De novo species delimitation in metabarcoding datasets using ecology and phylogeny

Authors: Caitlin Potter[1,7], Cuong Q. Tang[2,3], Vera G. Fonseca4, Delphine Lallias[5], John M. Gaspar[6], Kelley Thomas[6] and Simon Creer[1]

[1] Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Environment Centre Wales, Bangor University, Deiniol Road, Bangor, Gwynedd, LL57 2UW, United Kingdom.

[2] Department of Life Sciences, The Natural History Museum, Darwin Centre 2 - room 627, Cromwell Road, London, SW7 5BD.

[3] Department of Life Sciences, Imperial College London, Ascot, Berkshire SL5 7PY, UK.

[4] Zoological Research Museum Alexander Koenig (ZFMK), Centre for Molecular Biodiversity Research,Bonn, Germany, Adenauerallee 162, 53113 Bonn, Germany

[5]INRA, UMR 1313 GABI Génétique Animale et Biologie Intégrative, Domaine de Vilvert, 78350 Jouy-en-Josas, France.

[6]Department of Molecular, Cellular, & Biomedical Sciences, University of New Hampshire, Durham, New Hampshire, United States of America

[7]Institute for Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, UK.

Corresponding author: Dr. Simon Creer

Molecular Ecology and Fisheries Genetics Lab

Environment Centre Wales Building

School of Biological Sciences

Bangor University

Bangor

Gwynedd

LL57 2 UW

United Kingdom

Telephone: (+44) (0) 1248 382302

Fax: (+44) (0) 1248 370731

E-mail: s.creer@bangor.ac.uk

Abstract

Background: Metabarcoding studies allow a wide variety of taxa to be analysed simultaneously in a fraction of the time taken by morphological identification, but currently metabarcoding studies must rely on sequence similarity-based methodologies to delimit operational taxonomic units (OTUs). Similarity-based OTU clustering methodologies can lead to inaccurate estimates of diversity, species' distributions or responses to change, meaning that there is a critical need for methods to delimit species in metabarcoding datasets.

Methods: We introduce SNAPhy (Species delimitation using Niche And PHYlogeny), a novel approach which utilises ecological and phylogenetic information to delimit *de novo* OTUs in metabarcoding datasets and avoids the problems associated with current OTU clustering methods. Sequencing reads are first divided into ecological groups based on co-occurrence, thereby reducing data complexity and facilitating the use of evolutionary and phylogenetic models (e.g. BEAST and GMYC) to delimit species-level groupings within discrete ecologically informed phylogenies. The utility of SNAPhy is demonstrated using an 18S rDNA nuclear small subunit (nSSU) dataset representing replicated samples taken along the entire length of an estuarine salinity gradient, and SNAPhy is then compared to existing OTU clustering methods.

Results: All of the OTU clustering methods compared yielded different numbers of OTUs and a different taxonomic distribution of OTUs, which we suggest is due to the taxon differences that are known to exist in the degree of intraspecific divergence. SNAPhy and UCLUST (with a 98% similarity threshold) gave the most plausible numbers of OTUs, especially within the Nematoda. Additionally, the degree of variation within nematode OTUs delimited by SNAPhy lies within the range of variation in deeply metabarcoded individuals.

Discussion: SNAPhy avoids the static clustering threshold problems associated with current OTU clustering methods and instead focuses on genuine biological diversity delimited

according to a general lineage species concept. We suggest that the SNAPhy approach should play a crucial role in future sequencing-based biodiversity assessment by providing more accurate estimates of species diversity and distributions than current methods, thereby enabling more accurate impact assessments and better informing managerial decisions.

Introduction

As the natural world experiences increasing pressure from habitat loss, fragmentation and global environmental change, researchers progressively focus on the relationship between biodiversity and ecosystem processes (Loreau et al. 2001). Ecologists are interested in the interactions between organisms and their environment in relation to questions involving macroecology (Brown 1995), trophic linkages (Hagen et al. 2012) and the relationship between biodiversity and ecosystem services (Schröter et al. 2005). Conversely, regulators and stakeholders are interested in monitoring biological indicators to estimate environmental status, in association with thresholds for management action (Friberg et al. 2011). In all these fields there is an implicit need to identify community level biodiversity across many time points and geographical locations, creating a substantial volume of work for ecologists and taxonomists. Recent improvements in the throughput and cost of next-generation sequencing (Loman et al. 2012) have resulted in the increasing use of DNA sequence data to identify biodiversity *en masse*, shortcutting the need for traditional taxonomic identification (Caporaso et al. 2011; Bik et al. 2012).

A particularly useful approach for ecological studies is to assess biodiversity through *en masse* taxonomic classification of an environmental sample using high throughput sequencing of homologous gene markers (Creer et al. 2010; Hajibabaei 2012; Taberlet et al. 2012), termed metagenetics (Creer et al. 2010), metasystematics (Hajibabaei 2012) or metabarcoding (Taberlet et al. 2012 – adopted hereon). Inspired by the work of microbial ecologists using 16S rDNA gene markers (Caporaso et al. 2011), such studies use highly degenerate oligonucleotide primers situated either side of informative regions of the genome to delimit biodiversity across a broad range of taxonomic groupings by PCR amplifying the region of interest. Metabarcoding can quickly and objectively identify the majority of biodiversity in thousands of samples simultaneously and is now employed to identify

prokaryotes (Caporaso et al. 2011), microbial eukaryotes (Dumbrell et al. 2011; Pawlowski et al. 2012), meiofaunal (Fonseca et al. 2014; Lallias et al. 2015) and macrofaunal (Carew et al. 2013) size fractions, or from environmentally 'free' DNA (eDNA; Bohmann et al. 2014).

Nevertheless, the volume of reads resulting from contemporary DNA sequencers (Loman et al. 2012) cannot be easily incorporated into hypothesis testing without transformation into a smaller number of dependent variables that emulate genuine taxon diversity. The currently preferred transformation is to perform operational taxonomic unit (OTU) clustering. Once OTUs are constructed, a representative sequence (e.g. dominant/consensus) with associated frequency data forms the dependent variable for downstream analysis (Bik et al. 2012). Almost all current studies cluster metabarcoding datasets into OTUs using a static clustering threshold (although see e.g. Malviya et al. (2016)), despite the known problems with this approach: in particular, that a single clustering threshold cannot accurately delimit species, or any desired taxonomic level, because of the heterogeneous nature of intra-genomic, intraspecific and interspecific genetic diversity throughout the tree of life (Schloss & Westcott 2011). Therefore, OTUs do not accurately reflect species diversity in genuine biological communities; OTU construction may split some species and lump others. An alternative to OTU clustering is therefore critically needed in order to accurately delimit species-level diversity in metabarcoding studies. Several recent approaches have aimed to find solutions to the problems inherent in clustering with a static threshold. For example, Swarm uses a local clustering threshold, *d*, to generate clusters through an iterative process (Mahé et al. 2014). An alternative approach is to include distribution information in addition to sequence similarity in order to ensure OTUs are ecologically meaningful (Preheim et al. 2013). Information about read distributions has also been used to inform denoising approaches (Morgan et al. 2013; Tikhonov et al. 2014).

Here we describe and test SNAPhy (Species delimitation using Niche And PHYlogeny), an alternative framework to define genetic units in metabarcoding data under the general lineage species concept (de Queiroz 2007). Phylogenetic models are more powerful than simple metrics of sequence divergence (Barraclough et al. 2009), but are too computationally demanding for use on current metabarcoding datasets. Ecology here provides a convenient parsing mechanism: in the current approach we first divide the dataset based on ecological co-occurrence (or where this was not possible, based on taxonomy) in order to obtain subsets of data on which it is possible to apply phylogenetic models. In metabarcoding studies, reads are derived from species that are distributed in time and space according to ecological niches, environmental tolerances or neutral processes (Legendre & Fortin 1989; Vellend 2010). Importantly, variation caused by real intra-genomic and intra-specific diversity will also be accompanied by associated PCR and sequencing errors. If therefore, species delimitation is focused on co-occurring reads, the complexity of multiple sequence alignments can be reduced into a number of smaller tasks, according to niche or neutral occupancy models, based on genuine biological diversity (Chase & Myers 2011).

