonomic designations (from BLAST hits), there was no relationship between the sub-order of the LSU sequence and a positive COI PCR (Appendix 3.2). If COI barcodes are not universally recoverable from all specimens, then there is little to be gained by basing barcode system on COI. Specific primers are sometimes required to amplify COI depending on the taxa, for example spider mites (Ros and Breeuwer, 2007). If this is found to be true across many taxa, then using COI as a standard barcode would be unsuitable, where the specimen to be analysed had not already been classified to a taxonomic group. The LSU primers worked extremely well with a sequence generation success rate of over 90%. These primers have also worked well on other taxa (e.g. tardigrades, copepods (Mann, unpublished)). If a specimen was rare then it would be more appropriate to use a target that has a high probability of being recovered with PCR. Although this is not so much a problem for large organisms, which can be re-sampled easily, when the whole animal is degraded for lysis, as in meiofaunal barcoding, there is a limit to the number of PCRs one is able to perform. This may also be the case for larger organisms where obtaining DNA is problematic, for example when the specimen is old and DNA is degraded, or when it is not easily re-sampled because the specimen is rare.

MOTU_define.pl

For the congruent dataset, there were more COI MOTUs defined than LSU and SSU MOTUs. There were only similar numbers of MOTUs for each gene, *Chapter 3. Nematode barcoding* 99 when the cut-off value was between 10-11% (Figure 1). At low cut-off values (less than 1.5%), there were more SSU MOTUs than LSU. Increasing the cut-off value above 1.5% results in more LSU MOTUs than SSU MOTUs. If SSU has more MOTUs at low cut-off values, then it may be a more accurate predictor of individual variation. At higher values, the LSU barcodes exhibit more deep-seated differences, which may relate to differences between species.

DOTUR

Within each gene set, FN clustering generated more DOTUs than NN clustering. As with the MOTU data, there were more COI DOTUs than LSU or SSU DOTUs, up to a cut-off value of 12% and 15% for NN and FN clustering respectively. The number of COI nn-DOTUs and fn-DOTUs decreased by four OTUs over approximately 1% difference. This was a striking feature of the COI DOTUs only, and was not reproduced in either of the nuclear gene sets. In the absence of any morphological species identification, it is difficult to identify the root of this

feature. It is unlikely to be sequences segregating into species groups at the high cut-off value. It is more probable to be indicative of the grouping of species into genera. DOTUs generated from the LSU data exhibited minor differences between the FN and NN clustering and the membership of sequences to DOTUs was very similar. The decrease in the number of LSU DOTUs was smaller, and occurred over a larger range than COI DOTUs. This would indicate that different LSU lineages were more isolated than the COI sequences. If LSU lineages were similar, the number of DOTUs would be expected to fall more sharply as they would be assigned to the same DOTU at lower cut-off values. Whereas LSU DOTUs were similar both in the number generated and in the membership of sequences to a particular DOTU, SSU DOTUs differed in both aspects. FN clustering defined an average of 5 more DOTUs than NN clustering up until a 2% cut-off. This difference was greater than that seen between COI and LSU nn-DOTUs and fn-DOTUs. Moreover, membership of sequences differed in nn-DOTUs when compared to fn-DOTUs. It was 100 expected to have more fn-DOTUs than nn-DOTUs due to the stringent rules for assigning sequences. It was also expected that LSU and SSU would have similarities as both are nuclear genes, therefore subject to the same selection pressure, and are inherited in tandem.

MOTU vs. DOTUR

Results of MOTU and DOTUR were generally similar. The number of DOTUs for each gene set was more than the corresponding MOTUs. These differences are probably due to the multiple versus the pairwise alignment of DOTUR and MOTU_define.pl respectively. In a multiple alignment, differences in length can lead to sequences being defined in different DOTUs. Therefore, DOTUR is likely to overestimate the number of taxa at any one cut-off. Another issue with DOTUR is the way in which the cut-offs are reported only when the number of DOTUs changes. There is an obligate decrease in the number of DOTUs as the cut-off is increased. Where as MOTU_define.pl generates MOTUs at each cut-off, it is possible to see a plateau in the number MOTU, DOTUR does not. Working out the flattest part of the line for DOTUs could highlight a potential plateau, as during that phase the number and membership of DOTUs does not change. Different MOTUs could be generated depending on the addition order of sequences, though running multiple re-samplings could highlight robust MOTU. As DOTUs were ultimately based on the same multiple alignment, whatever clustering method was used, DOTUR should generate the equivalent DOTUs. In COI and LSU datasets, FN and NN clustering formed equivalent DOTUs, albeit at different cut-off values. However, fn- and nn-DOTUs were incongruent in the SSU data. If one clustering method produces different OTUs from another, both using the same alignment, without any means of investigating the robustness of either result, which one is to be believed?

Complete data comparisons

The inclusion of all the sequences in the analysis did not alter MOTUs defined using the congruent data set. New MOTUs were generated from the additional sequences in the LSU data set.

What gene(s) to use?

The most OTUs defined were from the COI gene set, regardless of which method is used. However, the overall sequence success rate was the lowest in this investigation. The LSU gene set had the greatest success rate. Moreover, LSU OTUs were very similar to the COI OTUs. This was unexpected as COI is a mitochondrial gene and LSU is a nuclear gene. If LSU OTUs are an accurate predictor of COI clustering behaviour, and sequences are more reliably recovered from samples, then LSU would be an improved proposition as a universal barcode target. As demonstrated by these results, it should not be used in isolation. Whilst the COI and LSU OTUs were broadly similar, both were different from the SSU OTUs generated. It is surprising that the SSU and LSU results are incongruent (as they are found on tandem repeats), but this case highlights the necessity to consider more than one target for barcoding.

Objective 3: Devise bioinformatic tools for storage and retrieval, and analysis, of molecular "barcode" sequences derived from multiple genes, and apply it to identification of anonymous specimens.

Thrips : identifying taxa within DNA barcode data

One of the problems encountered when trying to identify small organisms is their size, whereby key morphological features can be easily missed or misidentified. The vast majority of life on Earth has a body axis of less than 2 mm and many of the key morphological structures may be smaller than light microscopy resolution (De Ley and Blaxter, 2002). Even when a specimen is morphologically intact, it may be impossible to identify it to species in a quick, easy way. Invertebrate organisms are the most numerous animal taxa on Earth with many species poorly described. Those that have been recorded are likely to be a very small proportion of the true diversity (Blaxter, 2003).

Thrips belong to the order Thysanoptera (phylum Arthropoda, class Insecta) and are small winged insects ranging from 0.5 mm to 15 mm in length (Gullan and Cranston, 2005). Worldwide, there are over 5000 recorded species of thrips (Crespi *et al*., 1996; Brunner *et al.*, 2004; Inoue and Sakurai, 2007). Traditionally adult morphological characters and host plants have been used to classify two suborders; Terebrantia consisting of eight families, and Tubulifera with a single family (Brunner *et al.*, 2004; Crespi *et al.*, 1996).Although the monophyly of Thysanoptera is supported by morphological and molecular evidence (Brunner *et al.*, 2002; Inoue and Sakurai, 2007), the relationships among the nine families in the sub-orders are unresolved using morphological traits (Mound *et al.*, 1980). Around 100 species have been identified as pest-species (Brunner *et al.*, 2004) according to feeding behaviour, causing damage to plants, or as disease vectors (Toda and Murai, 2007). Easily over-looked, these small organisms have been transported across the world following trade routes for vegetables and ornamental flowers (Brunner *et al.*, 2004). Removing a species from its native habitat into a novel environment can lead to the species becoming invasive and damaging its new local environment (Brunner *et al.*, 2004; Toda and Murai, 2007). Although thrips can damage crops, different species are of varying importance to agriculture. In particular, *Thrips tabaci* Lindeman is a polyphagous species as well as being a vector of tomato spotted wilt virus (TSWV) and is therefore of economic importance (Brunner *et al.*, 2004; Toda and Murai, 2007). Different species vary in their importance to us, particularly pest species that are subject to quarantine regulations (Armstrong and Ball, 2005) and therefore need to be correctly identified. Traditionally, thrips identification uses keys mainly based on adult morphology from type specimens (Mound *et al.*, 1980). Identifying specimens this way relies on experts and takes time (Tautz *et al.*, 2003). If a specimen is a larva or damaged, important morphological structures used for identification are likely to be absent. When sampling, it is highly unlikely that all specimens will be adults (Hosseini *et al.*, 2007), making identification keys insufficient at identifying all specimens sampled. A quick and reliable method to identify specimens, regardless of condition, size or life stage, without the need for taxonomic experts, would facilitate routine identification and quarantining of pest thrips species. *Molecular Barcoding* The introduction of molecular diagnostic tools has aided the accuracy and speed of species identification. Molecular barcoding, using PCR to obtain short DNA sequences to identify specimens, has two applications. Firstly, it can be used to identify an unknown specimen by comparison of a short DNA sequence to a comprehensive data set of sequences from identified species. Alternatively, DNA barcoding can be used to aid species discovery (Meyer and Paulay, 2005). Initially molecular barcodes were used to identify particular species