In the current manuscript, we test SNAPhy on an estuarine dataset based on the 18S rDNA nuclear small subunit (nSSU) DNA marker derived from (Lallias et al. 2015) because: (a.) ecological heterogeneity is exemplified across an ecological cline, (b.) the 18S nSSU marker is predicted to display valuable, intragenomic and intraspecific diversity (Bik et al. 2013; Stage & Eickbush 2007) for phylogenetic species delimitation and (c.) we were able to deeply sequence individuals belonging to representative nematode species in order to validate the approach.

Materials and Methods

Metabarcoding dataset

The SNAPhy workflow was used to identify OTUs in an already well-characterised marine meiobenthic dataset described in Lallias et al. (2015). Briefly, three sediment cores were collected from each of twenty sites (n=60) along the full salinity range of the Thames Estuary (UK). Following community and DNA extraction, a 450bp region of the 18S nSSU region was amplified and sequenced on a 454 Roche GSFLX (454 Life Sciences, Roche Applied Science) sequencing platform (Lallias et al. 2015).

SNAPhy workflow

The SNAPhy workflow does not begin with trying to estimate and remove minor sequencing errors from the dataset. Such processes can be computationally intensive (Quince et al. 2009), restricted to either specific loci or sequencing chemistries (Quince et al. 2009), or unable to discriminate between errors and the intra-genomic/intra-specific genetic diversity characteristic of nuclear taxonomy markers (Bik et al. 2013; Stage & Eickbush 2007). Instead, SNAPhy focuses on identifying the ecological and genetic signal (including PCR/sequencing errors) derived from spatially and/or temporally dispersed individuals of different species using next-generation sequencing platforms. Nevertheless, the issue of DNA chimeras still persists in environmental DNA sequencing datasets (Fonseca et al. 2012) and their removal should be incorporated into emerging workflows as below. The workflow can be broadly broken down into the identification of unique reads, chimera detection and removal on a sample by sample basis; clustering reads into ecological co-occurrence networks and species delimitation based on a phylogenetic approach (Fig. 1). The first quality control step involves demultiplexing, length homogenisation and merging of identical reads into unique reads. These three processes were carried out simultaneously using the Perl script

"1_Filter_by_truncation.pl" from the Amplicon Pyrosequencing Denoising Program (APDP v1.1; Morgan et al. 2013). Reads were truncated at 225bp reflecting optimal quality (including removal of reads that were less than 225bp) and identical reads were binned together. Chimeras were removed from the dataset using the UCHIME algorithm run in *de novo* mode (Edgar et al. 2011) within USEARCH v6.0.307 with the default settings.

Singletons and reads that only occurred in one sample were removed. These reads could not be assigned to co-occurrence networks and reads that only occur in single samples have little comparative power in ecological studies and/or can represent sequencing artefacts. Read abundances were normalised by conversion into a proportion of total reads in a given sample. Further error removal steps (e.g. homopolymer correction; Quince et al. 2009; Caporaso et al. 2011) were not carried out because ecological co-occurrence networks should link sequence errors to the genuine genomic diversity from which they originated.

Following the above pre-processing steps, reads were clustered into ecological co-occurrence networks based on Pearson correlation using the CoNet package for Cytoscape (v3.01; Faust et al. 2012). Pairwise correlation coefficients were calculated for all read pairs, and an edge (connection) was drawn between each pair of reads (nodes) where R^2 was 0.95 or greater - this value of R^2 was found to give co-occurrence networks of appropriate size for downstream analysis while allowing for cases of incomplete co-occurrence. A given read was included in a network if it had at least one connection to another read in that network (nearest neighbour clustering/single-linkage clustering; Sun et al. 2011).

The next step of the workflow was to delimit species using a phylogenetic modelling approach. We tested several approaches on simulated data in order to select the most appropriate model for future applications of SNAPhy to 18S nSSU data. For testing, four artificial datasets containing between 19 and 60 reads (Table 1; Table S1) were generated

from 18S nSSU sequences downloaded from GenBank. Artificial datasets were generated using Grinder (v0.5.3; Angly et al. 2012) in order to mimic typical error patterns obtained using 454-Roche sequencing (homopolymer error model based on Balzer et al. (2010) and a uniform error rate of 0.1%). The species richness and evenness within each artificial dataset was based on the approximate richness and evenness within four co-occurrence networks within the real dataset.

Once the Grinder simulated datasets had been generated, two coalescent-based models for species delimitation were compared on the artificial datasets: Generalized Mixed Yule Coalescent (GMYC; Fujisawa & Barraclough 2013; implemented in R using package splits 1.0–11) and Poisson Tree Processes (PTP; Zhang et al. 2013; implemented using webserver found at http://species.h-its.org/). These methods combine coalescent theory with diversification models to infer the transition point between population and species-level processes on a gene tree; such a shift is indicative of the switch from between-species to within-species processes, expected if a sample comprises multiple individuals from a set of independently evolving species. Both methods delimit Evolutionarily Significant Units (ESUs) of diversity indicative of species (Barraclough et al. 2009) and require phylogenetic trees as input. These trees were reconstructed by first aligning reads using MAFFT (v7.147b; Katoh & Standley 2013) and then using both Bayesian Evolutionary Analysis by Sampling Trees (BEAST; v1.8.0; Drummond & Rambaut 2007) and Randomised Axelerated Maximum Likelihood (RAxML; Stamatakis et al. 2005), both of which were identified by Tang et al. (2014a) as being appropriate for these analyses.

Once an optimal phylogenetic method had been chosen, ESUs were delimited for each co-occurrence network from the estuarine dataset which contained at least 10 unique reads (i.e. adequate for accurate phylogenetic species delimitation). The results of the phylogenetic species delimitation were combined with the support for nodes on the phylogenetic tree,

which served two purposes. Initially, in cases where the phylogenetic model for a given tree was insignificant, combining the two methods gave a more discriminatory and phylogenetically plausible result. Secondly, the use of nodal support overcame a tendency of the GMYC to 'lump' reads into species with abnormally high intraspecific divergence. Where the phylogenetic model was significant at the 0.05 level, OTUs were further split at any node with a support value of 0.9 or greater. For trees which produced an insignificant species delimitation result, only OTUs which were supported by a posterior probability of 0.9 or greater were kept and unsupported OTUs were divided into singleton representatives of putative species - an example is given in Fig. 2. These units can be defined as species under the general lineage species concept (de Queiroz 2007).

'Orphan' Reads

A different workflow was adopted to assign 'orphan' reads to OTUs, i.e. reads that either did not belong to a co-occurrence network or belonged to a network that contained fewer than 10 reads. Orphan reads were extracted using a custom Perl script (Supplementary script 'Orphan_Sequence_Workflow.pl') and were partitioned into phyla (or higher taxon levels) following megablast (v2.2.28 with a minimum percentage ID of 90%; Camacho et al. 2009) and lowest common ancestor annotation using MEGAN (v4; Huson et al. 2007) and the SILVA 111 database (Quast et al. 2012). Therein, OTUs were delimited within the defined taxonomic groups using identical methods to those used for co-occurrence groupings (Table S2), thereby overcoming the lack of phylogenetic signal encountered in orphan groups.

Testing/validating the SNAPhy Workflow

Assessing read abundances and divergence within SNAPhy OTUs

Within each SNAPhy OTU, the majority of reads are expected to be variations of one or few dominant 18S nSSU reads (Bik et al. 2013), caused by a combination of

intragenomic/intraspecific variation and PCR or sequencing errors. In order to assess read frequencies within OTUs, five OTUs were chosen at random and used to generate neighbour joining trees in MEGA5.2 (Tamura et al. 2011; parameters chosen were phylogeny test: bootstrap with 1,000 replications; substitution type: nucleotide; model: p-dist; gaps: pairwise deletion). Abundances of each unique read were then mapped onto the SNAPhy OTU in order to test for the expected pattern (Bik et al. 2013). The percentage divergence within each OTU was calculated using "calc_distmx" command in USEARCH (Edgar 2010).