(Gasser *et al.*, 1994) and methods used were often restricted to the study. More recently, however, in the widening gap between diagnostic needs and trained taxonomists (Armstrong and Ball, 2005; Tautz *et al.*, 2003) molecular barcoding for species confirmation and species discovery has become widespread, and the need for a universal approach has long been recognised. Intraspecific (within species) variation in DNA sequences is expected to be small. Differences between species (interspecific variation) should be greater and relate to the length of time of divergence between species. Although the magnitude of both intra- and interspecific variation will vary depending on the study taxon, in a perfect barcoding world there should be no overlap between the two (Meyer and Paulay, 2005). This gap is referred to as the 'barcoding gap' and separates the coalescent of individual variation from the birth-death process of species' intraspecific divergences. Of course there are several situations where a barcoding gap may not exist. For example, where two distant populations of a species are genetically distinct due to limited gene flow, barcoding would incorrectly indicate a gap (Wiemers and Fiedler, 2007). In addition to this false positive problem, false negatives may be found in barcoding where no barcoding gap is found, e.g. when there is little sequence variation found in the barcoding gene. This could be true for very closely related species where ancestral polymorphism is still segregating or hybridisation is maintaining identity in both species (Trewick, 2008). Initial work by Johns and Avise (1998) on mitochondrial cytochrome *b* demonstrated that different vertebrate classes showed different levels of variation when genera within a class were compared. Amphibians, reptiles and fish showed large distances when compared to mammals and birds (which have the smallest distances) (Johns and Avise, 1998). The study suggested also that a mean difference of more than 2% between sequences would be sufficient to distinguish between vertebrate species. The mitochondrial gene cytochrome oxidase *c* subunit I (COI) has been proposed as the standard for molecular barcoding (Hebert *et al.*, 2003a). Early studies suggested that it could consistently and faithfully recover species based on differences in sequences (Hebert *et al.*, 2003b). Recent studies found a similar pattern to Johns and Avise (1998). Hebert et al. (2004a) used 437 COI sequences representing 260 bird species and found the average intraspecific difference was 0.43%. However, Vences *et al.* (2005) found intraspecific variation for mantillid frogs and salamanders was as high as 10- 14% and 7.8% respectively. At this level, intraspecific variation overlaps with interspecific variation so that species delineation was difficult. These studies all investigated well-defined and described vertebrate species. Invertebrates, on the other hand, are numerous, often have large effective population sizes and high speciation rates (Elias *et al.*, 2007). These two factors are likely to inhibit the usefulness of barcodes to identify invertebrate species with either too much or too little variation respectively. Initial work on invertebrate barcodes intimated that 2% difference in COI sequences would separate species. Work by Hebert *et al*. (2003a) on COI analysis of eight insect orders and 200 lepidopteran species suggested that 2% was capable of delimiting species. However, later work on other butterfly species (Elias *et al.*, 2007; Wiemers and Fiedler, 2007) failed to show such confidence in the ability of barcodes to identify species. Moreover, a recent investigation of New Zealand grasshoppers (order Orthoptera) failed to find any barcoding gap or matches between molecular and accepted morphological taxonomy (Trewick, 2008). Identification of Dipterans using DNA barcodes was also problematic (Meier *et al.*, 2006). The main issue of earlier studies has been the insufficient sampling of taxa (Trewick, 2008). If only one or two individuals were sampled within a species (Hebert *et al.*, 2003a), then it would not be possible to estimate the range of intraspecific variation and the likelihood of creating a false barcoding gap increased (Meyer and Paulay, 2005; Trewick, 2008). As well as the issues discussed

above, there are other potential drawbacks with COI. Primarily, it is maternally inherited and so can only ever reflect maternal evolution (Rubinoff, 2006). There is also evidence to suggest inherited symbionts can affect the variation of the mitochondrial genome within a species. Moreover, there are technical issues relating to the 'universal' nature of primers for the mitochondrion gene target. Where variability is high, multiple primers are required to recover COI targets from all specimens (contradicting results from Hebert *et al*., 2003a) and taxon specific primers must be designed. COI is also more difficult to amplify from some specimens and has a lower recovery rate when compared with other genes, such as the nuclear ribosomal large subunit (LSU or 28S). *Turning sequences into taxa* If COI was to be adopted as the standard for barcoding, there should also be a standard method for derivation of taxa from barcodes. Sequences may be compared by a simple BLAST search (Altschul *et al.*, 1997) using the similarity score to define taxa. Distances between sequences may also be used to construct phylogenetic trees to delineate taxa, using branch length as a measure of relatedness. This method lacks objective criteria to designate taxa as clades are defined by eye. Pons *et al*. (2006), delineated specimens of tiger beetles by identifying putative species based on branch lengths of clades, from a mtDNA phylogenetic tree. Using a molecular clock to 'best-fit' the data, a change in the rate of branching was assumed to be indicative of a species boundary. This method does not require species or populations to be defined a priori but it was assumed that each geographic sampling site represented a separate population unless morphological differences were observed and that the sampling regime was a good reflection of the total diversity (Pons *et al.*, 2006). However, only using mtDNA (in effect, a single locus) would fail to distinguish recently diverged lineages (Elias *et al.*, 2007 2008) or recently derived geographic populations (Pons *et al.*, 2006). If this method was to be used as a way to infer species from barcodes (Pons *et al.*, 2006) then taxa need to be extensively sampled. Pons *et al.* concede that "the extent of population sampling...may rarely be complete" (2006). Therefore, this method is unsuitable when trying to cluster high throughput barcode data that are likely to contain sequences from incompletely sampled species, isolated clades or populations that share gene flow (Lohse, 2009). *MOTU_define.pl* MOTU_define.pl (M. Blaxter, J. Mann and R. Floyd, unpublished, see http://www.nematode/org/bioinformatics/ for download) has been developed to cluster sequences into molecular operational taxonomic units (MOTUs) independent of phylogenetics. During MOTU_define.pl analysis, a sequence is compared pair-wise to all other sequences in the data set in a random order (the primary run). The program generates a local database of previously defined MOTU. MOTU_define.pl will ask if a query sequence matches any other sequences in the database with less than *x* b difference (the cut-off value) over more than a minimum overlap of *y* bases (Blaxter *et al.*, 2005). If the sequence matches a previously defined MOTU by less than the cut-off and along the overlap, then it is assigned to that MOTU. If it does not match, it forms a new MOTU and is given the next sequential MOTU identifier. A new sequence is then picked at random, and the process repeated. The user sets the cut-off and minimum overlap. Membership of a MOTU as defined by MOTU_define.pl, can be affected by the order in which the sequences are added (Blaxter *et al.*, 2005). Therefore, re-sampling is important to investigate the variability of MOTU classification at any cut-off. MOTU_define.pl does not establish the relatedness of MOTUs (although this can be investigated by observing changes in membership over different cutoffs). It can deal with isolated species, populations sharing gene flow and incomplete sampling of populations. Moreover, it is incremental, so that new data can be added to previously defined MOTUs without the need to start analysis from the beginning. *DOTUR* Defining Operational Taxonomic Units and Richness (DOTUR) is another method for assigning sequences to

defined operational taxonomic units (here called DOTUs) (Schloss and Handelsman, 2005). DOTUR defines taxa by clustering sequences based on distances derived from an alignment of the sequences. DOTUR uses three methods for clustering sequences, furthest, nearest and average neighbour. Furthest neighbour clustering only adds a sequence to a DOTU if it is sufficiently similar to all other member sequences in the DOTU, otherwise it will seed a new DOTU. Nearest neighbour adds a sequence to a DOTU if it is similar to any sequence in it. This means that if a DOTU has many sequences, the difference between the two most distant sequences within it could be quite large. Nearest neighbour clustering would be expected to define a similar number of DOTUs to MOTUs as both use a similar method to assign sequences to operational taxonomic units (OTU). Average neighbour clustering is an intermediate between the two methods. As well as different clustering methods, it is also possible to construct DNA distance matrices using different models of evolution. The Jukes-Cantor model assumes that there is no difference between transition and transversion rates when comparing sequences. COI is a protein-coding gene and thus there are likely to be differences between transversional and transitional rates, as modelled by the Kimura "2-Parameter" model.