Comparisons with UCLUST and Swarm

Results obtained using the SNAPhy workflow were compared to existing methods. First, data were quality checked and denoised using FlowClus (Gaspar & Thomas 2013), as described in (Lallias et al. 2015). Reads were then trimmed to 225bp in order to match the data which was input into the SNAPhy workflow. Next, OTU clustering was carried out using UCLUST (Edgar 2010) at two similarity thresholds (96% and 98% similarity,) and Swarm (Mahé et al. 2014). Both UCLUST and Swarm were implemented in QIIME v1.9.0 (Caporaso et al. 2010). To enable comparison, taxonomy was assigned to OTUs from all methods using the Silva 111 database using identical methods to those described in (Lallias et al. 2015).

Mapping individually metabarcoded estuarine nematode species 18S nSSU diversity onto SNAPhy OTUs

To ensure that variability within the OTUs obtained using the SNAPhy workflow was within the range expected for the intragenomic variability within a species, reads were compared to the results of "deep-metabarcoded" ecologically representative individuals (i.e. one amplicon library as above/individual nematode) of nematode worms co-extracted from the Thames Estuary, thereby creating an 18S nSSU genomic reference database of individual nematode worms.

Highly related matches between the 18S nSSU genomic database and the SNAPhy OTUs were obtained using megablast (parameters –D 2 –p 99 – m 7 –a 4 –b 1 –v 1 –F F). Where a deep metabarcoded individual 18S nSSU identity matched a read belonging to a SNAPhy OTU, that individual's deep-sequenced reads were combined with those within the SNAPHy OTU. The resulting set of reads was aligned using MUSCLE (Edgar 2004) and used to construct neighbour-joining trees, both in MEGA (v5.2; Tamura et al. 2011).

Results

SNAPhy Workflow

Sequencing yielded a total of 1,085,607 reads, which were collapsed into 10,699 unique reads by APDP. Chimera removal reduced the dataset to 10,529 reads, and removal of singletons (reads and ecological occurrences) further reduced this to 4,596 unique reads that were used as input for the SNAPhy workflow.

Based on the Grinder simulated datasets, the optimal method for species delimitation was found to be a combination of BEAST and GMYC with a single threshold (applied using splits 1.0–11; Ezard et al. 2009), which gave both the closest number of species to the 'true' value and the lowest number of erroneous species assignments (Table 1; Table S1). Application of GMYC to small BEAST trees was found to give unreliable results and so reads from co-occurrence networks with fewer than 10 reads were treated differently –see "'Orphan' reads".

Analysis of the estuarine dataset in CoNet yielded a total of 45 co-occurrence networks containing at least 10 unique reads, with an overall clustering coefficient of 0.769 (for a given node, the clustering coefficient is the proportion of neighbours that are connected). The largest network contained a total of 231 unique reads. However, the majority of networks were much smaller (Table S3). Altogether, the co-occurrence networks included a total of 2,331 reads.

BEAST and GMYC modelling alone gave a total of 589 OTUs belonging to co-occurrence networks, and further splitting GMYC units by highly supported clades (i.e. with posterior probabilities higher than 0.9) gave a total of 851 OTUs (Table S2).

'Orphan' Reads

A large number of 'orphan' reads either did not belong to a co-occurrence network (1,381 reads) or belonged to a co-occurrence network which was too small to be analysed by GMYC (884 reads). GMYC species delimitation thresholds were significant for orphan phylum groupings for Annelida, Mollusca, Fungi, Nematoda, Panarthropoda, Rhizaria, Platyhelminthes and Alveolata (Table S2) and were split into 206 OTUs by GMYC, and further split to give 478 OTUs once posterior probabilities were applied (see Fig. 2 for example).

Testing the SNAPhy Workflow

Of the five OTUs chosen at random, only two were present at high abundances (several hundred reads) and both of these show the expected pattern of a single dominant read with a number of rare variants present at much lower abundances (Fig. 3A; 3E). The remaining OTUs were present at low abundances, and lacked an obvious dominant read (Fig. 3B-3D).

Percentage similarity within SNAPhy-delimited OTUs varied greatly, ranging from 74.7% to 99.6%. However, percentage similarity was very high within the majority of OTUs: just under half (49%) of OTUs had mean intra-OTU similarity values of 99-100% and an additional one third (33%) had mean intra-OTU similarity values of 98-99% (Fig. 4).

Each of the methods compared delimited a different number of OTUs: 1,329 for SNAPhy, 1,005 for UCLUST with a 96% threshold, 2,021 for UCLUST with a 98% threshold, and 3,683 for Swarm. The taxonomic composition within taxa also varied between methods (Fig 5). For example, a higher proportion of SNAPhy OTUs belonged to the Metazoa and Fungi

compared to other methods, while UCLUST (96% threshold) detected the highest proportion of 'Unassigned' taxa.

Three SNAPhy OTUs were matched to nematode sequences from the deep-sequencing dataset and used to generate neighbour-joining trees. For each deep-sequenced individual the majority of reads belonged to a well-supported grouping of very similar reads (corresponding to the target individual, confirmed by chain termination sequencing), and this grouping included all reads belonging to the SNAPhy OTU (Fig. 6). A number of reads formed outlying clades, which belonged to non-target taxa.

Discussion

We have demonstrated a novel method for delimiting ecologically and phylogenetically informed species units in metabarcoding datasets using a combination of co-occurrence patterns and phylogenetic modelling. Unlike commonly used static OTU clustering methods, the SNAPhy workflow explicitly reflects the general lineage species concept.

SNAPhy Workflow

Relatively few chimeras were removed from the database (170 reads in total), probably as a result of trimming the reads to a length of 225bp, thereby reducing the opportunity to detect 3' PCR recombination events (Wintzingerode et al. 1997).

Grouping reads based on co-occurrence patterns vastly reduces the size of the dataset within which species can be delimited (e.g. here 4,596 reads to 45 networks), thereby allowing the use of computationally expensive species delimitations methods such as phylogeny-based approaches. Incorporating phylogeny-based delimitation methods is more powerful than relying on sequence divergence alone because it relies on a statistical model of branching rates that allow for optimisation, assignment of confidence limits and hypothesis testing

(Barraclough et al. 2009). Genetic sequence data is much more complex than static similarity thresholds take into account, and incorporating models of evolution (explicitly explored within the SNAPhy framework) gives a more nuanced perspective on how sequences differ.

Previous assessments of the GMYC and 18S nSSU, according to chain termination sequencing data, have been found to underestimate diversity owing to the lumping of separate species (Tang et al. 2012); the high degree of divergence within OTUs delimited by GMYC suggests this may also be true for the current dataset. Predicted lumping here was amended via the application of posterior probabilities (using an objective intervention of 0.9 that can be adapted by the user to suit specific datasets), where well supported clades within GMYC entities were further partitioned into potential OTUs. These units represent species under a general lineage species concept (de Queiroz 2007), wherein species are defined as "seperately evolving metapopulation lineages". The units defined by SNAPhy also have the potential to reflect species under evolutionary or monophyletic species concepts (de Queiroz 2007). In the current example posterior probabilities were applied in order to split clades within GMYC entities, meaning that the evolutionary species concept was not applicable: importantly, however, it is likely that if the SNAPhy workflow were applied to another marker gene (e.g. CO1), the greater ratio between intra- and interspecific genetic divergence would allow delimitation of species without posterior probabilities, representing species under the evolutionary species concept. Nevertheless, SNAPhy takes large next-generation sequencing datasets as input and returns robust OTU numbers that are defined following the general lineage species concept.

Community distribution patterns are affected by four key processes: selection, drift, dispersal, and speciation (Vellend 2010). When sampling along an environmental gradient (such as an estuary) selection plays a strong role in determining species distribution patterns (Ferrero et al. 2008; Fonseca et al. 2014; Lallias et al. 2015), with dispersal and potentially drift also

playing a role (Fonseca et al. 2014). In the existing dataset, it is interesting to see a breadth in sizes of co-occurrence networks that likely reflect varying levels of environmental tolerances, stochastic processes and/or niche breadth (Vellend 2010). Moreover, the SNAPHy workflow also yields robust, ecologically informed co-occurrence phylogenies for downstream "eco-evo" analyses.