What cut-off should be used to define taxa?

A difference of 2% between sequences has been proposed as the cut-off for defining species (Hebert *et al.*, 2003a). However this universal cut-off is not suitable for all animals as inter- and intraspecific distances vary among genera (DeSalle *et al.*, 2005; Vences *et al.*, 2005). Rather than assuming 2% is sufficient as a cut-off value, both MOTU_define.pl and DOTUR allow the user to investigate the clustering of sequences over multiple cut-offs, and so reveal a barcoding gap if one is present. Ideally, multiple lines of evidence, such as multiple genes with different modes of inheritance (Elias *et al.*, 2007), and more than one analysis method should be used to support the morphological and sequence-based definition of species. In this investigation, a large dataset of partial COI sequences was used to assess the ability of MOTU_define.pl and DOTUR to define thrips taxa based on sequence data. The dataset consisted of 332 specimens that had been morphologically identified and then sequenced for the 5' region of COI. The sequence data were analysed, independent of morphological designation, to define OTUs. The OTUs were then compared with morphological species designation (morpho-species) to test the ability of barcode OTU methods to recover morphologically identified taxa. A phylogenetic analysis was also performed as an independent check of OTUs defined.

Discussion

MOTU_define.pl

MOTU_define.pl clusters sequences based on similarity from pair-wise comparisons. As the cut-off increases, the probability that a sequence will match any member of a MOTU increases, and the number of MOTUs declines. Initially there was a sharp drop in the number of MOTU defined, but as the cut-off increased, the fall in the number of MOTU decreased (Figure 4). MOTU_define.pl generated a plateau in the number of MOTU defined over a range of cut-offs, from 4.4% to 6.6%, where the mean number of MOTU dropped from 44 to 43 (Figure 4). Importantly, this plateau is expected to indicate cut-offs at which the number (and members of) MOTUs were stable, i.e. a barcoding gap. Previous work by Morris and Mound (2001) shows that intraspecific distances of some thrips species can be higher than this. This barcoding gap is only visible if a range of cut-off values is used to investigate the clustering

behaviour of sequences. Re-sampling the data set allowed the robustness of those clusters to be verified. Using MOTU_define.pl, and using a range of cut-offs and multiple re-samplings, allows a user to interrogate the data fully.

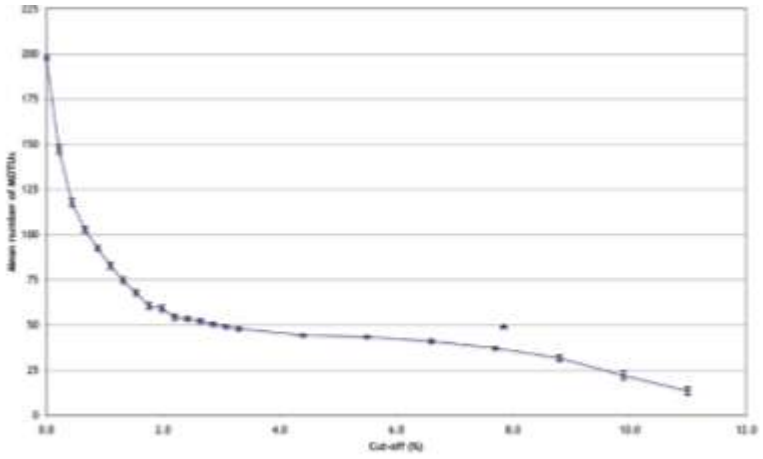


Figure 4: Mean number of MOTUs generated using MOTU_define.pl, cut-offs are given as a percentage of the mean sequence length. Means and standard deviation calculated from 100 re-samples.

Monospecificity of MOTUs

MOTU_define.pl is agnostic of any species designation. When species identifications were reunited with sequences, MOTUs defined tended to be made up of only one species. Whilst the maximum proportion of complete monospecific MOTUs was attained at the same values as the barcoding gap, it did not reach 1. Although the difficulties in identifying small organisms are well documented, it is worthwhile to note that there were only two species that never form monospecific molecular taxa. The ability of traditional taxonomists to define species based on morphological characteristics that has a high level of concordance with molecular taxa must be acknowledged. However, this dataset consists of only a small proportion of all thrips species and is not an exhaustive global sample. MOTU_define.pl can be applied to any level of a phylogenetic tree depending on the cut-off used. A low cut-off value may be suitable for identifying species although it may be confounded by intraspecific variation. However, genera within families are more difficult to delimit as they can be of different ages. Older genera will be segregated more easily than younger genera which are closely related and monophyly would be reached at different cut-off values.

DOTUR

DOTUR uses a DNA distance matrix generated from a multiple alignment of sequences to report the distance at which the number of DOTUs changes. Consistently, the number of DOTUs reduced with increasing distance and no clear plateau was formed (Figure 5) whichever distance or clustering method was used. There is no objective way to reveal a barcoding gap. Unless a suitable cut-off has previously been defined then DOTUR is less suitable for taxa identification than MOTU_define.pl. There were very minor differences between the Jukes-Cantor and Kimura “2- Parameter” results (Figure 5). The Kimura-“2-Parameter” models two rates of evolution based on the dataset. When genetic distances within the dataset are low, these two rates are approximately equal, so the model approaches the Jukes-Cantor model. When comparing the same gene across organisms, this is likely to be true as orthologous sequences are subject to similar constraints and therefore likely to be highly conserved. If this limits the phylogenetic usefulness of COI sequences, performing analysis based on complex models of evolution becomes redundant. Therefore it is valid to

use the Jukes-Cantor single-rate model in comparison with MOTU_define.pl results. Even though using furthest neighbour clustering is likely to overestimate the number of DOTUs in the data set, nevertheless this particular method is currently the most stringent as it requires all sequences in the DOTU to be within the similarity cut-off. Nearest neighbour clustering is more conservative in the estimation of DOTUs. This method defines OTUs using sequence similarity to any of the sequences in a DOTU. This is the most similar to the way MOTUs are defined.

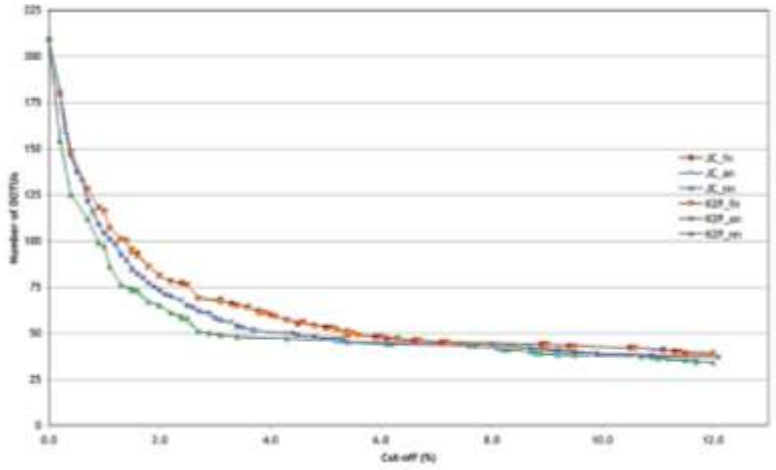


Figure 5: Number of DOTUs generated using Jukes-Cantor (JC) and Kimura "2-Parameter" (K2P) distance matrices for all three clustering methods (furthest, nearest and average neighbour).

Monospecificity of DOTUs

DOTUR did not fully sort sequences into monospecific taxa. As with MOTU results, *T. fuscipennis* and *T. sambuci* were not monospecific in DOTUR over the range of cut-offs investigated. However, these two species do not account for the missing monospecific taxa. The peak proportion of monospecific DOTU found was higher than for MOTU. DOTUs are not influenced by addition order (unlike MOTUs) so the fluctuations (at 0.7% and 2.7%) in the proportion of monospecific nn-DOTUs are not easily explained.

MOTU vs. DOTUR

The main difference between MOTU_define.pl and DOTUR is the method used to designate taxa. MOTU_define.pl uses BLAST to form pairwise comparisons between all sequences in the data set. DOTUR requires a full alignment of all the sequences, followed by a derived distance matrix, which uses two models of evolution to the data, first in the multiple-alignment and secondly in the construction of the DNA matrix. It may be assumed that these models accurately reflect the evolutionary history of the sequences. If the alignment is poor at the ends (where sequence quality can be lower) or sequences are of different lengths, DOTUR interprets these as real and generates DOTUs accordingly. For this data set the sequences were of varying lengths and this may explain the higher numbers of DOTUs versus MOTUs. Even when DOTUs were generated using nearest neighbour clustering, there were still more DOTUs. It is likely that using DOTUR for species identification from barcodes will overestimate the number of taxa, where as MOTU_define.pl will be more conservative. In the case of COI barcodes, it may be inappropriate to investigate taxon composition using methods driven by an evolutionary hypothesis. Although they can easily identify "species groups", it is not possible to resolve deep phylogeny with these sequences (Meyer and Paulay, 2005). DOTUR models the nodes of the phylogeny. The problems with using COI as a barcode are well reported. Not only is it unreliable to extract, the ability to resolve closely related species is under scrutiny (Hurst and Jiggins, 2005). When using barcodes for species identification, the sequence in question

needs only to be compared to an identified sequence and asked does it match a known sample or not? Species delimination from sequence is a more complex process as the differences between sequences are considered to have some sort of biological significance that relates to the difference between species, such as a difference in colour. In this case, it would be prudent to incorporate models of evolution into analysis where the hypothesis being tested is how one species relates to another. Both MOTU_define.pl and DOTUR grouped *T. fuscipennis* and *T. sambuci* in the same OTUs and neither found them to be monospecific. Although morphologically these have been designated has separate species, the molecular data suggested they belong a single OTU. The defining feature between the two species is antennae colouration, which is the final couplet of the identification key (D. Collins, personal communication). The colouration (or lack of it) may separate the two species morphologically, but does not show any molecular difference when looking at COI barcodes.

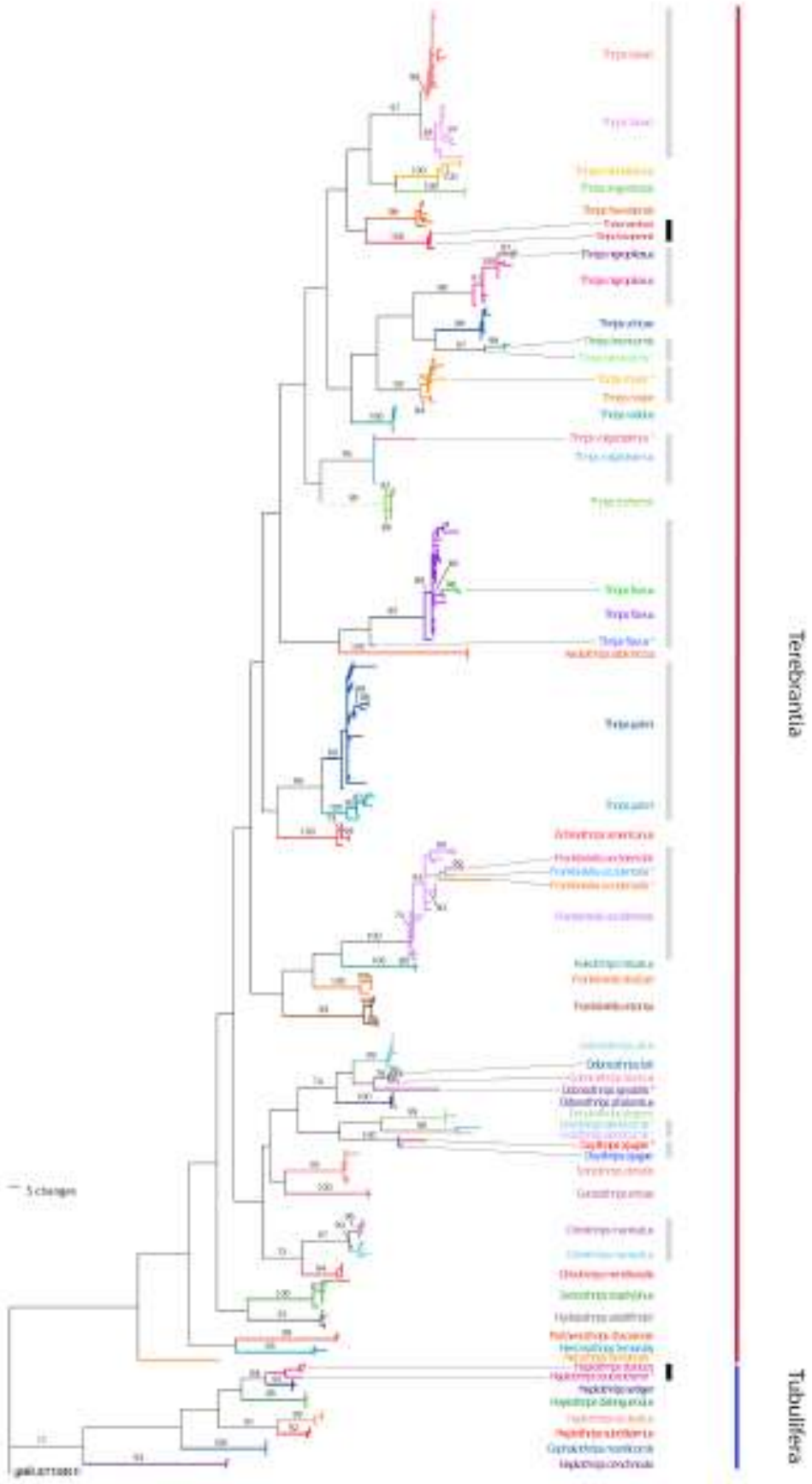
Is a cut-off of 2% sufficient to discriminate thrips species?

Some proponents of barcoding have suggested that 2% difference of COI sequences will be sufficient and reliable to discriminate species (Hebert *et al.*, 2003b). At this cut-off, all the sequences from a single morpho-species should be in a complete monospecific OTU. In addition, all OTUs should be monospecific, containing only one morpho-species. However, both DOTUR and MOTU_define.pl found maximal proportions of complete monospecific OTUs at higher cut-offs, 4.4% and 8.1% respectively. These results for interspecific divergences are lower than previous work on thrips COI sequences. Mound and Morris (2001) found that intraspecific differences were 6.1% and 8.4% for COI sequences of two thrips species, and divergence between the species was 14%. If these any of these divergence values hold for the majority of thrips species, then using 2% will illustrate variation associated with populations as species boundaries. It would be unrealistic to expect that any one cut-off would generate OTUs which were completely consistent with morpho-species.

Significance of barcodes for thrips

The sub-orders Terebrantia and Tubulifera were segregated by phylogenetic analysis (Figure 6). MOTU_define.pl and DOTUR also separated the two sub-orders. Within the Tubulifera, *H. statices* and *H. leucanthemi* were defined as belonging to the same OTU for the majority of cut-offs used. Other morpho-species with in the Tubulifera were well identified by OTUs as they tended to be completely monospecific for most of the cut-offs used. There were also morpho-species within the Terebrantia that were well identified such as *Kakothrips robustus* and *Ceratothrips ericae*. There is one example where two taxa are not clearly recognized. The five *T. sambuci* and *T. fuscipennis* sequences always intermingle, whether as OTUs or in a phylogenetic tree. This would indicate that these species designations need revision as the morphological feature, or features, used to currently distinguish the two, are not represented in the molecular data. There are several morpho-species which split into multiple OTUs and the corresponding clades have well supported branches on the phylogenetic tree. Two clusters of *Chirothrips manicatus* are seen in OTU results and in the tree. *Thrips palmi* and *T. tabaci* also show a deep split in the data set. It is unclear what the cause of these splits are. Populations from different geographic locations which are ecologically similar may maintain morphology but would experience genetic drift within the populations. Examining the *T. tabaci* clades, shows no simple correlation between MOTU or clades (Figure 6) and geography. Specimens from the same location cluster together, but not all specimens from the UK are in the same clade.