'Orphan' Reads

Almost half of the total reads did not belong to a co-occurrence network with more than 10 reads (again here, a parameter that can be adjusted by the user to facilitate phylogenetic modelling). This is unsurprising: most species exist at low abundances (Lim et al. 2012) and have few variants of the 18S nSSU gene (Ganley & Kobayashi 2007; Stage & Eickbush 2007). Also, the true distribution patterns of species may be obscured by incomplete sampling (for rare species) or the scale at which sampling was carried out (e.g. small species with microscopic niches). More surprising was the small number of OTUs delimited within the orphan reads. Despite similar numbers of unique reads being analysed as co-occurrence networks and orphans, the co-occurrence networks gave 851 OTUs while orphans gave only 478 OTUs. The discrepancy was due to a small number of very large OTUs within the orphan groupings, amongst a large numbers of smaller OTUs, most likely representing different sequence coverage focused on different species with unique occurrences or represented by smaller networks.

Testing the SNAPhy Workflow

As predicted, of the OTUs that contained a substantial number of reads (more than 400) (Fig. 3A; 3E), a single read was highly dominant amongst variants occurring at much lower abundances (Porazinska et al. 2010). While intragenomic rRNA variation is widespread amongst eukaryote taxa, in almost all cases examined so far a single variant is dominant, suggesting that concerted evolution is occurring (Bik et al. 2013; Ganley & Kobayashi 2007;

Stage & Eickbush 2007). Rarer OTUs represented by less than 100 reads did not show a clear pattern, likely due to low sequencing coverage for these genomes.

The percentage similarity within SNAPhy OTUs was high (Fig. 5) with the majority of OTUs having mean intra-OTU similarity of 98% or higher, as would be expected given the low divergence within the 18S nSSU gene (Tang et al. 2012; Wu et al. 2015). However, several OTUs had very low intra-OTU similarity values, with 18 OTUs containing mean divergence values of >10%. These OTUs were nonetheless strongly supported by either significant GMYC models and/or posterior probabilities, and therefore likely represent accurate groupings at higher taxonomic levels than species. These groups may in part be an artefact of low sequencing depths (meaning that there is not enough diversity within certain branches of the BEAST trees to distinguish species-level and genus-level differences). Alternatively, some may be a result of undetected chimeras or other errors, or may even be a result of extremely high levels of heterogeneity within the 18S nSSU region of some species (Lowe et al. 2005).

OTU clustering using a static similarity threshold (e.g. using UCLUST) is the most commonly-used method for OTU delimitation in sequencing datasets. Here, two similarity thresholds were chosen for comparison with the SNAPhy workflow: 96%, which has been shown to produce biologically plausible numbers of OTUs for nematode metabarcoding datasets (Fonseca et al. 2010), and 98%, which is closer to the average percentage similarity between SNAPhy OTUs (Fig. 5). An additional approach, Swarm, was included in the comparison as it avoids many of the pitfalls of clustering with a static threshold (Mahé et al. 2014).

Swarm yielded by far the highest number of OTUs: a total of 3,139, despite using a local clustering threshold (*d*) which was higher than recommended for most datasets

(https://github.com/torognes/swarm). This is potentially due to the relatively low sequencing depth in the current dataset: Swarm works in an iterative fashion, connecting reads to their 'neighbours' to form multi-branched chains (Mahé et al. 2014). If a given read is missing then the chain will be broken and an OTU may be split. More recent datasets have much higher coverage, e.g. due to the application of Illumina sequencing, and Swarm is therefore likely to perform better on these datasets.

The other three methods gave far fewer OTUs than Swarm: clustering in UCLUST yielded a total of 2,021 OTUs for the 98% similarity threshold and 1,002 for the 96% threshold, while SNAPhy yielded a total of 1,329 OTUs. Comparison of the four methods is difficult without knowing the 'true' number of species. However, focusing on the Nematoda suggests that Swarm overestimated the number of species present, giving a total of 802 nematode OTUs, while UCLUST with a 96% similarity threshold, underestimated the number with 149 OTUs. A previous study based on morphology (Ferrero et al. 2008) found a total of 153 nematode species along the Thames estuary, similar to the number detected by UCLUST with a 96% threshold, but the number in the current dataset would be expected to be considerably higher: the latter study included eight sites compared to 20 in the current work. In addition, molecular methods can detect cryptic species or eDNA (Bohmann et al. 2014), that studies based on morphospecies will not record. The WoRMS database (WoRMS Editorial Board 2015) recognises 416 Nematoda in UK marine habitats and so the OTU counts obtained by SNAPhy and by UCLUST with a 98% threshold (355 and 402, respectively) both seem reasonable given the wide range of conditions along the estuarine gradient (including freshwater environments, which are not featured in the WoRMS database) and the well-acknowledged hidden diversity in the phylum Nematoda (Fonseca et al. 2010).

As well as differences in the overall numbers of detected OTUs, different approaches differed considerably in the taxonomic distribution of OTUs. While Metazoa were the most abundant

phylum regardless of the OTU-delimitation method chosen, they made up a larger proportion of SNAPhy OTUs than they did of OTUs delimited by other methods. Conversely, protist groups (Alveolata, Rhizaria) and 'Unassigned' taxa made up a smaller proportion of SNAPhy OTUs than of OTUs delimited by other methods. The difference in taxonomic composition of OTUs between UCLUST and SNAPhy may result from inter-phylum differences in the degree of intraspecific variation found in the 18S nSSU region. For example, large intraspecies variation exists within the 18S nSSU region for many Alveolata and Rhizaria (Lowe et al. 2005; Caron et al. 2009; Weber & Pawlowski 2014) although other protists show much lower levels of intraspecies variation in 18S nSSU (Caron et al. 2009). Therefore, it is unclear whether protists and Metazoa consistently differ in the degree of variability within the 18S nSSU region. Another interesting feature of the SNAPhy OTUs was the low proportion of 'Unassigned' OTUs in comparison to standard OTU clustering. In standard OTU-clustering workflows undetected chimeras or erroneous reads may form OTUs based on similarity to one another. Since errors must always co-occur with parent sequences, SNAPhy is likely better able to link them to the true genomic sequences.

In each deep-sequenced nematode tree, reads belonging to the SNAPhy OTU fell within the clade formed by reads from the target organism, indicating that the range of variation within a SNAPhy OTU is well within the range of expected intragenomic variation. The deep-sequenced datasets also contained a number of reads that did not belong to the target nematode, most likely originating from stomach contents/contamination. The use of molecular methods to unravel food webs is a developing area of interest (Clare 2014) and the present data provides a glimpse into the potential of 18S nSSU metabarcoding to unravel trophic interactions in the meiofaunal biosphere (Pompanon et al. 2012).

A limited number of recent studies have demonstrated that co-occurrence patterns can be a powerful tool in the interpretation of microbial metabarcoding datasets, including a recently

447 described 16S rDNA denoising workflow, providing improved performance over error
448 model-based denoising algorithms (Tikhonov et al. 2014). Preheim et al. (2013) have also
449 incorporated distribution patterns into an OTU-calling method (Distribution-Based
450 Clustering, or DBC) that has been shown to outperform both *de novo* and closed reference
451 clustering methods on mock bacterial communities. However, DBC differs from SNAPhy in
452 several crucial ways. Firstly, DBC uses sequence similarity as the primary step in OTU
453 clustering despite the known disadvantages. Secondly, while SNAPhy clusters sequences
454 based on correlated occurrence patterns as a first step, DBC first matches reads based on
455 sequence similarity and merges the two as long as the two distributions are not significantly
456 different. While SNAPhy is currently limited to marker gene sequences, shotgun sequencing
457 is likely to become a more common tool in eukaryote ecology (Tang et al. 2015; Tang et al.
458 2014b; Zhou et al. 2013), and the use of co-occurrence patterns will become an even more
459 powerful approach to facilitate data analysis as the volume of sequence data increases. A
460 number of related approaches use co-occurrence patterns in order to bin metagenomics reads
461 into individual genomes (e.g. Albertsen et al. 2013; Alneberg et al. 2014).