Figure 6: Phylogram of thrips data set. Bootstrap values of over 70 (from 2000replicates) are reported. Different colours indicate different MOTUs defined using a 9 bpcut-off value. Black bars indicate multi-specific MOTUs, grey bars indicate multiple MOTUsdefined for a single morpho-species. The two sub-orders are also shown, Tubulifera (bluebar) and Terebrantia (red bar).



Alternatively, the divisions may be indicative of cryptic or incipient speciation. Whatever the cause, such cases of major partitioning of specimens warrant further investigations to reconnect morphological and molecular data. When the taxon is of major agricultural interest, as *T. tabaci* is, the validity of a species is particularly important. Whilst this study is by no means an exhaustive representation of thrips molecular diversity, using MOTUs to define taxa is a promising approach for thrips surveillance. There are issues regarding the current taxonomic status of some species which may benefit from further investigation. Increasing the breadth and depth of sampling of species, especially those which only had a one or two specimens (e.g. *M. floridensis* and *L. denticornis*) and from the suborder Tubulifera, would also increase the value of the data set. MOTU_define.pl can define thrips taxa from this data set and in the long term should provide identification of samples quicker than traditional morphological methods.

Objective 4: Use the "barcode" sequences to devise rapid bioassay methods (for example DNA oligonucleotide based chips) for field samples.

DNA Barcodes and Video Capture and Editing (VCE): Integrating morphological vouchering with molecular diversity surveys.

As DNA barcode projects become widespread, standard protocols have been developed so data from different projects can be integrated (such as ABBI8 and TOL9). A barcode sequence is generated from a target gene from a specimen which usually has been identified to species. This is then used as an exemplar for the species that other sequences (and specimens) are compared to. These projects mainly target charismatic megafauna, animals over 1 cm in length. These organisms are relatively easy to identify, as defining morphological characters are large and easily observed. There tends to be a lot of interest in these animals from a range of disciplines, such as behaviourists and ecologists as well as taxonomists. As human activities increase extinction rates, these animals tend to be at the centre of conservation campaigns and stimulate public interest. If megafaunal species disappeared, they would be missed, as they would no longer be seen. Obtaining DNA from type specimens of larger animals for barcoding is relatively simple and quickly done, without compromising the integrity of the DNA or the morphology. An individual is rarely destroyed in the process and can be re-sampled, for example another feather could be collected. When specimens are rare, can potentially go extinct, or re-sampling would prove difficult, then greater effort would be invested in the collection, documentation and preservation of a sample. Barcoding, and its applications, for megafaunal taxa tend to be straightforward to execute and fit within standard frameworks and definitions. 8 All Birds Barcoding Initiative, http://www.barcodingbirds.org 9 Tree of Life, USA, http://tolweb.org/tree/ Meiofauna, organisms with a body axis less than 2 mm, present different challenges for standardized barcoding. A lack of taxonomic expertise and species definitions inhibits species identification so specimens are unlikely to have been identified to species. At a cursory inspection, meiofauna can appear to be morphologically conserved, e.g. nematodes. However, there is morphological variation which is difficult to visualize using light microscopy. Scanning (SEM) and transmission electron microscopy (TEM) reveal true morphological diversity, hidden from the naked eye (De Ley and Blaxter, 2002). Species identification by SEM or TEM drastically increases the timescale and budget of a project, and becomes prohibitively expensive for large-scale surveys and excludes the possibility of DNA recovery (De Ley *et al.*, 2005). This means diversity surveys tend to be limited to genus level and the vast majority of meiofaunal specimens will not have a concrete species diagnosis or

type specimens, especially from large-scale environmental surveys (De Ley *et al.*, 2005). So for meiofaunal barcoding, we have to accept that we might not have type specimens or type sequences. In meiofaunal barcoding surveys, the process of obtaining and identifying specimens often means DNA is degraded beyond the minimum amount and length required. Moreover, the whole organism is digested to release DNA so maintaining a paratype specimen is not possible. Meiofaunal organisms are, by definition, small and such do not have large amounts of DNA. Moreover, specimens can be temperature sensitive such that leaving a sample at room temperature for a few hours will result in most specimens dying and rendering the DNA unusable for PCR. It is therefore critical to preserve DNA as quickly as possible to (preferably from live specimens) prevent the DNA breakdown by enzymes. In order to preserve morphology and DNA, samples (e.g. soil extractions) can be split into two where one sub-sample would be preserved sympathetically for morphology and the other for DNA. Morphological preservation methods (i.e. formalin) destroy DNA, and some DNA preservatives do not maintain morphology. There is also the potential for discrepancies between the sub-samples for rare taxa. In order to document morphological diversity as well as sequence variation, a new preservative method is needed. A promising solution is DESS. A solution of dimethylsuphoxide, EDTA, saturated with sodium chloride, which preserves morphology and DNA (Yoder *et al.*, 2006). Initially used as a preservative for avian blood (Seutin *et al.*, 1991), DESS also proved to be suitable for the preservation of morphology for up to six months (Dawson *et al.*, 1998). However, the preservative properties of DESS on either the morphology or DNA of meiofaunal organisms has not been extensively tested. In addition to using DESS, integrating morphological vouchering into meiofaunal surveys could allow identification posthumously. It is possible to preserve type megafaunal specimens as museum accessions. Meiofaunal taxa can be preserved permanently fixed on slides, but this prohibits DNA collection. These physical objects can be damaged and lost and can only exist in a single location at any one time. Making a digital voucher of a specimen, whether a video or still image, can make these issues obsolete. Documenting a large animal by photographs or video is relatively easy. For microscopic organisms, this is a little more challenging, as the specimen needs to mounted and magnified. The most efficient method, in terms of isolating and recording a single specimen, is to create a temporary slide. Previously this temporary slide was documented by taking digital photographs stepwise through a specimen forming an image stack. However, even taking photographs with the smallest possible distance between focal planes through a nematode, some detail will be lost. Using a digital format also means that even at high resolutions, some information will be lost between pixels. When the images are stacked, the final file is often very costly in terms of computer memory. As recording technology and equipment has advanced, the next step was to take multifocal video images instead of a series of stills. A video is recorded of the slide, whilst adjusting the focus by hand creating a virtual slide. By taking a video, structures can be followed through the body. Moreover, using high definition tape prevents the loss of detail than can be associated with using a digital, pixel based format. By using equipment and software which is publicly available, the raw video can be processed and compressed to smaller, manageable files (De Ley and Blaxter, 2002). These can be distributed across the web or as hardcopies and made available to anyone. In this way, morphological information is not confined to single physical location. A virtual slide has major advantages over a conventional collection such that the specimens cannot deteriorate or be damaged. Furthermore, they cannot be lost. There are issues associated with the protection and maintenance of digital information such as storage, curation and costs. These issues are being addressed by organisations such as CBOL where

metadata of specimens and projects are being stored and accessed electronically. However, generating morphological vouchers for large numbers of meiofaunal specimens is not yet standard practice and has only been tested on some nematode taxa (De Ley *et al.*, 2005). The investigations in this work were designed to explore the properties of DESS as a DNA and morphological preservative, the use of morphological vouchering with species identification keys, the integration of vouchering with small-scale meiofaunal diversity surveys and the performance of vouchering on other meiofauna taxa such as tardigrades, copepods and mites. Digital vouchering, if found to be suitable, could provide a method to document morphological diversity and posthumously identify taxa once the specimen had been destroyed for PCR (De Ley *et al.*, 2005). This would enable congruence, or discord, between morphological information and sequence data to be explored. For barcoding projects, COI has been suggested as a standard target (Hebert *et al.*, 2003b). However there are concerns regarding the universal nature of the COI primer sets available (particularly with nematode taxa (De Ley *et al.*, 2005), so in addition to COI, LSU and SSU were also used as barcode targets.