462 Unlike other OTU clustering based approaches, SNAPhy presents a totally novel approach to
463 the delimitation of *de novo* species units in eukaryotic metabarcoding datasets, informed by
464 ecology and phylogeny. While we have tested SNAPhy on an 18S nSSU metabarcoding
465 dataset generated using 454-Roche pyrosequencing, our approach is easily adapted to other
466 sequencing technologies (e.g. Illumina, Pacific Biosciences) or genetic markers (such as
467 mtDNA) and will only be enhanced by increasing read lengths and increased genetic
468 variation (Tang et al. 2012). We envisage that broader scale testing will signal a move away
469 from computationally intensive quality control algorithms and static OTU-clustering and
470 towards an ecologically informed approach for delimiting species level biodiversity in
471 metabarcoding datasets. Once species can be effectively delimited in metabarcoding datasets,

accurate estimates of taxon diversity can be more effectively integrated into ecological studies, biomonitoring programs, with consequent benefits to ecologists and stakeholders.

Acknowledgements

We thank Matthew J. Morgan for his contributions to the development of the workflow and associated scripts. We thank Dale Falgate for his investigations into an early version of the SNAPhy workflow and for Tim Ferrero/Natalie Barnes for morphological identification of deep-sequenced individuals.

Data Accessibility

The Thames 18S nSSU metabarcoding data and the single nematode deep metabarcoding data can be found under the study numbers SRP043457 and SRP007674 respectively at the NCBI short read archive. Further details of the Thames sampling is given in Lallias *et al.* (2015).

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, and Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* 31:533-538.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, and Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods* 11:1144-1146.

Angly FE, Willner D, Rohwer F, Hugenholtz P, and Tyson GW. 2012. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*:gks251-gks251.

Balzer S, Malde K, Lanzén A, Sharma A, and Jonassen I. 2010. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics* 26:i420--i425.

Barraclough TG, Hughes, M, Ashford-Hodges, , Natalie, and Fujisawa T. 2009. Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data. *Biology Letters* rsbl-2009.

Bik HM, Fournier D, Sung W, Bergeron RD, and Thomas WK. 2013. Intra-genomic variation in the ribosomal repeats of nematodes. *PloS One* 8:e78230-e78230. 10.1371/journal.pone.0078230

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, and Thomas WK. 2012. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* 27:233-243.

Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, Yu DW, and de Bruyn M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution* 29:358-367.

Brown JH. 1995. *Macroecology*: University of Chicago Press.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421-421.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, and others. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7:335-336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, and Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions

of sequences per sample. *Proceedings of the National Academy of Sciences* 108:4516-4522.

Carew ME, Pettigrove VJ, Metzeling L, and Hoffmann AA. 2013. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology* 10:45-45.

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD, Dennett MR, Moran DM, and Jones AC. 2009. Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Applied and Environmental Microbiology* 75:5797-5808.

Chase JM, and Myers JA. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B: Biological sciences* 366:2351-2363.

Clare EL. 2014. Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. *Evolutionary Applications* 7:1144-1157.

Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lambshead PJD, and others. 2010. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology* 19:4-20.

De Queiroz K. 2007. Species Concepts and Species Delimitation. *Systematic Biology* 56:879-886.

Drummond AJ, and Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7:214-214.

Dumbrell AJ, Ashton PD, Aziz N, Feng G, Nelson M, Dytham C, Fitter AH, and Helgason T. 2011. Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *The New Phytologist* 190:794-804.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113-113.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, and Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194-2200.

Ezard T, Fujisawa T, and Barraclough TG. 2009. SPLITS: species' limits by threshold statistics. *R package version* 1.

549 Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, and Huttenhower C.
550 2012. Microbial co-occurrence relationships in the human microbiome. *PLoS*
551 *Computational Biology* 8:e1002606--e1002606.

552 Ferrero TJ, Debenham NJ, and Lambshead PJD. 2008. The nematodes of the Thames estuary:
553 Assemblage structure and biodiversity, with a test of Attrill's linear model. *Estuarine,*
554 *Coastal and Shelf Science* 79:409-418.

555 Fonseca VG, Carvalho GR, Nichols B, Quince C, Johnson HF, Neill SP, Lambshead JD,
556 Thomas WK, Power DM, and Creer S. 2014. Metagenetic analysis of patterns of
557 distribution and diversity of marine meiobenthic eukaryotes. *Global Ecology and*
558 *Biogeography* 23:1293-1302.

559 Fonseca VG, Carvalho GR, Sung W, Johnson HF, Power DM, Neill SP, Packer M, Blaxter
560 ML, Lambshead PJD, Thomas WK, and Creer S. 2010. Second-generation
561 environmental sequencing unmasks marine metazoan biodiversity. *Nature*
562 *Communications* 1:98-98.

563 Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, and Creer S. 2012.
564 Sample richness and genetic diversity as drivers of chimera formation in nSSU
565 metagenetic analyses. *Nucleic Acids Research* 40:e66--e66.

566 Friberg N, Bonada N, Bradley DC, Dunbar MJ, Edwards FK, Grey J, Hayes RB, Hildrew
567 AG, Lamouroux N, Trimmer M, and others. 2011. Biomonitoring of human impacts
568 in freshwater ecosystems: the good, the bad and the ugly. *Advances in Ecological*
569 *Research* 44:1-68.

570 Fujisawa T, and Barraclough TG. 2013. Delimiting Species Using Single-locus Data and the
571 Generalized Mixed Yule Coalescent (GMYC) Approach: A Revised Method and
572 Evaluation on Simulated Datasets. *Systematic Biology*:syt033.

573 Ganley ARD, and Kobayashi T. 2007. Highly efficient concerted evolution in the ribosomal
574 DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun
575 sequence data. *Genome Research* 17:184-191.

576 Gaspar JM, and Thomas WK. 2013. Assessing the consequences of denoising marker-based
577 metagenomic data. *PloS One* 8:e60458-e60458.

578 Hagen M, Kissling WD, Rasmussen C, Carstensen DW, Dupont YL, Kaiser-Bunbury CN,
579 O'Gorman EJ, Olesen JM, De Aguiar MAM, Brown LE, and others. 2012.
580 Biodiversity, species interactions and ecological networks in a fragmented world.
581 *Advances in Ecological Research* 46:89-120.

582 Hajibabaei M. 2012. The golden age of DNA metasystematics. *Trends in Genetics* 28:535-
583 537.

584 Huson DH, Auch AF, Qi J, and Schuster SC. 2007. MEGAN analysis of metagenomic data.
585 *Genome Research* 17:377-386.

586 Katoh, K, and Standley, DM. 2013. MAFFT Multiple Sequence Alignment Software Version
587 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*
588 30:772-780.

589 Lallias D, Hiddink JG, Fonseca VG, Gaspar JM, Sung W, Neill SP, Barnes N, Ferrero T, Hall
590 N, Lambshead PJD, and others. 2015. Environmental metabarcoding reveals
591 heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine
592 ecosystems. *The ISME Journal* 9:1208-1221.

593 Legendre P, and Fortin MJ. 1989. Spatial pattern and ecological analysis. *Vegetatio* 80:107-
594 138.

595 Lim GS, Balke M, and Meier R. 2012. Determining Species Boundaries in a World Full of
596 Rarity: Singletons, Species Delimitation Methods. *Systematic Biology* 61:165-169.

597 Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, and Pallen MJ.
598 2012. Performance comparison of benchtop high-throughput sequencing platforms.
599 *Nature Biotechnology* 30:434-439.

600 Loreau M, Naeem S, Inchausti P, Bengtsson J, Grime JP, Hector A, Hooper DU, Huston MA,
601 Raffaelli D, Schmid B, and others. 2001. Biodiversity and ecosystem functioning:
602 current knowledge and future challenges. *Science* 294:804-808.