Discussion

DESS is a preservative Barcode surveys of meiofaunal organisms have been a compromise between preserving morphological detail for the traditional taxonomists and maintaining DNA integrity for barcoding. Morphological preservation techniques are not able to preserved DNA and vice versa. Splitting a sample to preserve both morphology and DNA is not ideal as rare taxa, represented by few individuals may be in one sub-sample but not the other. Moreover, traditional preserving methods for morphology and DNA use hazardous materials (such as formalin and ethanol respectively), can require special conditions and are problematic to ship using postal or courier services (Yoder *et al.*, 2006). The results from this investigation have shown that DESS is a good DNA preservative. It is easy to prepare, samples can be stored at ambient temperature and transported without specialist equipment or precautions. These qualities make DESS an ideal preservative not only in the lab, where freezer space can be limited, but also for fieldwork. Moreover DNA can be amplified from specimens that have been stored for up to three years. In addition, DESS preserves morphology. Although the high salt concentration of the solution can distort morphology, particularly of organisms with soft cuticles, rinsing with ddH2O allows recovery of body shape. This step also removes salt crystals which may obscure internal morphological detail and is necessary for successful PCR. DESS works well as a morphological and a DNA preservative is cheap and safe.

Utility of VCE for identifying known species VCE clips can retain sufficient morphological detail to identify some nematode specimens to genus level and can identify laboratory cultures. In environmental surveys, VCE clips can distinguish between nematode orders, at worst, and at best identify genera. Some specimens can be distorted due to preservation methods which could hinder identification. A greater obstacle to species identification, especially for meiofaunal taxa, is the lack of concrete species descriptions and definitions. Initially, VCE is unlikely to cement species descriptions but will highlight the true extent of morphological variation, and when used in conjunction with barcoding will provide a molecular framework for species definitions. VCE should prove very effective for documenting morphological variation in meiofaunal surveys where ‘type’ specimens are likely to be rare as the whole animal is destroyed in the process of barcoding. It also circumvents issues with slide preservation. Online databases of VCE clips that are linked to sequence information could provide major assets for meiofaunal taxonomy. They can provide a permanent, mobile record of type specimens. For new taxa, it could provide a record of type features that would enable

direct comparisons with identified taxa. It could also provide a record of known morphological variations within a species which would prevent multiple descriptions for a single species being recorded and taxonomic conflicts. This would also encourage communication between international taxonomists. Finally, it could provide users with a method of comparing unknown taxa. NemATOL (http:// nematol.unh.edu) currently provides an interactive database containing molecular, morphological, ecological and phylogenetic information for the nematode community, but there are other phyla which constitute meiofauna. It should be a simple process to replicate the NemATOL format and expand it to cover other phyla such as Tardigrada and Rotifera. The number of unclassified taxa has been estimated to be several millions. With an increase in extinction rates, finding a system that can integrate molecular and morphological data would help to measure the range of diversity on Earth. VCE vouchering could provide a quick, cheap method which can be easily integrated into current barcoding protocols which may help to catalogue life on Earth. The VCE method does require practice. There were significant improvements, both in efficiency and quality of specimens and images, the more specimens that were processed. The surveys described in this work, were carried out solely by the author. Although there is a standard VCE protocol, differences between investigators could produce slight variations between clips. In larger scale surveys where multiple investigators produce clips, care should be taken to ensure consistency.

Integration of VCE into meiofaunal surveys

VCE has been demonstrated to be sufficient for documenting morphological variation of nematodes (De Ley *et al.*, 2005; Yoder *et al.*, 2006). When used as part of a barcoding survey, it can provide a putative identification which can be linked with molecular data. Key to integrating VCE into a barcode survey is using DESS, preserving both morphology and DNA so both can be obtained from a single specimen. Initially developed for nematode taxa, this investigation has demonstrated that VCE is easily applicable to other meiofaunal taxa. Any organism which can be made into a temporary slide should be suitable for VCE. Tardigrades often contract when put into DESS and it can be difficult to plainly see the claws and pharyngeal bulb. Making the temporary slide causes tardigrades to be flattened under the coverslip, thus allowing structures to be seen more clearly. Single specimen PCRs from environmental surveys are quick to perform: the specimen is picked from the substrate or extract and transferred to an individual tube before being lysed. If the VCE protocol is followed, a specimen is picked, temporarily slide mounted and clips generated before lysis and PCR. Although this does take longer, with practice it is possible to process 32 specimens in a day. If the sample has been preserved in DESS, then there is no rush to process everything. The quickest environmental barcode surveys use a bulk extract from the substrate and a single PCR reaction which generates sequences from multiple taxa. There is no way to record morphology using a bulk extraction method as everything is collected and then pooled. In addition, setting up a VCE system within a laboratory is quick and simple. It is also possible to check initial morphological designation once molecular results have been generated. In the meiofaunal survey of Disko Island, specimens D04_10 and D05_11 were originally recorded as Cephalobidae specimens. However molecular results indicated they were similar to other *Plectus* specimens collected in the survey. Re-examination of the VCE clips confirmed specimens D04_10 and D05_11 were morphologically most similar to D04_07 and D05_10 respectively. The specimens did not resemble other Cephalobidae specimens and the most probable cause for the incorrect identification being recorded was an error during data entering. There appear to be mis-matches between morphological identification and some MOTUs in the tardigrade survey. This may be due to the process of identifying specimens from

VCE clips and over-estimating the amount of morphological variation within the specimens. VCE clips were easily sufficient to differentiate between genera of tardigrade specimens and these were supported by defined MOTUs. However, the features used to differentiate among Macrobiotus TI, TII, TIII and TV do not seem to relate to molecular differences. There was an interesting result from the Disko Island meiofaunal survey. In the SSU MOTU analysis, where as the cut-off approached 10%, the mean number of MOTU defined approached one. As barcoding surveys are normally targeted to specific taxa, this would not be an issue. However, in this survey specimens derived from multiple phyla. At a cut-off of 9.9%, approximately two SSU MOTU were defined which included specimens from Nematoda, Tardigrada and Arthropoda. If this result is found in other meiofaunal surveys, then it may be necessary to re-assess the utility of the SSU gene to act as a barcode marker, depending on the aims of the survey.

General Discussion

Meiofaunal organisms present unique challenges for molecular barcoding due to unresolved questions of species definitions, species identification and genetic variation. The key issues for meiofaunal barcoding are the assumptions that protocols developed for large organisms will work for small organisms and a lack of standardisation across different surveys. Initial barcoding surveys sampled a broad range of taxa, where specimens were easily identified and readily available. As surveys have move towards intensive sampling of a few species across greater ranges, barcode data has highlighted distinct molecular taxa (MOTUs). Many of these MOTUs can be associated with different geographical locations, but not always. Sometimes, the molecular taxa are sympatric and can only be segregated by subtle ecological differences (e.g. Hebert *et al.*, 2004a). If barcodes had not been used to investigate the molecular taxonomy of the group, then we would have been none the wiser as to the complexity of the situation. As well as being surprising, these types of cases raise a delicate issue within the traditional taxonomic community. How do we know that morphological designations are correct? In reference to larger animals, the answer is “probably”. Although it requires time and occasionally specialist equipment, it is generally easy to identify whom mates with whom, what food species are consuming and where specimens are in their natural environment. With meiofauna, this is not the case. In comparison with megafauna, we know very little about meiofaunal organisms and due to their size, it is very difficult to study these organisms in their natural habitat. We don’t always know whom mates with whom (if a species even requires sex for reproduction), what species are feeding on, or if a species is native to a particular site. Although it is possible to maintain some species as laboratory cultures, not all taxa will survive and the constant laboratory conditions will remove the natural variation of environmental abiotic factors. These factors alone cause problems within traditional meiofaunal morphological taxonomy when describing species. Adding to this situation that we know only a small proportion of meiofaunal organisms have been described, and that there is a deficit of taxonomists, this means that it is unlikely that meiofauna can continue to be described based solely on morphology. If meiofaunal identifications through barcoding are to be accepted across the global taxonomic community, there needs to be clearer communication between morphological and molecular taxonomists. There seems to have been misinterpretations of the undertakings of both groups, particularly circa 2003, when the number of barcoding surveys dramatically increased. During this period, there was no consistency between surveys as protocols, targets and analysis methods were being developed. Molecular barcoding has now been in routine use for almost a decade, and whilst there are standard protocols for sequence generation, a universal target and analysis method have yet to be adopted.

Is there a 'universal' target?

Initially, COI was proposed as a 'universal' target for barcoding (Hebert *et al.*, 2003a). Whilst this gene has been successfully used in a number of surveys, nematologists have shied away from COI, favouring instead the LSU or SSU genes for molecular surveys (Floyd *et al.*, 2002; Meldal *et al.*, 2007; Subbotin *et al.*, 2007; Ye *et al.*, 2007). Results from the investigations in this thesis suggest that, in some cases, LSU results may be used as a proxy measure for COI. Given the poor success rates of generating COI barcodes from meiofaunal taxa, the LSU should be targeted first. In addition, targeting the SSU will allow results from future surveys to be integrated with previous investigations, as there are more SSU sequences available in public databases. In October 2009, GenBank contained 7668 SSU entries compared with 3927 LSU and 1206 COI entries for the phylum Nematoda. For Tardigrada there were 753, 49 and 203 entries for SSU, LSU and COI respectively. Major barcoding initiatives may continue to insist on COI as the target, so attempts may be made to generate them. However, there may need to be a change in attitude regarding the virtues of COI.

Is there a 'universal' cut-off?

A value of 2% was designated as sufficient for species designation (Hebert *et al.*, 2003b). Again, this was based on results from large animals. Moreover this value was derived from analyses of the mitochondrial *cytB* gene, which shares the same mode of inheritance as COI, from a small survey of taxa (Johns and Avise, 1998). Although 2% may be sufficient for COI in some taxa, results from this thesis suggest that it is not suitable for LSU or SSU datasets. For some LSU data, a cut-off of 2% seems to be too high for species discrimination and results in clustering seen between genera (e.g. in the meiofaunal survey of Disko Island, Section 4.3.3). In other LSU data sets, 2% maybe considered as too low and is reporting individual variation (e.g. in the tardigrade survey, Section 4.3.4). These conflicting results are also seen in the SSU data. Rather than assuming 2% is a suitable cut-off, it would be better practice to first generate the sequences, investigate the clustering behaviour, and then decide on a cut-off. Different surveys are likely to find different cut-offs which best describe the data. Ultimately, it is unlikely that there is a universal cut-off which can be used across multiple taxa. In studies where species identification is the key aim, this should not be an issue as the sequence data may define the cut-off. In bulk environmental surveys, where the diversity is being investigated, a cut-off may have to be middling value from the different taxa found within the sample, or different taxa need to be considered separately.

Can taxa be defined by sequence similarity?

Results from the thrips COI data demonstrated the ability of MOTU_define.pl and DOTUR to recover morphologically identified taxa at varying cut-offs. When MOTU_define.pl was challenged with anonymous sequences in, taxa were defined, some of which were supported by morphological identifications. Where barcodes are being used as a confirmational tool, the level of sequence similarity varies depending on the cut-off used and therefore so do the taxa defined. Molecular barcodes are short DNA sequences which generally do not contain sufficient information to reconstruct phylogenetic processes. Using a taxon defining method based on an evolutionary model will overestimate taxa when sequences are of different lengths, as in DOTUR. Using a trimmed data set removes variation. If the evolutionary history of a set of samples is required, multiple, longer targets should be used. If molecular groups are sought, taxa can be adequately defined using simple comparisons of sequence similarity.

Can DNA barcodes identify meiofaunal taxa?

In meiofaunal taxa, it is unknown how morphological variation relates to molecular differences. Thus there is a tendency to perhaps overestimate the amount of morphological variation of meiofauna. The microscopic size of meiofauna means morphological identification is difficult and can lead to misidentification of specimens. DNA barcodes are a promising tool for the identification of any specimen, regardless of size. They can separate morphologically cryptic specimens and can reduce continuous morphological variation to robust molecular taxa. Where there is no morphological identification for a specimen, barcodes can provide some level of taxonomic assignment.

The future of meiofaunal barcoding

Previously, sequence generation had been limited to single individuals. It became possible to PCR from environmental bulk extracts but this required purification steps to remove PCR inhibitors and cloning in order to separate individual sequences. With the advent of next-generation sequencing platforms, there is the potential to barcode every individual within a sample; environmental metagenetic sequencing (Creer *et al.*, 2009). The amount of data generated by these surveys is vast, but are agnostic to any morphological information. Moreover, methods have not been optimised to enable relative abundances of individuals from a single taxon to be calculated from sequence frequencies (Porazinska *et al.*, 2009). Meiofaunal organisms are a paraphyletic collection of taxa. Bulk extractions and sequencing may bias results through extraction methods, primer binding, DNA amplification and types of sequencing. In addition, accurately identifying species post sequencing requires a comprehensive database for comparisons (Machida *et al.*, 2009). Even so, it is normally possible to assign some level of taxonomic information to sequence data. Ultrasequencing platforms will readily allow molecular variation to be recorded and assessed, but can make no assessment of the morphological diversity. Without similar input into measuring morphological diversity, there is likely to be a widening gap between the two. We will only be able to say what sequences exist in an environment, which may or may not be related to the biology of the sample. Integrating digital vouchering of specimens into surveys is possible and will enhance molecular results by expanding the range of taxa included in databases. Capturing morphology detail by VCE will allow meiofaunal diversity to be matched to molecular variation. There may be certain situations where VCE is not practical, but it should not be overlooked in the favour of molecular techniques and should be considered a priority for any sampling regime. DNA taxonomy was proposed as an integration of molecular data and traditional taxonomy, when it became apparent that traditional taxonomy would be unable to complete the catalogue of life, especially the vast numbers of meiofauna taxa. However descriptions using DNA barcodes were felt to be lacking the rigorous and detailed study required by traditional taxonomy. Combining sequence generation with VCE would provide a permanent bridge between the demands of morphological and DNA taxonomy. In its current state, meiofaunal barcoding is performing sub-optimally. Whilst exciting advances have been made, and the volumes of barcodes generated have increased exponentially, there are still gaps in the foundations. These need to be filled before we are able to generate an anonymous sequence and with certainty assign it to a single meiofaunal species and fully assess the meiofaunal diversity within a habitat.

References to published material

9. This section should be used to record links (hypertext links where possible) or references to other published material generated by, or relating to this project.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zang, Z. Zang, W. Miller and D. J. Lipman. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Armstrong, K. F. and S. L. Ball (2005). "DNA barcodes for biosecurity: invasive species identification." Philosophical Transactions of the Royal Society B-Biological Sciences **360**(1462): 1813-1823.

Blaxter, M. (2003). "Molecular systematics: Counting angels with DNA." Nature **421**(6919): 122-124.

Blaxter, M. and R. Floyd (2003). "Molecular taxonomics for biodiversity surveys: already a reality." Trends in Ecology & Evolution **18**(6): 268-269.

Blaxter, M., J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd and E. Abebe (2005). "Defining operational taxonomic units using DNA barcode data." Philos Trans R Soc Lond B Biol Sci **360**(1462): 1935-1943.

Brunner, P. C., E. K. Chatzivassiliou, N. I. Katis and J. E. Frey (2004). "Hostassociated genetic differentiation in Thrips tabaci (Insecta; Thysanoptera), as determined from mtDNA sequence data." Heredity **93**(4): 364-370. Brunner, P. C., C. C. Fleming and J. E. Frey (2002). "A molecular identification key for economically important thrips species (Thysanoptera: Thripidae) using direct sequencing and a PCR-RFLP-based approach." Agricultural and Forest Entomology **4**: 127-136.

Creer, S., V. Fonseca, D. L. Porazinska, R. M. Giblin-Davis, W. Sung, D. M. Power, M. Packer, G. R. Carvalho, M. Blaxter, P. J. D. Lambshead, and W. K. Thomas. (2009). "Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises." Molecular Ecology ***In press***.