603 Lowe CD, Day A, Kemp SJ, and Montagnes DJS. 2005. There are high levels of functional
604 and genetic diversity in *Oxyrrhis marina*. *Journal of Eukaryotic Microbiology*
605 52:250-257.

606 Mahé F, T. R, C. Q, de Vargas C, and Dunthorn M. 2014. Swarm: robust and fast clustering
607 method for amplicon-based studies. *PeerJ* 2.

608 Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone D,
609 de Vargas C, Bittner L, Zingone A, and Bowler C. 2016. Insights into global diatom
610 distribution and diversity in the world's ocean. *Proceedings of the National Academy*
611 *of Sciences* 113:E1516-E1525.

612 Morgan MJ, Chariton AA, Hartley DM, Hardy CM, and others. 2013. Improved inference of
613 taxonomic richness from environmental DNA. *PLoS One* 8:e71974-e71974.

614 Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle
615 J, Dunthorn M, and others. 2012. CBOL protist working group: barcoding eukaryotic
616 richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology* 10:e1001419.

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, and Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology* 21:1931-1950.

Porazinska DL, Giblin-Davis RM, Sung W, and Thomas WK. 2010. Linking operational clustered taxonomic units (OCTUs) from parallel ultra sequencing (PUS) to nematode species. *Zootaxa* 2427:55-63.

Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, and Alm EJ. 2013. Distribution-Based Clustering: Using Ecology to Refine the Operational Taxonomic Unit. *Applied and Environmental Microbiology* 79:6593-6603.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, and Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*:gks1219-gks1219.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, and Sloan WT. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* 6:639-641.

Schloss PD, and Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology* 77:3219-3226.

Schröter D, Cramer W, Leemans R, Prentice IC, Araújo MB, Arnell NW, Bondeau A, Bugmann H, Carter TR, Gracia CA, and others. 2005. Ecosystem service supply and vulnerability to global change in Europe. *Science* 310:1333-1337.

Stage DE, and Eickbush TH. 2007. Sequence variation within the rRNA gene loci of 12 Drosophila species. *Genome Research* 17:1888-1897.

Stamatakis A, Ludwig T, and Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456-463.

Taberlet P, Coissac E, Pompanon F, Brochmann C, and Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21:2045-2050.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28:2731-2739.

Tang CQ, Humphreys AM, Fontaneto D, and Barraclough TG. 2014a. Effects of phylogenetic reconstruction method on the robustness of species delimitation using single-locus data. *Methods in Ecology and Evolution* 5:1086-1094.

Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, and Fontaneto D. 2012. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proceedings of the National Academy of Sciences* 109:16208-16212.

Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce C, Nevard T, Potts SG, Zhou X, and Yu DW. 2015. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution* 6:1034-1043.

Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A, and others. 2014b. Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research* 42:e166--e166.

Tikhonov M, Leach RW, and Wingreen NS. 2014. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME Journal* 9:68-80.

Vellend M. 2010. Conceptual synthesis in community ecology. *The Quarterly review of Biology* 85:183-206.

Weber AA-T, and Pawlowski J. 2014. Wide occurrence of SSU rDNA intragenomic polymorphism in Foraminifera and its implications for molecular species identification. *Protist* 165:645-661.

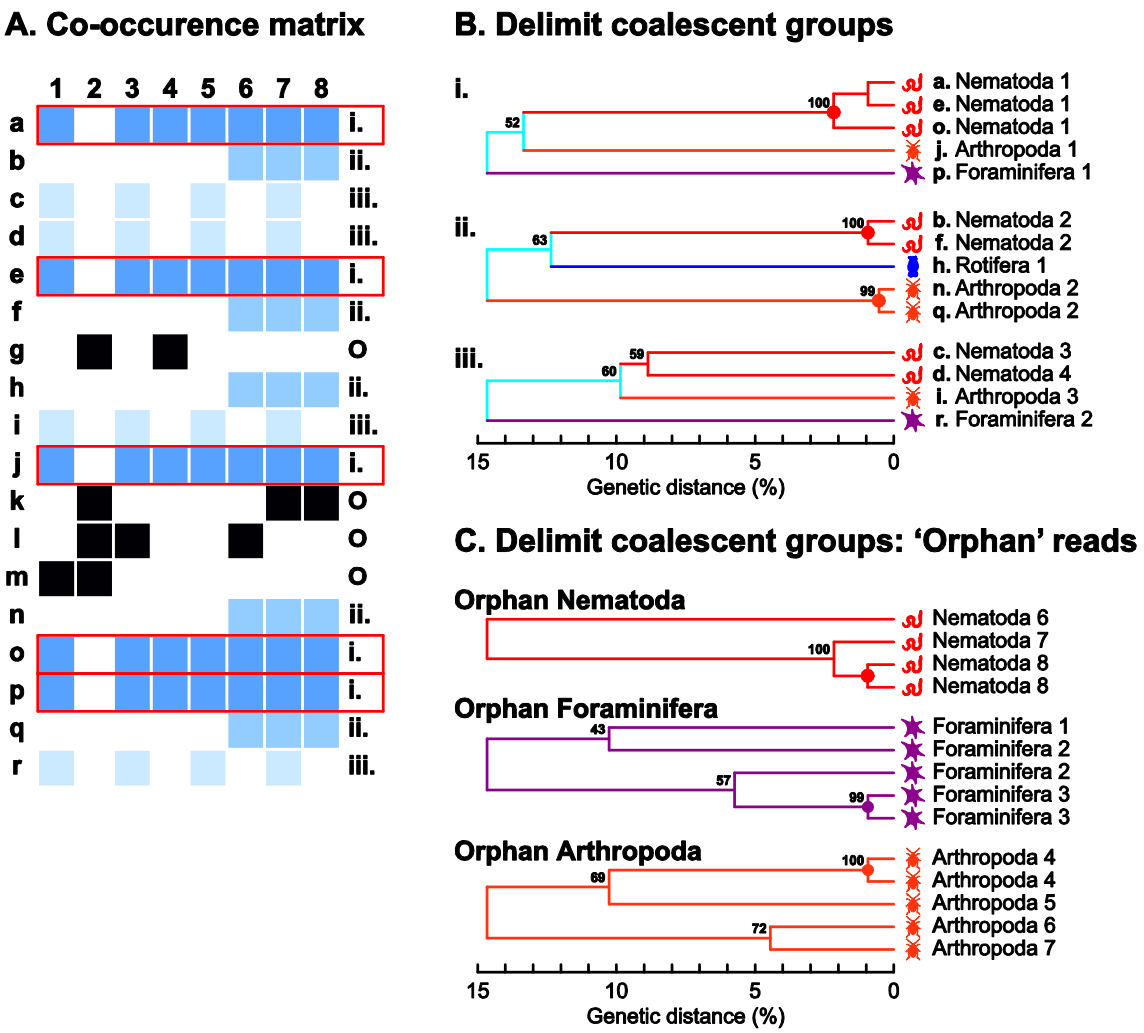
Wintzingerode FV, Göbel UB, and Stackebrandt E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews* 21:213-229.

Wu S, Xiong J, and Yu Y. 2015. Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass Copepoda. *PLoS one* 10:e0131498.