Dawson, M. N., K. A. Raskoff and D. K. Jacobs (1998). "Field preservation of marine invertebrate tissue for DNA analyses." Molecular Marine Biology and Biotechnology **7**(2): 145-152. De Ley, P. and W. Bert (2002). "Video Capture and Editing as a Tool for the Storage, Distribution, and Illustration of Morphological Characters of Nematodes." Journal of Nematology **34**(4): 296-302.

De Ley, P. and M. Blaxter (2002). Systematic Position and Phylogeny. The Biology of Nematodes. D. L. Lee. London, Taylor & Francis**:** 1-30.

De Ley, P., I. T. De Ley, K. Morris, E. Abebe, M. Mundo-Ocampo, M. Yoder, J. Heras, D. Waumann, A. Rocha-Olivares, A. H. Jay Burr, J. G. Baldwin and W. K. Thomas (2005). "An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding." Philos Trans R Soc Lond B Biol Sci **360**(1462): 1945-1958.

DeSalle, R., M. G. Egan and M. Siddall (2005). "The unholy trinity: taxonomy, species delimitation and DNA barcoding." Philosophical Transactions of the Royal Society B-Biological Sciences **360**(1462): 1905-1916.

Elias, M., R. I. Hill, K. R. Willmott, K. K. Dasmahapatra, A. V. Z. Brower, J. Mallet and C. D. Jiggins (2007). "Limited performance of DNA barcoding in a diverse community of tropical butterflies." Proceedings of the Royal Society B-Biological Sciences **274**(1627): 2881-2889.

Floyd, R., E. Abebe, A. Papert and M. Blaxter (2002). "Molecular barcodes for soil nematode identification." Mol Ecol **11**(4): 839-850.

Gasser, R. B., N. B. Chilton, H. Hoste and L. A. Stevenson (1994). "Species identification of trichostrongyle nematodes by PCR-linked RFLP." Int J Parasitol **24**(2): 291-293. Gewin, V. (2002). "Taxonomy - All living things, online." Nature **418**(6896): 362-363. Godfray, H. C. J. (2002). "Challenges for taxonomy - The discipline will have to reinvent itself if it is to survive and flourish." Nature **417**(6884): 17-19.

Gullan, P. J. and P. S. Cranston (2005). The Insects: An Outline of Entomology, Blackwell Publishing Ltd.

Hebert, P. D., A. Cywinska, S. L. Cywinska S. L. Ball and J. R. deWaard (2003a). "Biological identifications through DNA barcodes." Proc Biol Sci **270**(1512): 313-321.

Hebert, P. D., S. Ratnasingham and J. R. deWaard (2003b). "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species." Proc Biol Sci **270 Suppl 1**: S96-99.

Acad Sci U S A **101**(41): 14812-14817.

Hebert, P. D., E. H. Penton, J. M. Burns, D. H. Janzen and W. Hallwachs (2004a). "Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*." Proc Natl Acad Sci U S A **101**(41): 14812-14817.

Hosseini, R., M. A. Keller, O. Schmidt and V. W. Framenau (2007). "Molecular identification of wolf spiders (Araneae : Lycosidae) by multiplex polymerase chain reaction." Biological Control **40**(1): 128-135.

Hunt, D. J. (1993). Aphelenchida, Longidoridae and Trichodoridae: Their Systematics and Bionomics. Cambridge, CAB International. Hurst, G. D. and F. M. Jiggins (2005). "Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts." Proc Biol Sci **272**(1572): 1525-1534.

Inoue, T. and T. Sakurai (2007). "The phylogeny of thrips (Thysanoptera : Thripidae) based on partial sequences of cytochrome oxidase I, 28S ribosomal DNA and elongation factor-1 alpha and the association with vector competence of tospoviruses." Applied Entomology and Zoology **42**(1): 71-81.

Johns, G. C. and J. C. Avise (1998). "A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene." Molecular Biology and Evolution **15**(11): 1481-1490.

Lohse, K. (2009). "Can mtDNA Barcodes Be Used to Delimit Species? A Response to Pons et al. (2006)." Systematic Biology **58**(4): 439-441.

Machida, R. J., Y. Hashiguchi, M. Nishida and S. Nishida (2009). "Zooplankton diversity analysis through single-gene sequencing of a community sample." BMC Genomics **10**: 438-444.

Meier, R., K. Shiyang, G. Vaidya and P. K. L. Ng (2006). "DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success." Syst Biol **55**(5): 715-728.

Meldal, B. H. M., N. J. Debenham, P. De Ley, I. T. De Ley, J. R. Vanfleteren, A. R. Vierstraete, W. Bert, G. Borgonie, T. Moens, P. A. Tyler, M. C. Austen, M. L. Blaxter, A. D. Rogers and P. J. D. Lambshead (2007). "An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa." Molecular Phylogenetics and Evolution **42**(3): 622-636.

Meyer, C. P. and G. Paulay (2005). "DNA barcoding: error rates based on comprehensive sampling." PLoS Biol **3**(12): e422. Minelli, A. (2003). "The status of taxonomic literature." Trends in Ecology & Evolution **18**(2): 75-76. *References* 205

Mound, L. A., B. S. Heming and J. M. Palmer (1980). "Phylogenetic relationships between the families of recent Thysanoptera (Insecta)." Zoological Journal of the Linnean Society **69**: 111-141.

Mound, L. A. and D. C. Morris (2001). "Domicile constructing phlaeothripine Thysanoptera from Acacia phyllodes in Australia: Dunatothrips Moulton and Sartrithrips gen.n., with a key to associated genera." Systematic Entomology **26**(4): 401-419.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin and A. P. Vogler (2006). "Sequence-based species delimitation for the DNA taxonomy of undescribed insects." Syst Biol **55**(4): 595-609. *References* 206

Porazinska, D. L., R. M. Giblin-Davis, L. Faller, W. Farmerie, N. Kanzaki, K. Morris, T. Powers, A. Tucker, W. Sung and W. K. Thomas (2009). "Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity." Molecular Ecology Resources **9**: 1439-1450.

Rubinoff, D. (2006). "Utility of mitochondrial DNA barcodes in species conservation." Conservation Biology **20**(4): 1026-1033.

Schloss, P. D. and J. Handelsman (2005). "Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness." Appl Environ Microbiol **71**(3): 1501-1506.

Seutin, G., B. N. White and P. T. Boag (1991). "Preservation of avian blood and tissue samples for DNA analyses." Can. J. Zool. **69**: 82-90.

Subbotin, S. A., D. Sturhan, N. Vovlas, P. Castillo, J. T. Tambe, M. Moens and J. G. Baldwin (2007). "Application of the secondary structure model of rRNA for phylogeny: D2-D3 expansion segments of the LSU gene of plant-parasitic nematodes from the family Hoplolaimidae

Tautz, D., P. Arctander, M. Alessandro, R. H. Thomas and A. P. Vogler (2003). "A plea for DNA taxonomy." Trends Ecol Evol **18**(2): 70-74.

Toda, S. and T. Murai (2007). "Phylogenetic analysis based on mitochondrial COI gene sequences in Thrips tabaci Lindeman (Thysanoptera : Thripidae) in relation to reproductive forms and geographic distribution." Applied Entomology and Zoology **42**(2): 309-316.

Trewick, S. A. (2008). "DNA Barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera : Acrididae)." Cladistics **24**(2): 240-254.

Vences, M., M. Thomas, R. M. Bonett and D. R. Vieites (2005). "Deciphering amphibian diversity through DNA barcoding: chances and challenges." Philosophical Transactions of the Royal Society B-Biological Sciences **360**(1462): 1859-1868.

Wiemers, M. and K. Fiedler (2007). "Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae)." Frontiers in Zoology **4**(8). Wilson, E. O. (2003). "The encyclopedia of life." Trends Ecol Evol **18**(2): 77-80.

Yoder, M., I. Tandingan De Ley, I. W. King, M. Mundo-Ocampo, J. Mann, M. Blaxter, L. Poiras and P. De Ley (2006). "DESS: a versatile solution for preserving morphology and extractable DNA of nematodes." Nematology **8**(3): 367-376.