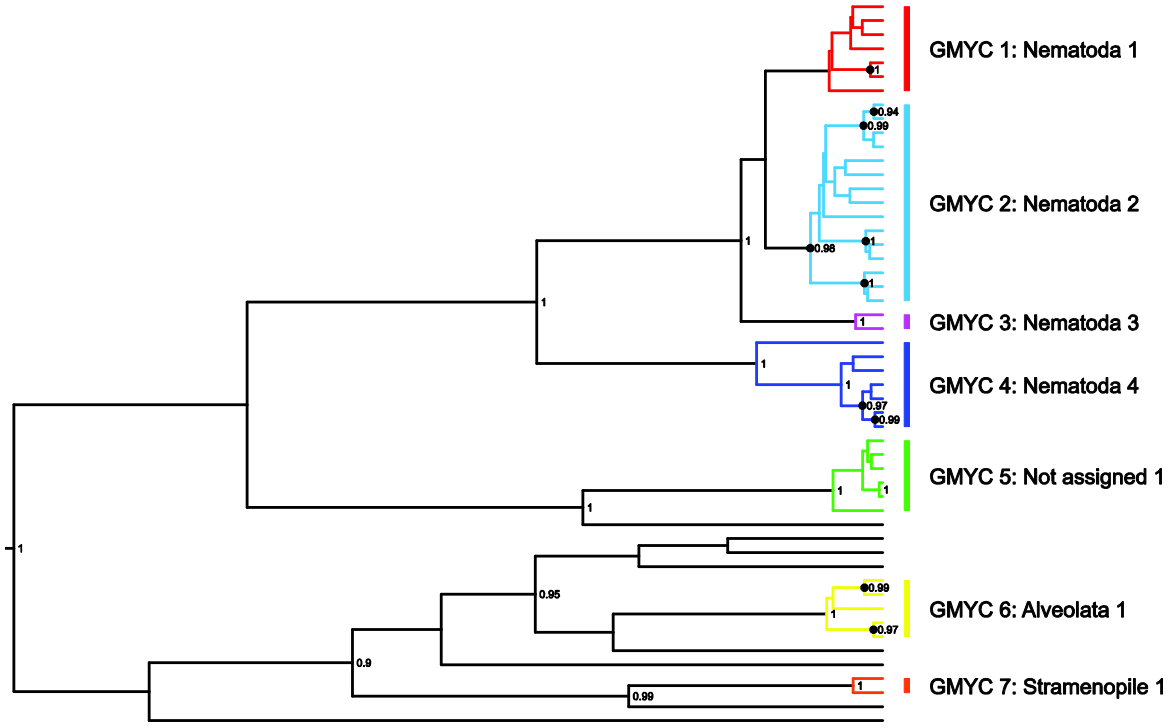
Zhang J, Kapli P, Pavlidis P, and Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29:2869-2876.

Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, and Huang Q. 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience* 2:4-4.

Figure 1: Summary of the key steps in the SNAPhy process. A. Quality-controlled sequences are clustered based on ecological co-occurrences. In the depicted co-occurrence matrix, columns represent samples and rows represent sequences. Different shades of blue cells represent occurrences of different species/ESUs. This gives clusters of reads, which co-occur in a subset of samples (e.g. ESU 'ii' contains reads which occur together in samples 6, 7 & 8). B. Species delimitation and phylogenetic modelling is applied to co-occurrence clusters. Numbered nodes on phylogenetic trees in B and C represent branch support. C. The reads that do not form co-occurrence clusters ('orphans', marked 'O' on A) are grouped based on taxonomy and species delimitation analysis proceeds as in B.

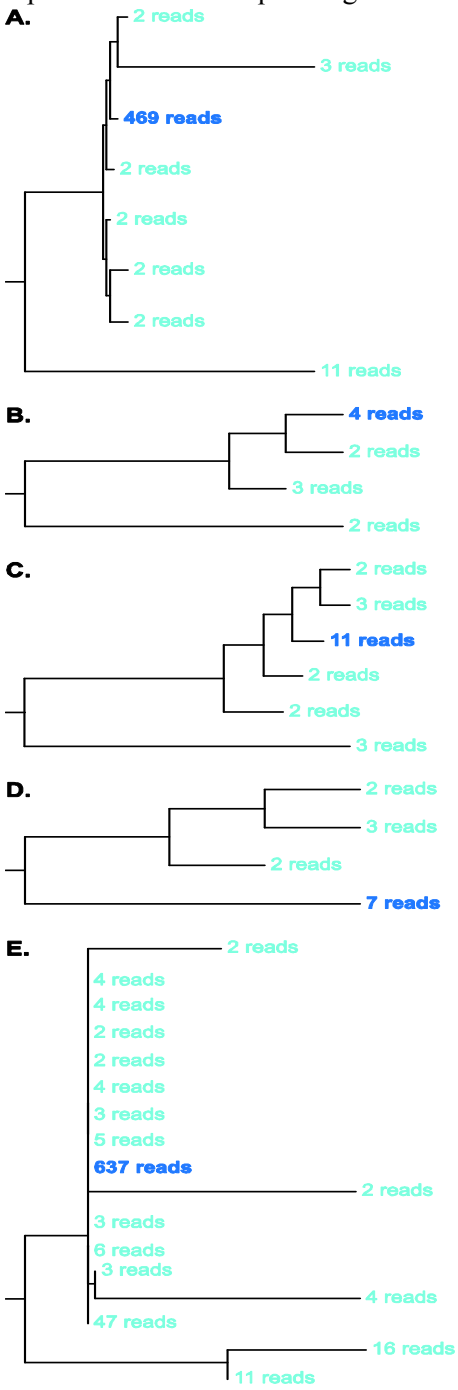


695 **Figure 2:** Example BEAST tree demonstrating a single co-occurrence network (network 290). Each
696 multi-sequence OTU delimited by OTU is shown as a different colour, while black dots show nodes at
697 which OTUs were split according to posterior probabilities.



698

699 **Figure 3:** Neighbour-joining trees representing five randomly chosen OTUs (generated using default
700 settings in MEGA), including abundances of each unique read within a given OTU. Read counts in
701 bold represent dominant reads - these are expected to be the "true" sequence, while other reads
702 represent errors in sequencing/PCR or intraspecific variants.



703

Figure 4: Histogram showing percentage divergence within all SNAPhy OTUs containing more than one sequence. Percentage divergence was calculated using "calc_distmx" command in USEARCH (Edgar 2010).

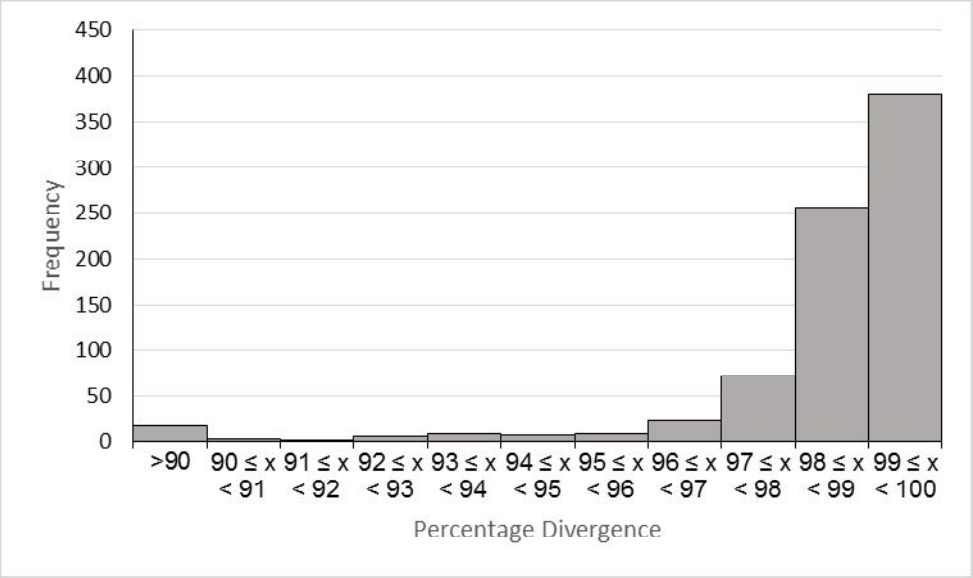


Figure 5: Proportion of Thames meiofaunal OTUs belonging to detected phyla using Swarm, UCLUST (at 96% and 98% thresholds) and SNAPhy. Taxonomic annotation was assigned using UCLUST within QIIME and the Silva 111 database for both SNAPhy and UCLUST OTUs.

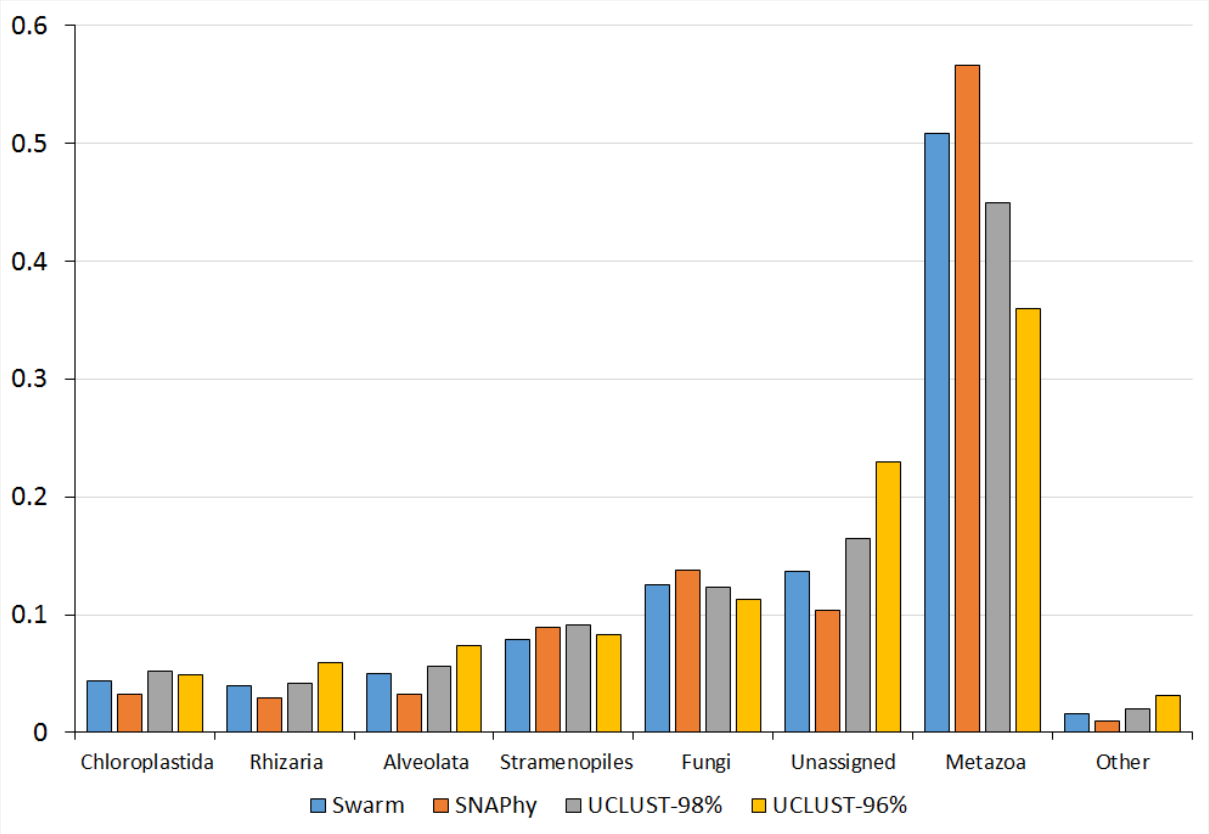


Figure 6: Neighbour-joining trees showing reads from deep metabarcoded nematodes with reads from a matching SNAPhy OTU. In each case, reads belonging to the SNAPhy OTU are located within reads belonging to the target nematode (red triangles). Branches were collapsed if divergence was less than 3%.

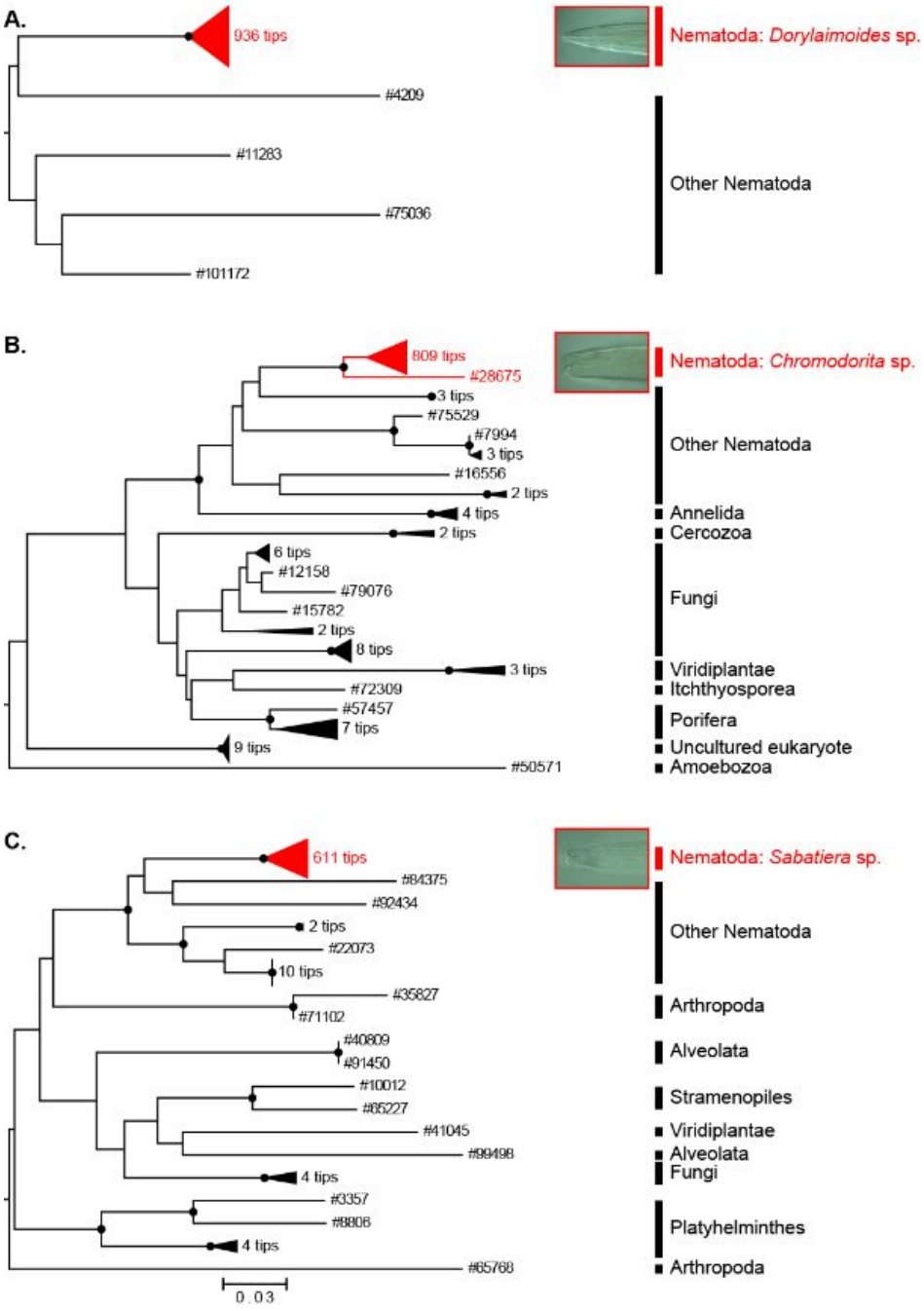


Table 1: Total OTU counts identified within the Grinder simulated datasets using different combinations of the tree reconstruction methods and phylogenetic delimitation models. ML = maximum likelihood solution; BI = most supported Bayesian inference; BI mean = average Bayesian inference; ST = single threshold; MT = multiple threshold; † = P value could not be calculated due to polytomous nodes; * = not significant; • = webserver could not analyse.

	RAxML				BEAST					
	PTP		GMYC		PTP		GMYC		Total Read	Expected OTU
Alignment	ML	BI	ST	MT	ML	BI	ST	MT	Count	Count
Mock 1	19	20	52†	29	NA•	NA•	22	**29**	**60**	**32**
Mock 2	28	28	8†	13	NA•	NA•	**14**	24*	**44**	**15**
Mock 3	**9**	**9**	12†	2	8	**9**	**9**	6	**14**	**10**
Mock 4	**3**	3	6†	NA	**3**	3	**3**	7*	**19**	**3**